ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Εφαρμογή και Υλοποίηση Αλγορίθμων Μηχανικής Μάθησης στην Ανάλυση Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΑΥΡΙΚΗ ΣΤΑΜΑΤΗ

Επιβλέπων : Μέντζας Γρηγόριος
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας ΜΑΘΗΜΑΤΙΚΩΝ

Εφαρμογή και Υλοποίηση Αλγορίθμων Μηχανικής Μάθησης στην Ανάλυση Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΜΑΥΡΙΚΗ ΣΤΑΜΑΤΗ

Επιβλέπων : Μέντζας Γρηγόριος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Νοεμβρίου 2011.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Μέντζας Γρηγόριος	Λουλάκης Μηχάλης	Ασκούνης Δημήτριος
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2011

(Υπογραφή)

.....

ΜΑΥΡΙΚΗΣ ΣΤΑΜΑΤΗΣ

Διπλωματούχος της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

ΚΕΦΑΛΑΙΟ 1
Εισαγωγή

1.1 Ανάλυση κειμένων

Η ανάλυση κειμένων (text mining), αφορά την διαδικασία εξαγωγής πληροφοριών υψηλής ποιότητας από κείμενα. Στο πεδίο ανάλυσης κειμένων περιέχονται διαδικασίες όπως διαδικασίες δόμησης (συνήθως διαπέραση κειμένου, με την πρόσθεση κάποιων λεξιλογικών χαρακτηριστικών και την αφαίρεση κάποιων όρων, και στην συνέχεια την αποθήκευση σε μία βάση δεδομένων), την εξαγωγή δομών που έχουν προέρθει από δομημένα δεδομένα, και τέλος την αξιολόγηση και ερμηνεία των αποτελεσμάτων. Η τυπική ανάλυση κειμένων περιλαμβάνει "κατηγοριοποίηση κειμένων", "ομαδοποίηση κειμένων", "περίληψη κειμένων" και "ανάλυση σχέσης οντοτήτων (π.χ. την εκμάθηση σχέσεων μεταξύ κάποιων συγκεκριμένων λέξεων)", και άλλα.

1.2 Αντικείμενο Διπλωματικής

Αντικειμενικός στόχος της παρούσας εργασίας είναι η ανάλυση και η εξέταση του Latent Dirichlet Allocation (LDA), ενός αλγορίθμου ανάλυσης κειμένων, και διάφορων παραλλαγών του. Στην Στατιστική, ο LDA είναι ένα γενετικό μοντέλο, που επιτρέπει την ερμηνεία κάποιων δεδομένων μέσα από κάποια σύνολα παρατηρήσεων. Για παράδειγμα, αν οι παρατηρήσεις είναι λέξεις που έχουν συλλεχθεί από κάποια κείμενα, ο αλγόριθμος θεωρεί ότι κάθε κείμενο είναι μία ανάμειξη από έναν μικρό αριθμών θεμάτων. Ο LDA είναι ένα παράδειγμα μοντέλου θεμάτων και παρουσιάστηκε για πρώτη φορά ως γενετικό μοντέλο ανακάλυψης θεμάτων από τους David Blei, Andrew Ng και Michael Jordan το 2002.

Για να υπάρξει μία πιο ακριβής σημασιολογική ανάλυση ενός κειμένου, ο αλγόριθμος έχει εξελιχθεί σε διάφορες παραλλαγές. Το ποια παραλλαγή θα χρησιμοποιηθεί εξαρτάται άμεσα από τη συλλογή δεδομένων που έχουμε στη διάθεσή μας. Έτσι θα δούμε και θα αναλύσουμε τον Ιεραρχικό LDA, μία μορφή του LDA που θεωρεί ότι ανάμεσα στα θέματα υπάρχει ιεραρχία, για παράδειγμα, αν αντικείμενο ενός κειμένου είναι η καρδιολογία σίγουρα θα υπάρχουν όροι που θα αναφέρονται στην ιατρική, δεν συμβαίνει όμως το αντίθετο, π.χ. ένα αντικείμενο ενός κειμένου ιατρικής μπορεί να είναι η παθολογία. Ακόμα θα δούμε τον Correlated LDA, μία παραλλαγή του LDA που λαμβάνει υπόψη της την σχετικότητα μεταξύ θεμάτων, λόγου χάρη, ένα κείμενο που μιλάει για ιατρική είναι πιθανότερο να μιλάει για την καρδιολογία πάρα για την αστρονομία. Θα ασχοληθούμε επίσης με την περιγραφή του Δυναμικού LDA, ο οποίο θεωρεί ότι τα θέματα εξελίσσονται στο βάθος του χρόνου, για παράδειγμα η ορολογία που χρησιμοποιείται στην νευρολογία ήταν πολύ διαφορετική το 1920, που τότε ήταν μία νέα σχετικά επιστήμη σε σχέση με το σήμερα. Τέλος θα δούμε και τον Επιβλεπόμενο LDA που θεωρεί ότι κάποιες συγκεκριμένες λέξεις έχουν μεγαλύτερη επιρροή σε ένα θέμα από όλες τις υπόλοιπες.

Ο LDA έχει εφαρμοστεί για την επίλυση πολλών προβλημάτων που αφορούν την ανάλυση κειμένων. Στα κοινωνικά δίκτυα, για παράδειγμα, ένας χρήστης μπορεί με την βοήθεια του LDA να κάνει μία πιο αποτελεσματική αναζήτηση για την εύρεση άλλων χρηστών με παρόμοια ενδιαφέροντα ή παρόμοιο χαρακτήρα. Ο αλγόριθμος βρίσκει, επίσης, εφαρμογή και στην ηλεκτρονική αγορά όπου μπορεί να συστήσει στον υποψήφιο αγοραστή μία σειρά από προϊόντα παρόμοια με αυτό που τον ενδιαφέρει να αγοράσει. Ο αλγόριθμος έχει ακόμα εφαρμογές και στην ανάλυση βιογραφικού υλικού, στην σύσταση ταινιών,

ηλεκτρονική παιχνιδιών και άρθρων.

Κύριο πλεονέκτημα του LDA, καθώς και των παραλλαγών του, είναι ο μεγάλος βαθμός προσαρμογής τους. Έτσι μπορούν να εφαρμοστούν και σε διάφορα άλλα προβλήματα που αντικείμενα απασχόλησης δεν είναι κείμενα λέξεων. Για παράδειγμα ο αλγόριθμος έχει χρησιμοποιηθεί στο πεδίο της μηχανική όρασης για ανάλυση εικόνας, στο πεδίο της βιοπληροφορικής για ανάλυση και ανάκτηση γενετικού κώδικα σε δεδομένα ερευνών κ.α.

Μία υλοποίηση του απλού LDA καθώς και του Ιεραρχικού LDA έχει γίνει στο πρόγραμμα MALLET στην γλώσσα JAVA. Στην παρούσα εργασία θα αναλυθούν οι κλάσεις και οι μέθοδοι των δύο αυτών υλοποιήσεων καθώς και θα τους εφαρμόσουμε στο σώμα *Polarity Dataset Version 2.0*. Μία παρόμοια ανάλυση και εφαρμογή θα γίνει στη υλοποίηση του Correlated LDA, που έχει δημιουργηθεί από τον συγγραφέα αυτής της εργασίας.

1.3 Οργάνωση κειμένου

Η παρούσα εργασία είναι οργανωμένη ως εξής: Στο κεφάλαιο 2 παραθέτονται κάποιες βασικές γνώσεις που είναι απαραίτητες για την περαιτέρω κατανόηση της εργασίας. Εδώ θα δοθούν οι βασικοί ορισμοί των αλγορίθμων εκμάθησης, των μοντέλων θεμάτων, των μοντέλων διανυσματικών χώρων. Στην συνέχεια θα αναλυθεί η Λανθάνουσα Σημασιολογική Ανάλυση μετεξέλιξη της οποίας είναι ο LDA και τέλος δοθεί μία σύντομη εισαγωγή στα Δίκτυα Πιθανοτήτων. Στο κεφάλαιο 3 αναλύουμε σε βάθος τον LDA και άλλες τέσσερις εκδοχές του. Πιο συγκεκριμένα θα εξετάσουμε τον απλό LDA, τον Ιεραρχικό LDA, τον Correlated LDA, τον Δυναμικό LDA (καθώς και τον Δυναμικό LDA Συνεχούς Χρόνου) και τέλος τον Επιβλεπόμενο LDA. Στο κεφάλαιο 4 θα δώσουμε μερικά παραδείγματα εφαρμογής του αλγορίθμου καθώς και το πώς προσαρμόστηκε σε αυτά. Συγκεκριμένα θα αναλύσουμε την εφαρμογή αλγορίθμου στην σύσταση ετικετών (λ.χ. σύσταση ταινιών), στην σύσταση άρθρων, στο microblogging, στην ηλεκτρονική αγορά, και στην ανάλυση οντοτήτων. Στο κεφάλαιο 5 γίνεται μία σύντομη εισαγωγή στο Mallet, στη συνέχεια αναλύονται οι κλάσεις ParallelTopicModel και HierachicalLDA (υλοποιήσεις του απλού LDA και του Ιεραρχικού LDA αντίστοιχα), ο Correlated LDA καθώς και οι βασικότερες κλάσεις του Mallet που παίρνουν μέρος στην διεργασία επεξεργασίας του κειμένου. Στο κεφάλαιο 6 παρατίθενται μερικές μέθοδοι αξιολόγησης του αλγορίθμου και γίνεται οι εφαρμογή των τριών υλοποιήσεων του αλγορίθμου (απλού LDA, Ιεραρχικού LDA, Correlated LDA) στο σώμα *Polarity Dataset Version 2.0*. Τέλος στο κεφάλαιο 7 γίνεται μία τελική σύνοψη του αλγορίθμου και των παραλλαγών του, το που έχει συνεισφέρει και ποια είναι τα πλεονεκτήματά του.

Κεφάλαιο 2
Υπόβαθρο

Στο κεφάλαιο αυτό δίνονται κάποιοι ορισμοί που η κατανόηση τους είναι απαραίτητη για την συνέχεια της παρούσας διπλωματικής εργασίας.

2.1 Αλγόριθμοι Εκμάθησης

Η μηχανική εκμάθηση, ένας υποκλάδος της τεχνητής νοημοσύνης, είναι μια επιστήμη που έχει σαν αντικείμενο τον σχεδιασμό και την ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να εξελίσσουν συμπεριφορά που βασίζεται σε εμπειρικά δεδομένα ή βάσεις δεδομένων. Ένα πρόγραμμα με την ικανότητα να μαθαίνει μπορεί να μάθει από παραδείγματα για να συλλάβει σημαντικά χαρακτηριστικά καθώς και την άγνωστη κατανομή πιθανότητας. Τα δεδομένα μπορούν να χαρακτηριστούν ως παραδείγματα που δείχνουν σχέσεις μεταξύ των παρατηρούμενων μεταβλητών. Ένας σημαντικός στόχος της έρευνας μηχανικής εκμάθησης είναι να μπορέσει να αναγνωρίζει αυτόματα πολύπλοκα σχήματα και να παίρνει αποφάσεις βασισμένη σε δεδομένα. Η δυσκολία βρίσκεται στο γεγονός ότι το σύνολο των πιθανών συμπεριφορών δεδομένων όλων των δυνατών εισόδων είναι πάρα πολύ μεγάλος για να καλυφθεί από τα παρατηρούμενα παραδείγματα. Έτσι ο αλγόριθμος εκμάθησης πρέπει να γενικεύσει από τα δεδομένα παραδείγματα, για να είναι ικανός να παράγει ένα ασφαλές αποτέλεσμα σε νέες περιπτώσεις.

Η υπολογιστική ανάλυση των αλγορίθμων εκμάθησης και η λειτουργία τους είναι ένας κλάδος της θεωρητικής επιστήμης υπολογιστών γνωστή ως θεωρία υπολογιστικής εκμάθησης. Επειδή τα σύνολα που χρησιμοποιούνται στην εκπαίδευση είναι περιορισμένα και το μέλλον είναι αβέβαιο, η θεωρία θεωρείται αποτελεσματική αν μπορεί να υλοποιηθεί σε πολυωνυμικό χρόνο. Υπάρχουν δύο ειδών αποτελέσματα πολυπλοκότητας χρόνου. Τα θετικά αποτελέσματα δείχνουν ότι μία κλάση συναρτήσεων μπορεί να διδαχθεί σε πολυωνυμικό χρόνο. Τα αρνητικά αποτελέσματα δείχνουν ότι δεν υπάρχουν κλάσεις που να μπορούν να διδαχθούν σε πολυωνυμικό χρόνο. Υπάρχουν ομοιότητες μεταξύ της θεωρίας υπολογιστικής εκμάθησης και της στατιστικής, παρόλο που χρησιμοποιούν διαφορετικούς όρους.

Οι αλγόριθμοι εκμάθησης μπορούν να ταξινομηθούν σύμφωνα με τα επιθυμητά αποτελέσματα του κάθε αλγόριθμου:

•Ένας Supervised learning αλγόριθμος δημιουργεί μία συνάρτηση που σχεδιάζει τις εισόδους σε επιθυμητές εξόδους π.χ. σε ένα classification πρόβλημα ο αλγόριθμος υπολογίζει την συνάρτηση σχεδιάζοντας ένα διάνυσμα σε κλάσεις παρατηρώντας παραδείγματα εισόδου-εξόδου της συνάρτησης.
•Ένας Unsupervised learning αλγόριθμος μοντελοποιεί ένα σύνολο εισόδων, όπως την ομαδοποίηση.
•Ένας Semi-supervised learning συνδυάζει και δύο παραδείγματα για παράγει την κατάλληλη συνάρτηση.
•Ένας Reinforcement learning μαθαίνει πως να συμπεριφέρεται σύμφωνα με αυτά που παρατηρεί στον κόσμο. Κάθε ενέργεια ασκεί κάποια επιρροή στο περιβάλλον, και το περιβάλλον παρέχει πληροφορίες που τον οδηγούν αλγόριθμοι
•Ένας Transduction προσπαθεί να προβλέψει νέα αποτελέσματα βασισμένα σε εισόδους, εξόδους και πειραματικές εισόδους.

•Ένας Learning to learn αλγόριθμος μαθαίνει από μόνος του επαγωγικά από προηγούμενες εμπειρίες.

2.1.1 Μοντέλα Θεμάτων

Στην μηχανική εκμάθηση και στην επεξεργασία φυσικής γλώσσας, ένα μοντέλο θεμάτων (topic model) είναι ένας τύπος στατιστικού μοντέλου για την ανακάλυψη θεμάτων που υπάρχουν σε μία συλλογή κειμένων. Τα μοντέλα θεμάτων βασίζονται στην ιδέα ότι τα κείμενα είναι μείγματα θεμάτων όπου ένα θέμα είναι μια πιθανότητα κατανομής λέξεων. Ένα μοντέλο θεμάτων είναι ένα γενετικό μοντέλο για κείμενα. Καθορίζει μια απλή πιθανοτική διαδικασία με την οποία μπορούν να παραχθούν νέα κείμενα. Για να φτιάξει ένα νέο κείμενο, διαλέγει ένα θέμα τυχαία δεδομένης της κατανομής και γράφει μια λέξη αυτού του θέματος. Κλασσικές στατιστικές τεχνικές μπορούν να χρησιμοποιηθούν για να αντιστρέψουν την διαδικασία, ανακαλύπτοντας το σύνολο των θεμάτων που ήταν υπεύθυνα για αυτή την συλλογή κείμενων.

2.2 Ανάκτηση Πληροφορίας

Η ανάκτηση πληροφορίας (Information Retrieval (IR)) είναι μια περιοχή μελέτης που αφορά την αναζήτηση κειμένων για πληροφορίες και για μεταδεδομένα (metadata) κειμένων, καθώς επίσης την αναζήτηση σε συσχετισμένες βάσεις δεδομένων και στο Διαδικτύου. Πολλοί ορισμοί είναι κοινοί μεταξύ ανάκτησης κειμένου και ανάκτησης δεδομένων αλλά η ανάκτηση κειμένων έχει τις δικές της ορολογίες καθώς και την δική της τεχνολογία. Η ανάκτηση πληροφορίας είναι ένας τομέας που βασίζεται στην επιστήμη υπολογιστών, τα μαθηματικά, στην επιστήμη πληροφορίας, στην αρχιτεκτονική πληροφορίας, στη ψυχολογία αντίληψης, στη γλωσσολογία και την στατιστική.

Τα αυτόματα συστήματα ανάκτησης πληροφορίας χρησιμοποιούνται για να μειώσουν αυτό που λέγεται “υπερφόρτωση δεδομένων (information overload)”. Πολλά πανεπιστήμια και δημόσιες βιβλιοθήκες χρησιμοποιούν τέτοια συστήματα για παρέχουν πρόσβαση σε βιβλία, άρθρα και άλλα κείμενα. Οι μηχανές αναζήτησης του Διαδικτύου είναι οι πιο προφανείς IR εφαρμογές.

Ιστορία. Η ιδέα της χρησιμοποίησης υπολογιστών για συσχετισμένα κομμάτια πληροφορίας έγινε γνωστή με το άρθρο *As we may Thing* του Vannerar Bush το 1945. Τα πρώτα αυτόματα IR συστήματα δημιουργήθηκαν στις δεκαετίες του '50 και '60. Το 1970 πολλές διαφορετικές τεχνικές έδειχναν να λειτουργούν καλά σε μικρές συλλογές κειμένων, όπως η συλλογή Cranfield. Μεγαλύτερων διαστάσεων συστήματα ανάκτησης, όπως το σύστημα Lockheed Dialog αναδείχθηκαν στην αρχή της δεκαετίας του '70.

Το 1992, το υπουργείο άμυνας των ΗΠΑ σε συνεργασία με το National Institute of Stantards and Technology (NIST), επιχορήγησαν το Text Retrieval συνέδριο σαν κομμάτι του προγράμματος κειμένων TIPSER. Ο στόχος αυτού του συνεδρίου ήταν η αναζήτηση διαμέσου της IR κοινότητας για μια δομή αξιολόγησης μεθοδολογιών σε μεγάλες συλλογές κειμένων. Κάτι που επέφερε ένα καταλυτικό αποτέλεσμα στην έρευνα μεθόδων μεγάλων συλλογών. Η δημιουργία μηχανών αναζήτησης Διαδικτύου προώθησε την ανάγκη για μεγάλου μεγέθους συστήματα ανάκτησης ακόμα περισσότερο.

Η χρησιμοποίηση ψηφιακών μεθόδων για αποθήκευση και ανάκτηση δεδομένων οδήγησε στο φαινόμενο ψηφιακής απαρχαίωσης, όπου η ψηφιακή πηγή σταματάει να είναι αναγνώσιμη λόγω των φυσικών μέσων. Η πληροφορία είναι ευκολότερο να ανακτηθεί από ότι αν ήταν γραμμένη σε χαρτί, αλλά μετά χάνεται.

2.3 Μοντέλο Διανυσματικού Χώρου

Το μοντέλο διανυσματικού χώρου είναι ένα αλγεβρικό μοντέλο που αντιστοιχεί σε κείμενα. Χρησιμοποιείται στο φιλτράρισμα πληροφορίας, στην ανάκτηση δεδομένων και στην αξιολόγηση συσχετίσεων. Χρησιμοποιήθηκε για πρώτη φορά στο σύστημα *SMART Information*.

Τα κείμενα αντιπροσωπεύονται από διανύσματα του τύπου:

$d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$

Κάθε διάσταση αντιστοιχεί σε έναν ξεχωριστό όρο. Αν ο όρος υπάρχει μέσα στο κείμενο, η τιμή του διανύσματος δεν είναι μηδενική. Υπάρχουν πολλοί τρόποι για τον υπολογισμό αυτών των τιμών, που έχουν αναπτυχθεί. Ένας από τους πιο γνωστούς είναι ο tf-idf που θα αναλύσουμε παρακάτω.

Ο ορισμός των όρων εξαρτάται από την εφαρμογή. Τυπικά οι όροι είναι απλές λέξεις ή φράσεις. Αν οι λέξεις είναι επιλεγμένες να είναι οι όροι, τότε η διάσταση του διανύσματος είναι ο αριθμός των λέξεων σε ένα λεξικό (Ο αριθμός των ξεχωριστών λέξεων που έχουν παρατηρηθεί σε ένα σώμα).

2.3.1 Tf-idf

Το tf-idf βάρος είναι ένα μέτρο που χρησιμοποιείται συχνά στην ανάκτηση δεδομένων και στην εξόρυξη κειμένων(text mining). Αυτό το βάρος είναι ένα στατιστικό μέτρο που χρησιμοποιούμε για να αξιολογήσουμε πόσο σημαντική είναι μια λέξη σε μια συλλογή κειμένων. Η σημαντικότητα ανεβαίνει αν η αναλογία των φορών που εμφανίζεται μια λέξη σε ένα κείμενο είναι αντισταθμισμένη από την συχνότητα της λέξης μέσα στη συλλογή. Διαφορές μορφές της μεθόδου tf-idf χρησιμοποιούνται συχνά από μηχανές αναζήτησης σαν το κύριο εργαλείο για την αξιολόγηση συσχέτισης κειμένων. Η μέθοδος αυτή χρησιμοποιεί αποτελεσματικά το φιλτράρισμα τετριμμένων λέξεων σε ποικίλες περιοχές συμπεριλαμβανομένου της περίληψης κειμένων και της αρχικοποίησης.

Μία από της πιο απλές συναρτήσεις αξιολόγησης υπολογίζεται αθροίζοντας το tf-idf βάρος για κάθε όρο. Πολλές άλλες, πιο πολύπλοκες, συναρτήσεις αξιολόγησης είναι μορφές αυτού του απλού μοντέλου.

Ας υποθέσουμε λοιπόν ότι έχουμε ένα σύνολο από αγγλικά κείμενα και θέλουμε να ορίσουμε ποιο κείμενο είναι πιο σχετικό με την φράση "the brown cow". Ένας απλός τρόπος να αρχίσουμε είναι να αποκλείσουμε τα κείμενα που δεν περιέχουν τις τρείς αυτές λέξεις "the", "brown" και "cow", άλλα έχουν ακόμα απομείνει πολλά κείμενα. Για να τα ξεχωρίσουμε, θα πρέπει να μετρήσουμε τον αριθμό που κάθε ένας από αυτούς τους όρους υπάρχει σε κάθε κείμενο και να τους αθροίσουμε. Ο αριθμός των φορών που υπάρχει ένας όρος σε ένα κείμενο ονομάζεται συχνότητα όρου. Όμως, επειδή ο όρος "the" είναι πιο συχνός δεν θεωρείται μια καλή λέξη-κλειδί για να χωρίσουμε τα κείμενα σε σχετικά ή μη σχετικά με την φράση. Αντίθετα, οι λέξεις "brown" και "cow" που συναντώνται πολύ σπάνια είναι καλές λέξεις-κλειδιά για να χωρίσουν τα κείμενα σε σχετικά ή μη. Έτσι ένας παράγοντας αντίστροφης συχνότητας κειμένων χρειάζεται για να μειώσει το βάρος των όρων που απαντώνται πολύ συχνά στην συλλογή και αυξάνει το βάρος αυτών που συναντώνται σπάνια.

Μαθηματικές λεπτομέρειες. Η *μέτρηση όρου* είναι απλώς ο αριθμός των φορών που εμφανίζεται ένας όρος σε ένα κείμενο. Αυτή η μέτρηση κανονικοποιείται συνήθως για να αποκλείσει την προδιάθεση σε μεγαλύτερα κείμενα (που μπορεί να έχουν μια υψηλότερη συχνότητα όρου ασχέτως της πραγματικής σημαντικότητας του όρου στο κείμενο) να

δώσουν ένα μέτρο σημαντικότητας του όρου t σε ένα συγκεκριμένο κείμενο d. Έτσι έχουμε την *συχνότητα όρου* tf(*t*,*d*), που ορίζεται στην απλούστερη περίπτωση ως ο αριθμός εύρεσης ενός όρου σε ένα κείμενο.

Η αντίστροφη συχνότητα κειμένων είναι ένα μέτρο για την γενική σημασία του όρου που έχει αποκτηθεί διαιρώντας τον συνολικό αριθμό των κειμένων με τον αριθμό των κειμένων που περιέχουν τον όρο, και μετά παίρνοντας τον όρο του πηλίκου.

$$\mathrm{idf}(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

όπου

- $|D|$: ο συνολικός αριθμός των κειμένων στην συλλογή
- $|\{d : t \in d\}|$:ο αριθμός το κειμένων που εμφανίζεται ο όρος(δηλ, $\mathrm{tf}(t, d) \neq 0$). Αν ο όρος δεν είναι στη συλλογή, αυτό θα οδηγήσει σε διαίρεση με το 0. Έτσι μετατρέπουμε την φόρμουλα στην μορφή $1 + |\{d : t \in d\}|$

Έπειτα,

$$\text{tf-idf}(t, d) = \mathrm{tf}(t, d) \times \mathrm{idf}(t)$$

Ένα υψηλό βάρος tf-idf χαρακτηρίζεται από μια υψηλή συχνότητα όρου στο δεδομένο κείμενο και μια χαμηλή συχνότητα όρου στην υπόλοιπη συλλογή κειμένων. Τα βάρη έτσι τείνουν να φιλτράρουν τους τετριμμένους όρους. Η tf-idf τιμή ενός όρου θα είναι μεγαλύτερή του μηδενός αν και μόνο αν η αναλογία μέσα στην λογαριθμική συνάρτηση του idf είναι μεγαλύτερή από 1. Ανάλογα με το αν το 1 προστεθεί ή όχι στον παρονομαστή, ένας όρος θα έχει τιμή idf είτε 0 είτε αρνητική, και αν το 1 έχει προστεθεί στον παρονομαστή ένας όρος που δεν υπάρχει σε κανένα κείμενο εκτός από ένα θα έχει idf τιμή ίση με 0.

Πολλές μορφές του tf-idf βάρους μπορούν να δημιουργηθούν από ένα πιθανοτικό μοντέλο ανάκτησης που μιμείται την ανθρώπινη συμπεριφορά στη λήψη αποφάσεων.

2.4 Λανθάνουσα Σημασιολογική Ανάλυση (LSA)

2.4.1 Εισαγωγή

Η Λανθάνουσα Σημασιολογική Ανάλυση (LSA) είναι μια θεωρία και μέθοδος που συλλαμβάνει το νόημα των λέξεων με στατιστικούς υπολογισμούς πάνω σε μια μεγάλη συλλογή κειμένων. Η ιδέα πίσω από αυτή την θεωρία είναι να βρεθεί ο λόγος των κείμενων στα οποία μία συγκεκριμένη λέξη εμφανίζεται προς αυτά που δεν εμφανίζονται παρέχοντας ένα σύνολο αμοιβαίων περιορισμών που καθορίζουν την ομοιότητα των σημασιών των λέξεων ή συνόλων αυτών. Η ικανότητα της LSA στην απεικόνιση της ανθρώπινης γνώσης έχει εδραιωθεί με ποικίλους τρόπους. Ένα παράδειγμα είναι η απόκτηση γνώσης που συμπίπτει με αυτή των ανθρώπων σε ένα συνηθισμένο λεξιλόγιο, όπως και η μίμηση της ανθρώπινης συμπεριφοράς στην σύνταξη των λέξεων.

2.4.2 Περιγραφή Αλγόριθμου

Ο αλγόριθμος της Λανθάνουσας Σημασιολογικής Ανάλυσης (LSA) είναι μια πλήρως αυτοματοποιημένη μαθηματική/στατιστική τεχνική για την εύρεση συσχετίσεων μεταξύ λέξεων. Δεν είναι μια παραδοσιακή διαδικασία φυσικής γλώσσας ή ένα πρόγραμμα τεχνητής νοημοσύνης. Χρησιμοποιεί κατασκευασμένα λεξικά, βάσεις πληροφοριών, σημασιολογικά δίκτυα, γραμματικές, συντακτικούς διαπερραστές ή μορφολογίες και παίρνει σαν είσοδο μόνο απλό κείμενο χωρισμένο σε λέξεις, ορισμένες μοναδικά η καθεμία και χωρισμένες σε κομμάτια όπως προτάσεις και παραγράφους.

Το πρώτο βήμα είναι να αντιστοιχίσουμε ένα κείμενο σε έναν πίνακα όπου κάθε σειρά δηλώνει μια μοναδική λέξη και κάθε στήλη ένα κείμενο. Κάθε κελί περιέχει την πιθανότητα με την οποία κάθε λέξη εμφανίζεται σε ένα κείμενο που ορίζεται από την στήλη. Στην συνέχεια, οι είσοδοι των κελιών υπόκεινται σε μια προκαταρκτική μεταμόρφωση, στην οποία κάθε συχνότητα κελιού ζυγίζεται από μια συνάρτηση που αντιστοιχεί στην σημαντικότητα της λέξης στο συγκεκριμένο κείμενο και στον βαθμό που η λέξη αυτή παρέχει πληροφορία.

Μετά, ο LSA εφαρμόζει μια ανάλυση τιμών(SVD) στον πίνακα. Αυτή είναι η μορφή ανάλυσης παραγόντων ή πιο κατάλληλα η μαθηματική γενίκευση του κατά την οποία η ανάλυση παραγόντων είναι μια ειδική περίπτωση. Με τη διαδικασία SVD, ένας ορθογώνιος πίνακας μετατρέπεται σε γινόμενο τριών άλλων πινάκων. Ο ένας πίνακας περιγράφει τις αυθεντικές οντότητες σειράς σαν διανύσματα παραγόμενα από έναν ορθοκανονικό παράγοντα τιμών, ο άλλος περιγράφει τις αυθεντικές οντότητες στήλης με παρόμοιο τρόπο, και ο τρίτος είναι ένας διαγώνιος πίνακας που περιέχει αύξουσες τιμές έτσι ώστε όταν και οι τρεις πίνακες πολλαπλασιαστούν, ξαναδημιουργείται ο πρώτος πίνακας. Υπάρχει μαθηματική απόδειξη ότι κάθε πίνακας μπορεί να αποσυνδεθεί τέλεια, χρησιμοποιώντας όχι περισσότερους παράγοντες από την μικρότερη διάσταση του αρχικού πίνακα. Όταν λιγότεροι από τον απαραίτητο αριθμό παραγόντων χρησιμοποιούνται, ο ανακατασκευασμένος πίνακας είναι καλύτερα φιξαρισμένος. Μπορούμε να μειώσουμε την διάσταση της λύσης απλά διαγράφοντας συντελεστές του διαγώνιου πίνακα, συνήθως αρχίζοντας από τον μικρότερο. (Πρακτικά, για υπολογιστικούς λόγους, για μια πολύ μεγάλη συλλογή μόνο ένα μικρός αριθμός διαστάσεων της τάξεως χιλιάδων μπορεί να κατασκευαστεί.)

2.5 Δίκτυα Πιθανοτήτων

2.5.1 Εισαγωγή

Τα μοντέλα θεμάτων παίζουν, όπως είδαμε στην εισαγωγή, ένα σημαντικό ρόλο στην ανάλυση κειμένων. Πριν όμως αρχίσουμε την ανάλυσή τους, είναι σημαντικό να επεκτείνουμε τις γνώσεις μας πάνω στις διαγραμματικές απεικονίσεις των πιθανοτικών κατανομών, γνωστές ως πιθανοτικά γραφικά μοντέλα, τα οποία έχουν τα εξής πλεονεκτήματα:

1) Παρέχουν ένα απλό τρόπο για την κατανόηση των ιδιοτήτων ενός μοντέλου

2) Βοηθούν στην εμβάθυνση των ιδιοτήτων του μοντέλου, συμπεριλαμβανομένων και των ιδιοτήτων δεσμευμένης ανεξαρτησίας, που μπορούν να αποκτηθούν από την ανάλυση του γράφου.

3)Οι πολύπλοκοι υπολογισμοί που απαιτούνται για την έκβαση συμπερασματολογίας και εκμάθησης μοντέλων, μπορούν να εκφραστούν πλέον με γραφικούς όρους, στους οποίους υπόκεινται οι υπονοούμενες μαθηματικές εκφράσεις

Ένας γράφος αποτελείται από κόμβους και γραμμές. Στα πιθανοτικά μοντέλα, κάθε κόμβος απεικονίζει μια τυχαία μεταβλητή, και οι γραμμές εκφράζουν τις πιθανοτικές σχέσεις μεταξύ αυτών των μεταβλητών. Ο γράφος έτσι συλλαμβάνει τον τρόπο με τον οποίο οι από κοινού κατανομές ως προς τις τυχαίες μεταβλητές μπορούν να αναλυθούν σε ένα γινόμενο παραγόντων, ο καθένας από τους οποίους εξαρτάται μόνο από ένα υποσύνολο μεταβλητών. Υπάρχουν δύο ειδών γραφικά μοντέλα τα κατευθυνόμενα γραφικά μοντέλα, των οποίων οι συνδέσεις έχουν ένα βέλος που δείχνει την κατεύθυνση και τα μη κατευθυνόμενα όπου αντίστοιχα οι γραμμές τους δεν δηλώνουν κατεύθυνση. Λόγω του ότι η παρούσα εργασία αφορά μοντέλα θεμάτων τα οποία είναι κατευθυνόμενα, θα ακολουθήσει μια σύντομη ανάλυση και περιγραφή των κατευθυνόμενων μοντέλων γνωστά και ως δίκτυα πιθανοτήτων.

2.5.2 Δίκτυα Πιθανοτήτων

Για να γίνει αντιληπτό, το πως χρησιμοποιούνται τα κατευθυνόμενα μοντέλα , σκεφτείτε πρώτα αφηρημένη από κοινού κατανομή *p(a,b,c)* ως προς τρεις μεταβλητές α, b και c. Με εφαρμογή του κανόνα παραγοντοποίησης, μπορούμε να γράψουμε την από κοινού κατανομή ως:

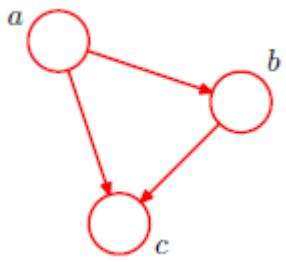
$$p(a, b, c) = p(c|a, b)p(a, b). \qquad (1)$$

Μια δεύτερη εφαρμογή του κανόνα παραγοντοποίησης,αυτή την φορά στον δεύτερο όρο της (1) δίνει:

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \qquad (2)$$

Τώρα μπορούμε να απεικονίσουμε την δεξιά πλευρά της (2) με τους όρους ενός απλού γραφικού μοντέλου όπως ακολουθεί. Πρώτα εισαγάγουμε έναν κόμβο για κάθε μία από τις

τυχαίες μεταβλητές a,b και c και έπειτα συσχετίζουμε κάθε κόμβο με την αντίστοιχη δεσμευμένη κατανομή της αριστερής πλευράς της (2). Μετά για κάθε δεσμευμένη κατανομή προσθέτουμε κατευθυνόμενες ακμές στο γράφημα από τους κόμβους που αντιστοιχούν στις μεταβλητές στις οποίες η κατανομή



Εικόνα 2.1: Ένα γραφικό μοντέλο που απεικονίζει την από κοινού κατανομή τριών μεταβλητών α,b και c, αντιστοιχεί στην ανάλυση την δεξίας πλευράς της

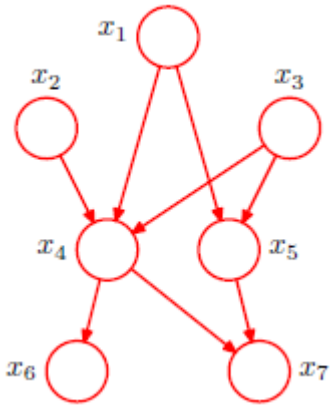
είναι δεσμευμένη. Έτσι από τον παράγοντα $p(c|\alpha,b)$, θα υπάρχουν συνδέσεις από τους κόμβους α και b
στον κόμβο c, όπου ο παράγοντας $p(\alpha)$ δεν θα έχει καμία εισερχόμενη γραμμή. Το αποτέλεσμα δίνεται στην Εικόνα 2.1. Αν υπάρχει μια συνδετική γραμμή που κατευθύνεται από τον κόμβο α στον b θα λέμε ότι ο κόμβος α είναι γονέας του κόμβου b, και ότι ο κόμβος b είναι παιδί του κόμβου α. Σημειώστε ότι δεν θα κάνουμε καμία διάκριση μεταξύ ενός κόμβου και της μεταβλητής στην οποία αντιστοιχεί αλλά απλά θα χρησιμοποιήσουμε το ίδιο σύμβολο και για τα δύο.

Προς στιγμήν ας αναλύσουμε την Εικόνα 2.1 θεωρώντας μία από κοινού κατανομή ως προς K μεταβλητές δεδομένων την $p(x_1,.....x_K)$. Από την επαλαμβανόμενη εφαρμογή του κανόνα παραγοντοποίησης της πιθανότητας, αυτή η από κοινού κατανομή μπορεί να γραφτεί ως γινόμενο δεσμευμένων κατανομών, μια για κάθε μεταβλητή

$$p(x_1,\ldots,x_K) = p(x_K|x_1,\ldots,x_{K-1})\ldots p(x_2|x_1)p(x_1). \qquad (3)$$

Για μία δεδομένη επιλογή του K, μπορούμε ξανά να την απεικονίσουμε ως κατευθυνόμενο γράφημα έχοντας Κ κόμβους, έναν για κάθε δεσμευμένη κατανομή στην δεξιά πλευρά της (3), με κάθε κόμβο να έχει εισερχόμενα διανύσματα από όλους τους κόμβους χαμηλότερου αριθμού. Λέμε ότι αυτό το γράφημα είναι πλήρως συνδεδεμένο επειδή υπάρχει μία σύνδεση για κάθε ζευγάρι κόμβων.

Μέχρι στιγμής, δουλέψαμε με πλήρως γενικές από κοινού κατανομές, έτσι ώστε οι αναλύσεις τους και οι απεικονίσεις τους ως πλήρως συνδεδεμένα γραφήματα, να είναι πλήρως εφαρμόσιμες σε κάθε επιλογή κατανομής. Όπως θα δούμε σε λίγο, είναι η απουσία των γραμμών που δίνει ενδιαφέρουσα πληροφορία σχετικά με τις ιδιότητες τις κλάσης των κατανομών που ένα γράφημα απεικονίζει. Σκεφτείτε το γράφημα της Εικόνας 2.2. Δεν είναι ένα πλήρως συνδεδεμένο γράφημα διότι, για παράδειγμα δεν υπάρχει σύνδεση από το x_1 στο x_2 ή από το x_3 στο x_7.



Εικόνα 2.2: Παράδειγμα ενός κατευθυνόμενου ακυκλικού γραφήματος που περιγράφει την από κοινού κατανομή των μεταβλητών x_1, ... , x_7. Η αντίστοιχη αναλυτική εξίσωση δίνεται στην (4).

Τώρα θα μεταβούμε από αυτό το γράφημα στην αντίστοιχη απεικόνισή της από κοινού κατανομής γραμμένης με τους όρους παραγοντοποίησης ενός συνόλου δεσμευμένων κατανομών, μια από κάθε κόμβο του γραφήματος. Κάθε τέτοια δεσμευμένη κατανομή θα είναι δεσμευμένη μόνο ως προς τους γονείς του αντίστοιχου κόμβου στο γράφημα. Για παράδειγμα, το x_5 είναι δεσμευμένο ως προς τα x_1 και x_3. Η από κοινού κατανομή όλων των 7 μεταβλητών δίνεται από την

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5). \qquad (4)$$

Καλό θα ήταν ο αναγνώστης να αφιερώσει λίγο χρόνο για την κατανόηση της αντιστοιχία της (4) με την Εικόνα 2.2.

Μπορούμε τώρα να ορίσουμε με γενικούς όρους της σχέση μεταξύ ενός δεδομένου κατευθυνόμενου γραφήματος και της αντιστοίχησης κατανομής των μεταβλητών. Η από κοινού κατανομή ορισμένη από ένα γράφημα δίνεται από το γινόμενο, ως προς όλους τους κόμβους του γραφήματος, της δεσμευμένης κατανομής για κάθε κόμβο δεσμευμένο από τις μεταβλητές που αντιστοιχούν στους γονείς αυτού του κόμβου του γραφήματος. Έτσι, για ένα γράφημα με K κόμβους, η από κοινού κατανομή δίνεται από

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \qquad (5)$$

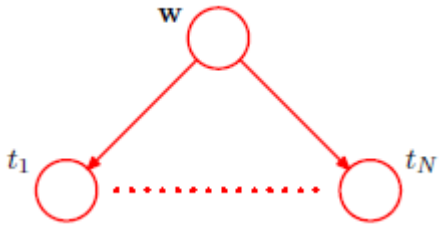
όπου το pa_k δηλώνει το σύνολο των γονέων του x, και $\mathbf{x} = \{x_1,.....,x_K\}$. Αυτή η εξίσωση κλειδί εκφράζει τις ιδιότητες παραγοντοποίησης της από κοινού κατανομής ενός κατευθυνόμενου γραφικού μοντέλου. Παρόλο που έχουμε θεωρήσει κάθε κόμβο να αντιστοιχεί σε μία μεταβλητή, μπορούμε επίσης να σχετίσουμε σύνολα μεταβλητών και διανυσματικές μεταβλητές με ένα κόμβο. Είναι εύκολο να δείξουμε ότι η απεικόνιση της δεξιάς πλευράς της (5) είναι πάντα σωστά κανονικοποιημένη, δεδομένου ότι κάθε δεσμευμένη κατανομή είναι κανονικοποιημένη.

Τα κατευθυνόμενα γραφήματα που θεωρούμε, υπόκεινται σε έναν περιορισμό, δεν πρέπει

να σχηματίζουν κατευθυνόμενους κύκλους, με άλλα λόγια δεν υπάρχουν κλειστά μονοπάτια μέσα σε ένα γράφημα που να μπορούμε να μεταφερόμαστε από κόμβο σε κόμβο και να καταλήξουμε στον αρχικό. Αυτά τα γραφήματα λέγονται και ακυκλικά κατευθυνόμενα διανύσματα.

2.5.3 Παράδειγμα: Πολυωνυμική Παλινδρόμηση

Σαν ένα παράδειγμα κατευθυνόμενων γράφων, μπορούμε να σκεφτούμε ένα μοντέλο Bayesian πολυωνυμικής παλινδρόμησης της Εικόνας 2.3. Οι τυχαίες μεταβλητές αυτού του μοντέλου είναι ένα διάνυσμα από πολυωνυμικές μεταβλητές **w** από τα παρατηρούμενα δεδομένα $\mathbf{t}=(t_1,.....,t_N)^T$.



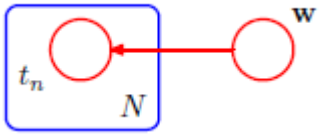
Εικόνα 2.3: Κατευθυνόμενο γραφικό μοντέλο που απεικονίζει στην από κοινού κατανομή της (6) που αντιστοιχεί σε ένα Bayesian μοντέλο πολυωνιμικής παλινδρόμησης.

Επιπλέον, αυτό το μοντέλο περιέχει τα δεδομένα εισαγωγής $\mathbf{x} = (x_1,.....x_N)^T$, τον θόρυβο διασποράς σ^2, και την υπερπαράμετρο pa που απεικονίζει την ακρίβεια της εκ των προτέρων Gaussian κατανομής πάνω στην **w**, όλες αυτές είναι παράμετροι του μοντέλου παρά τυχαίες μεταβλητές. Συγκεντρώνοντας την προσοχή μας στις τυχαίες μεταβλητές, βλέπουμε ότι η από κοινού κατανομή δίνεται από την παράγωγο της εκ των προτέρων $p(w)$ και των N δεσμευμένων κατανομών $p(t_n|w)$ για $n = 1, \ldots, N$ έτσι ώστε

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}). \qquad (6)$$

Αυτή λοιπόν η κατανομή απεικονίζεται στο μοντέλο της Εικόνας 2.3.

Για τα πιο πολύπλοκα όμως μοντέλα είναι άβολο να σχεδιάζουμε πολλούς κόμβους $t_1, \ldots, t_N$ ξεχωριστά όπως στην Εικόνα 2.3. Για αυτό εισαγάγουμε μια καινούρια ορολογία γραφημάτων που επιτρέπει τέτοιου είδους κόμβοι να απεικονίζονται πιο συνοχικά. Σχεδιάζουμε λοιπόν έναν κόμβο t_n και μετά τον περικυκλώνουμε με ένα τετράγωνο που περιέχει έναν δείκτη N, ο οποίος υποδηλώνει τον αριθμό των κόμβων. Ξανασχεδιάζοντας το μοντέλο της Εικόνας 2.3 με αυτόν τον τρόπο, παίρνουμε το γράφημα της Εικόνας 2.4.

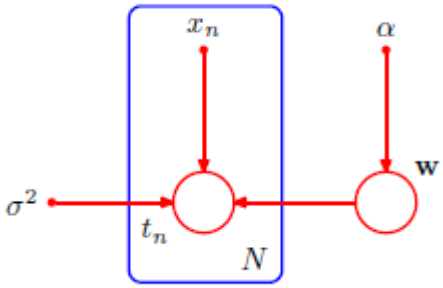


Εικόνα 2.4: Μια εναλλακτική, πιο συμπαγής, απεικόνιση του γραφήματος που δίνεται στην Εικόνα 2.3, όπου έχουμε εισαγάγει το πλαίσιο που απεικονίζει N κόμβους εκ των οποίων μόνο ένας, ο t_n, απεικονίζεται.

Θα βρούμε κάποιες φορές χρήσιμο να κάνουμε τις παραμέτρους ενός μοντέλου, καθώς και τις στοχαστικές του μεταβλητές, σαφέστερες. Σε αυτή την περίπτωση η (6) γίνεται

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$

Συνεπώς, μπορούμε να κάνουμε τα **x** και α πιο σαφή στην γραφική απεικόνιση. Για να το κάνουμε αυτό θα υιοθετήσουμε ένας συμβολισμό όπου οι τυχαίες μεταβλητές θα απεικονίζονται ως ανοιχτοί κύκλοι, και οι ντετερμινιστικές παράμετροι θα ορίζονται από μικρότερους συμπαγείς κύκλους. Αν πάρουμε την εικόνα 2.4, συμπεριλαμβανομένων και των ντετερμινιστικών παραμέτρων, παίρνουμε το γράφημα στην εικόνα 2.5.



Εικόνα 2.5: Εδώ απεικονίζεται το ίδιο μοντέλο της Εικόνα 2.4 αλλά οι νετερμινιστικοί παράγοντες απεικονίζονται αναλυτικότερα από μικρότερους συμπαγείς κόμβους

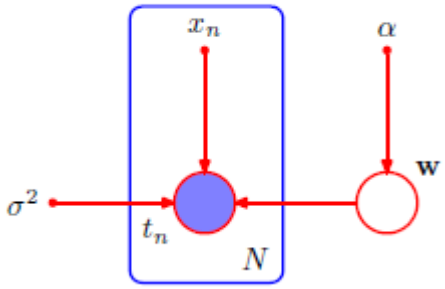
Όταν εφαρμόζουμε ένα γραφικό μοντέλο σε ένα πρόβλημα σημ μηχανική εκμάθηση ή στην αναγνώριση σχημάτων, θα θέσουμε σε ένα σύνολο από μερικές τυχαίες μεταβλητές, συγκεκριμένες γνωστές τιμές. Σε ένα γραφικό μοντέλο, θα ορίζουμε τις γνωστές μεταβλητές σκιάζοντας τους αντίστοιχους κόμβους. Έτσι το αντίστοιχο γράφημα της Εικόνας 2.5 στο οποίο οι μεταβλητές $\{t_n\}$ είναι γνωστές δίνεται στην Εικόνα 2.6. Σημειώστε ότι τιμή του **w** δεν είναι γνωστή, και έτσι το **w** είναι ένα παράδειγμα κρυφής μεταβλητής. Τέτοιες μεταβλητές παίζουν έναν κρίσιμο ρόλο στα πιθανοτικά μοντέλα.

Γνωρίζοντας τις τιμές $\{t_n\}$ μπορούμε, αν επιθυμούμε, να υπολογίσουμε την εκ των υστέρων κατανομή των πολυωνυμικών συντελεστών **w**. Αυτό αποτελεί μια άμεση εφαρμογή του θεωρήματος Bayes.

$$p(\mathbf{w}|\mathbf{T}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

Σε γενικές γραμμές, οι παράμετροι του μοντέλου όπως η **w** μεταβλητή είναι μικρού ενδιαφέροντος, επειδή ο τελικός μας στόχος είναι να κάνουμε προβλέψεις για νέες τιμές εισόδου. Υποθέστε ότι μας έχει δοθεί μια νέα τιμή εισόδου $\hat{x}$ και επιθυμούμε να βρούμε την αντίστοιχη πιθανοτική κατανομή για την $\hat{t}$ που είναι δεσμευμένη ως προς τα γνωστά μας δεδομένα. Το γραφικό μοντέλο που περιγράφει το συγκεκριμένο πρόβλημα δίνεται στην Εικόνα 2.7 και η αντίστοιχη από κοινού κατανομή για όλες τις τυχαίες μεταβλητές σε αυτό το μοντέλο, δεσμευμένη ως προς τις ντετερμινιστικές παραμέτρους, δίνεται από την

$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2). \tag{7}$$

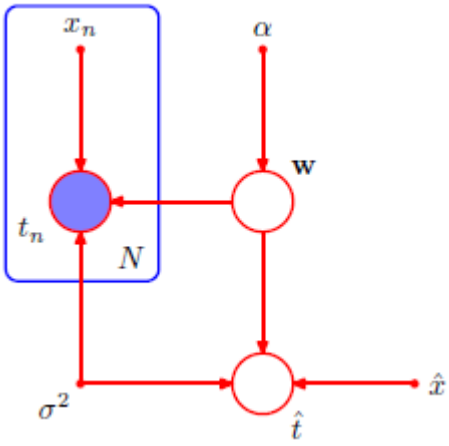


Εικόνα 2.6: Όπως και στην Εικόνα 2.5 μονό που εδώ οι κόμβοι $\{t_n\}$ έχουν σκιαστεί για να δείξουν τις αντίστοιχες τυχαίες μεταβλητές που τεθεί σύμφωνα με τις γνωστές τιμές.

Η απαιτούμενη κατανομή πρόβλεψης για το $\hat{t}$ υπολογίζεται , από το κανόνα αθροίσματος πιθανοτήτων, ολοκληρώνοντας ως προς τις παραμέτρους του μοντέλου

$$p(\widehat{t}|\widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2)\, \mathrm{d}\mathbf{w} \tag{8}$$

όπου θέτουμε τις τυχαίες μεταβλητές του t στις συγκεκριμένες τιμές που παρατηρήθηκαν στο σύνολο δεδομένων.

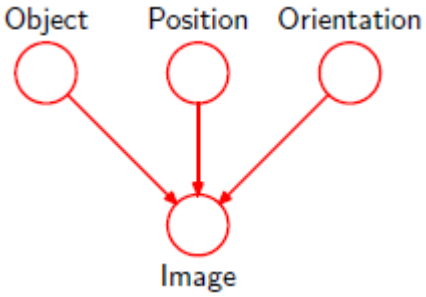


Εικόνα 2.7: Το μοντέλο πολυωνιμικής παλινδρόμησης, που αντιστοιχεί στην Εικόνα 2.6, δείχνει επίσης μία νέα μεταβλητή εισόδου μαζί με την αντίστοιχη πρόβλεψη μοντέλου.

2.5.4 Γενετικά Μοντέλα

Υπάρχουν πολλές περιπτώσεις όπου επιθυμούμε να επιλέξουμε δείγματα από μία δεδομένη πιθανοτική κατανομή. Σε αυτό το σημείο είναι καλό να περιγράψουμε περιληπτικά μια τεχνική που λέγεται ancestral sampling, όπου είναι επιμερώς σχετική με τα γραφικά μοντέλα. Σκεφτείτε μία από κοινού κατανομή $p(x_1,....,x_K)$ σε K μεταβλητές που παρογοντοποιείται σύμφωνα με την (5) απεικονισμένη σε ένα κατευθυνόμενο ακυκλικό γράφημα. Θα υποθέσουμε ότι οι μεταβλητές έχουν ταξινομηθεί έτσι που να μη υπάρχουν συνδέσεις από κόμβο σε έναν άλλο χαμηλότερου βαθμού, με άλλα λόγια κάθε κόμβος έχει μεγαλύτερο βαθμό από τους γονείς του. Στόχος μας είναι να επιλέξουμε ένα δείγμα $\hat{x_1}, ..., \hat{x_K}$ από την από κοινού κατανομή.

Για να το κάνουμε αυτό αρχίζουμε από τον κόμβο χαμηλότερου βαθμού και σχεδιάζουμε ένα δείγμα από την κατανομή $p(x_1)$, και το καλούμε $\hat{x_1}$. Στην συνέχεια δουλεύουμε για κάθε κόμβο, έτσι ώστε για τον κόμβο n να επιλέξουμε ένα δείγμα από την κατανομή $p(x_n|p_{an})$ στην οποία οι γονείς μεταβλητές έχουν τεθεί στις επιλεγμένες τους τιμές. Σημειώστε ότι σε κάθε επίπεδο, αυτές οι τιμές των γονέων θα είναι πάντα διαθέσιμες γιατί αντιστοιχούν σε χαμηλότερου βαθμού κόμβους που έχουν ήδη επιλεχθεί οι τιμές τους. Όταν πλέον έχουμε επιλέξει και την τελική μεταβλητή x_k, θα πετύχουμε τον στόχο μας λαμβάνοντας δείγμα από την από κοινού κατανομή. Για να πάρουμε δείγμα από μερικές περιθωριακές κατανομές αντίστοιχες σε ένα υποσύνολο των μεταβλητών, παίρνουμε απλά τις ήδη επιλεγμένες τιμές από τους κόμβους που απαιτείται και αγνοούμε τις επιλεγμένες τιμές των κόμβων που έχουν απομείνει. Για παράδειγμα, για να επιλέξουμε ένα δείγμα από την κατανομή $p(x_1,x_4)$, επιλέγουμε απλά τιμές από την πλήρη από κοινού κατανομή και μετά κρατάμε τις τιμές $\hat{x_2}, \hat{x_4}$ και αφαιρούμε τις τιμές που απέμειναν.



Εικόνα 2.8: Ένα γραφικό μοντέλο που απεικονίζει την διαδικασία με την οποία οι εικόνες των αντικειμένων δημιουργούνται, και στο οποίο η ταυτότητα ενός αντικειμένου (μια διακριτή μεταβλητή) και η θέση και ο προσανατολισμός του αντικειμένου (συνεχόμενες μεταβλητές) έχουν ανεξάρτητες εκ των προτέρων μεταβλητές. Η εικόνα έχει πιθανοτική κατανομή που είναι εξαρτημένη από την ταυτότητα του αντικειμένου όπως επίσης και από τη θέση και τον προσανατολισμό του.

Για πρακτικές εφαρμογές των πιθανοτικών μοντέλων, είναι τυπικό οι υψηλότερου βαθμού μεταβλητές να αντιστοιχούν σε τερματικούς κόμβους του γραφήματος που απεικονίζουν παρατηρήσεις, και οι χαμηλότερου βαθμού κόμβοι να αντιστοιχούν σε κρυφές μεταβλητές. Ο κύριος ρόλος των κρυφών μεταβλητών είναι να επιτρέπουν σε μια πολύπλοκη κατανομή ως προς τις γνωστές μεταβλητές να απεικονίζεται με όρους ενός μοντέλου κατασκευασμένου από απλούστερες δεσμευμένες κατανομές.

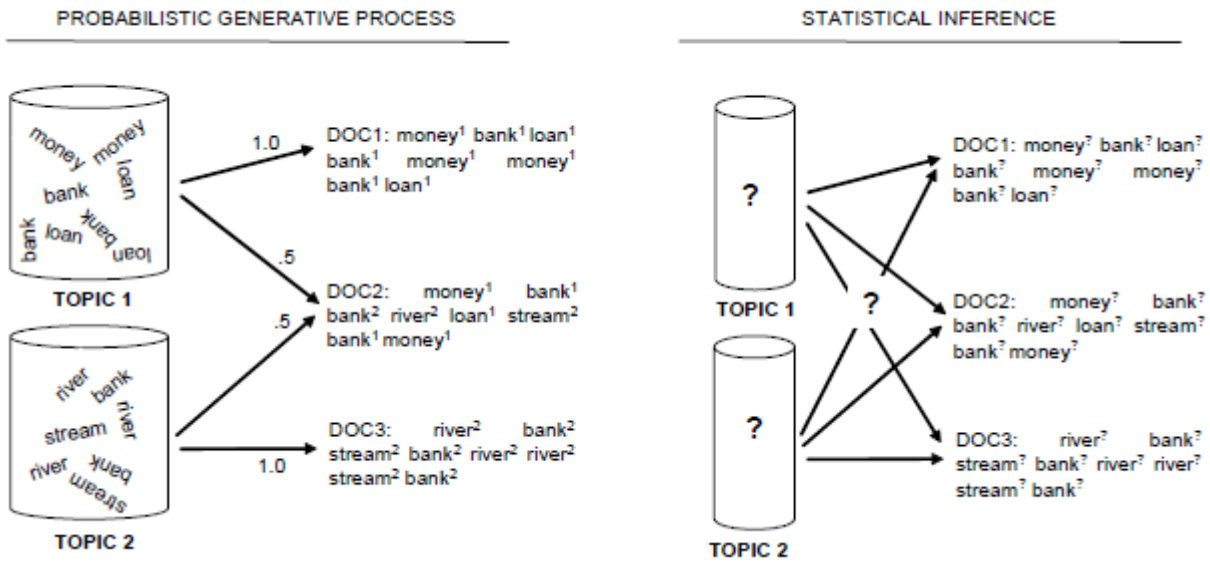
Μπορούμε να ερμηνεύσουμε τέτοια μοντέλα εκφράζοντας τις διαδικασίες με τις οποίες τα γνωστά δεδομένο αναδύονται. Για παράδειγμα, σκεφτείτε μια εργασία αναγνώρισης αντικείμενου στην οποία ένα γνωστό σημείο δεδομένων αντιστοιχεί σε μια εικόνα για ένα από τα αντικείμενα. Δεδομένης μίας γνωστής εικόνας, στόχος μας είναι να βρούμε την εκ των υστέρων κατανομή ως προς τα αντικείμενα, ως προς τα οποία θα ολοκληρώσουμε πάνω σε όλες τις πιθανές θέσεις και προσανατολισμούς. Μπορούμε να αναπαραστήσουμε το πρόβλημα χρησιμοποιώντας το σχήμα της Εικόνας 2.8.

Το γραφικό μοντέλο δείχνει την *αιτιολογική διαδικασία* (Pearl, 1988) με την οποία τα γνωστά δεδομένα δημιουργούνται. Για αυτό τον λόγο, τέτοια μοντέλα λέγονται συχνά *γενετικά μοντέλα*. Αντίθετα το μοντέλο πολυωνυμικής παλινδρόμησης που περιγράφεται από την Εικόνα 2.5 δεν είναι γενετικό γιατί δεν υπάρχει πιθανοτική κατανομή που να σχετίζεται με μια μεταβλητή εισόδου x, και έτσι δεν είναι πιθανό να δημιουργήσουμε συνθετικά σημεία δεδομένων από αυτό το μοντέλο. Θα μπορούσαμε να το κάνουμε γενετικό εισάγοντας μια κατάλληλη εκ των προτέρων κατανομή $p(x)$, κατασκευάζοντας έτσι ένα πιο πολύπλοκο μοντέλο.

Οι κρυφές μεταβλητές σε ένα πιθανοτικό μοντέλο δεν χρειάζονται, όμως, καμία φυσική ερμηνεία, απλά εισάγονται για να επιτρέψουν σε μια πολύπλοκη από κοινού κατανομή να κατασκευαστεί από απλούστερα συστατικά. Σε κάθε περίπτωση, η τεχνική του ancestral sampling που εφαρμόζεται σε ένα γενετικό μοντέλο μιμείται την δημιουργία των γνωστών δεδομένων και έτσι δίνει "φανταστικά" δεδομένα όπου η πιθανοτική τους κατανομή είναι ίδια με τα γνωστά δεδομένα. Πρακτικά, η δημιουργία συνθετικών παρατηρήσεων από ένα γενετικό μοντέλο μπορεί να αποδειχθεί χρήσιμη στην κατανόηση του σχήματος της πιθανοτικής κατανομής που απεικονίζεται στο μοντέλο.

2.5.5 Γενετικά μοντέλα κειμένων

Ένα γενετικό μοντέλο κειμένων βασίζεται στους απλούς πιθανοτικούς κανόνες επιλογής (sampling) που περιγράφουν πως οι λέξεις μέσα στα κείμενα μπορεί να δημιουργηθούν μέσω τυχαίων κρυφών μεταβλητών. Όταν προσαρμόζουμε ένα γενετικό μοντέλο, ο στόχος είναι να βρεθεί το καταλληλότερο σύνολο κρυφών μεταβλητών που μπορεί να εξηγήσει τα παρατηρούμενα δεδομένα υποθέτοντας ότι το μοντέλο δημιουργήθηκε από τα δεδομένα. Η Εικόνα 2.9 δείχνει την προσέγγιση μοντέλων θεμάτων με δύο τρόπους: σαν γενετικό μοντέλο και σαν πρόβλημα μεταβολικής συμπερασματολογίας. Στα αριστερά, η γενετική διαδικασία παρουσιάζεται με δύο θέματα (topics) . Τα Topic 1 και 2 είναι θεματικά συσχετισμένα με το 'money' και τα 'rivers' και απεικονίζονται σαν τσάντες που περιέχουν διαφορετικές κατανομές ως προς τις λέξεις. Διαφορετικά κείμενα μπορούν να παραχθούν επιλέγοντας λέξεις από ένα θέμα δεδομένου του βάρους που έχει. Για παράδειγμα , τα κείμενα Doc1 και Doc3 έχουν δημιουργηθεί με την διαδικασία sampling μόνο από το topic 1 και το topic 2 αντίστοιχα, ενώ το κείμενο Doc2 δημιουργήθηκε από την ίση μείξη των δύο θεμάτων. Με τον τρόπο όπου το μοντέλο ορίστηκε, δεν υπάρχει κάποια αμοιβαία αποκλειστικότητα που να περιορίζει τις λέξεις να είναι μέρος ενός μόνο θέματος. Αυτό επιτρέπει στο μοντέλο να καταλαβαίνει την πολυσημία, όπου η ίδια η λέξη μπορεί να έχει πολλές σημασίες ταυτόχρονα. Για παράδειγμα, και τα δύο θέματα 'money' και ΄rivers' δίνουν υψηλή πιθανότητα στην λέξη BANK, που είναι λογικό δεδομένης της πολυσημικής φύσης της λέξης.



Εικόνα 2.9: Απεικόνιση γενετικής διαδικασίας και του προβλήματος στατιστικής συμπερασματολογίας που λύνεται με μοντέλα θεμάτων.

Η γενετική διαδικασία εδώ δεν κάνει κάποιες υποθέσεις για την σειρά των λέξεων μέσα στα κείμενα. Η μόνη πληροφορία σχετική με το μοντέλο είναι ο αριθμός των φορών που μια λέξη έχει δημιουργηθεί. Αυτό είναι γνωστό ως η υπόθεση “της τσάντας με τις λέξεις (bag-of-words)”, και είναι συνήθης σε πολλά στατιστικά μοντέλα γλώσσας συμπεριλαμβανομένου και του LDA(Latent Dirischlet Allocation) που θα αναλύσουμε στο επόμενο κεφάλαιο. Φυσικά η πληροφορία για την σειρά των λέξεων μπορεί να περιέχει νύξεις του περιεχομένου ενός κειμένου, αυτή η πληροφορία δεν υλοποιείται από το

μοντέλο. Το δεξί πάνελ την Εικόνας 2.9 απεικονίζει το πρόβλημα μεταβολικής συμπερασματολογίας. Δεδομένου των παρατηρούμενων λέξεων σε ένα σύνολο από κείμενα, θα θέλαμε να ξέρουμε ποιο μοντέλο θεμάτων είναι πιο πιθανό να έχει δημιουργήσει τα δεδομένα. Αυτό συμπεριλαμβάνει την διεξαγωγή συμπερασματολογίας της πιθανοτικής κατανομής στις λέξεις που συσχετίζονται με κάθε θέμα, της κατανομής των θεμάτων για κάθε κείμενο και, συχνά, την κατανομή του θέματος που είναι υπεύθυνο για την δημιουργία κάθε λέξης.

Κεφάλαιο 3
Ο αλγόριθμος LDA
και οι παραλλαγές του

3.1 Latent Dirichlet Allocation(LDA)

Περιγράφουμε τον αλγόριθμο Latent Dirichlet Allocation (LDA) ως ένα γενετικό πιθανοτικό μοντέλο διακεκριμένων δεδομένων όπως μία συλλογή κειμένων. Το LDA είναι ένα ιεραρχικό πιθανοτικό μοντέλο τριών επιπέδων, στο οποίο κάθε αντικείμενο της συλλογής μοντελοποιείται ως ένα πεπερασμένο μείγμα από ένα σύνολο θεματικών πιθανοτήτων.

3.1.1 Εισαγωγή

Σε αυτό το κεφάλαιο θεωρούμε το πρόβλημα της μοντελοποίησης συλλογών κειμένων και άλλων συλλογών διακεκριμένων δεδομένων. Στόχος μας είναι να βρούμε σύντομες περιγραφές των μελών της συλλογής που επιτρέπουν μια αποτελεσματική επεξεργασία μεγάλων συλλογών διατηρώντας τις απαραίτητες στατιστικές σχέσεις που είναι χρήσιμες για βασικές διεργασίες όπως αρχικοποίηση, ο εντοπισμός νεωτερισμών, περίληψη κειμένου, ομοιότητα και συσχετισμός κρίσεων.

Σημαντική επεξεργασία έχει γίνει πάνω σε αυτό το πρόβλημα από ερευνητές στο αντικείμενο του πεδίου ανάκτησης πληροφοριών(IR)(Baeza-Yates και Ribeiro-Neto, 1999). Η βασική μεθοδολογία που προτάθηκε από τους IR ερευνητές για συλλογές κειμένων – μια μεθοδολογία η οποία εφαρμόστηκε με επιτυχία στις μοντέρνες μηχανές αναζήτησης του Διαδικτύου - μετατρέπει κάθε κείμενο σε ένα διάνυσμα πραγματικών αριθμών, καθένα από τα οποία αντιπροσωπεύει μία αναλογία μετρήσεων. Στο δημοφιλές tf-idf σχήμα (Salton και McGill, 1983), το βασικό λεξιλόγιο των “λέξεων” και των “όρων” επιλέγεται, και, για κάθε κείμενο της συλλογής, μία μέτρηση σχηματίζεται από το αριθμό των φορών που έχει παρουσιαστεί μια λέξη. Μετά από κατάλληλη κανονικοποίηση, ο δείκτης συχνότητας συγκρίνεται με το αντίστροφο δείκτη συχνότητας κειμένων, που μετράει τον αριθμό που έχει βρεθεί μια λέξη σε όλη την συλλογή (γενικά σε λογαριθμική κλίμακα και μετά μοντελοποιείται κατάλληλα). Το τελικό αποτέλεσμα είναι ένας πίνακας όρων ανά κειμένων X του οποίου οι στήλες περιέχουν τις td-idf αξίες για κάθε κείμενο της. Έτσι το td-idf σχήμα μειώνει τα κείμενα αυθαίρετου μήκους σε φορμαρισμένου μήκους λίστες αριθμών.

Καθώς η td-idf μείωση έχει κάποιες δελεαστικές ιδιότητες η προσέγγιση παρέχει επίσης μια σχετικά μικρή μείωση του μήκους περιγραφής και φανερώνει λίγα για την στατιστική μορφή των κειμένων. Για την αντιμετώπιση των παραπάνω προβλημάτων, οι IR ερευνητές πρότειναν πολλές άλλες τεχνικές μείωσης του πλήθους διαστάσεων, η πιο αξιοσημείωτη εκ των οποίων είναι η Latent Semantic Indexing (LSI) (Deerwester et al., 1990)). Η LSI εφαρμόζει αποσύνθεση τιμών στον X πίνακα για να ορίσει έναν γραμμικό υποχώρο στον χώρο των tf-idf ιδιοτήτων, που πιάνει ένα μεγάλο μέρος της διακύμανσης της συλλογής. Αυτή η προσέγγιση μπορεί να πετύχει μια σημαντική ελαχιστοποίηση μεγάλων συλλογών. Επιπλέον, ο Deerwester υποστηρίζει ότι τα παραγόμενα χαρακτηριστικά του LSI , που είναι γραμμικοί συνδυασμοί των αυθεντικών tf-idf χαρακτηριστικών, μπορούν να πιάσουν κάποιες πτυχές των βασικών γλωσσολογικών ιδεών όπως η συνωνυμία και η πολυσημία.

Ένα σημαντικό βήμα έκανε ο Hoffman, που παρουσίασε το πιθανοτικό LSI (pLSI) μοντέλο, γνωστό επίσης και σαν μοντέλο συμπερασμάτων. Η pLSI προσέγγιση μοντελοποιεί κάθε λέξη ενός κειμένου ως ένα δείγμα ενός mixture μοντέλου, όπου τα αναμεμιγμένα στοιχεία είναι τυχαίες πολυωνυμικές μεταβλητές που μπορούν να κατανοηθούν ως παρουσιάσεις θεμάτων. Έτσι κάθε λέξη δημιουργείται από ένα απλό θέμα, και διαφορετικές λέξεις σε ένα κείμενο μπορεί να δημιουργηθούν από διαφορετικά θέματα. Κάθε κείμενο απεικονίζεται ως μία λίστα αναμεμειγμένων αναλογιών για κάθε mixture συστατικό και έτσι μειώνεται σε μια πιθανοτική κατανομή φορμαρισμένων συνόλων θεμάτων. Αυτή η κατανομή είναι μια "ελαχιστοποιημένη περιγραφή" συσχετισμένη με ένα κείμενο

Αν και η δουλειά του Hoffman είναι ένα χρήσιμο βήμα πάνω στην θεματική μοντελοποίηση κειμένων δεν είναι πλήρης, γιατί δεν παρέχει κανένα πιθανοτικό μοντέλο για την επεξεργασία σε επίπεδο κειμένων. Στο pLSI, κάθε κείμενο παρουσιάζεται ως λίστα αριθμών και δεν υπάρχει κανένα γενετικό πιθανοτικό μοντέλο για αυτούς τους αριθμούς. Αυτό επιφέρει αρκετά προβλήματα: (1) ο αριθμός των παραμέτρων στο μοντέλο μεγαλώνει γραμμικά με το μέγεθος της συλλογής, κάτι που οδηγεί σε σοβαρά προβλήματα σχετικά με το overfitting και (2) δεν είναι ξεκάθαρο πως διανέμεται η κατανομή σε ένα κείμενο εκτός του εκπαιδευτικού συνόλου.

Για να καταλάβετε πως θα προχωρήσουμε πέρα από το LSI, ας θεωρήσουμε τις βασικές πιθανοτικές υποθέσεις υπογραμμίζοντας την κλάση των μεθόδων μείωσης διαστάσεων που περιέχει το LSI και το pLSI. Όλες αυτές οι μέθοδοι βασίζονται στην 'bag-of-words' υπόθεση – ότι η σειρά των λέξεων μπορεί να αγνοηθεί. Στην γλώσσα της πιθανοτικής θεωρίας, αυτή είναι η υπόθεση της ανταλλαξιμότητας των λέξεων σε ένα κείμενο. Αυτές οι μέθοδοι υποθέτουν επίσης ότι τα κείμενα είναι ανταλλάξιμα: η σειρά των κείμενων σε μια συλλογή μπορεί να αλλάξει.

Ένα κλασσικό θεώρημα του de Finetti (1990) τεκμηριώνει ότι κάθε συλλογή από ανταλλάξιμες τυχαίες μεταβλητές έχει μια απεικόνιση σαν mixture κατανομή. Έτσι, αν επιθυμούμε να θεωρήσουμε ανταλλάξιμες απεικονίσεις κειμένων για κείμενα και λέξεις, πρέπει να θεωρήσουμε mixture μοντέλα που λαμβάνουν υπόψη τους την ανταλλαξιμότητα και των λέξεων αλλά και των κειμένων. Αυτό οδήγησε στην δημιουργία του μοντέλου Latent Dirichlet Allocation (LDA) που θα παρουσιάσουμε στο παρών κεφάλαιο.

Είναι σημαντικό να αναφέρουμε ότι η υπόθεση την ανταλλαξιμότητας είναι ίδια με την υπόθεση ότι οι τυχαίες μεταβλητές είναι ανεξάρτητες και ιδανικά κατανεμημένες. Τέλος πρέπει να ειπωθεί ότι υπάρχει ένας μεγάλος αριθμός γενικεύσεων της βασικής ιδέας της ανταλλαξιμότητας, συμπεριλαμβανομένου ποικίλων μορφών μερικής ανταλλαξιμότητας, και ότι τα θεωρήματα είναι διαθέσιμα σε κάθε περίπτωση(Diaconis 1988). Έτσι, ενώ εμείς θα συγκεντρώσουμε όλη μας την προσοχή σε απλά 'bag-of-words' μοντέλα, που οδηγούν σε αναμεμειγμένες κατανομές απλών λέξεων, οι μέθοδοι μας είναι επίσης εφαρμόσιμες και σε πιο πολύπλοκα μοντέλα που περιλαμβάνουν μεγαλύτερες δομές όπως τα n-διαγράμματα και παραγράφους.

3.1.2 Ορολογία

Θα χρησιμοποιήσουμε την γλώσσα των συλλογών κειμένων σε όλη την εργασία, αναφερόμενοι σε οντότητες όπως "λέξεις", "κείμενα" και "σώματα". Αυτό είναι χρήσιμο σαν οδηγός, κυρίως όταν εισαγάγουμε κρυφές μεταβλητές που στοχεύουν στην κατανόηση αφηρημένων εννοιών όπως τα "θέματα". Είναι σημαντικό να σημειωθεί ότι το LDA δεν είναι συνδεδεμένο μόνο με ανάλυση κειμένων αλλά έχει και άλλες εφαρμογές όπως προβλήματα που συμπεριλαμβάνουν συλλογές δεδομένων όπως τα collaborative filtering, ανάκτηση εικόνων βασισμένη σε κείμενο και βιοπληροφορική.

Επίσημα ορίζουμε τους παρακάτω όρους:

•Η λέξη είναι η βασική μονάδα των διακριτών δεδομένων, η οποία ορίζεται σαν ένα αντικείμενο από ένα λεξιλόγιο {1,....,V}. Απεικονίζουμε τις λέξεις ως διανύσματα οπού έχουν ένα και μόνο στοιχείο ίσο με 1 και όλα τα υπόλοιπα ίσα με 0. Έτσι η $v\eta$ λέξη απεικονίζεται στο λεξιλόγιο ως ενα V-διάνυσμα **w** όπου $w^v=1$ και $w^u=0$ για $u \neq v$.

•Ένα κείμενο είναι μια σειρά από *N* λέξεις ορισμένο ως $\mathbf{w}=(w_1,w_2,....,w_N)$, όπου το w_n είναι η n-οστή λέξη της σειράς.

•Ένα σώμα είναι μια συλλογή από *M* κείμενα ορισμένα ως $D=\{\mathbf{w_1},\mathbf{w_2},....\mathbf{w_M}\}$.

Επιθυμούμε να βρούμε ένα πιθανοτικό μοντέλο σε ένα σώμα που όχι μόνο δίνει υψηλή πιθανότητα στο σώμα, αλλά επίσης και σε αλλά "παρόμοια" κείμενα.

3.1.3 Latent Dirichlet Allocation

Το Latent Dirischlet allocation (LDA) είναι ένα γενετικό πιθανοτικό μοντέλο ενός σώματος. Η βασική ιδέα είναι ότι τα κείμενα αντιπροσωπεύονται από τυχαίες προσμείξεις κρυφών θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μία κατανομή ως προς τις λέξεις.

Το LDA υποθέτει την παρακάτω γενετική διαδικασία για κάθε κείμενο w σε ένα σώμα D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Έχουν γίνει πολλές υποθέσεις απλοποίησης στο βασικό μοντέλο. Πρώτον, η διάσταση k της Dirichlet κατανομής (και έτσι η διάσταση της μεταβλητής θέματος *z*) θεωρείται γνωστή και φιξαρισμένη. Δεύτερον, οι πιθανότητες λέξεων είναι παραμετροποιημένες από ένα $K \times V$ πίνακα β όπου $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, που προς το παρόν θα θεωρείται μια φιξαρισμένη ποσότητα που πρέπει να υπολογιστεί. Τελικώς, η υπόθεση Poisson δεν χρησιμοποιείται πουθενά και πιο ρεαλιστικές κατανομές μήκους κειμένου μπορούν να χρησιμοποιηθούν αν χρειαστεί. Επιπλέον, πρέπει να σημειωθεί ότι το N είναι ανεξάρτητο από όλα τα άλλα δεδομένα που δημιουργούν μεταβλητές (θ και *z*). Είναι έτσι μία βοηθητική

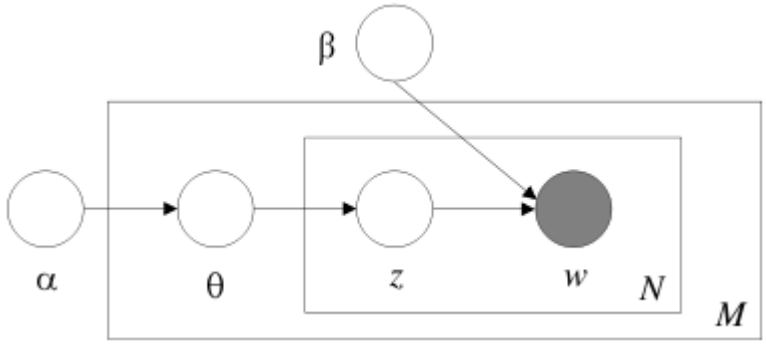
μεταβλητή. Γενικά θα αγνοήσουμε την τυχαιότητα στην παρακάτω ανάπτυξη.

Μία k-διάστασης Dirichlet τυχαία μεταβλητή u μπορεί να πάρει τιμές σε ένα (k-1)-simplex. Ένα k-διάνυσμα θ βρίσκεται στο (k-1)-simplex αν $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$ και την ακόλουθη πυκνότητα πιθανότητας πάνω στο συγκεκριμένο simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_{1.}^{\alpha_1 - 1} \ldots . \theta_k^{\alpha_k - 1} \quad (1)$$

όπου η παράμετρος α είναι ένα k-διάνυσμα με στοιχεία $\alpha_i > 1$, και όπου το Γ(x) είναι η Γάμμα συνάρτηση. Η Dirichlet είναι μια βολική κατανομή στο simplex – είναι μια εκθετική οικογένεια, και είναι συζευγμένη με την πολυωνυμική κατανομή. Δεδομένων των α και β, και την από κοινού κατανομή ενός θέματος μείξης θ, ένα σύνολο απο N θέματα **z**, και ένα σύνολο από N λέξεις **w** ισχύει:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta), \quad (2)$$



Εικόνα 3.1: Γραφικό μοντέλο απεικόνισης του LDA. Τα πλαίσια απεικονίζουν τις επαναλήψεις. Το εξωτερικό πλαίσιο απεικονίζει τα κείμενα ενώ το εσωτερικό την επαναληπτική επιλογή των θεμάτων και των λέξεων μέσα σε ένα κείμενο.

όπου το $p(z_n|\theta)$ είναι το θ_i για ένα μοναδικό i έτσι ώστε $z^i_n = 1$. Ολοκληρώνοντας ως προς θ και αθροίζοντας ως προς z, παίρνουμε την περιθωριακή πιθανότητα ενός κειμένου:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta. \quad (3)$$

Τέλος παίρνοντας το παράγωγο των περιθωριακών πιθανοτήτων των κειμένων, λαμβάνουμε την πιθανότητα του σώματος:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d.$$

Το LDA μοντέλο απεικονίζεται ως πιθανοτικό γραφικό μοντέλο στην εικόνα 3.1. Όπως κάνει η εικόνα κατανοητό, υπάρχουν τρία επίπεδα στο LDA. Οι παράμετροι α και β είναι επιπέδου σώματος παράμετροι, που έχουν συλλεχθεί από την διαδικασία δημιουργίας του σώματος. Οι μεταβλητές θ_d είναι επιπέδου κειμένου μεταβλητές, που έχουν λειφθεί από ένα κείμενο. Τέλος, οι μεταβλητές z_n και w_{dn} είναι επιπέδου λέξεων μεταβλητές και έχουν συλλεχθεί για κάθε λέξη κάθε κειμένου.

Είναι απαραίτητο να ξεχωρίσουμε το LDA από το απλό Dirichlet πολυωνυμικό μοντέλο ομαδοποίησης. Ένα κλασσικό μοντέλο ομαδοποίησης, θα ήταν ένα δύο επιπέδων μοντέλο όπου η Dirichlet θα υπολογιζόταν μία φόρα για ένα σώμα, μια πολυωνυμική μεταβλητή ομαδοποίησης επιλέγεται για κάθε κείμενο του σώματος και ένα σύνολο λέξεων επιλέγονται για ένα κείμενο της μεταβλητής ομάδας. Όπως και με πολλά αλλά μοντέλα ομαδοποίησης, ένα τέτοιο μοντέλο περιορίζει ένα κείμενο στο να συσχετίζεται μόνο με ένα θέμα. Αντιθέτως το LDA, είναι τρία επίπεδων μοντέλο, και ο θεματικός κόμβος επιλέγεται επαναλαμβανόμενα από κάθε κείμενο. Με αυτό το μοντέλο, τα κείμενα μπορούν να συσχετιστούν με πολλά θέματα.

3.1.4 Αλγόριθμος εύρεσης θεμάτων

Οι κύριες μεταβλητές που μας ενδιαφέρουν σε ένα μοντέλο είναι οι μεταβλητές θεμάτων-λέξεων φ και οι κατανομές θεμάτων θ για κάθε κείμενο. Ο Hoffman(1999) χρησιμοποίησε τον Expectation-Maximization(EM) αλγόριθμο για να υπολογίσει τα φ και θ. Αυτή η προσέγγιση παρουσιάζει προβλήματα εύρεσης των τοπικών μέγιστων στην συνάρτηση πιθανότητας, κάτι που παρακίνησε στην εύρεση καλύτερων αλγόριθμων. Αντί για τον κατευθείαν υπολογισμό των κατανομών θεμάτων-λέξεων φ και των θεματικών κατανομών θ για κάθε κείμενο, μια άλλη προσέγγιση είναι να υπολογίσουμε κατευθείαν την εκ των υστέρων κατανομή ως προς z, δεδομένου των παρατηρούμενων λέξεων **w,** καθώς περιθωριοποιούμε τα φ και θ. Κάθε z_i δίνει μια ακέραια τιμή [1....T] για κάθε θέμα που το δείγμα λέξης i έχει αντιστοιχηθεί. Επειδή πολλές συλλογές κειμένων περιέχουν εκατομμύρια τέτοιων δειγμάτων, ο υπολογισμός της εκ των υστέρων κατανομής ως προς το z απαιτεί αποτελεσματικές διαδικασίες υπολογισμού. Θα περιγράψουμε έναν αλγόριθμο που χρησιμοποιεί Gibbs sampling, μια μορφή Μαρκοβιανών αλυσίδων Monte Carlo(MCMC), που είναι εύκολο να εφαρμοστούν και παρέχει μια σχετικά εύκολη μέθοδο εύρεσης του συνόλου των θεμάτων για ένα μεγάλο σώμα. Μια Μαρκοβιανή αλυσίδα Monte Carlo αναφέρεται σε ένα σύνολο κατά προσέγγιση επαναληπτικών τεχνικών που έχουν σχεδιαστεί να συλλέγουν τιμές για πολύπλοκες κατανομές. Το Gibbs sampling, μια συγκεκριμένη μορφή των MCMC, προσομοιώνει μια πολυδιάστατη κατανομή συλλέγοντας υποσύνολα λιγότερων διαστάσεων από τις μεταβλητές, όπου κάθε υποσύνολο είναι δεσμευμένο ως προς τις τιμές των άλλων. Η διαδικασία γίνεται διαδοχικά και δεν σταματάει αν δεν υπολογίσει τις τιμές κατά προσέγγιση ίσες με την κατανομή-στόχο. Καθώς η Gibbs διαδικασία που θα περιγράψουμε, δεν παρέχει τους απευθείας υπολογισμούς των φ και θ, θα δείξουμε πως τα φ και θ μπορούν να υπολογιστούν χρησιμοποιώντας τους εκ των

υστέρων υπολογισμούς του *z*.

Ο Gibbs Sampling αλγόριθμος. Αναπαριστούμε την συλλογή κειμένων από ένα σύνολο δεικτών λέξεων w_i και δεικτών κειμένων d_i, για κάθε λέξη δείγμα *i*. Η Gibbs sampling διαδικασία θεωρεί γνωστή κάθε λέξη-δείγμα στην συλλογή κειμένων και υπολογίζει την πιθανότητα να πάρουμε την παρούσα λέξη από κάθε θέμα, δεδομένων των εναποθέσεων θεμάτων στις άλλες λέξεις-δείγματα. Γράφουμε αυτή την δεσμευμένη κατανομή ως $P(z_i=j| z_{-i}, w_i, d_i, .)$, όπου το $z_i=j$ αναφέρεται στην εναπόθεση του θέματος του δείγματος i στο θέμα *j*, *το* z_{-i} αναφέρεται στις εναποθέσεις θεμάτων των άλλων λέξεων-δειγμάτων, και το '.' αναφέρεται σε όλες τις άλλες γνωστές ή παρατηρηθέντες πληροφορίες όπως όλοι οι άλλοι δείκτες λέξεων και κειμένων w_{-i} και d_{-i} και οι υπερπαράμετροι α και β. Οι Griffiths και Steyvers (2004) έδειξαν πως αυτή η κατανομή μπορεί να υπολογιστεί:

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d,t} + T\alpha} \quad (4)$$

όπου C^{WT} και C^{DT} είναι πίνακες των μετρήσεων με διάσταση W x T και D x T αντίστοιχα: Το στοιχείο C^{WT}_{wj} είναι ο αριθμός των φορών που η λέξη *w* έχει εναποτεθεί στο θέμα *j*, μην συμπεριλαμβανομένης της παρούσας τιμής *i* και το C^{DT}_{dj} είναι ο αριθμός των φορών που το θέμα *j* έχει εναποτεθεί σε ένα δείγμα λέξης στο δείγμα d, μη συμπεριλαμβάνοντας την παρούσα τιμή *i*. Σημειώστε ότι η εξίσωση (4) δίνει την μη κανονικοποιημένη κατανομή. Η πραγματική κατανομή εναπόθεσης μιας λέξης-δείγματος σε ένα θέμα *j* υπολογίζεται διαιρώντας την ποσότητα της εξίσωσης (4) για το θέμα *t* με το άθροισμα όλων των θεμάτων.

Οι παράγοντες που επηρεάζουν τις εναποθέσεις θεμάτων για μία συγκεκριμένη λέξη-δείγμα μπορούν να γίνουν αντιληπτοί εξετάζοντας τα δύο μέρη της εξίσωσης (4). Το αριστερό κομμάτι είναι η πιθανότητα την λέξης *w* στο θέμα *j* και το δεξί κομμάτι είναι η πιθανότητα του θέματος *j* στη παρούσα κατανομή θέματος του κειμένου *d*. Καθώς θα έχουν εναποτεθεί πολλά δείγματα μια λέξης σε ένα θέμα *j*, θα αυξηθεί η πιθανότητα να εναποτεθεί ένα συγκεκριμένο δείγμα της λέξης στο θέμα *j*. Τη ίδια στιγμή, αν το θέμα *j* έχει χρησιμοποιηθεί πολλές φορές σε ένα κείμενο, θα αυξηθεί η πιθανότητα κάθε λέξη του κειμένου να εναποτεθεί στο θέμα *j*. Ως εκ τούτου, η εναπόθεση των λέξεων στα θέματα εξαρτάται από το πόσο είναι πιθανό να ανήκει η λέξη στο θέμα, καθώς επίσης και από το πόσο κυρίαρχο είναι ένα θέμα σε ένα κείμενο.

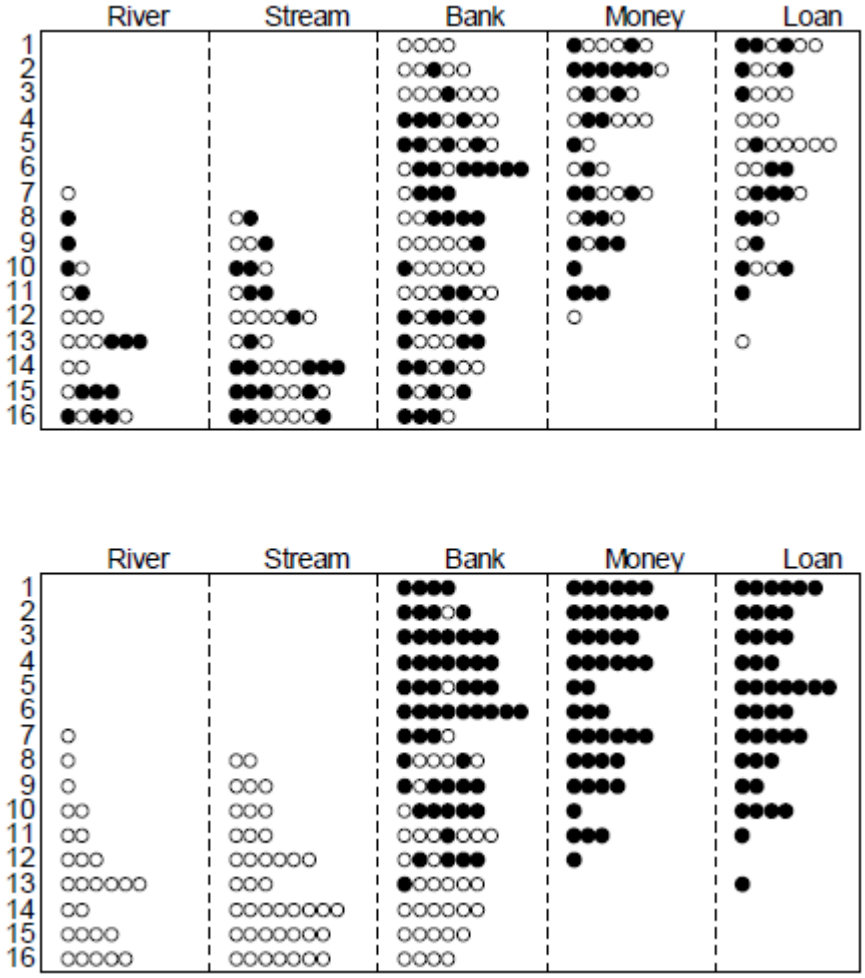
Ο Gibbs sampling αλγόριθμος αρχίζει εναποθέτοντας τις λέξεις-δείγματα σε ένα τυχαίο θέμα [1....T]. Για κάθε λέξη-δείγμα οι πίνακες C^{WT} και C^{DT} πρώτα μειώνονται κατά 1 για τις εισόδους που αφορούν την παρούσα εναπόθεση θέματος. Μετά, ένα νέο θέμα δημιουργείται από την κατανομή της εξίσωσης (4) και οι πίνακες C^{WT} και C^{DT} προσαυξάνονται με την νέα εναπόθεση θέματος. Κάθε Gibbs δείγμα αποτελείται από ένα σύνολο εναποθέσεων θεμάτων σε όλες τις N λέξεις-δείγματα του σώματος. Καθ' όλη την αρχική φάση της sampling διαδικασίας(γνωστή επίσης και ως burning περίοδος), τα Gibbs δείγματα πρέπει να απορριφθούν γιατί δεν βρίσκονται κοντά στον υπολογισμό της εκ των υστέρων κατανομής. Μετά την burning περίοδο, τα επιτύχοντα Gibbs δείγματα αρχίζουν να υπολογίζουν την κατανομή-στόχο. Σε αυτό το σημείο, για να πάρουμε ένα αντιπροσωπευτικό σύνολο δειγμάτων από την κατανομή, ένας αριθμός από Gibbs δείγματα σώζονται ανά τακτά χρονικά διαστήματα, για να αποφευχθούν οι συσχετίσεις μεταξύ των δειγμάτων.

Υπολογίζοντας τα φ και θ. Ο sampling αλγόριθμος υπολογίζει απευθείας το *z* για κάθε λέξη. Όμως, πολλές εφαρμογές του μοντέλου απαιτούν τους υπολογισμούς του θ και του φ των κατανομών λέξεων θεμάτων και των κατανομών θεμάτων κειμένων αντίστοιχα. Αυτές μπορούν να ληφθούν από τους τύπους όπως ακολουθούν:

$$\phi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad \theta'^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \qquad (5)$$

Αυτές οι τιμές αντιστοιχούν στις προβλεπόμενες κατανομές του sampling μια λέξης-δείγματος *i* από το θέμα *j*, και του sampling ενός νέου δείγματος σε ένα κείμενο *d* από το θέμα *j*, και είναι επίσης οι εκ των υστέρων μέσοι αυτών τον ποσοτήτων δεσμευμένων σε ένα συγκεκριμένο δείγμα z.

Ένα παράδειγμα. Ο Gibbs sampling αλγόριθμος μπορεί να κατανοηθεί δημιουργώντας τεχνητά δεδομένα από γνωστά θεματικά μοντέλα και εφαρμόζοντας τον αλγόριθμο για να ελέγξουμε αν είναι δυνατό να επαληθεύσουμε την αυθεντική γενετική κατασκευή. Το παράδειγμα στην Εικόνα 2.9 του προηγούμενου κεφαλαίου είναι χαρακτηριστικό. Υποθέστε πως το topic 1 δίνει ίση πιθανότητα στις λέξεις MONEY, LOAN, και BANK, έτσι $\varphi_{MONEY}^{(1)}=\varphi_{LOAN}^{(1)}=\varphi_{BANK}^{(1)}=1/3$, καθώς το topic 2 δίνει ίση πιθανότητα στις λέξεις $\varphi_{RIVER}^{(2)}=\varphi_{STREAM}^{(2)}=\varphi_{BANK}^{(2)}=1/3$. Η Εικόνα 3.2 δείχνει πως 16 κείμενα μπορούν να δημιουργηθούν από μια αφηρημένη μείξη των δύο θεμάτων. Κάθε κύκλος αντιστοιχεί σε μια συγκεκριμένη λέξη-δείγμα και κάθε σειρά σε ένα κείμενο. Στην Εικόνα 3.2, το χρώμα των κύκλων υποδηλώνει τις εναποθέσεις θεμάτων(μαύρο= topic 1,άσπρο = topic 2). Στην αρχή του sampling, οι εναποθέσεις δεν δίνουν καμία δομή ακόμα, απλώς είναι τυχαίες. Το κάτω πάνελ δείχνει την δομή ύστερα από 64 επαναλήψεις. Βασισμένη σε αυτές τις εναποθέσεις, η εξίσωση 4 δίνει τα ακόλουθα αποτελέσματα $\varphi_{MONEY}'^{(1)}=0.32$ $\varphi_{LOAN}'^{(1)}=0.29$ $\varphi_{BANK}'^{(1)}=0.39$ και $\varphi_{RIVER}'^{(2)} = 0.25$ $\varphi_{STREAM}'^{(2)}=0.4$ $\varphi_{BANK}'^{(2)}=0.35$. Δεδομένου του μεγέθους του συνόλου δεδομένων, αυτά τα αποτελέσματα είναι λογικές επανακατασκευές των παραμέτρων που χρησιμοποιήθηκαν για να δημιουργηθούν τα δεδομένα.



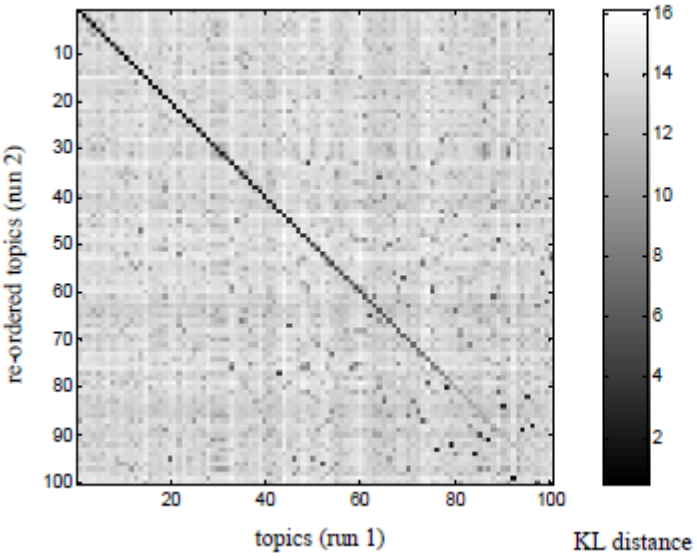
Εικόνα 3.2: Ένα παράδειγμα της Gibbs sampling διαδικασίας.

Ανταλλαξιμότητα των θεμάτων. Δεν υπάρχει καμία εκ των προτέρων σειρά θεμάτων που θα κάνει τα θέματα αναγνωρίσιμα ή ακόμα που θα τρέχει τον αλγόριθμο. Το θέμα *j* σε ένα Gibbs δείγμα δεν είναι θεωρητικά περιορισμένο σε ένα άλλο δείγμα ασχέτως αν τα δείγματα έρχονται από διαφορετικές Μαρκοβιανές αλυσίδες. Έτσι, σε διαφορετικά δείγματα δεν μπορεί να βρεθεί ο μέσος όρος σε επίπεδο θεμάτων. Όμως, όταν τα θέματα χρησιμοποιούνται για να υπολογίσουν κάτι στατιστικό που είναι αναλλοίωτο ως προς την σειρά των θεμάτων, το παραπάνω γίνεται δυνατό.

Σταθερότητα θεμάτων. Σε μερικές εφαρμογές, είναι επιθυμητό να συγκεντρωθούμε σε μια απλή λύση για να ερμηνεύσουμε κάθε θέμα. Σε αυτή την περίπτωση, είναι σημαντικό να ξέρουμε ότι τα θέματα είναι σταθερά και θα επανεμφανιστούν στα δείγματα και στα οποία τα θέματα έχουν μια ιδιοσυγκρασία για μια συγκεκριμένη λύση. Στην Εικόνα 3.3, απεικονίζεται η ανάλυση με την όποια δύο λύσεις μπορούν να ευθυγραμμιστούν μεταξύ των θεμάτων από διαφορετικές Μαρκοβιανές αλυσίδες. Η είσοδος του TASA σώματος είναι (W=26,414; D=37,651; N=5,628,867; T=100; α=50/T; β=0.1) και ένα Gibbs δείγμα επιλέχθηκε ύστερα από 2000 επαναλήψεις για δύο διαφορετικές αρχικές τυχαίες επαναλήψεις. Το αριστερό πάνελ δείχνει την ομοιότητα του πίνακα για τις δύο λύσεις. Η διαφορετικότητα των j_1 και j_2 μετρήθηκε από την συμμετρική Kullback-Liebler(KL) απόσταση μεταξύ των δύο θεματικών κατανομών:

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W} \phi'^{(j_1)}_k \log_2 \phi'^{(j_1)}_k / \phi''^{(j_2)}_k + \frac{1}{2}\sum_{k=1}^{W} \phi''^{(j_2)}_k \log_2 \phi''^{(j_2)}_k / \phi'^{(j_1)}_k$$

όπου τα *φ′* και *φ″ αντιστοιχούν* στις υπολογισμένες κατανομές θεμάτων λέξεων. Τα θέματα του δεύτερου τρεξίματος έχουν επαναδιαταχθεί για να αντιστοιχούν όσο το καλύτερο δυνατόν με τα θέματα του πρώτου τρεξίματος του αλγόριθμου. Η ακρίβεια της αντιστοίχησης μετρήθηκε από το άθροισμα των KL αποστάσεων. Το δεξί πάνελ δείχνει το χειρότερο ζευγάρι ευθυγραμμισμένο θεμάτων με KL απόσταση 9.4. Και τα δύο θέματα συσχετίζονται με το money αλλά είναι διαφορετικά. Συνοψίζοντας, αυτά τα αποτελέσματα δείχνουν πρακτικά ότι οι λύσεις διαφορετικών δειγμάτων θα δώσουν διαφορετικά αποτελέσματα αλλά πολλά θέματα παραμένουν σταθερά.



Run 1		Run 2	
MONEY	.094	MONEY	.086
GOLD	.044	PAY	.033
POOR	.034	BANK	.027
FOUND	.023	INCOME	.027
RICH	.021	INTEREST	.022
SILVER	.020	TAX	.021
HARD	.019	PAID	.016
DOLLARS	.018	TAXES	.016
GIVE	.016	BANKS	.015
WORTH	.016	INSURANCE	.015
BUY	.015	AMOUNT	.011
WORKED	.014	CREDIT	.010
LOST	.013	DOLLARS	.010
SOON	.013	COST	.008
PAY	.013	FUNDS	.008

Εικόνα *3: Σταθερότητα θεμάτων μεταξύ διαφορετικών εκτελέσεων.*

Καθορίζοντας τον αριθμό των θεμάτων. Η επιλογή του αριθμού των θεμάτων μπορεί να επηρεάζει την ερμηνεία των αποτελεσμάτων. Μία λύση με πολύ λίγα θέματα θα δώσει γενικευμένα θέματα και μια λύση με πάρα πολλά θέματα θα δώσει θέματα που δεν μπορούν να ερμηνευθούν. Υπάρχει ένας αριθμός αντικειμενικών μεθόδων για να διαλέξουμε τον αριθμό των θεμάτων. Οι Griffiths και Steyvers (2004) έδωσαν μια προσέγγιση επιλογής πιθανοτικών μοντέλων. Η ιδέα είναι να υπολογίσουμε την εκ των υστέρων πιθανότητα του μοντέλου καθώς ολοκληρώνουμε ως προς τις δυνατές ρυθμίσεις παραμέτρων.

3.1.5 Πολυσημία των θεμάτων

Πολλές λέξεις στην φυσική γλώσσα είναι πολυσημικές, έχοντας πολλά νοήματα. Η σημασιολογική τους αβεβαιότητα μπορεί να λυθεί από τις άλλες λέξεις του κειμένου. Το παράδειγμα στην εικόνα 3.4 δείχνει 3 θέματα που έχουν επιλεχθεί από μια 300-θεμάτων λύση από το TASA σώμα. Σε κάθε από αυτά τα θέματα, στη λέξη PLAY δίνεται σχετικά μεγάλη πιθανότητα συσχετισμένη με διαφορετικά νοήματα της λέξης.

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Εικόνα 3.4: Τρία θέματα που συσχετίζονται με την λέξη PLAY

Document #29795

Bix beiderbecke, at age[060] fifteen[207], sat[174] on the slope[071] of a bluff[055] overlooking[027] the mississippi[137] river[137]. He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He showed[002] promise[134] on the piano[077], and his parents[035] hoped[268] he might consider[118] becoming a concert[077] pianist[077]. But bix was interested[268] in another kind[050] of music[077]. He wanted[268] to play[077] the cornet. And he wanted[268] to play[077] jazz[077]...

Document #1883

There is a simple[050] reason[106] why there are so few periods[078] of really great theater[082] in our whole western[046] world. Too many things[300] have to come right at the very same time. The dramatists must have the right actors[082] the actors[082] must have the right playhouses, the playhouses must have the right audiences[082]. We must remember[288] that plays[082] exist[143] to be performed[077], not merely[050] to be read[254]. (even when you read[254] a play[082] to yourself, try[288] to perform[062] it, to put[174] it on a stage[078], as you go along.) as soon[028] as a play[082] has to be performed[082], then some kind[126] of theatrical[082]...

Document #21359

Jim[296] has a game[166] book[254]. Jim[296] reads[254] the book[254]. Jim[296] sees[081] a game[166] for one. Jim[296] plays[166] the game[166]. Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296]. Don[180] comes[040] into the house[038] Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] play[166] the game[166]. The boys[020] play[166] the game[166] for two. The boys[020] like the game[166]. Meg[282] comes[040] into the house[282]. Meg[282] and don[180] and jim[296] read[254] the book[254]. They see a game[166] for three. Meg[282] and don[180] and jim[296] play[166] the game[166]. They play[166]...

Εικόνα 3.5: Τρία TASA κείμενα με την λέξη PLAY

Σε ένα νέο κείμενο έχοντας παρατηρήσει μόνο την λέξη PLAY, θα υπάρχει αβεβαιότητα ανάμεσα στα θέματα που μπορεί να είχαν δημιουργήσει την λέξη. Αυτή η αβεβαιότητα μπορεί να μειωθεί παρατηρώντας άλλες λέξεις του κειμένου με λιγότερη αμφιβολία. Αυτή η διαδικασία μπορεί να περιγραφεί σαν μία διαδικασία επαναληπτικού sampling όπως το περιγράψαμε παραπάνω, όπου η εναπόθεση της κάθε λέξης-δείγματος σε ένα θέμα εξαρτάται από τις εναποθέσεις των άλλων λέξεων του κειμένου. Στην Εικόνα 3.5, βλέπουμε τρία κείμενα από το TASA που χρησιμοποιούν την λέξη PLAY σε τρεις διαφορετικές εκδοχές. Οι εκθετικοί αριθμοί δείχνουν τις εναποθέσεις θεμάτων για κάθε λέξη-δείγμα. Οι γκρι λέξεις είναι τετριμμένες λέξεις ή λέξεις χαμηλής συχνότητας και δεν έχουν χρησιμοποιηθεί στην ανάλυση. Η sampling διαδικασία των λιγότερο αβέβαιων λέξεων ισχυροποιεί ένα συγκεκριμένο θέμα στο κείμενο. Όταν μια λέξη έχει αβεβαιότητα, η κατανομή θέματος που αναπτύσσεται για το κείμενο είναι ο βασικό παράγοντας για τον προσδιορισμό της λέξης.

3.2 Ιεραρχικό Μοντέλο Θεμάτων

3.2.1 Εισαγωγή

Τα πολύπλοκα πιθανοτικά μοντέλα κυριαρχούν όλο και περισσότερο σε περιοχές όπως της βιοπληροφορικής, ανάκτησης πληροφορίας και όρασης μηχανών. Αυτές οι περιοχές δημιουργούν βασικές προκλήσεις λόγω της open-ended κλίσης φύσης τους – τα σύνολα δεδομένων συχνά μεγαλώνουν καθώς ο χρόνος περνάει, και καθώς μεγαλώνουν φέρνουν καινούριες οντότητες και δομές. Τα σημερινά στατιστικά εργαλεία μοντέλων συχνά δεν συμπεριφέρονται και τόσο λειτουργικά. Συγκεκριμένα, οι κλασσικές τεχνικές συλλογής μοντέλων, βασισμένες στην υπόθεση του πειραματισμού, δύσκολα μπορούν να προσαρμοστούν σε προβλήματα που τα δεδομένα συνεχίζουν να αυξάνονται και απεριόριστα σύνολα συχνά δυσανάλογων δομών πρέπει να ορίζονται σε κάθε βήμα.

Ένα σημαντικό παράδειγμα αυτών των προκλήσεων παρέχεται από το πρόβλημα εκμάθησης ιεραρχίας θεμάτων από τα δεδομένα. Δεδομένης μιας συλλογής από κείμενα το καθένα από τα οποία περιέχει "λέξεις", επιθυμούμε να ανακαλύψουμε χρήσιμες δομές ή "θέματα" σε κείμενα και να οργανώσουμε αυτά τα θέματα σύμφωνα με μια ιεραρχία. Στο κεφάλαιο αυτό, αναπτύσσουμε αποτελεσματικές στατιστικές τεχνικές για την κατασκευή μιας τέτοιας ιεραρχίας που μπορεί να μεγαλώνει και να αλλάζει καθώς συσσωρεύονται όλο και περισσότερα δεδομένα.

Προσεγγίζουμε αυτό το πρόβλημα επιλογής μοντέλων ορίζοντας ένα γενικό πιθανοτικό μοντέλο ιεραρχικών δομών και διαλέγοντας την Baeysian προοπτική στο πρόβλημα εκμάθησης αυτών των δομών από τα δεδομένα. Έτσι οι ιεραρχίες μας είναι τυχαίες μεταβλητές. Επίσης, αυτές οι τυχαίες μεταβλητές είναι διαδικαστικά ορισμένες, σύμφωνα με έναν αλγόριθμο που κατασκευάζει την ιεραρχία καθώς τα δεδομένα εισέρχονται. Το πιθανοτικό αντικείμενο που οδηγεί αυτή την προσέγγιση είναι μια κατανομή πάνω σε διαχωρίσεις ακέραιων αριθμών γνωστή ως Chinese restaurant διαδικασία. Θα δείξουμε πως να επεκτείνουμε την Chinese restaurant διαδικασία σε μια ιεραρχία διαμερίσεων, και θα δείξουμε πως χρησιμοποιείται αυτή η διαδικασία για την απεικόνιση των εκ των προτέρων και των εκ των υστέρων κατανομών για ιεραρχίες θεμάτων.

Υπάρχουν πολλές πιθανές προσεγγίσεις στην μοντελοποίηση ιεραρχιών θεμάτων. Στην δική μας προσέγγιση, κάθε κόμβος είναι η ιεραρχία που συσχετίζεται με ένα θέμα, όπου ένα θέμα είναι μια κατανομή από λέξεις. Ένα κείμενο παράγεται διαλέγοντας ένα μονοπάτι από την ρίζα σε ένα φύλλο, συνεχώς επιλέγοντας θέματα καθώς διασχίζει το μονοπάτι, και επιλέγοντας λέξεις από αυτά. Έτσι η οργάνωση των θεμάτων σε μια ιεραρχία στοχεύει στο να συλλάβει το ποσοστό χρησιμοποίησης των θεμάτων στο σώμα. Αυτή η προσέγγιση διαφέρει από μοντέλα ιεραρχιών θεμάτων που είναι κατασκευασμένα πάνω στην υπόθεση ότι οι κόμβοι-γονείς είναι συσχετισμένοι με τους κόμβους παιδιά. Εμείς δεν κάνουμε μια τέτοια υπόθεση – για παράδειγμα, ο κόμβος-ρίζα μπορεί να τοποθετήσει όλη την μάζα πιθανότητας σε συναρτήσεις λέξεων, με κανέναν από τους απογόνους του να μην έχει τοποθετήσει μάζα πιθανότητας σε συναρτήσεις λέξεων.

3.2.2 Chinese restaurant διαδικασία

Θα αρχίσουμε με μια σύντομη περιγραφή της Chinese restaurant διαδικασίας και στην συνέχεια θα δείξουμε πως αυτή η διαδικασία μπορεί να επεκταθεί και σε άλλες ιεραρχίες. Η Chinese restaurant διαδικασία (CRP) είναι μια κατανομή διαμερίσεων ακεραίων που έχει αποκτηθεί, φανταζόμενοι μια διαδικασία στην οποία M πελάτες κάθονται σε ένα εστιατόριο με άπειρο αριθμό τραπεζιών. Η βασική διαδικασία ορίζεται όπως ακολουθεί. Ο πρώτος πελάτης κάθεται στο πρώτο τραπέζι. Ο m-οστός πελάτης κάθεται σε ένα τραπέζι σύμφωνα με την παρακάτω κατανομή:

$$\begin{aligned} p(\text{previously occupied table } i) &= \frac{m_i}{\gamma+m-1} \\ p(\text{the next unoccupied table}) &= \frac{\gamma}{\gamma+m-1}, \end{aligned} \qquad (1)$$

όπου το m_i είναι ο αριθμός των προηγούμενων πελατών στο τραπέζι i, και γ μια παράμετρος. Αφού καθίσουν M πελάτες, η συνάρτηση δίνει μια διαμέριση M αντικειμένων. Αυτή η κατανομή δίνει την ίδια δομή διαμερίσεων όπως η Dirichlet διαδικασία. Όμως, η CRP επιτρέπει επίσης πολλές ποικιλομορφίες στον βασικό κανόνα της εξίσωσης (1), συμπεριλαμβανομένης και μιας επιλογής του γ εξαρτώμενης από τα δεδομένα και μίας πιο γενικής αποτελεσματικής εξάρτησης στην παρούσα διαμέριση. Αυτή η ευκαμψία θα αποδειχθεί πολύ χρήσιμη.

Η CRP έχει χρησιμοποιηθεί για να απεικονίσει την αβεβαιότητα σε ένα αριθμό στοιχείων στο mixture μοντέλο. Σε ένα μείγμα συλλογής ειδών, κάθε τραπέζι στο Κινέζικο εστιατόριο συσχετίζεται με μια επιλογή από την $p(\beta|\eta)$ όπου το β είναι μια παράμετρος στοιχείων μείγματος. Κάθε σημείο δεδομένων παράγεται με το να διαλέξουμε ένα τραπέζι i από την εξίσωση (1) και μετά να σχηματίσουμε μια τιμή από την κατανομή παραμετροποιημένη από το β_i (την παράμετρο που συσχετίζεται με αυτό το τραπέζι). Δεδομένου ενός συνόλου δεδομένων, η εκ των υστέρων κατανομή κάτω από αυτό το μοντέλο έχει δύο συστατικά. Το πρώτο, είναι η κατανομή των σχεδιαγραμμάτων θέσεων. Ο αριθμός των στοιχείων του μείγματος καθορίζεται από τον αριθμό των τραπεζιών που καταλαμβάνουν τα δεδομένα. Το δεύτερο, δεδομένου ενός σχεδιαγράμματος θέσεων, είναι τα συγκεκριμένα δεδομένα που υπάρχουν σε κάθε τραπέζι και δείχνουν την κατανομή που συσχετίζεται με την παράμετρο β σε αυτό το τραπέζι, Η εκ των υστέρων κατανομή μπορεί να υπολογιστεί χρησιμοποιώντας μία Μαρκοβιανή Monte Carlo αλυσίδα. Εφαρμόγες ποικίλων ειδών mixture μοντέλων έχουν αρχίσει να εμφανίζονται. Τα παραδείγματα περιλαμβάνουν Gaussian mixture μοντέλα και λανθάνοντα Markov μοντέλα.

3.2.3 Επεκτείνοντας την CRP σε ιεραρχίες

Η CRP είναι υπαγόμενη στην mixture μοντελοποίηση γιατί μπορούμε να εδραιώσουμε μια ένα-προς-ένα σχέση μεταξύ των τραπεζιών και των αναμεμειγμένων στοιχείων και ένα-προς-πολλά σχέση ανάμεσα στα αναμεμιγμένα στοιχεία και στα δεδομένα. Στα μοντέλα που θα θεωρήσουμε, όμως, κάθε σημείο δεδομένων είναι συσχετισμένο με πολλαπλά αναμεμιγμένα στοιχεία που βρίσκονται σε ένα μονοπάτι μιας ιεραρχίας. Θα αναπτύξουμε μια ιεραρχική εκδοχή της CRP για να χρησιμοποιήσουμε ώστε να βρούμε την εκ των προτέρων κατανομή αυτών των μοντέλων.

Μια nested Chinese restaurant διαδικασία μπορεί να οριστεί υποθέτοντας το εξής σενάριο. Ας φανταστούμε λοιπόν ότι υπάρχει ένας άπειρος αριθμός Κινέζικων εστιατορίων με άπειρα τραπέζια σε μια πόλη. Ένα εστιατόριο ορίζεται να είναι το εστιατόριο-ρίζα, και πάνω σε καθένα από τα άπειρα τραπέζια του, υπάρχει μαι κάρτα με το όνομα ενός άλλου

εστιατόριου. Σε κάθε ένα από τα τραπέζια αυτών των εστιατορίων υπάρχουν κάρτες που αναφέρονται σε άλλα εστιατόρια, και αυτή η δομή επαναλαμβάνεται απείρως. Κάθε εστιατόριο αναφέρεται ακριβώς μια φορά. Έτσι τα εστιατόρια της πόλης είναι οργανωμένα σε ένα δέντρο με άπειρα κλαδιά. Σημειώστε ότι κάθε εστιατόριο συσχετίζεται με ένα επίπεδο αυτού του δέντρου.

Ένας τουρίστας φτάνει στην πόλη για να κάνει τις διακοπές του. Το πρώτο βράδυ, πηγαίνει στο εστιατόριο-ρίζα και διαλέγει ένα τραπέζι σύμφωνα με την εξίσωση (1). Το δεύτερο βράδυ πηγαίνει στο εστιατόριο που ορίστηκε από το τραπέζι της πρώτης νύχτας και διαλέγει ένα άλλο τραπέζι, πάλι από την εξίσωση (1). Επαναλαμβάνει την διαδικασία για L ημέρες. Στο τέλος της εκδρομής, ο τουρίστας έχει επισκεφτεί L εστιατόρια που συγκροτούν ένα μονοπάτι μέχρι το L-οστό επίπεδο του δέντρου που περιγράψαμε πριν. Μετά από τις διακοπές L-ημερών M τουριστών, μια συλλογή μονοπατιών σχηματίζει ένα υποδένδρο L-επιπέδων.

Αυτή η εκ των προτέρων κατανομή μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει ιεραρχίες θεμάτων. Όπως ακριβώς και η κλασσική CPR μπορεί να χρησιμοποιηθεί για να εκφράσει την αβεβαιότητα των πιθανών αριθμών των στοιχείων, έτσι και η nested CRP μπορεί να χρησιμοποιηθεί για να εκφράσει την αβεβαιότητα πιθανών δένδρων L-επιπέδων.

3.2.4 Επίλυση με Gibbs Sampling

Σε αυτή την ενότητα, θα περιγράψουμε έναν Gibbs Sampling αλγόριθμο εφαρμοσμένο πάνω στην εκ των υστέρων nested CRP και στα αντίστοιχα θέματα στο hLDA μοντέλο. Ο Gibbs sampler παρέχει μια μέθοδο για την ταυτόχρονη εξερεύνηση του χώρου παραμέτρων (του σώματος) και του χώρου μοντέλου (του L-επιπέδων δέντρου).

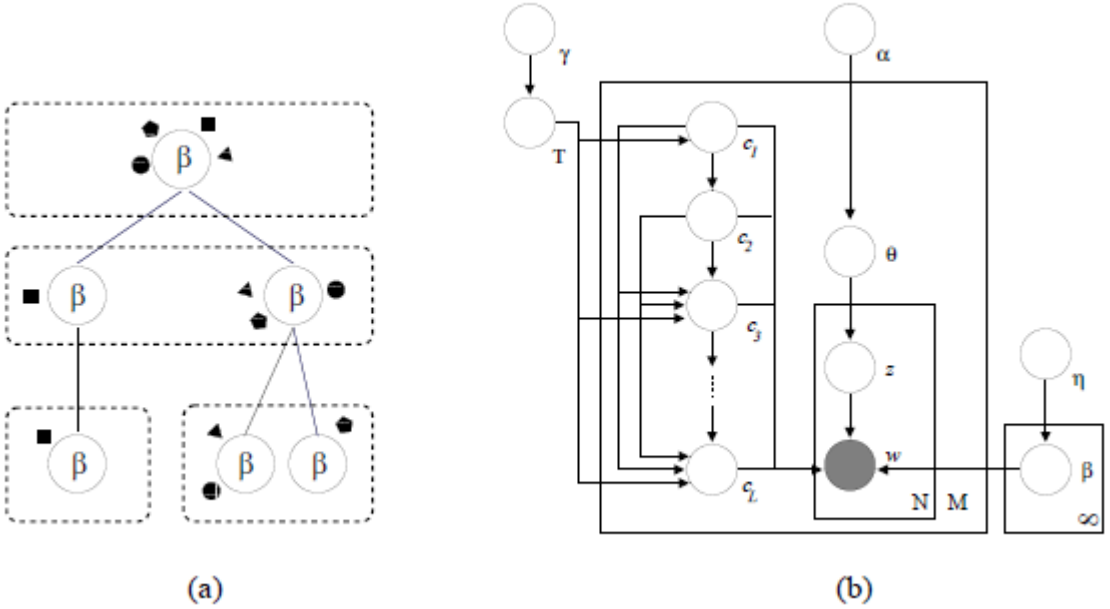
Οι μεταβλητές που χρειάζονται από τον αλγόριθμο είναι: $w_{m,n}$ η n-οστή λέξη στο m-οστό κείμενο (οι μόνες γνωστές μεταβλητές στο μοντέλο), το $c_{m,l}$, το αντίστοιχο εστιατόριο στο l-οστό θέμα στο κείμενο m, και το $z_{m,n}$, η απεικόνιση της n-οστής λέξης στο m-οστό κείμενο σε ένα από τα L διαθέσιμα θέματα. Όλες οι άλλες μεταβλητές του μοντέλου – θ και β – ενοποιούνται. Ο Gibbs επιλογέας, με αυτό τον τρόπο, εκτιμά τις τιμές $z_{m,n}$ και $c_{m,l}$.

Συνεπώς, χωρίζουμε τον Gibbs επιλογέα σε δύο μέρη. Στο πρώτο, δεδομένης της εκάστοτε κατάστασης της CRP, επιλέγουμε τις $z_{m,n}$ μεταβλητές του LDA μοντέλου. Στο δεύτερο, δεδομένων των τιμών των κρυμμένων LDA τιμών, επιλέγουμε τις $c_{m,l}$ μεταβλητές που συσχετίζονται με την εκ των προτέρων κατανομή CRP . Η δεσμευμένη πιθανότητα για το c_m, όπου τα L θέματα συσχετίζονται με το κείμενο m, είναι:

$$p(\mathbf{c}_m \mid \mathbf{w}, \mathbf{c}_{-m}, \mathbf{z}) \propto p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) p(\mathbf{c}_m \mid \mathbf{c}_{-m}),$$

όπου το $\mathbf{w}_{-m}$ και το $\mathbf{c}_{-m}$ δηλώνουν τις $\mathbf{w}$ και $\mathbf{c}$ μεταβλητές για όλα τα κείμενα εκτός του m. Αυτή η έκφραση είναι μια απόρροια του κανόνα Bayes με $p(\mathbf{w}_m|\mathbf{c},\mathbf{w}_{-m},\mathbf{z})$ σαν πιθανότητα των δεδομένων δεδομένης μιας συγκεκριμένης επιλογής του c_m και σαν την εκ των προτέρων κατανομή $p(c_m|c_{-m})$ του c_m εφαρμοσμένη από τη nested CRP. Η πιθανότητα λαμβάνεται ολοκληρώνοντας ως προς τις παραμέτρους β, που είναι:

$$p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) = \prod_{\ell=1}^{L} \left(\frac{\Gamma(n^{(\cdot)}_{c_{m,\ell},-m} + W\eta)}{\prod_w \Gamma(n^{(w)}_{c_{m,\ell},-m} + \eta)} \frac{\prod_w \Gamma(n^{(w)}_{c_{m,\ell},-m} + n^{(w)}_{c_{m,\ell},m} + \eta)}{\Gamma(n^{(\cdot)}_{c_{m,\ell},-m} + n^{(\cdot)}_{c_{m,\ell},m} + W\eta)} \right),$$



Εικόνα 3.8: (α)Τα μονοπάτια των τεσσάρων τουριστών δια μέσου του άπειρου δέντρου των Κινέζικων εστιατορίων (L=3). Οι συνεχείς γραμμές συνδέουν κάθε εστιατόριο με τα εστιατόρια που αναφέρονται στα τραπέζια του. Τα επιλεγμένα μονοπάτια των τεσσάρων τουριστών περιγράφουν ένα συγκεκριμένο υποδένδρο του άπειρου δένδρου. Αυτό απεικονίζει ένα παράδειγμα από των χώρο καταστάσεων της εκ των υστέρων nested CRP της Εικόνας b για τα τέσσερα κείμενα. (b) Η απεικόνιση του γραφικού μοντέλου ιεραρχικού LDA με την εκ των προτέρων nested CRP. Έχουμε ξεχωρίσει την nested Chinese restaurant διαδικασία από τα θέματα. Κάθε ένα από τα άπειρα β αντιστοιχεί σε ένα από τα εστιατόρια.

όπου το $n^{(w)}_{c_{m,l},-m}$ είναι ο αριθμός των παρατηρήσεων της λέξης *w* που έχει εναποτεθεί στο θέμα δηλωμένο ως $c_{m,l}$ συμπεριλαμβανομένου του παρόντος κειμένου, το W είναι ολόκληρο το μέγεθος του λεξιλογίου, και το Γ(.) δηλώνει την κλασσική συνάρτηση Γάμμα. Όταν το c περιέχει ένα εστιατόριο που το έχουν επισκεφτεί προηγουμένως, το $n^{(w)}_{c_{m,l},-m}$ είναι 0.

Το σύνολο των δυνατών τιμών για το $\mathbf{c_m}$ αντιστοιχεί στην ένωση του συνόλου των υπάρχοντων μονοπατιών διαμέσου του δένδρου, ίσου με τον αριθμό των φύλλων, και με το σύνολο των δυνατών νέων μονοπατιών, ίσο με τον αριθμό των εσωτερικών κόμβων. Αυτό το σύνολο μπορεί να απαριθμηθεί και να αξιολογηθεί χρησιμοποιώντας την εξίσωση (1) και τον ορισμό μιας nested CRP από την προηγούμενη ενότητα.

3.2.5 Ένα ιεραρχικό μοντέλο θεμάτων

Ας θεωρήσουμε ένα σύνολο δεδομένων επιλεγμένο από ένα σώμα κειμένων. Κάθε κείμενο είναι μια συλλογή λέξεων, όπου μια λέξη είναι αντικείμενο ενός λεξικού. Η βασική μας υπόθεση είναι ότι οι λέξεις σε ένα κείμενο δημιουργούνται σύμφωνα με ένα mixture μοντέλο όπου οι αναμεμειγμένες αναλογίες είναι τυχαίες και ορισμένες από τα κείμενα. Θεωρούμε μια πολυωνυμική μεταβλητή *z*, και ένα συσχετισμένο σύνολο κατανομών λέξεων *p(w| z, β)*, όπου *β* είναι μια παράμετρος. Αυτά τα θέματα είναι τα βασικά αναμεμιγμένα στοιχεία στο μοντέλο μας. Οι αναμεμειγμένες αναλογίες είναι συσχετισμένες με στοιχεία που δηλώνονται από ένα διάνυσμα *θ*. Προσωρινά υποθέτοντας ότι υπάρχουν K θέματα σε ένα σώμα, μια υπόθεση που αργότερα θα αφεθεί, το *z* επεκτείνεται σε K τιμές και το *θ* ένα K-διαστάσεων διάνυσμα. Η κατανομή είναι $p(w|\theta) = \sum_{i=1}^{K} \theta_i \, p(w|z=i, \beta_i)$ που είναι μια τυχαία κατανομή αφού το *θ* είναι τυχαίο.

Ας ορίσουμε τώρα την ακόλουθη δύο-επιπέδων γενετική πιθανοτική διαδικασία για την δημιουργία ενός κειμένου: (1) Διάλεξε ένα K-διάνυσμα *θ* αναλογιών θεμάτων από μια κατανομή p(*θ*|*α*),
όπου το α είναι μία επιπέδου-σώματος παράμετρος, (2) επέλεγε συνεχώς λέξεις από την mixture κατανομή p(*w*|*θ*) για την επιλεγμένη τιμή του *θ*. Αν η κατανομή *p(θ|α)* επιλεχθεί να είναι μια Dirichlet κατανομή, χρησιμοποιούμε το LDA μοντέλο. Το LDA , όπως το έχουμε αναλύσει σε προηγούμενο κεφάλαιο, είναι μια δύο-επιπέδων γενετική διαδικασία στην οποία τα κείμενα συσχετίζονται με αναλογίες θεμάτων, και το σώμα μοντελοποιείται ως μια Dirichlet κατανομή σε αυτές τις αναλογίες.

Στην συνέχεια θα περιγράψουμε μια επέκταση αυτού του μοντέλου στην οποία τα θέματα βρίσκονται σε ιεραρχία. Προς στιγμήν, ας υποθέσουμε ότι μας έχει δοθεί ένα L-διαστάσεων δέντρο και κάθε κόμβος είναι συσχετισμένος με ένα θέμα. Ένα κείμενο δημιουργείται όπως ακολούθως: (1) διάλεξε ένα μονοπάτι απ την ρίζα του δέντρου σε ένα φύλλο, (2) σχεδίασε ένα διάνυσμα αναλογιών θεμάτων *θ* από μια L-διαστάσεων Dirichlet, (3) δημιούργησε λέξεις στο κείμενα από ένα μείγμα θεμάτων κατά την διάρκεια του μονοπατιού στο φύλλο, με αναμεμειγμένες αναλογίες *θ*.

Τέλος, θα χρησιμοποιήσουμε την nested CRP για να χαλαρώσουμε την υπόθεση της φιξαρισμένης δομής δέντρου. Τοποθετούμε επίσης μια εκ των προτέρων κατανομή στα θέματα β_i, καθένα από τα οποία συσχετίζεται με ένα εστιατόριο στο άπειρο δέντρο. Ένα κείμενο δημιουργείται επιλέγοντας πρώτα, ένα L-επιπέδων μονοπάτι διαμέσου των εστιατορίων και μετά παίρνοντας λέξεις από L θέματα που συσχετίζονται με τα εστιατόρια διαμέσου του μονοπατιού. Σημειώστε ότι όλα τα κείμενα μοιράζονται το θέμα που συσχετίζεται με το εστιατόριο-ρίζα.

1. Let c_1 be the root restaurant.
2. For each level $\ell \in \{2, \ldots, L\}$:
   (a) Draw a table from restaurant $c_{\ell-1}$ using Eq. (1). Set c_ℓ to be the restaurant referred to by that table.
3. Draw L-dimensional topic proportions θ from Dir(α).
4. For each word $n \in \{1, \ldots, N\}$:
   (a) Draw $z \in \{1, \ldots, L\}$ from Mult(θ).
   (b) Draw w_n from the topic distribution associated with restaurant c_z.

Αυτό το μοντέλο, το ιεραρχικό LDA (hLDA), δίνεται στην Εικόνα 3.8(b). Ο κόμβος με το όνομα *T* αναφέρεται σε μια συλλογή από έναν άπειρο αριθμό L-επιπέδων μονοπατιών και σχεδιάζεται από την nested CRP. Δεδομένου του *T*, οι $c_{m,l}$ μεταβλητές είναι ντετερμινιστικές – απλά κοιτάξτε το *l*-οστό επίπεδο και το *m*-οστό μονοπάτι στην άπειρη συλλογή μονοπατιών. Όμως, μην έχοντας παρατηρήσει το T, η κατανομή του $c_{m,l}$ θα ορίζεται από την nested CRP, περιθωριοποιημένων όλων των $c_{q,l}$ για q < m.

Τώρα ας υποθέσουμε ότι μας έχει δοθεί ένα σώμα M κειμένων, $\mathbf{w_1,....,w_M}$. Η εκ των υστέρων κατανομή των c μεταφέρεται ουσιαστικά, σε μια εκ των υστέρων κατανομή των πρώτων M μονοπατιών στο *T*. Θεωρείστε ένα νέο κείμενο $\mathbf{w_{M+1}}$. Η εκ των υστέρων κατανομή του μονοπατιού του θα εξαρτάται από το μη γνωστό *T*, των εκ των υστέρων μονοπατιών όλων των κειμένων του αρχικού σώματος. Επακολούθως τα νέα κείμενα θα εξαρτώνται επίσης από το αρχικό σώμα και όλων των άλλων καινούριων κειμένων που είχαν παρατηρηθεί πριν από αυτά. Σημειώστε ότι, μέσω της εξίσωσης (1), κάθε νέο κείμενο μπορεί να επιλέξει ένα προηγούμενο εστιατόριο, που ήταν μέχρι στιγμής μη επισκέψιμο, σε οποιοδήποτε από τα επίπεδα του δέντρου.

Σε μια άλλη εκδοχή αυτού του μοντέλου, μπορούμε να βρούμε μια διαδικασία που να μετατρέπει την nested CPR σε κλασσική CPR, αλλά διατηρεί την ιδέα ότι ο ένας τουρίστας θα γευματίσει *L* φορές. Ο τουρίστας τρώει *L* φορές σε ένα εστιατόριο υπό το περιορισμό ότι δεν μπορεί να διαλέξει το ίδιο τραπέζι δυο φορές. Παρόλο που οι διακοπές είναι λιγότερο ενδιαφέρουσες, αυτό το μοντέλο παρέχει μια ενδιαφέρουσα εκ των προτέρων κατανομή. Πιο συγκεκριμένα, μπορεί να χρησιμοποιηθεί σαν εκ των προτέρων κατανομή ενός επίπεδου LDA μοντέλου στο οποίο κάθε κείμενο μπορεί να χρησιμοποιήσει το πολύ *L* θέματα από το εν δυνάμει άπειρο συνολικό σύνολο των θεμάτων.

3.3 Correlated Μοντέλο Θεμάτων (CTM)

3.3.1 Εισαγωγή

Το Latent Dirichlet Allocation (LDA) μοντέλο, μπορεί να αποδειχθεί ένα χρήσιμο εργαλείο για την στατιστική ανάλυση συλλογών κειμένων και άλλων διακριτών δεδομένων. Το LDA μοντέλο υποθέτει ότι οι λέξεις παράγονται από ένα μείγμα θεμάτων, το καθένα από τα οποία είναι μια κατανομή σε ένα λεξιλόγιο. Το LDA όμως περιορίζεται στο ότι δεν έχει την ικανότητα να συσχετίζει θέματα μεταξύ τους, για παράδειγμα, ένα κείμενο για την γενετική είναι περισσότερο πιθανό να μιλάει και για την ιατρική πάρα για την αστρονομία. Ο περιορισμός αυτός προέρχεται από την χρησιμοποίηση της Dirichlet κατανομής για να μοντελοποιηθεί η ποικιλομορφία ανάμεσα στις αναλογίες θεμάτων. Σε αυτό το κεφάλαιο θα δούμε το Correlated μοντέλο θεμάτων(CTM), όπου οι αναλογίες στα θέματα υποδεικνύουν συσχέτιση μέσα από μία λογαριθμική κανονική κατανομή. Θα δείξουμε ένα γρήγορο αλγόριθμο μεταβολικής συμπερασματολογίας για τον υπολογισμό της εκ των υστέρων κατανομής , ο οποίος είναι πολύπλοκος κάτι που οφείλεται στο ότι η κανονική λογαριθμική δεν είναι συζευγμένη με την πολυωνυμική.

Το CTM είναι ένα ιεραρχικό μοντέλο από συλλογές κειμένων. Το CTM μοντελοποιεί τις λέξεις του κάθε κειμένου από ένα mixture μοντέλο. Οι mixture παράγοντες μοιράζονται από όλα τα κείμενα της συλλογής: οι mixture αναλογίες είναι τυχαίες μεταβλητές. Το CTM αφήνει κάθε κείμενο να πάρει πολλά θέματα με διαφορετικές αναλογίες. Έτσι μπορεί να καταλάβει την ετερογένεια των ομαδοποιημένων αρχείων που παίρνουν ποικίλες κρυφές δομές.

3.3.2 Ορολογία

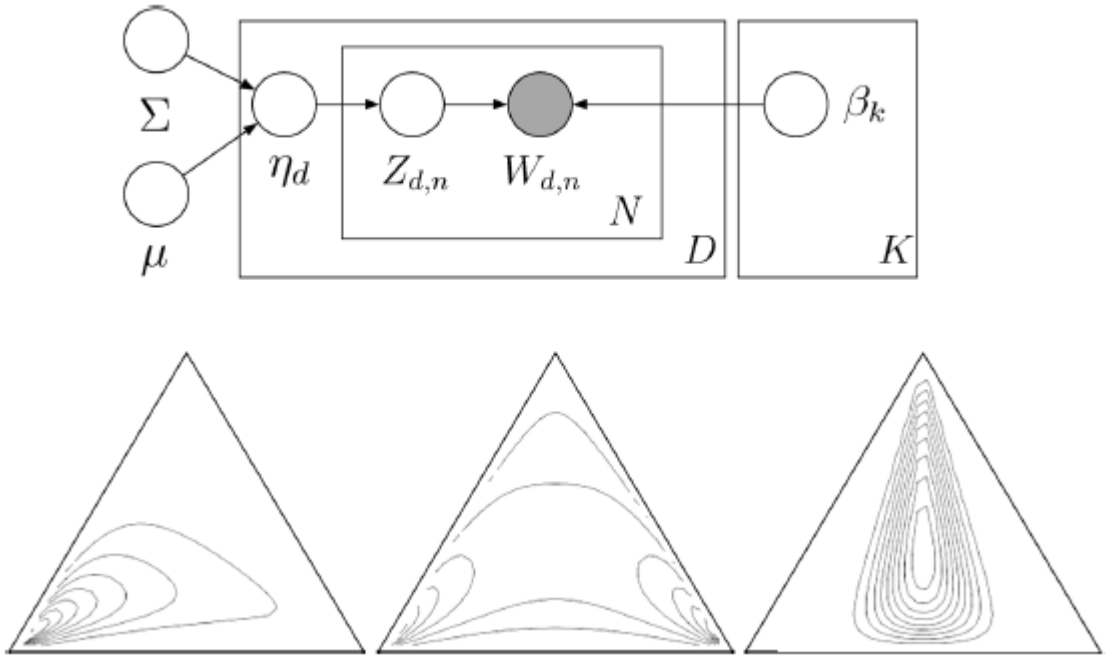
Χρησιμοποιούμε την παρακάτω ορολογία για να περιγράψουμε τα δεδομένα, τις κρυφές μεταβλητές και τις παραμέτρους στο CTM:

•*Λέξεις και δεδομένα. Οι* μόνες γνωστές παρατηρούμενες μεταβλητές που θεωρούμε είναι οι λέξεις που οργανώνονται σε κείμενα. Ορίζουμε το $w_{d,n}$ να είναι η n-οστή λέξη του d-οστού κειμένου, όπου είναι ένα στοιχείο σε ένα λεξιλόγιο V όρων. Το $\mathbf{w_d}$ ορίζει ένα διάνυσμα από N_d λέξεις που συσχετίζονται με κείμενο d.

•*Θέματα.* Ένα θέμα **β** είναι μια κατανομή σε ένα λεξιλόγιο, που δείχνει σε ένα V-1 simplex. Το μοντέλο θα περιέχει K θέματα $\boldsymbol{\beta_{1:K}}$.

•*Εναποθέσεις θεμάτων.* Κάθε λέξη θεωρείται ότι παράγεται από ένα θέμα των K συνολικά. Η εναπόθεση θέματος $z_{d,n}$ συσχετίζεται με την n-οστή λέξη και το d-οστό κείμενο.

•*Αναλογίες θεμάτων.* Τέλος, κάθε κείμενο συσχετίζεται με ένα σύνολο αναλογιών θεμάτων θ_d, που δείχνουν στο K-1 simplex. Έτσι το $\boldsymbol{\theta_d}$ είναι μια διανομή από δείκτες θεμάτων, και αντικατοπτρίζει τις πιθανότητες με τις οποίες οι λέξεις επιλέγονται από κάθε θέμα της συλλογής. Τυπικά, θα το θεωρούμε μια παραμετροποίηση της πολυωνυμικής μεταβλητής $\eta=\log(\theta_i|\theta_K)$.



Εικόνα 3.7: Κορυφή: Το γραφικό μοντέλο απεικόνισης του Correlated μοντέλου θεμάτων. Η λογαριθμική κανονική κατανομή μπορεί να απεικονίσει τις συσχετίσεις μεταξύ των θεμάτων που είναι αδύνατο να συμπεριληφθούν υπόψη από μία απλή Dirichlet. Κάτω: Παραδείγματα των πυκνοτήτων της λογικής κανονικής κατανομής σε ένα 2-simplex. Από τα αριστερά: Διαγονική συνδιασπορά του μη μηδενικού μέσου, αρνητική συσχέτιση μεταξύ των συστατικών 1 και 2, θετική συσχέτιση μεταξύ των συστατικών 1 και 2.

3.3.3 Ο αλγόριθμος

Το CTM θεωρεί ότι ένα N-λέξεων κείμενο παράγεται από την εξής γενετική διαδικασία. Δεδομένων των θεμάτων $\beta_{1:K}$, ένα K-διάνυσμα μ, και έναν K x K πίνακα διασποράς Σ:

1. Draw $\eta_d|\{\mu, \Sigma\} \sim \mathcal{N}(\mu, \Sigma)$.
2. For $n \in \{1, \ldots, N_d\}$:

(a) Draw topic assignment $Z_{d,n}|\eta_d$ from Mult($f(\eta_d)$).
(b) Draw word $W_{d,n}|\{z_{d,n}, \beta_{1:K}\}$ from Mult($\beta_{z_{d,n}}$),

Όπου το f(η) είναι μια φυσική παραμετροποίηση των θεματικών αναλογιών στην παραμετροποίηση του μέσου,

$$\theta = f(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}}.$$

Το CTM βασίζεται στο Latent Dirichlet μοντέλο (LDA). Το LDA θεωρεί μια σχεδόν ιδανική γενετική διαδικασία, που όμως οι αναλογίες θεμάτων δημιουργούνται μέσω μιας Dirichlet κατανομής. Στο LDA, η Dirichlet είναι υπολογιστικά μια εύκολη κατανομή αναλογιών θεμάτων επειδή είναι συζευγμένη με τις εναποθέσεις θεμάτων. Άλλα η Dirichlet θεωρεί σχεδόν ανεξάρτητα τα στοιχεία των αναλογιών. Στην πραγματικότητα, μπορεί κάποιος να προσομοιώσει ένα γράφημα της Dirichlet σχεδιάζοντας K ανεξάρτητες Γάμμα κατανομές και κανονικοποιώντας το διάνυσμα που προέκυψε.

Αντί να χρησιμοποιήσει την Dirichlet, το CTM σχεδιάζει ένα διάνυσμα πραγματικών τυχαίων αριθμών από την πολυωνυμική Gaussian και την σχεδιάζει στον simplex για να πάρει μια πολυωνυμική παράμετρο. Αυτό είναι το καθοριστικό χαρακτηριστικό της λογαριθμικής κανονικής κατανομής. Η διασπορά της Gaussian προκαλεί εξαρτήσεις ανάμεσα στα συστατικά του απλοποιημένου αλλαγμένου τυχαίου διανύσματος, επιτρέποντας έτσι μια γενική δομή ποικιλομορφίας ανάμεσα στα συστατικά. Η λογαριθμική κανονική κατανομή μελετήθηκε αρχικά για την ανάλυση σύνθετων δεδομένων, όπως για παράδειγμα η ανάλυση ορυκτών από γεωλογικά δείγματα. Στο CTM, χρησιμοποιούμε την *κρυφή* σύνθεση των θεμάτων που συσχετίζονται σε κάθε κείμενο.

Το μειονέκτημα χρησιμοποίησης της λογαριθμικής κανονικής είναι ότι δεν είναι συζευγμένη με την πολυωνυμική, κάτι που κάνει πολύπλοκο τον υπολογισμό της εκ των υστέρων κατανομής. Το πλεονέκτημα, όμως, είναι ότι παρέχει ένα πιο περιγραφικό μοντέλο κειμένων. Η ισχυρή υπόθεση ανεξαρτησίας που επιβάλλεται από την Dirichlet δεν είναι ρεαλιστική όταν αναλύουμε συλλογές κειμένων στον αληθινό κόσμο, όπου μπορούν να βρεθούν ισχυρές συσχετίσεις μεταξύ κρυφών θεμάτων. Για παράδειγμα, ένα κείμενο σχετικά με την γεωλογία είναι πιο πιθανό να μιλάει επίσης για την αρχαιολογία παρά για την γενετική. Στόχος μας είναι να χρησιμοποιήσουμε τον πίνακα διασποράς της λογαριθμικής κανονικής για να δούμε αυτές τις σχέσεις.

3.3.4 Υπολογισμός του CTM

Έχουμε δύο υπολογιστικά προβλήματα όταν χρησιμοποιούμε το CTM για να αναλύσουμε δεδομένα. Πρώτον δεδομένης μίας συλλογής θεμάτων και μιας κατανομής θεματικών αναλογιών $\{\boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ πρέπει να υπολογίσουμε την εκ των υστέρων κατανομή των κρυφών μεταβλητών δεσμευμένων ως προς τις λέξεις ενός κειμένου $p(\boldsymbol{\eta}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Αυτό μας αφήνει να ενσωματώσουμε τα νέα κείμενα σε έναν θεματικό χώρο λιγότερων διαστάσεων που αναπαριστά το μοντέλο. Χρησιμοποιούμε ένα γρήγορο αλγόριθμο μεταβολικής συμπερασματολογίας για να υπολογίσουμε την εκ των υστέρων κατανομή, κάτι που μας επιτρέπει να αναλύσουμε γρήγορα μεγάλες συλλογές δεδομένων κάτω από αυτές τις πολύπλοκες προϋποθέσεις.

Δεύτερον, δεδομένης μια συλλογής κειμένων $\{\mathbf{w}_1, ..., \mathbf{w}_D\}$, βρίσκουμε τις μέγιστες πιθανοτικές τιμές των και την λογαριθμική κανονική κατανομή κάτω από τις υποθέσεις του CTM. Θα χρησιμοποιήσουμε μια μεταβλητή του Expectation-Maximization αλγόριθμου, όπου το E-βήμα είναι ένα ανά-κείμενο πρόβλημα μεταβολικής συμπερασματολογίας. Επιπλέον θα ψάξουμε να βρούμε αραιές λύσεις του αντίστροφου πίνακα διασποράς μεταξύ των θεμάτων, και θα προσαρμόσουμε έναν l_1-κανονικοποιημένο υπολογισμό για αυτό τον σκοπό.

3.1 Εκ των υστέρων συμπερασματολογία με στατιστικές μεθόδους. Δεδομένου ενός κειμένου w και ενός μοντέλου $\{\boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, η εκ των υστέρων κατανομή των ανά-κείμενο κρυφών μεταβλητών είναι

$$p(\eta, z|\boldsymbol{w}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \frac{p(\eta|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} p(z_n|\eta) p(w_n|z_n, \boldsymbol{\beta}_{1:K})}{\int p(\eta|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\eta) p(w_n|z_n, \boldsymbol{\beta}_{1:K}) d\eta},$$

που είναι αδύνατο να υπολογίσουμε λόγω του ολοκληρώματος στον παρονομαστή, δηλαδή την περιθωριακή πιθανότητα του κειμένου που δεσμεύουμε. Υπάρχουν δύο λόγοι για αυτήν την αδυναμία. Ο πρώτος είναι ότι το άθροισμα ως προς τις K τιμές του z_n βρίσκεται μέσα στο παράγωγο των λέξεων, περιέχοντας έναν συνδυαστικό αριθμό από όρους. Ο δεύτερος, είναι ότι αν το K^N μείνει υπό την επήρεια δυνατότητας υπολογισμού, η κατανομή των θεματικών αναλογιών $p(\boldsymbol{\eta},z|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ δεν είναι συζευγμένη με την κατανομή των εναποθέσεων θεμάτων $p(z_n|\boldsymbol{\eta})$. Έτσι, δεν μπορούμε αναλυτικά να υπολογίσουμε τα ολοκληρώματα κάθε όρου.

Η μη συζευξιμότητα αποκλείει περαιτέρω την χρησιμοποίηση πολλών Monte Carlo Markov Chain(MCMC) sampling τεχνικών που έχουν αναπτυχθεί με Dirichlet-βασισμένα mixture μοντέλα μέλους. Αυτές οι MCMC μέθοδοι βασίζονται όλες στο Gibbs sampling, που η σύζευξη μεταξύ των κρυφών μεταβλητών μας αφήνει να υπολογίσουμε τις εκ των υστέρων αναλυτικά. Για να εφαρμόσουμε MCMC σε μια λογαριθμική κανονική, πρέπει να βασιστούμε σε μια Metropolis-Hastings λύση. Μία τέτοια τεχνική δεν θα είχε τις ίδιες δυνατότητες σύγκλισης και ταχύτητας ενός Gibbs sampler, γεγονός που ουσιαστικά εμποδίζει τον στόχο μας για την ανάλυσή συλλογών που περιέχουν εκατομμύρια λέξεων.

Έτσι για να υπολογίσουμε την εκ των υστέρων κατανομή, βασιζόμαστε σε μεταβολικές μεθόδους όπως ένα ντετερμινιστικό αντίστοιχο των MCMC. Η ιδέα πίσω από τις στατιστικές μεθόδους είναι να βελτιστοποιήσουμε τις ελεύθερες παραμέτρους ως προς τις κρυφές μεταβλητές έτσι ώστε η κατανομή να είναι όσο το δυνατό πιο κοντά στην Kullback-Leibler κατανομή της αληθινής εκ των υστέρων κατανομής. Η φορμαρισμένη στατιστική κατανομή χρησιμοποιείται σαν υποκατάστατο της εκ των υστέρων κατανομής όπως ακριβώς η εμπειρική κατανομή των δειγμάτων που χρησιμοποιείται από τις MCMC. Οι μεταβολικές μέθοδοι έχουν διαδεδομένη χρήση στο πεδίο μηχανικής μάθησης, η δυνατότητά τους στην εφαρμοσμένη στατιστική Bayesian τώρα αρχίζει να συνειδητοποιείται.

Στα μοντέλα που συγκροτούνται από ζευγάρια και μείγματα μίας οικογένειας εκθετικής-κλίσης, ο αλγόριθμος μεταβολικής συμπερασματολογίας μπορεί αυτόματα να δημιουργηθεί από τις υπολογιστικές εκτιμήσεις των φυσικών παραμέτρων στην πιθανοτική κατανομή. Όμως, το μη συζυγές ζευγάρι των μεταβλητών στο CTM απαιτεί την παραγωγή του αλγορίθμου μεταβολικής συμπερασματολογίας από τις αρχικές συνθήκες.

Θα ξεκινήσουμε χρησιμοποιώντας την ανισότητα του Jensen για να βρούμε το όριο της λογαριθμικής πιθανότητας στο κείμενο,

$$\log p(w_{1:N} \mid \mu, \Sigma, \beta) \geq$$
$$\mathrm{E}_q[\log p(\eta \mid \mu, \Sigma)] + \sum_{n=1}^{N} (\mathrm{E}_q[\log p(z_n \mid \eta)] + \mathrm{E}_q[\log p(w_n \mid z_n, \beta)]) + \mathrm{H}(q) \quad (1)$$

όπου η εκτίμηση υπολογίζεται παραγωγίζοντας ως προς το q, την στατιστική κατανομή των κρυφών μεταβλητών, και το H(q) δηλώνει την εντροπία αυτής της κατανομής. Επειδή είναι μία στατιστική κατανομή χρησιμοποιούμε ένα πλήρως παραγωγίσιμο μοντέλο, όπου οι μεταβλητές ελέγχονται ανεξάρτητα από διαφορετικές κατανομές,

$$q(\eta_{1:K}, z_{1:N} \mid \lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} q(\eta_i \mid \lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n \mid \phi_n).$$

Οι στατιστικές κατανομές των διακριτών εναποθέσεων θεμάτων $z_{1:N}$ είναι καθορισμένες από τις K-διαστάσεων πολυωνυμικές παραμέτρους $\varphi_{1:K}$ (αυτοί οι παράμετροι μέσου είναι πολυωνυμικές). Η στατιστική κατανομή των συνεχών μεταβλητών $\eta_{1:K}$ είναι μιας μεταβλητής Gaussians $\{\lambda_i,\nu_i\}$. Αφού οι στατιστικές παράμετροι φιξάρονται χρησιμοποιώντας ένα απλό κείμενο w1:N δεν υπάρχει πλεονέκτημα στην εισαγωγή μη διαγώνιων πινάκων διασποράς.

Ο αλγόριθμος μεταβολικής συμπερασματολογίας βελτιστοποιεί την εξίσωση (1) ως προς τις στατιστικές παραμέτρους, με το να στενεύει το όριο στην περιθωριακή πιθανότητα των παρατηρήσεων όπως ακριβώς η δομή των στατιστικών κατανομών επιτρέπει. Αυτό ισοδυναμεί με το να βρούμε την στατιστική κατανομή που ελαττώνει την τιμή KL(q||p), όπου το *p* είναι η αληθινή εκ των υστέρων κατανομή.

Ας σημειώσουμε εδώ ότι οι μεταβολικές μέθοδοι δεν έχουν την ίδια θεωρητική εγγύηση που έχουν οι MCMC, που περιορίζουν την κατανομή της αλυσίδας ακριβώς στην εκ των υστέρων που μας ενδιαφέρει. Όμως, οι μετατβολικές μέθοδοι παρέχουν γρήγορους αλγόριθμους και ένα κριτήριο καθαρής σύγκλισης, όπου οι MCMC μέθοδοι δεν μπορούν να είναι υπολογιστικά αποτελεσματικές καθώς επίσης είναι δύσκολο να καταλάβουμε πότε μια Μαρκοβιανή αλυσίδα έχει συγκλίνει.

3.2Υπολογισμός παραμέτρων. Δεδομένης μια συλλογής κειμένων, επιχειρούμε να να μεγιστοποιήσουμε την πιθανότητα ενός σώματος δεδομένων ως μια συνάρτηση θεμάτων $\beta_{1:K}$ και των πολυωνυμικών Gaussian($\boldsymbol{\mu,\Sigma}$).

Όπως σε πολλά μοντέλα κρυφών μεταβλητών, δεν μπορούμε να υπολογίσουμε την περιθωριακή πιθανότητα των δεδομένων γιατί η κρυφή δομή χρειάζεται να περιθωριοποιηθεί. Για να λύσουμε αυτό το πρόβλημα χρησιμοποιούμε τον στατιστικό Expectation-Maximization αλγόριθμο (EM). Στο E-βήμα του παραδοσιακού EM, υπολογίζεται η εκ των υστέρων κατανομή των κρυφών μεταβλητών δεδομένων και των παρόντων παραμέτρων του μοντέλου Στον μεταβολικό EM, υπολογίζουμε την εκ των υστέρων όπως περιγράψαμε στο προηγούμενο υποκεφάλαιο. Σημειώστε ότι είναι παρόμοιος με τον Monte Carlo EM, όπου εκεί το E-βήμα υπολογίζεται από τον Monte Carlo υπολογισμό της εκ των υστέρων κατανομής.

Πιο συγκεκριμένα, η ζητούμενη συνάρτηση του μεταβολικού EM είναι το όριο πιθανότητας αθροίζοντας την εξίσωση (3) ως προς την συλλογή κειμένων $\{\mathbf{w_1,.....,w_D}\}$,

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}_{1:K}; \boldsymbol{w}_{1:D}) \geq \sum_{d=1}^{D} \mathrm{E}_{q_d}[\log p(\boldsymbol{\eta}_d, \boldsymbol{z}_d, \boldsymbol{w}_d | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}_{1:K})] + \mathrm{H}(q_d).$$

Ο μεταβολικός αλγόριθμος είναι αύξων στην ζητούμενη συνάρτηση. Στο E-βήμα, μεγιστοποιούμε το όριο ως προς τις μεταβολικές παραμέτρους υπολογίζοντας την μεταβολική συμπερασματολογία για κάθε κείμενο. Στο M-βήμα, μεγιστοποιούμε το όριο ως προς τις παραμέτρους του μοντέλου. Αυτό αντιστοιχεί στην μεγαλύτερη πιθανότητα των θεμάτων και τον πολυμεταβλητών Gaussian, οπού η εκτίμηση υπολογίζεται ως προς τις μεταβολικές κατανομές στο E-βήμα,

$$\widehat{\boldsymbol{\beta}}_i \propto \sum_d \phi_{d,i} \mathbf{n}_d,$$

$$\widehat{\boldsymbol{\mu}} = \frac{1}{D} \sum_d \boldsymbol{\lambda}_d,$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{D} \sum_d I\nu_d^2 + (\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})(\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})^T,$$

όπου το $\mathbf{n_d}$ είναι το διάνυσμα του πλήθους των λέξεων που βρέθηκαν στο κείμενο d.

3.4 Δυναμικό μοντέλο θεμάτων

3.4.1 Εισαγωγή

Η αύξηση των ηλεκτρονικών δεδομένων επιβάλει την εύρεση νέων εργαλείων για την αυτοματοποιημένη οργάνωση, την εύρεση και την αναζήτηση σε μεγάλες συλλογές. Μια πρόσφατη έρευνα πάνω στη μηχανική μάθηση και τη στατιστική ανέπτυξε νέες τεχνικές για την εύρεση δομών λέξεων σε συλλογές κειμένων χρησιμοποιώντας ιεραρχικά πιθανοτικά μοντέλα.

Σε ένα ανταλλάξιμο μοντέλο θεμάτων, οι λέξεις κάθε κειμένου θεωρούνται ότι έχουν επιλεχθεί ανεξάρτητα, από ένα μείγμα πολυωνυμικών μεταβλητών. Οι αναμεμειγμένες αναλογίες επιλέγονται τυχαία για κάθε κείμενο. Τα αναμεμιγμένα στοιχεία ή θέματα χρησιμοποιούνται από όλα τα κείμενα. Έτσι, κάθε κείμενο αντιστοιχεί σε στοιχεία με διαφορετικές αναλογίες. Αυτά τα μοντέλα είναι μια αρκετά καλή μέθοδος για την μείωση διαστάσεων σε μεγάλες συλλογές άναρχων δεδομένων. Επιπλέον, η εκ των υστέρων συμπερασματολογία, σε επίπεδο κειμένων, είναι χρήσιμη για την ανάκτηση δεδομένων, την αρχικοποίηση και την θεματική αναζήτηση.

Το να χρησιμοποιούμε τις λέξεις σαν να ήταν ανταλλάξιμες είναι μια απλοποίηση που είναι απαραίτητη για την επίτευξη του στόχου μας, την εύρεση δηλαδή σημασιολογικών σχημάτων μέσα σε κάθε κείμενο. Για πολλές συλλογές, όμως, η υπόθεση της ανταλλαξιμότητας είναι ακατάλληλη. Συλλογές δεδομένων όπως επιστημονικά άρθρα, email, νέα άρθρα εξελίσσονται διαρκώς. Για παράδειγμα το επιστημονικό άρθρο “The Brain of Professor Laborde” μπορεί να ανήκει στο ίδιο επιστημονικό μονοπάτι με το “Reshaping the Cortical Motor Mab by Unmasking Latend intracortical Connections”, αλλά η μελέτη της νευρολογίας έδειχνε πολύ διαφορετική το 1903 σε σχέση με το 1991. Τα θέματα μιας συλλογής κειμένων εξελίσσονται διαρκώς, και είναι ο σαφής στόχος του δυναμικού μοντέλου να βρει τα θέματα που κρύβονται.

Σε αυτό το κεφάλαιο, θα αναπτύξουμε ένα δυναμικό μοντέλο που συλλαμβάνει την εξέλιξη των θεμάτων που είναι οργανωμένα σε ένα σώμα κειμένων. Κάτω από αυτό το μοντέλο τα κείμενα μπορούν να ομαδοποιηθούν με βάση τα έτη, και κάθε κείμενο ενός έτους να προέρχεται από ένα σύνολο θεμάτων που έχουν εξελιχθεί από το αμέσως προηγούμενο έτος.

Θα αναπτύξουμε κλασσικά μοντέλα χώρων για να ορίσουμε ένα μοντέλο εξέλιξης θεμάτων. Στην συνέχεια θα αναπτύξουμε αποτελεσματικές τεχνικές εκ των υστέρων συμπερασματολογίας για να υπολογίσουμε τα εξελισσόμενα θέματα από μια οργανωμένη συλλογή δεδομένων.

3.4.2 Δυναμικό Μοντέλο Θεμάτων

Ενώ η κλασσική μοντελοποίηση χρονοσειρών συγκεντρώνεται στην ανάλυση συνεχών δεδομένων, τα μοντέλα θεμάτων σχεδιάστηκαν για κατηγορικά δεδομένα. Η προσέγγισή μας είναι η μοντελοποίηση καθορισμένου χώρου στον χώρο φυσικών παραμέτρων των κρυμμένων πολυωνυμικών μεταβλητών θεμάτων, καθώς επίσης και των φυσικών παραμέτρων για τις λογαριθμικές κανονικές κατανομές που χρησιμοποιούνται στην μοντελοποίηση των ορισμένων από κείμενα αναλογιών θεμάτων.

Πρώτα, θα εξετάσουμε τις υποκείμενες στατιστικές υποθέσεις του στατικού μοντέλου θεμάτων, όπως ο αλγόριθμος LDA. Ορίζουμε τα $\beta_{1:K}$ να είναι K θέματα, το καθένα από τα οποία είναι μια κατανομή πάνω σε ένα φιξαρισμένο λεξιλόγιο. Σε ένα στατιστικό μοντέλο, κάθε κείμενο δημιουργείται από την παρακάτω διαδικασία:

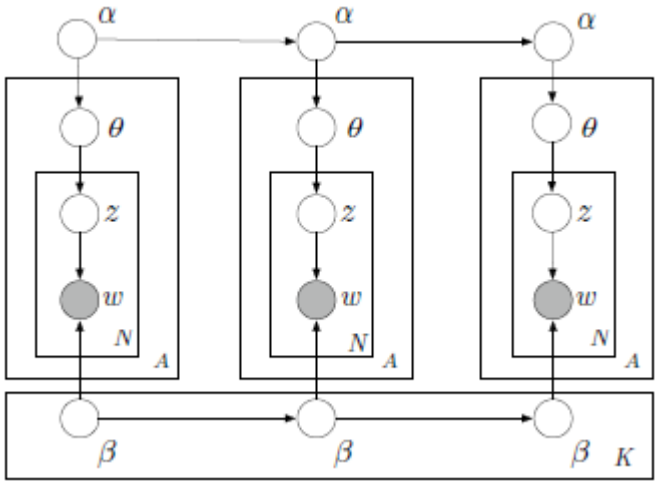
1. Choose topic proportions θ from a distribution over the $(K-1)$-simplex, such as a Dirichlet.
2. For each word:
   (a) Choose a topic assignment $Z \sim Mult(\theta)$.
   (b) Choose a word $W \sim Mult(\beta_z)$.

Αυτή η υπονοούμενη διαδικασία υποθέτει ότι τα κείμενα είναι ανταλλάξιμα σε ένα σύνολο θεμάτων. Για πολλές συλλογές, όμως, η σειρά των κειμένων αντικατοπτρίζει ένα εξελισσόμενο σύνολο θεμάτων. Σε ένα δυναμικό μοντέλο θεμάτων, θεωρούμε ότι τα δεδομένα χωρίζονται σε περιόδους χρόνου, για παράδειγμα ανά έτος. Μοντελοποιούμε τα κείμενα κάθε περιόδου με ένα K-στοιχείων μοντέλο θεμάτων, όπου τα θέματα που αφορούν την περίοδο t εξελίσσονται από τα θέματα που αφορούν την περίοδο t-1.

Για ένα K-στοιχείων μοντέλο με V όρους, ορίζουμε ως $\beta_{t,k}$ να είναι ένα V-διάνυσμα φυσικών παραμέτρων του θέματος k στην περίοδο t. Η συνηθισμένη απεικόνιση μια πολυωνυμικής κατανομής γίνεται με την παραμετροποίηση του μέσου της. Αν ορίσουμε την παράμετρο μέσου της V-διαστάσεων πολυωνυμικής κατανομής ως π, το i-οστό στοιχείο της φυσική παραμέτρου δίνεται σχεδιάζοντας το $\beta_i = log(\pi_i/\pi_V)$. Στις τυπικές εφαρμογές γλωσσικής μοντελοποίησης, οι Dirichlet κατανομές χρησιμοποιούνται για να μοντελοποιήσουν την αβεβαιότητα των κατανομών πάνω στις λέξεις. Όμως, η Dirichlet κατανομή δεν είναι υπεύθυνη για την διαδοχική μοντελοποίηση. Αντί αυτής, δένουμε τις φυσικές παραμέτρους κάθε θέματος $\beta_{t,k}$ σε ένα μοντέλο κατάστασης χώρου που εξελίσσεται με θόρυβο Gaussian. Η απλούστερη εκδοχή ενός τέτοιου μοντέλου είναι

$$\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)\,. \qquad (1)$$

Η προσέγγισή μας λοιπόν είναι να μοντελοποιήσουμε σειρές σύνθετων τυχαίων μεταβλητών δένοντας Gaussian κατανομές σε ένα δυναμικό μοντέλο και σχεδιάζοντας τις προκύπτουσες τιμές του σχήματος. Αυτή είναι μια επέκταση των λογαριθμικών κανονικών κατανομών σε δεδομένα χρονοσειρών.



Εικόνα 3.8: Γραφική απεικόνιση του δυναμικού μοντέλου θεμάτων(για τρεις χρονικές περιόδους). Οι παράμετροι κάθε θέματος $\beta_{t,k}$ εξελίσσονται καθώς περνάει ο χρόνος, μαζί με τις παραμέτρους μέσου α_t και την λογαριθμική κανονική κατανομή των αναλογιών θεμάτων.

Στον LDA, οι αναλογίες θεμάτων δημιουργούνται από μία Dirichlet κατανομή. Στο δυναμικό μοντέλο θεμάτων, χρησιμοποιούμε μια λογαριθμική κανονική με μέσο α για να εκφράσουμε την αβεβαιότητα στις αναλογίες. Η σειριακή δομή μεταξύ των μοντέλων λαμβάνεται ξανά υπόψη με ένα δυναμικό μοντέλο

$$\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)\,. \qquad (2)$$

Συσχετίζοντας τα θέματα μαζί με κατανομές θεματικών αναλογιών, έχουμε κατ' επέκταση δέσει μια συλλογή μοντέλων θεμάτων. Η γενετική διαδικασία για περίοδο t για ένα σειριακό σώμα είναι όπως ακολουθεί:

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim \textbf{Mult}(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim \textbf{Mult}(\pi(\beta_{t,z}))$.

Σημειώστε ότι το π σχεδιάζει τις πολυωνυμικές φυσικές παραμέτρους ως προς τις

παραμέτρους μέσου,

$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}.$$

Το γραφικό μοντέλο για αυτή την γενετική διαδικασία δίνεται στην Εικόνα 3.8 Αν τα οριζόντια βέλη διαγραφούν, σπάει η δυναμικότητα του χρόνου και το γραφικό μοντέλο μετατρέπεται σε ένα σύνολο ανεξάρτητων μοντέλων θεμάτων. Με την δυναμικότητα του χρόνου, το *k*-οστό θέμα την περίοδο *t* έχει *ομαλά εξελιχθεί από το k-οστό θέμα της περιόδου t-1*.

Τώρα θα συγκεντρωθούμε σε ένα μοντέλο με K δυναμικά μοντέλα που εξελίσσονται όπως στην (1), και όπου η αναλογία θεμάτων είναι φιξαρισμένη σε μία Dirichlet κατανομή. Τα τεχνικά ζητήματα που συσχετίζονται με την μοντελοποίηση των θεματικών αναλογιών στις χρονοσειρές είναι τα ίδια με εκείνα των θεμάτων που αλληλοδένονται.

3.4.3 Υπολογισμός

Δουλεύοντας με χρονοσειρές πάνω σε φυσικές παραμέτρους μας επιτρέπει την χρήση Gaussian μοντέλων για την δυναμική χρόνου. Όμως, εξαιτίας της μη συζευξιμότητας της Gaussian και των φυσικών παραμέτρων, η εκ των υστέρων συμπερασματολογία δεν μπορεί να υπολογιστεί. Σε αυτή την ενότητα, παρουσιάζουμε μια μεταβολική μέθοδο για τον υπολογισμό της εκ των υστέρων συμπερασματολογίας. Χρησιμοποιούμε μεταβολικές μεθόδους σαν μία ντετερμινιστική εναλλακτική επιλογή στην στοχαστική απεικόνιση, για να μπορέσουμε να διαχειριστούμε πολυπληθή σύνολα δεδομένων, κάτι συνηθισμένο στην ανάλυση κειμένων. Ενώ η Gibbs μέθοδος έχει εφαρμοστεί με επιτυχία στα στατικά μοντέλα θεμάτων, η μη συζευξιμότητα καθιστά αυτή την μέθοδο δύσκολα εφαρμόσιμη για το δυναμικό μοντέλο.

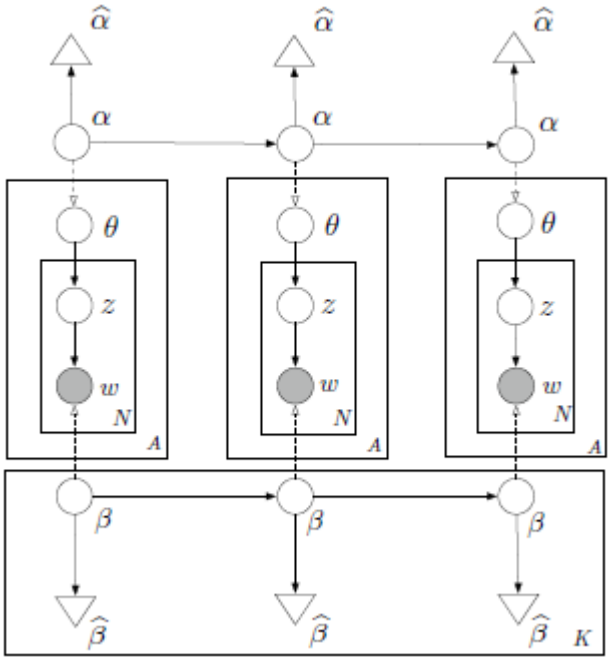
Η ιδέα πίσω από τις μεταβολικές μεθόδους είναι να βελτιστοποιήσουμε τις ελεύθερες παραμέτρους μιας κατανομής ως προς τις κρυφές μεταβλητές έτσι ώστε η κατανομή να είναι όσο το δυνατόν πιο κοντά στην Kullback-Liebler (KL) σύγκλιση της αληθινής εκ των υστέρων κατανομής. Αυτή η κατανομή μπορεί συνεπώς να χρησιμοποιηθεί για την εύρεση της αληθινής εκ των υστέρων κατανομής. Στο δυναμικό μοντέλο θεμάτων, οι κρυφές μεταβλητές είναι θέματα $\beta_{t,k}$, αναμεμειγμένων αναλογιών $\theta_{t,d}$, και δεικτών $z_{t,d,n}$. Η μεταβολική κατανομή αντικατοπτρίζει την ομαδική δομή των κρυφών μεταβλητών. Υπάρχουν μεταβολικές παράμετροι για κάθε σειρά θεμάτων πολυωνυμικών παραμέτρων, και μεταβολικές παράμετροι για κάθε μία από τις επιπέδου-κειμένων κρυφές μεταβλητές. Η υπολογισμένη μεταβολική εκ των υστέρων κατανομή είναι

$$\prod_{k=1}^{K} q(\beta_{k,1},\ldots,\beta_{k,T} \mid \hat{\beta}_{k,1},\ldots,\hat{\beta}_{k,T}) \times \prod_{t=1}^{T}\left(\prod_{d=1}^{D_t} q(\theta_{t,d} \mid \gamma_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} \mid \phi_{t,d,n})\right). \quad (3)$$

Στον συχνά χρησιμοποιούμενο mean-field υπολογισμό, κάθε κρυφή μεταβλητή θεωρείται ανεξάρτητη από τις άλλες. Στην στατιστική κατανομή των $\{\beta_{k,1},.....,\beta_{k,T}\}$, όμως, διατηρούμε την σειριακή δομή του θέματος θέτοντας ένα δυναμικό μοντέλο Gaussian "μεταβολικών παρατηρήσεων"

$$\{\hat{\beta}_{k,1}, \ldots, \hat{\beta}_{k,T}\}.$$

Αυτές οι παράμετροι έχουν κατασκευαστεί για να ελαχιστοποιήσουν την KL σύγκλιση μεταξύ της προκύπτουσας εκ των υστέρων κατανομής, η οποία είναι Gaussian, και της αληθινής εκ των υστέρων κατανομής, που δεν είναι Gaussian.



Εικόνα 3.9: Μία γραφική απεικόνιση του μεταβολικού υπολογισμού για τις χρονοσειρές του θεματικού μοντέλου στην Εικόνα 3.8. Οι μεταβολικές παράμετροι $\hat{a}$ και τύπος $\hat{\beta}$ θεωρούνται η έξοδος του Kalman φίλτρου.

Κάθε διάνυσμα αναλογίας $\theta_{t,d}$ εμπλουτισμένο με μια ελεύθερη Dirichlet παράμετρο $\gamma_{t,d}$, κάθε θεματικός δείκτης $z_{t,d,n}$ είναι εμπλουτισμένος με μια ελεύθερη πολυωνυμική παράμετρο $\varphi_{t,d,,n}$ και η βελτιστοποίηση είναι αύξουσα. Οι ενημερώσεις για τις επιπέδου-κειμένων μεταβλητές έχουν ένα κλειστό σχήμα. Χρησιμοποιούμε μια συζευγμένη βαθμωτή μέθοδο για να βελτιστοποιήσουμε τις επιπέδου-κειμένων μεταβολικές παρατηρήσεις. Ο προκύπτων στατιστικός υπολογισμός για τις φυσικές θεματικές παραμέτρους $\{\beta_{k,1},.....\beta_{k,T}\}$ ενσωματώνει την δυναμικότητα του χρόνου. Στη συνέχεια θα περιγράψουμε ένα υπολογισμό βασισμένο στο φίλτρο Kalman, και το δεύτερο βασισμένο στην wavelet παλινδρόμηση.

3.4.4 Μεταβολικό Kalman Φιλτράρισμα

Η εικόνα των μεταβολικών παραμέτρων ως αποτελέσματα βασίζεται στην συμμετρία των τιμών της Gaussian πυκνότητας, $f_{\mu,\Sigma(x)}=f_{x,\Sigma(\mu)}$, που επιτρέπει την χρησιμοποίηση forward-backward υπολογισμών για μοντέλα γραμμικής χρονικής κατάστασης. Το γραφικό μοντέλο και ο μεταβολικός υπολογισμός του δίνεται στην Εικόνα 3.9. Εδώ τα τρίγωνα δηλώνουν τις μεταβολικές παραμέτρους, Μπορείτε να τα σκεφτείτε σαν τα "υποθετικά αποτελέσματα" του φίλτρου Kalman που διευκολύνουν την επίλυση.

Για να εξηγήσουμε την κύρια ιδέα που κρύβεται πίσω από αυτή την τεχνική, σκεφτείτε ένα μοντέλο όπου unigram μοντέλα β_t εξελίσσονται στον χρόνο. Σε αυτό το μοντέλο δεν υπάρχουν θέματα, μόνο αναμεμειγμένες παράμετροι. Οι υπολογισμοί είναι απλούστερες εκδοχές εκείνων που χρειαζόμαστε για πιο γενικά μοντέλα κρυφών μεταβλητών, αλλά έχουν τα βασικά χαρακτηριστικά. Το μοντέλο χωρικής κατάστασης είναι

$$\begin{aligned} \beta_t \,|\, \beta_{t-1} &\sim \mathcal{N}(\beta_{t-1}, \sigma^2 I) \\ w_{t,n} \,|\, \beta_t &\sim \textbf{Mult}(\pi(\beta_t)) \end{aligned}$$

και σχηματίζουμε ένα στατιστικό μοντέλο χωρικής κατάστασης όπου

$$\hat{\beta}_t \,|\, \beta_t \sim \mathcal{N}(\beta_t, \hat{\nu}_t^2 I)$$

Οι μεταβολικές παράμετροι είναι το $\hat{\beta}_t$ και το $\hat{\nu}_t$. Χρησιμοποιώντας τους κλασσικούς υπολογισμούς του φίλτρου Kalman (Kalman 1960), Η μέσος και η διασπορά της εκ των υστέρων κατανομής δίνονται από

$$\begin{aligned} m_t &\equiv \mathbb{E}(\beta_t \,|\, \hat{\beta}_{1:t}) = \\ &\left(\frac{\hat{\nu}_t^2}{V_{t-1}+\sigma^2+\hat{\nu}_t^2}\right) m_{t-1} + \left(1 - \frac{\hat{\nu}_t^2}{V_{t-1}+\sigma^2+\hat{\nu}_t^2}\right) \hat{\beta}_t \end{aligned}$$

$$\begin{aligned} V_t &\equiv \mathbb{E}((\beta_t - m_t)^2 \,|\, \hat{\beta}_{1:t}) \\ &= \left(\frac{\hat{\nu}_t^2}{V_{t-1}+\sigma^2+\hat{\nu}_t^2}\right)(V_{t-1}+\sigma^2) \end{aligned}$$

με αρχικές συνθήκες m_0 και V_0. Η παλινδρόμηση υπολογίζει τον περιθωριακό μέσο και διασπορά του β_t δεδομένων των $\beta_{1:T}$ ως

$$\tilde{m}_{t-1} \equiv \mathbb{E}(\beta_{t-1} \mid \hat{\beta}_{1:T}) = \left(\frac{\sigma^2}{V_{t-1}+\sigma^2}\right) m_{t-1} + \left(1 - \frac{\sigma^2}{V_{t-1}+\sigma^2}\right) \tilde{m}_t$$

$$\tilde{V}_{t-1} \equiv \mathbb{E}((\beta_{t-1} - \tilde{m}_{t-1})^2 \mid \hat{\beta}_{1:T}) = V_{t-1} + \left(\frac{V_{t-1}}{V_{t-1}+\sigma^2}\right)^2 \left(\tilde{V}_t - (V_{t-1}+\sigma^2)\right)$$

με αρχικές συνθήκες $\tilde{m_T} = m_T$ και $\tilde{V_T} = V_T$. Υπολογίζουμε την εκ των υστέρων κατανομή $p(\beta_{1:T}|\boldsymbol{w}_{1:T})$ χρησιμοποιώντας την εκ των υστέρων κατανομή χωρικής κατάστασης $q(\beta_{1:T}| \hat{\beta_{1:T}})$. Από την ανισότητα του Jensen, η λογαριθμική πιθανότητα είναι μεγαλύτερη από

$$\log p(d_{1:T}) \geq \int q(\beta_{1:T} \mid \hat{\beta}_{1:T}) \log \left(\frac{p(\beta_{1:T})\, p(d_{1:T} \mid \beta_{1:T})}{q(\beta_{1:T} \mid \hat{\beta}_{1:T})}\right) d\beta_{1:T} \quad (4)$$

$$= \mathbb{E}_q \log p(\beta_{1:T}) + \sum_{t=1}^{T} \mathbb{E}_q \log p(d_t \mid \beta_t) + H(q)$$

3.4.5 Μεταβολική Wavelet Παλινδρόμηση

Αντί του φίλτρου Kalman μπορεί να χρησιμοποιηθεί η μεταβολική wavelet παλινδρόμηση. Αναδιαμορφώνουμε τον χρόνο ώστε να είναι μεταξύ 0 και 1. Για να είμαστε συνεπείς με την προηγούμενη ορολογία μας, υποθέτουμε ότι

$$\hat{\beta}_t = \tilde{m}_t + \hat{\nu}\epsilon_t$$

όπου $\varepsilon_t \sim N(0,1)$. Ο αλγόριθμος της μεταβολικής wavelet παλινδρόμησης υπολογίζει τα $\{\hat{\beta}_\tau\}$, που είναι τα παρατηρούμενα δεδομένα όπως ακριβώς στην μέθοδο του Kalman φίλτρου, όπως επίσης και τον θόρυβο formula $\hat{\nu}$.

Πιο συγκεκριμένα θα δείξουμε αυτή την τεχνική χρησιμοποιώντας την Haar wavelet βάση. Οι Daubechies wavelets χρησιμοποιούνται στα παραδείγματά μας. Το μοντέλο λοιπόν είναι

$$\hat{\beta}_t = \alpha\phi(x_t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} D_{jk}\psi_{jk}(x_t)$$

όπου $x_t = t/n$. $\varphi(x) = 1$ για $0 \le x \le 1$,

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \le x \le \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < x \le 1 \end{cases}$$

και $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. Ο μεταβολικός υπολογισμός για τον εκ των υστέρων μέσο γίνεται

$$\tilde{m}_t = \hat{\alpha}\phi(x_t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{D}_{jk}\psi_{jk}(x_t).$$

όπου $\hat{\alpha} = n^{-1}\sum_{t=1}^{n}\hat{\beta}_t$, και το $\hat{D}_{jk}$ υπολογίζονται από τους συντελεστές

$$Z_{jk} = \frac{1}{n}\sum_{t=1}^{n}\hat{\beta}_t\psi_{jk}(x_t).$$

Για να υπολογίσουμε το $\hat{\beta}_t$ θα χρησιμοποιήσουμε μια βαθμωτή αύξουσα, όπως και στο φιλτράρισμα Kalman, απαιτώντας τα παράγωγα $\theta\tilde{m}/\theta\tilde{\beta}_t$. Αν απαλύνουμε το threshholding που χρησιμοποιείται, θα έχουμε

$$\frac{\partial \tilde{m}_t}{\partial \hat{\beta}_s} = \frac{\partial \hat{\alpha}}{\partial \hat{\beta}_s}\phi(x_t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\frac{\partial \hat{D}_{jk}}{\partial \hat{\beta}_s}\psi_{jk}(x_t).$$

με $\theta\tilde{\alpha}/\theta\tilde{\beta}_s = n^{-1}$ και

$$\partial \hat{D}_{jk}/\partial \hat{\beta}_s = \begin{cases} \frac{1}{n}\psi_{jk}(x_s) & \text{if } |Z_{jk}| > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

Σημειώστε επίσης ότι $|Z_{jk}| > \lambda$ αν και μόνο αν $|\hat{D}_{jk}| > 0$. Αυτά τα παράγωγα μπορούν να υπολογιστούν χρησιμοποιώντας έτοιμα προγράμματα για την μετατροπή της wavelet σε μια οποιωνδήποτε από τις κλασσικές wavelet βάσεις.

3.4.6 Δυναμικό Θεματικό Μοντέλο Συνεχούς Χρόνου

Το δυναμικό DTM χρησιμοποιεί ένα μοντέλο κατάστασης χώρου στις φυσικές παραμέτρους των πολυωνυμικών κατανομών που απεικονίζουν τα θέματα. Αυτό απαιτεί ο χρόνος να διαμερίζεται σε πολλές περιόδους, και μέσα σε κάθε περίοδο το LDA χρησιμοποιείται για να μοντελοποιήσει τα κείμενα. Παρόλο που το δυναμικό DTM είναι ένα ισχυρό μοντέλο, η επιλογή της διαμέρισης επηρεάζει τις απαιτήσεις μνήμης και την υπολογιστική πολυπλοκότητα της εκ των υστέρων συμπερασματολογίας. Αυτό καθορίζει σε μεγάλο βαθμό την ανάλυση, όπου θα προσαρμόσουμε το μοντέλο.

Για να λύσουμε το πρόβλημα της διαμέρισης, θεωρούμε τον χρόνο ως συνεχόμενο. Το δυναμικό μοντέλο θεμάτων συνεχούς χρόνου (cDTM) αντικαταστεί το διακριτό μοντέλο κατάστασης χώρου του dDTM με την γενίκευση του, σε κίνηση Brown. Το cDTM γενικεύει το dDTM διαμερίζοντας μόνο την ανάλυσή με την οποία ο χρόνος διατρέχει τα κείμενα που μετρούνται.

Το cDTM μοντέλο εισαγάγει γενικώς πολύ περισσότερες κρυφές μεταβλητές από το dDTM. Όμως αυτό το φαινομενικά πολύπλοκο μοντέλο είναι απλούστερο και πιο αποτελεσματικό. Όπως θα δούμε παρακάτω, κάτω από αυτό το μοντέλο η διαδικασία της μεταβολικής εκ των υστέρων συμπερασματολογίας μπορεί να έχει ως πλεονέκτημα την φυσική αραιότητα του κειμένου, του γεγονότος, δηλαδή, ότι οι λέξεις χρησιμοποιούνται σε κάθε μετρούμενο βήμα χρόνου. Στην πραγματικότητα, όσο η ανάλυση γίνεται καλύτερη, όλο και λιγότερες λέξεις χρησιμοποιούνται.

Αυτό παρέχει μια συμπερασματολογική επιτάχυνση που κάνει το μοντέλο δυνατό να προσαρμοστεί σε διάφορους βαθμούς ανάλυσης. Σαν παραδείγματα, τα δημοσιογραφικά άρθρα μπορεί να είναι ανταλλάξιμα μέσα σε μία έκδοση, μία υπόθεση που μπορεί να είναι πιο ρεαλιστική από μία που υποθέτει ότι είναι ανταλλάξιμα σε ένα χρόνο. Άλλα δεδομένα, όπως οι ειδήσεις, μπορεί να διατρέχουν χρονικές περιόδους χωρίς κάποια αλλαγή. Ενώ το dDTM απαιτεί την απεικόνιση όλων των θεμάτων σε διακριτές στιγμές αυτών των περιόδων, το cDTM μπορεί να αναλύσει τέτοια δεδομένα χωρίς να θυσιάσει μνήμη ή

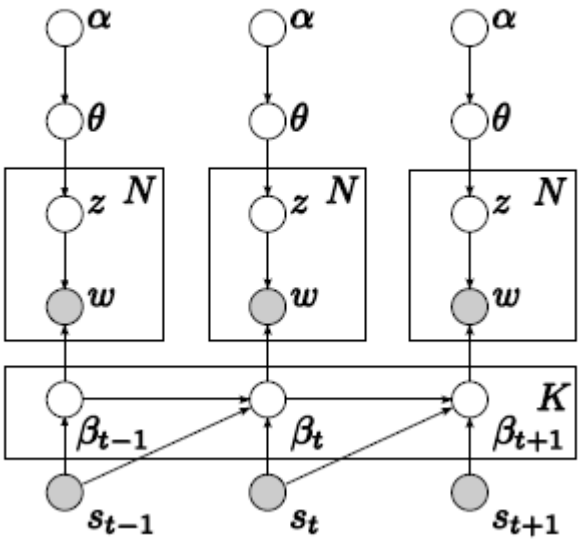
ταχύτητα. Με το cDTM, ο βαθμός ανάλυσης, μπορεί να επιλεχθεί ώστε να μεγιστοποιήσει την προσαρμογή του μοντέλου από το να περιορίσει την υπολογιστική του πολυπλοκότητα.

Σημειώνουμε ότι τα cDTM και dDTM δεν είναι τα μόνα θεματικά μοντέλα που λαμβάνουν τον χρόνο υπόψη τους. Τα Θέματα μοντέλων χρόνου (TOT) και τα δυναμικά mixture μοντέλα (DMM) συμπεριλαμβάνουν στιγμιότυπα στην ανάλυση των κειμένων. Το TOT μοντέλο αντιμετωπίζει τις χρονοσφραγίδες ως απεικονίσεις κρυφών θεμάτων, και το DMM υποθέτει ότι οι θεματικές mixture αναλογίες κάθε κειμένου εξαρτώνται από τις προηγούμενες θεματικές mixture αναλογίες. Και στα δύο μοντέλα TOT και DMM, τα θέματα είναι σταθερά, και η πληροφορία χρόνου χρησιμοποιείται για να τα ανακαλύψει.

3.4.7 Περιγραφή cDTM

Σε μία συλλογή κειμένων χρονικά διαμερισμένη, θα θέλαμε να μοντελοποιήσουμε τα κρυφά θέματα που αλλάζουν την κατεύθυνση της συλλογής. Σε δεδομένα ειδήσεων, για παράδειγμα, ένα απλό θέμα θα αλλάξει, καθώς οι ιστορίες που σχετίζονται με αυτό εξελίσσονται. Το δυναμικό θεματικό μοντέλο διακριτού χρόνου (dDTM) κατασκευάζεται πάνω σε ένα ανταλλάξιμο μοντέλο θεμάτων για να παρέχει έναν τέτοιο μηχανισμό. Στο dDTM, τα κείμενα χωρίζονται σε σειριακές ομάδες, και τα θέματα κάθε περιόδου εξελίσσονται από τα θέματα της προηγούμενης περιόδου. Τα κείμενα σε μία ομάδα θεωρούνται ανταλλάξιμα.

Πιο συγκεκριμένα, ένα θέμα απεικονίζεται ως μια απεικόνιση σε ένα φιξαρισμένο λεξιλόγιο της συλλογής. Το dDTM υποθέτει ότι ένα μοντέλο κατάστασης χώρου διακριτού χρόνου ελέγχει την εξέλιξη των φυσικών παραμέτρων των πολυωνυμικών κατανομών που απεικονίζουν τα θέματα. Αυτή είναι μία επέκταση χρονοσειρών στην λογαριθμική κανονική κατανομή.



Εικόνα 3.10: Απεικόνιση γραφικού μοντέλου του cDTM. Η εξέλιξη των θεματικών παραμέτρων β_t ελέγχεται από την κίνηση Brown. Η μεταβλητή s_t είναι το παρατηρούμενο στιγμιότυπο του κειμένου d_t.

Ένα μειονέκτημα του dDTM είναι ότι ο χρόνος είναι διαμερισμένος. Αν η ανάλυση έχει επιλεχθεί να είναι πολύ τραχιά, τότε η υπόθεση ότι τα κείμενα μέσα σε ένα χρονικό βήμα είναι ανταλλάξιμα δεν θα είναι αληθινή. Αν η ανάλυση είναι πολύ καλή, τότε ο αριθμός των μεταβολικών παραμέτρων θα εκτινάσσεται καθώς περισσότερα στιγμιότυπα προστίθενται. Η επιλογή διαμέρισης πρέπει να είναι μία απόφαση που θα βασίζεται στις υποθέσεις σχετικά με τα δεδομένα. Όμως, οι υπολογιστικές απαιτήσεις μπορεί να εμποδίσουν την ανάλυση στο κατάλληλη κλίμακα χρόνου.

Στο cDTM, θα απεικονίζουμε τα θέματα στην φυσική τους παραμετροποίηση, αλλά θα χρησιμοποιήσουμε την κίνηση Brown, για να μοντελοποιήσουμε την εξέλιξη δια μέσου του χρόνου. Θεωρούμε τα *i,j* (*j > i > 0*) να είναι δύο αφηρημένοι δείκτες χρόνου, τα s_i και s_j να είναι οι χρονοσφραγίδες, και $\Delta_{sj,si}$ να είναι ο παρερχόμενος χρόνος ανάμεσα τους. Σε ένα cDTM μοντέλο K θεμάτων, η κατανομή της k-οστής $(1 \le k \le K)$ θεματικής παραμέτρου στον όρο *w* είναι:

$$\begin{aligned} \beta_{0,k,w} &\sim \mathcal{N}(m, v_0) \\ \beta_{j,k,w}|\beta_{i,k,w}, s &\sim \mathcal{N}\left(\beta_{i,k,w}, v\Delta_{s_j,s_i}\right), \end{aligned} \tag{5}$$

όπου η διασπορά μεγαλώνει γραμμικά με την καθυστέρηση.

Αυτή η κατασκευή χρησιμοποιείται σαν συστατικό της όλης γενετικής διαδικασίας. (Σημειώστε ότι: Αν j = i + 1, γράφουμε $\Delta_{sj,si}$ σαν Δ_{sj} για συντομία.

1. For each topic $k, 1 \le k \le K$,
   (a) Draw $\beta_{0,k} \sim \mathcal{N}(m, v_0 I)$.
2. For document d_t at time s_t $(t > 0)$:
   (a) For each topic $k, 1 \le k \le K$,
      i. From the Brownian motion model, draw $\beta_{t,k}|\beta_{t-1,k}, s \sim \mathcal{N}(\beta_{t-1,k}, v\Delta_{s_t} I)$.
   (b) Draw $\theta_t \sim \text{Dir}(\alpha)$.
   (c) For each word,
      i. Draw $z_{t,n} \sim \text{Mult}(\theta_t)$.
      ii. Draw $w_{t,n} \sim \text{Mult}(\pi(\beta_{t,z_{t,n}}))$.

Η συνάρτηση π σχεδιάζει τις πολυωνυμικές φυσικές παραμέτρους, οι οποίες είναι απεριόριστες, στις παραμέτρους μέσου, όπου βρίσκονται στο simplex,

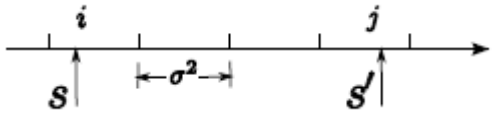
$$\pi(\beta_{t,k})_w = \frac{\exp(\beta_{t,k,w})}{\sum_w \exp(\beta_{t,k,w})}. \tag{6}$$

Το cDTM απεικονίζεται στην Εικόνα 3.10.

Το cDTM μπορεί να κατανοηθεί και ως μία γενίκευση του dDTM. Και τα δύο μοντέλα η λογαριθμική πιθανότητα ενός όρου εκθέτει τη διασπορά ενός χρονικού διαστήματος μεταξύ των παρατηρήσεων. Στο dDTM, αυτό το διάστημα χωρίζεται ισάξια σε διακριτά κομμάτια. Μία παράμετρος ελέγχει την διασπορά για κάθε χρονικό κομμάτι, και την διασπορά διαμέσου ενός ολόκληρου διαστήματος είναι αυτή η παράμετρος πολλαπλασιασμένη με τον αριθμό των κομματιών. Ως συνέπεια αυτής της απεικόνισης, το θέμα, δηλαδή, η πλήρης απεικόνιση των όρων είναι ρητά απεικονισμένη σε κάθε κομμάτι. Για fine-grained χρονοσειρές, οδηγούμαστε σε υψηλής απαίτησης μνήμη για την εκ των υστέρων συμπερασματολογία, ακόμα και αν οι παρατηρήσεις είναι αραιά διαμερισμένες σε όλο τον χρόνο.

Στο cDTM, όμως, η διασπορά είναι μια συνάρτηση της καθυστέρησης μεταξύ των παρατηρήσεων, και οι πιθανότητες στα διακριτά βήματα αυτών των παρατηρήσεων δεν χρειάζεται να ληφθούν υπόψη. Τα συμπεράσματα όπως θα δούμε και παρακάτω, μπορούμε να τα διαχειριστούμε σποραδικά. Ένα dDTM μπορεί να παραχθεί από ένα cDTM μετρώντας τα στιγμιότυπα των κειμένων στον επιθυμητό βαθμό ανάλυσης.

Παρόμοια με την κίνηση Brown ως η διαδικασία εύρεσης ορίου του Gaussian τυχαίου περιπάτου διακριτού χρόνου, το cDTM είναι η διαδικασία εύρεσης ορίου του dDTM. Σημειώστε ότι η ανά χρονικό κομμάτι διασπορά στο dDTM είναι σ^2, και ότι είναι μία συνάρτηση του βαθμού ανάλυσης. Το cDTM είναι ένα μοντέλο εύρεσης ορίου καθώς το σ^2 πλησιάζει το μηδέν. Είναι σημαντικό να ειπωθεί ότι με το cDTM , δεν χρειάζεται να απεικονίσουμε τις λογαριθμικές πιθανότητες των χρονικών στιγμών ανάμεσα στα γνωστά κείμενα. Αυτή η προοπτική δίνεται στην Εικόνα 3.11



Εικόνα 3.11: Τα κείμενα είναι διαθέσιμα μόνο στις στιγμές s και s' , και δεν δίνεται κανένα κείμενο ανάμεσα σε αυτές. Όταν $\sigma^2 \rightarrow 0$, το dDTM γίνεται ένα cDTM, και δεν χρειαζόμαστε πλέον να απεικονίσουμε τα βήματα ανάμεσα στο i και στο j.

3.4.8 Σποραδική μεταβολική συμπερασματολογία

Το κύριο πρόβλημα στην μοντελοποίηση θεμάτων είναι η εκ των υστέρων συμπερασματολογία, δηλαδή, η εύρεση της κατανομής της κρυφής δομής θεμάτων δεδομένων των γνωστών κειμένων. Στα σειριακά θεματικά μοντέλα, αυτή η δομή περιέχει τις ανά κείμενο θεματικές αναλογίες θd, τις ανά λέξη θεματικές εναποθέσεις $z_{d,n}$ και τις K σειρές των θεματικών κατανομών $\beta_{t,k.}$ Η αληθινή εκ των υστέρων κατανομή δεν είναι υπολογίσιμη. Πρέπει να περιοριστούμε σε μία προσέγγισή.

Πολλές προσεγγιστικές μέθοδοι συμπερασματολογίας έχουν αναπτυχθεί για τα θεματικά μοντέλα. Οι πιο ευθέως γνωστές είναι η μεταβολική συμπερασματολογία και το collapsed Gibbs sampling. Στα σειριακά μοντέλα το collapsed Gibbs sampling δεν είναι μία σωστή επιλογή γιατί η κατανομή των λέξεων για κάθε θέμα δεν είναι συζυγής με τις πιθανότητες των λέξεων. Για αυτόν το λόγο, θα χρησιμοποιήσουμε μεταβολικές μεθόδους.

Η κύρια ιδέα πίσω από τις μεταβολικές μεθόδους είναι να θέσουμε μία οικογένεια κατανομών ως προς τις κρυφές μεταβλητές, που να απεικονίζονται ως μεταβολικές

παράμετροι, και να βρούμε το μέλος αυτής της οικογένειας που είναι πιο κοντά στην Kullback-Leibler απόκλιση της αληθινής εκ των υστέρων κατανομής.

Για το cDTM που περιγράψαμε παραπάνω, προσαρμόζουμε το μεταβολικό Kalman φιλτράρισμα με ρυθμίσεις συνεχόμενου χρόνου. Για απλοποίηση, υποθέτουμε ότι ένα κείμενο υπάρχει για κάθε στιγμιότυπο. Στον αλγόριθμο, η μεταβολική κατανομή ως προς τις κρυφές μεταβλητές είναι:

$$q(\beta_{1:T}, z_{1:T,1:N}, \theta_{1:T} \mid \hat{\beta}, \phi, \gamma) = \prod_{k=1}^{K} q(\beta_{1,k}, \ldots, \beta_{T,k} | \hat{\beta}_{1,k}, \ldots, \hat{\beta}_{T,k}) \times \prod_{t=1}^{T} \left(q(\theta_t | \gamma_t) \prod_{n=1}^{N_t} q(z_{t,n} | \phi_{t,n}) \right). \quad (7)$$

Οι μεταβολικές παράμετροι είναι μία Dirichlet γ_t, για τις ανά κείμενο θεματικές αναλογίες, οι πολυωνυμικές φ για κάθε εναπόθεση λέξης σε ένα θέμα, και οι $\hat{\beta}$ μεταβλητές που είναι οι "παρατηρήσεις" σε ένα μεταβολικό Kalman φίλτρο.

Αυτές οι μεταβλητές προσαρμόζονται έτσι ώστε η προσεγγιστική εκ των υστέρων κατανομή να είναι κοντά στην αληθινή. Από το μεταβολικό Kalman φίλτρο, οι $\beta_{k,t}$, $1 \le t \le T$ μεταβλητές διατηρούν την αλυσιδωτή δομή στην μεταβολική κατανομή. Η μεταβολική συμπερασματολογία προχωράει αυξητικά, ανανεώνοντας κάθε μία από αυτές τις παραμέτρους για να ελαχιστοποιήσει την KL απόσταση ανάμεσα στην αληθινή εκ των υστέρων κατανομή και την μεταβολική εκ των υστέρων κατανομή.

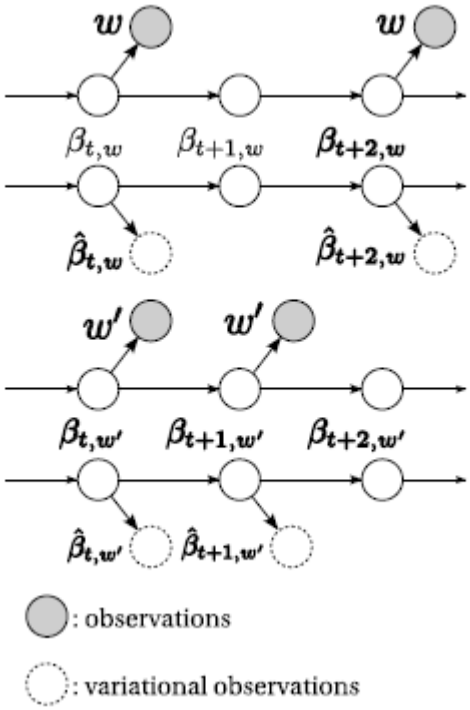
Για απλοποίηση, θεωρούμε ένα μοντέλο με ένα μόνο θέμα. Αυτοί οι υπολογισμοί είναι απλούστερες εκδοχές αυτών που χρειαζόμαστε για το πιο γενικό μοντέλο κρυφών μεταβλητών, δείχνουν όμως τα βασικά χαρακτηριστικά του αλγορίθμου. Για το cDTM, υποθέτουμε ότι υπάρχει μία παρόμοια μεταβολική κατανομή, με τις ίδιες μεταβολικές Dirichlet και μεταβολικές πολυωνυμικές, για τις ανά κείμενο, μεταβλητές.

Σε γενικές γραμμές, μπορούμε άμεσα να χρησιμοποιήσουμε τον αλγόριθμο μεταβολικού Kalman φιλτραρίσματος για το cDTM αντικαθιστώντας το μοντέλο κατάστασης χώρου με την κίνηση Brown. Θεωρούμε το V να είναι το μέγεθος του λεξιλογίου. Αυτό θα κρατάει VT μεταβολικές παραμέτρους στα διανύσματα $\beta_{1:T}$. Όταν τα T και V είναι μεγάλα, όπως σε ένα υψηλής ανάλυσης μοντέλο, η εκ των υστέρων συμπερασματολογία θα απαιτεί μεγάλες ποσότητες χρόνου και μνήμης. Έτσι, αναπτύσσουμε μία διαδικασία σποραδικής μεταβολικής συμπερασματολογίας, όπου βελτιώνει σημαντικά την πολυπλοκότητα χωρίς να θυσιάζει ακρίβεια.

Η κύρια ιδέα είναι ότι ο αλγόριθμος του σποραδικού Kalman φιλτραρίσματος είναι ότι αν οι σίγουρες $\beta_{t,w}$ δεν περιγράφουν εκδοχές όρων, δηλαδή, δεν υπάρχουν παρατηρήσεις των w και r, τότε η αληθινή εκ των υστέρων κατανομή των $\beta_{t,w}$ δεν καθορίζεται μόνο από τις παρατηρήσεις των άλλων λέξεων στην συγκεκριμένη χρονική στιγμή. Έτσι, δεν χρειάζεται να απεικονίσουμε τα $\hat{\beta}_{t,w}$ για εκείνα τα w που δεν είναι γνωστά.

Η Εικόνα 3.12 δείχνει την ιδέα πίσω από την σποραδική μεταβολική συμπερασματολογία για το cDTM. Στην Εικόνα 3.12, η μεταβολική εκ των υστέρων κατανομή είναι η λογαριθμική πιθανότητα για μία λέξη $\beta_{t,w}$ καθορίζεται από τις μεταβολικές μεταβλητές των ήδη παρακρατηθέντων λέξεων. Η πιθανότητα της μεταβολικής παρατήρησης τύπος $\hat{\beta}_{t,w}$ δεδομένου του $\beta_{t,w}$ είναι μία Gaussian:

$$\hat{\beta}_{t,w}|\beta_{t,w} \sim \mathcal{N}(\beta_{t,w}, \hat{v}_t). \quad (8)$$



Εικόνα 3.12: Ένα απλοποιημένο γραφικό μοντέλο που δείχνει πως η σποραδική μεταβολική συμπερασματολογία δουλεύει με ένα μόνο θέμα. Σημειώστε ότι η διαδικασία γενίκευσης χρειάζεται κανονικοποίηση στο β_t σύμφωνα με την εξίσωση 6, αυτό βέβαια δεν επηρεάζει την σποραδική λύση. Για τον όρο w, δεν υπάρχουν παρατηρήσεις στον χρονικό δείκτη t+1, οι ανταποκρινόμενες μεταβολικές παρατηρήσεις δεν εμφανίζονται στον χρόνο t+1. Για τον όρο w', δεν υπάρχουν παρατηρήσεις στον χρόνο t+2, οι ανταποκρινόμενες μεταβολικές παρατηρήσεις δεν εμφανίζονται στον χρόνο t+2.

Στην συνέχεια περιγράφουμε τον forward-backward αλγόριθμο για το σποραδικό Kalman φίλτρο, που χρειάζεται για να υπολογίσουμε τις προσδοκώμενες τιμές για την ανανέωση των μεταβολικών παραμέτρων. Για κάθε γνωστό όρο w, η μεταβολική forward κατανομή $p(\beta_{t,w}|\hat{\beta}_{i,i\leq t,w})$ είναι μία Gaussian και μπορεί να χαρακτηριστεί όπως ακολουθεί.

$$\begin{aligned} \beta_{t,w}|\hat{\beta}_{i,i\leq t,w} &\sim \mathcal{N}(m_{t,w}, V_{t,w}) \\ m_{t,w} &= \mathbb{E}(\beta_{t,w}|\hat{\beta}_{i,i\leq t,w}) \\ V_{t,w} &= \mathbb{E}((\beta_{t,w} - m_{t,w})^2|\hat{\beta}_{i,i\leq t,w}). \end{aligned} \quad (9)$$

Αν το w δεν είναι γνωστό στο βήμα χρόνου t τότε

$$\begin{aligned} \beta_{t,w}|\hat{\beta}_{i,i\leq t,w} &\sim \mathcal{N}(m_{t,w}, V_{t,w}) \\ m_{t,w} &= \mathbb{E}(\beta_{t,w}|\hat{\beta}_{i,i\leq t,w}) \\ V_{t,w} &= \mathbb{E}((\beta_{t,w} - m_{t,w})^2|\hat{\beta}_{i,i\leq t,w}). \end{aligned} \quad (10)$$

κάτι που σημαίνει ότι ο forward μέσος παραμένει ο ίδιος με το προηγούμενο βήμα. Διαφορετικά,

$$\begin{aligned} m_{t,w} &= \frac{\hat{\beta}_{t,w} P_{t,w} + \hat{v}_t m_{t-1,w}}{P_{t,w} + \hat{v}_t} \\ V_{t,w} &= \hat{v}_t \frac{P_{t,w}}{P_{t,w} + \hat{v}_t} \\ \hat{\beta}_{t,w} | \hat{\beta}_{i,i \le t-1,w} &\sim \mathcal{N}(m_{t-1,w}, P_{t,w} + \hat{v}_t). \end{aligned} \tag{11}$$

Παρομοίως, η μεταβολική backward κατανομή τύπος $p(\beta_{t,w}|\hat{\beta}_{i,i\le T,w})$ είναι επίσης μία Gaussian:

$$\begin{aligned} \beta_{t,w} | \hat{\beta}_{i,i \le T,w} &\sim \mathcal{N}(\widetilde{m}_{t,w}, \widetilde{V}_{t,w}) \\ \widetilde{m}_{t,w} &= \mathbb{E}(\beta_{t,w} | \hat{\beta}_{i,i \le T,w}) \\ \widetilde{V}_{t,w} &= \mathbb{E}((\beta_{t,w} - \widetilde{m}_{t,w})^2 | \hat{\beta}_{i,i \le T,w}). \\ \widetilde{m}_{t-1,w} &= m_{t-1,w} \frac{f_t v}{P_{t,w}} + \widetilde{m}_{t,w} \frac{V_{t-1,w}}{P_{t,w}} \\ \widetilde{V}_{t-1,w} &= V_{t-1,w} + \frac{V_{t-1,w}^2}{P_{t,w}^2} (\widetilde{V}_{t,w} - P_{t,w}) \end{aligned} \tag{12}$$

Γνωρίζοντας τον forward-backward υπολογισμό, στρέφουμε την προσοχή μας στην βελτιστοποίηση των μεταβολικών παρατηρήσεων τύπος $\hat{\beta}_{w,k}$. Ισοδύναμη με την ελαχιστοποίηση της KL απόστασης είναι η προσέγγισή του ορίου της πιθανοφάνειας των παρατηρήσεων που δίνεται από την ανισότητα του Jensen

$$\mathcal{L}(\hat{\beta}) \ge \sum_{t=1}^{T} \mathbb{E}_q \left[(\log p(\mathbf{w}_t | \beta_t) + \log p(\beta_t | \beta_{t-1}) \right] + H(q), \tag{13}$$

όπου το H(q) είναι η εντροπία. Αυτό απλοποιείται στην

$$\begin{aligned} \mathcal{L}(\hat{\beta}) \quad \ge \quad & \sum_{t=1}^{T} \mathbb{E}_q \left[\log p(\mathbf{w}_t | \beta_t) - \log q(\hat{\beta}_t | \beta_t) \right] \\ & + \sum_{t=1}^{T} \log q(\hat{\beta}_t | \hat{\beta}_{i,i \le t-1}), \end{aligned} \tag{14}$$

Χρησιμοποιούμε τύπος $\hat{\delta_{t,w}}=1$ ή 0 για να δηλώσουμε αν το τύπος $\hat{\beta_{t,w}}$ είναι μέσα στις μεταβολικές παρατηρήσεις ή όχι. Έτσι οι όροι είναι

$$\begin{aligned}
\mathbb{E}_q \log q(\mathbf{w}_t|\beta_t) &\geq \sum_w n_{t,w}\tilde{m}_{t,w} \\
&\quad -n_t \log \sum_w \exp(\tilde{m}_{t,w} + \tilde{V}_{t,w}/2) \\
\mathbb{E}_q \log p(\hat{\beta}_t|\beta_t) &= \sum_w \delta_{t,w}\mathbb{E}_q \log q(\hat{\beta}_{t,w}|\beta_{t,w}) \\
\log q(\hat{\beta}_t|\hat{\beta}_{i,i\leq t-1}) &= \sum_w \delta_{t,w} \log q(\hat{\beta}_{t,w}|\hat{\beta}_{i,i\leq t-1,w}).
\end{aligned}$$

ο αριθμός *w* στο κείμενο d_t είναι $n_{t,w}$ και $n_t = \sum_w n_{t,w}$.

Έτσι για να βελτιστοποιήσουμε τις μεταβολικές παρατηρήσεις, χρειάζεται να υπολογίσουμε $\partial\mathcal{L}/\partial\hat{\beta}_{t,w}$ παράγωγο για δt,w = 1. Η βασική απαίτηση $\mathcal{O}(\sum_t \sum_w \delta_{t,w})$ που είναι το άθροισμα των μοναδικών όρων σε κάθε στιγμιότυπο, είναι συνήθως μικρότερη από $\mathcal{O}(VT)$, την απαίτηση μνήμης, δηλαδή, του πυκνά απεικονισμένου αλγορίθμου. Επίσημα, μπορούμε να ορίσουμε την σποραδικότητα ενός συνόλου δεδομένων να είναι

$$\text{σποραδικότητα} = 1 - \left(\sum_t \sum_w \delta_{t,w}\right)/(VT), \qquad (15)$$

Σαν παράδειγμα της επιτάχυνσης που επιφέρει η σποραδική μεταβολική συμπερασματολογία, σκεφτείτε το σώμα Science από το 1880 μέχρι το 2002, το οποίο περιέχει 6243 άρθρα περιοδικών. Στο dDTM τα κείμενα χωρίστηκαν ανά έτος. Για να αναλυθούν στην βέλτιστη κλίμακα, δηλαδή άρθρο ανα άρθρο, χρειάζονται 6243 στιγμιότυπα. Με ένα λεξιλόγιο μεγέθους 5000, για εξαγωγή 10 θεμάτων, ο cDTM απαιτεί 0.8G μνήμη ενώ ο dDTM απαιτεί 2.3G μνήμη, σχεδόν 3 φορές μεγαλύτερη. Η σποραδικότητα του σώματος Science είναι 0.65. Αυτό σημαίνει ότι ένας όρος εμφανίζεται στο ένα τρίτο περίπου των συνολικών στιγμιότυπων.

3.5 Επιβλεπόμενο Μοντέλο Θεμάτων

3.5.1 Εισαγωγή

Τα περισσότερα μοντέλα, όπως το LDA, είναι μη επιβλεπόμενα: μόνο οι λέξεις των κειμένων μοντελοποιούνται. Ο στόχος μας είναι να βρούμε θέματα που μεγιστοποιούν την πιθανότητα να αντιπροσωπεύουν μια συλλογή. Σε αυτό το κεφάλαιο, αναπτύσσουμε επιβλεπόμενα μοντέλα θεμάτων, όπου κάθε κείμενο συνδυάζεται με μια απόκριση. Ο στόχος είναι να βρούμε τα κρυφά θέματα που προβλέπουν αυτή την απόκριση. Δεδομένου ενός κείμενου χωρίς ετικέτες, συμπεραίνουμε την θεματική του δομή χρησιμοποιώντας ένα κατάλληλο μοντέλο, και μετά σχηματίζουμε την πρόβλεψή του. Σημειώστε ότι η απόκριση δεν είναι περιορισμένη σε κατηγορίες κειμένων. Άλλα είδη κειμένων-αποκρίσεων σωμάτων εμπεριέχουν εκθέσεις με βαθμούς, επισκοπήσεις ταινιών με βαθμολογία, και σελίδες διαδικτύου με μετρήσεις του πόσα μέλη της διαδικτυακής κοινότητας βρήκαν αυτή την ιστοσελίδα ενδιαφέρουσα

Το απλό LDA κατασκευάζει μια αρχικοποίηση. Η ελπίδα ήταν ότι τα LDA θέματα θα ήταν χρήσιμα για κατηγοριοποίηση, αφού λειτουργούν μειώνοντας την διάσταση των δεδομένων. Όμως, όταν ο στόχος είναι η πρόβλεψη, η χρησιμοποίηση μη επιβλεπόμενων θεμάτων μπορεί να μην είναι μια καλή επιλογή. Σκεφτείτε ότι θέλουμε να προβλέψουμε την βαθμολογία μια ταινίας με λέξεις στην επισκόπησή της. Ενστικτωδώς, τα καλά θέματα πρόβλεψης θα ξεχώριζαν λέξεις όπως “άριστη”, “απαίσια”, και “μέτρια”, χωρίς να δώσουν βάση στο είδος της ταινίας. Αλλά τα θέματα που έχουν υπολογιστεί από ένα μη επιβλεπόμενο μοντέλο δεν θα ανταποκρίνονται σωστά σε είδη ταινιών, αν αυτό είναι η κυρίαρχη δομή του σώματος.

Η διάκριση μεταξύ μη επιβλεπόμενων και επιβλεπόμενων μοντέλων καθρεφτίζεται στις υπάρχουσες τεχνικές μείωσης των διαστάσεων. Για παράδειγμα, σκεφτείτε να εφαρμοστεί παλινδρόμηση σε μη επιβλεπόμενα βασικά στοιχεία έναντι μεθόδου ελαχίστων τετραγώνων, που και τα δύο ψάχνουν για συσχετιζόμενους γραμμικούς συνδυασμούς και είναι οι πιο αποτελεσματικές μέθοδοι στις πρόβλεψη μεταβλητών απόκρισης. Αυτές οι γραμμικές επιβλεπόμενες μέθοδοι έχουν μη παραμετρικά αναλογικά, όπως μια προσέγγιση βασισμένη στον πυρήνα ICA.

3.5.2 Επιβλεπόμενο LDA

Στα θεματικά μοντέλα, αντιμετωπίζουμε τις λέξεις ενός κειμένου ως προερχόμενες από ένα σύνολο κρυφών θεμάτων, το οποίο είναι ένα σύνολο, στη ουσία, άγνωστων κατανομών πάνω σε ένα λεξιλόγιο. Τα κείμενα σε ένα σώμα μοιράζονται το ίδιο σύνολο K θεμάτων, αλλά κάθε κείμενο χρησιμοποιεί ένα μείγμα από θέματα το καθένα από τα οποία είναι μοναδικό. Έτσι, τα μοντέλα θεμάτων είναι μια πιο απλή μορφή των κλασσικών mixture μοντέλων κειμένων, όπου συσχετίζουν κάθε θέμα με ένα άγνωστο θέμα.

Εδώ θα κατασκευάσουμε το μοντέλο μας σύμφωνα με το LDA. Στο LDA, αντιμετωπίζουμε τις θεματικές αναλογίες για ένα κείμενο σαν μια επιλογή από την Dirichlet κατανομή. Αποκτούμε τις λέξεις του κειμένου με τον να διαλέγουμε συνεχώς ένα θέμα από αυτές τις αναλογίες, και μετά διαλέγοντας της λέξη από το συγκεκριμένο θέμα.

Στον επιβλεπόμενο LDA (sLDA), προσθέτουμε στον LDA μια μεταβλητή απόκρισης που συσχετίζεται με κάθε κείμενο. Όπως είπαμε και πριν, αυτή η μεταβλητή μπορεί να είναι ένας αριθμός από αστέρια που δίνονται σε μια ταινία, ο αριθμός των χρηστών μια on-line κοινότητας που βρήκε μια σελίδα ενδιαφέρουσα, ή η κατηγορία ενός κειμένου. Θα μοντελοποιήσουμε τα κείμενα μαζί με τις αποκρίσεις, με σκοπό να βρούμε κρυφά θέματα

που προβλέπουν καλύτερα τις μεταβλητές απόκρισης για μελλοντικά κείμενα χωρίς ετικέτα.

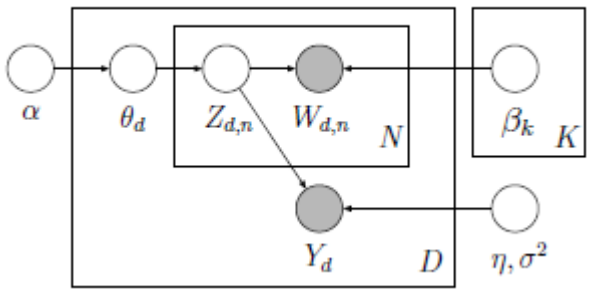
Θα δώσουμε έμφαση στο ότι ο sLDA περιέχει διάφορους τύπους απόκρισης: μη εξαναγκασμένες πραγματικές τιμές, πραγματικές τιμές εξαναγκασμένες στο να είναι θετικά, ταξινομημένες ή μη ταξινομημένες ετικέτες κλάσης, με αρνητικούς ακεραίους και άλλους τύπους. Όμως, ο μηχανισμός που χρησιμοποιείται για να επιτύχουμε την γενικοποίηση κάνει την κατάσταση λίγο πολύπλοκη, έτσι πρώτα θα δώσουμε τον ορισμό του sLDA για την ειδική περίπτωση μίας μη εξαναγκασμένης πραγματικής απόκρισης.

Ας συγκεντρωθούμε στην περίπτωση που $y \in \mathbb{R}$. Και προς στιγμήν ας φτιάξουμε τις παραμέτρους του μοντέλου: τα K θέματα β1:K, την Dirichlet παράμετρο α, και τις παραμέτρους απόκρισης ξ και σ^2. Κάτω από το sLDA μοντέλο, κάθε κείμενο και κάθε απόκριση δημιουργείται από την παρακάτω διαδικασία:

1. Draw topic proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \text{N}(\eta^\top \bar{z}, \sigma^2)$.

Εδώ ορίζουμε $\bar{z} := (1/N) \sum_{n=1}^{N} z_n$. Η οικογένεια των πιθανοτικών κατανομών που αντιστοιχεί σε αυτή την γενετική διαδικασία απεικονίζεται ως γραφικό μοντέλο στην Εικόνα 3.13.

Σημειώστε ότι η απόκριση έρχεται από ένα κανονικό γραμμικό μοντέλο. Οι μεταβλητές διασποράς του μοντέλου είναι οι εμπειρικές συχνότητες των θεμάτων στο κείμενο. Οι παράγοντες παλινδρόμησης αυτών των συχνοτήτων σχηματίζουν το η. Σημειώστε ότι το γραμμικό μοντέλο εμπεριέχει έναν όρο ανακοπής, που μετράει την προστιθέμενη διασπορά α που είναι πάντα ίση με 1. Εδώ, ένας τέτοιος όρος είναι περιττός, γιατί τα στοιχεία του $\bar{z}$ έχουν πάντα άθροισμα ίσο με 1.



Εικόνα 3.13: Απεικόνιση γραφικού μοντέλου του επιβλεπόμενου LDA αλγορίθμου.

Παλινδρομώντας την απόκριση στις εμπειρικές θεματικές συχνότητες, αντιμετωπίζουμε την απόκριση σαν μη ανταλλάξιμη ως προς τις λέξεις. Το κείμενο δημιουργείται πρώτα, κάτω από πλήρη ανταλλαξιμότητα λέξεων. Έπειτα, βασισμένη στο κείμενο, η μεταβλητή απόκρισης δημιουργείται. Σε αντίθεση, θα μπορούσε να σχηματιστεί ένα μοντέλο όπου το y έχει παλινδρομηνθεί σε θεματικές αναλογίες θ. Αυτό καθιστά την απόκριση και όλες τις λέξεις ανταλλάξιμες. Αλλά πρακτικά, ο επιλεγμένος σχηματισμός φαίνεται πιο λογικός: Η απόκριση εξαρτάται περισσότερο από τις θεματικές αναλογίες που πραγματικά υπάρχουν

στο κείμενο, παρά από τον μέσο της κατανομής που δημιουργεί τα θέματα.

Αντιμετωπίζουμε τα α, $\beta_{1:K}$, η και σ^2 σαν άγνωστες σταθερές που πρέπει να βρεθούν, παρά σαν τυχαίες μεταβλητές. Θα βρούμε λοιπόν έναν μέγιστης πιθανότητας υπολογισμό με την μια μεταβολική Expectation-Maximization διαδικασία, παρόμοια με αυτή που ακολουθήσαμε στον μη επιβλεπόμενο LDA.

3.5.3 Μεταβολικό Ε-βήμα

Δεδομένων ενός κειμένων και μια απόκρισης, η εκ των υστέρων κατανομή των κρυφών μεταβλητών είναι

$$p(\theta, z_{1:N} \mid w_{1:N}, y, \alpha, \beta_{1:K}, \eta, \sigma^2) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})\right) p(y \mid z_{1:N}, \eta, \sigma^2)}{\int d\theta\, p(\theta \mid \alpha) \sum_{z_{1:N}} \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})\right) p(y \mid z_{1:N}, \eta, \sigma^2)}. \quad (1)$$

Η τιμή κανονικοποίησης είναι είναι η περιθωριακή πιθανότητα των παρατηρούμενων δεδομένων, του κείμενου $w_{1:N}$ και της απόκρισης y. Αυτός ο κανονικοποιητής είναι γνωστός ως πιθανοφάνεια. Όπως και με τον LDA, δεν είναι εύκολα υπολογίσιμος. Έτσι, ανατρέχουμε στις μεταβολικές μεθόδους για να υπολογίσουμε την εκ των υστέρων κατανομή.

Μεταβολική αντικειμενική συνάρτηση. Μεγιστοποιούμε το χαμηλότερο όριο πιθανοφάνειας L(.) (ELBO), όπου για ένα κείμενο έχει την μορφή

$$\log p(w_{1:N}, y \mid \alpha, \beta_{1:K}, \eta, \sigma^2) \geq \mathcal{L}(\gamma, \phi_{1:N}; \alpha, \beta_{1:K}, \eta, \sigma^2) = \mathrm{E}[\log p(\theta \mid \alpha)] + \sum_{n=1}^{N} \mathrm{E}[\log p(Z_n \mid \theta)] + \sum_{n=1}^{N} \mathrm{E}[\log p(w_n \mid Z_n, \beta_{1:K})] + \mathrm{E}[\log p(y \mid Z_{1:N}, \eta, \sigma^2)] + \mathrm{H}(q). \quad (2)$$

Εδώ η προσδοκώμενη λύση βρίσκεται με παραγώγιση ως προς μια μεταβολική κατανομή q. Διαλέγουμε την ολοκληρωτικά παραγωγισμένη κατανομή,

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n), \quad (3)$$

όπου το γ είναι ένα K-διαστάσεων Dirichlet διάνυσμα παραμέτρου και κάθε φ_n παραμετροποιεί μια κατηγορική κατανομή ως προς K στοιχεία. Σημειώστε ότι $E[Z_n] = \varphi_n$.

Οι πρώτοι τρεις όροι και η εντροπία της μεταβολικής κατανομής είναι ιδανικοί ως προς τους ανταποκρινόμενους όρους του ELBO για τον επιβλεπόμενο LDA. Ο τέταρτος όρος είναι η προσδοκώμενη λογαριθμική πιθανότητα της μεταβλητής απόκρισης δεδομένων των κρυφών εναποθέσεων θεμάτων,

$$\mathrm{E}[\log p(y \mid Z_{1:N}, \eta, \sigma^2)] = = -\frac{1}{2}\log(2\pi\sigma^2) - \left(y^2 - 2y\eta^\top \mathrm{E}[\bar{Z}] + \eta^\top \mathrm{E}[\bar{Z}\bar{Z}^\top]\eta\right) / 2\sigma^2 .$$

Η πρώτη προσδοκώμενη τιμή είναι $\mathrm{E}[\bar{Z}] = \bar{\phi} := (1/N)\sum_{n=1}^{N}\phi_n$, και η δεύτερη είναι,

$$\mathrm{E}[\bar{Z}\bar{Z}^\top] = (1/N^2)\left(\sum_{n=1}^{N}\sum_{m\neq n}\phi_n\phi_m^\top + \sum_{n=1}^{N}\mathrm{diag}\{\phi_n\}\right). \quad (5)$$

Για να καταλάβετε $m \neq n$, $\mathrm{E}[Z_n Z_m^\top] = \mathrm{E}[Z_n]\mathrm{E}[Z_m]^\top = \phi_n\phi_m^\top$ για επειδή η μεταβολική κατανομή είναι πλήρως παραγωγισμένη. Από την άλλη $\mathrm{E}[Z_n Z_n^\top] = \mathrm{diag}(\mathrm{E}[Z_n]) = \mathrm{diag}(\phi_n)$ επειδή το Z_n είναι ένας δείκτης-διάνυσμα.

<u>Βελτιστοποίηση ως προς γ.</u> Οι όροι που περιέχουν την μεταβολική Dirichlet γ είναι ιδανικά ως προς αυτά ενός μη επιβλεπόμενου LDA, δηλαδή δεν περιέχουν την μεταβλητή απόκρισης y. Έτσι η συντεταγμένη αύξουσα είναι

$$\gamma^{\text{new}} \leftarrow \alpha + \sum_{n=1}^{N}\phi_n. \quad (6)$$

<u>Βελτιστοποίηση ως προς φ_j.</u> Ορίζουμ$\phi_{-j} := \sum_{n\neq j}\phi_n$. $j \in \{1,\ldots,N\}$. Στη (3), μεγιστοποιούμε την Λαγκρανιανή του ELBO, που περιέχει τον περιορισμό ότι τα στοιχεία του φ_j έχουν άθροισμα ίσο με 1, και έτσι παίρνουμε την καινούρια τιμή,

$$\phi_j^{\text{new}} \propto \exp\left\{\mathrm{E}[\log\theta\,|\,\gamma] + \mathrm{E}[\log p(w_j\,|\,\beta_{1:K})] + \left(\frac{y}{N\sigma^2}\right)\eta - \frac{\left[2(\eta^\top\phi_{-j})\eta + (\eta\circ\eta)\right]}{2N^2\sigma^2}\right\}. \quad (7)$$

Η κύρια διαφορά μεταξύ LDA και sLDA βρίσκεται σε αυτή την καινούρια εξίσωση. Όπως και στον LDA, η μεταβολική κατανομή της j-οστής λέξης ως προς τα θέματα εξαρτάται από τις θεματικές πιθανότητες της λέξης κάτω από το πραγματικό μοντέλο. Άλλα η μεταβολική κατανομή w_j , και αυτές όλων των άλλων λέξεων, επηρεάζουν την πιθανότητα της απόκρισης, διαμέσου του αθροίσματος υπολοίπων των τετραγώνων (RSS), όπου είναι ο δεύτερος όρος στην (4). Το τελικό αποτέλεσμα της καινούριας τιμής ενθαρρύνει την μείωση της προσδοκώμενης RSS.

Η καινούρια συνάρτηση εξαρτάται από τις μεταβολικές παραμέτρους φ_{-j} όλων των άλλων λέξεων. Έτσι, αντιθέτως από τον LDA, το φ_j δεν μπορεί να ενημερώνεται παράλληλα. Διαφορετικές τιμές αυτού του όρου αντιμετωπίζονται ξεχωριστά.

3.5.4 Μ-βήμα και πρόβλεψη

Το επιπέδου-σώματος ELBO φράζει την από κοινού λογαριθμική πιθανοφάνεια μέσα από τα κείμενα, όπου είναι το άθροισμα των ανα-κείμενο λογαριθμικών πιθανοφανειών. Στο Ε-βήμα, υπολογίζουμε την προσδοκώμενη εκ των υστέρων κατανομή για κάθε ζευγάρι κειμένου-απόκρισης χρησιμοποιώντας τον αλγόριθμο μεταβολικής συμπερασματολογίας. Στο Μ-βήμα μεγιστοποιούμε το επιπέδου-σώματος ELBO ως προς της παραμέτρους $\beta_{1:K}$, η, και σ^2. Σε αυτή την ενότητα προσθέτουμε δείκτες κειμένου στις ποσότητες της προηγούμενης ενότητας,έτσι το y γίνεται y_d και το $\bar{Z}$ γίνεται formula $\bar{Z}_d$.

<u>Υπολογίζοντας τα θέματα.</u> Οι Μ-βήματος καινούριες τιμές των θεμάτων $\beta_{1:K}$ είναι ίδιες ακριβώς όπως και στον μη επιβλεπόμενο LDA, όπου η πιθανότητα μια λέξης κάτω από ένα

θέμα είναι ανάλογη του προσδοκώμενου αριθμού των φορών που αυτή έχει εναποτεθεί στο συγκεκριμένο θέμα,

$$\hat{\beta}_{k,w}^{\text{new}} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = w)\phi_{d,n}^{k}. \tag{8}$$

Εδώ ξανά η αναλογικότητα σημαίνει ότι κάθε $\hat{\beta}_k^{new}$ κανονικοποιείται.

Υπολογίζοντας τις παραμέτρους παλινδρόμησης. Οι μόνοι όροι του επιπέδου-σώματος ELBO που περιέχουν το η και το σ^2 έρχονται από το επιπέδου-σώματος ανάλογο του (4).

Ορίζουμε το $y=y_{1:D}$ ως το διάνυσμα των τιμών απόκρισης μέσα από τα κείμενα. Ας ορίσουμε το A να είναι ο $D\times(K+1)$ πίνακας του οποίου οι σειρές είναι τα διανύσματα formula $\bar{Z}_d^T$. Έτσι η επιπέδου-σώματος εκδοχή (4) είναι

$$\mathrm{E}[\log p(y \mid A, \eta, \sigma^2)] = -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathrm{E}\left[(y - A\eta)^\top (y - A\eta)\right]. \tag{9}$$

Εδώ ο προσδοκώμενος υπολογισμός γίνεται πάνω στον A, χρησιμοποιώντας παραμέτρους μεταβολικής κατανομής επιλεγμένες από το προηγούμενο E-βήμα. Επεκτείνοντας στο εσωτερικό γινόμενο, χρησιμοποιώντας γραμμικότητα στην προσδοκία, φτάνουμε σε μια εκδοχή κανονικών εξισώσεων:

$$\mathrm{E}\left[A^\top A\right]\eta = \mathrm{E}[A]^\top y \qquad \Rightarrow \qquad \hat{\eta}_{\text{new}} \leftarrow \left(\mathrm{E}\left[A^\top A\right]\right)^{-1} \mathrm{E}[A]^\top y\,. \tag{10}$$

Τώρα χρησιμοποιώντας first-order προϋπόθεση για το σ2 στην (9) και εκτιμώντας το $\hat{\eta_{new}}$ παίρνουμε:

$$\hat{\sigma}^2_{\text{new}} \leftarrow (1/D)\{y^\top y - y^\top \mathrm{E}[A]\left(\mathrm{E}\left[A^\top A\right]\right)^{-1} \mathrm{E}[A]^\top y\}\,. \tag{11}$$

Πρόβλεψη. Ο λόγος που εφαρμόζουμε sLDA είναι η πρόβλεψη. Συγκεκριμένα, επιθυμούμε να υπολογίσουμε την προσδοκώμενη τιμή απόκρισης, δεδομένου ενός καινούριου κειμένου $w_{1:N}$ σε ένα φιξαρισμένο μοντέλο $\{\alpha, \beta_{1:K}, \eta, \sigma^2\}$:

$$\mathrm{E}[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top \mathrm{E}[\bar{Z} \mid w_{1:N}, \alpha, \beta_{1:K}]. \tag{12}$$

Το ταίριασμα δημιουργείται εύκολα από την επαναληπτική προσδοκία. Υπολογίζουμε την εκ των υστέρων κατανομή του μέσου formula $\bar{Z}$ χρησιμοποιώντας την διαδικασία μεταβολικής συμπερασματολογίας της προηγούμενης ενότητας. Αλλά εδώ, οι όροι που εξαρτώνται από την y απομακρύνονται από την καινούρια τιμή φj (7). Σημειώστε ότι είναι ακριβώς η ίδια διαδικασία μεταβολικής συμπερασματολογίας όπως και στον LDA.

Έτσι, δεδομένου ενός νέου κειμένου, υπολογίζουμε πρώτα το $E_q[Z_{1:N}]$, την μεταβολική εκ των υστέρων κατανομή των κρυφών μεταβλητών Z_n.. Έτσι καταλήγουμε στην τιμή

$$E[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] \approx \eta^\top E_q[\bar{Z}] = \eta^\top \bar{\phi}. \quad (13)$$

ΚΕΦΑΛΑΙΟ 4
Εφαρμογές αλγορίθμου LDA

4.1 Σύσταση με ετικέτες

4.1.1 Εισαγωγή

Τα συστήματα tagging όπως το Flickr[11], Last.fm[2], Delicious[3] έχουν γίνει μία βασική εφαρμογή στο Διαδίκτυο. Αυτά τα συστήματα αφήνουν του χρήστες να δημιουργούν και να διαχειρίζονται ετικέτες, να σχολιάζουν και να κατηγοριοποιούν το κείμενο. Στα κοινωνικά tagging συστήματα όπως το Delicious ο χρήστης μπορεί όχι μόνο να σχολιάζει τα δικά του κείμενα αλλά επίσης και τα κείμενα των άλλων χρηστών. Η υπηρεσία που προσφέρουν αυτά τα συστήματα είναι διπλή: Αφήνουν του χρήστες να δημοσιεύουν κείμενο και να ψάχνουν κείμενο. Έτσι επιτυγχάνονται δύο στόχοι για τον χρήστη:

1. Οι ετικέτες βοηθούν στην οργάνωση και στην διαχείριση του κειμένου του χρήστη, και
2. Βρίσκουν σχετικά κείμενα με άλλους χρήστες.

Η σύσταση με ετικέτες (εναλλακτικές προτάσεις δηλαδή που δίνει ένα λογισμικό σε ένα χρήστη ανάλογα με αυτές που έχει διαλέξει) μπορεί να συγκεντρωθεί σε μία από τις δύο δυνατότητες. Η προσωποποιημένη σύσταση με ετικέτες βοηθάει τους χρήστες να σχολιάζουν το κείμενό τους με σκοπό να διαχειρίζονται και να ανακτούν τις δικές τους πηγές. Η συλλεκτική σύσταση στοχεύει στο να κάνει τις πηγές πιο ορατές στους άλλους χρήστες συστήνοντας ετικέτες που διευκολύνουν την αναζήτηση.

Όμως, αφού οι ετικέτες δεν είναι περιορισμένες σε ένα συγκεκριμένο λεξιλόγιο, οι χρήστες μπορούν να διαλέξουν οποιαδήποτε ετικέτα για να περιγράψουν τις πηγές τους. Έτσι αυτές οι ετικέτες μπορεί να είναι ασταθείς και ιδιοσυγκρατικές, εξαιτίας της προσωπικής ορολογίας κάθε χρήστη καθώς επίσης και εξαιτίας των διαφορετικών στόχων που αυτές οι ετικέτες προσπαθούν να επιφέρουν. Αυτό μειώνει την χρησιμότητα των ετικετών και ιδιαίτερα των πηγών που σχολιάζονται από πολύ λίγους μόνο χρήστες, όπου για δημοφιλείς πηγές, το tagging που γίνεται από πολλούς χρήστες μειώνει το πρόβλημα σε έναν αρκετά μεγάλο βαθμό, για παράδειγμα η ακρίβεια ορισμένων ετικετών μεγαλώνει καθώς σχολιάζουν όλο και περισσότεροι χρήστες μία πηγή.

Ένας από τους στόχους στο πεδίο αυτό, είναι να ξεπεραστεί το λεγόμενο “cold start” πρόβλημα για το tagging νέων πηγών. Για να το λύσουμε αυτό, μπορούμε να χρησιμοποιήσουμε τον LDA, ώστε να εξάγουμε τα κρυφά θέματα από τις πηγές και με ένα αρκετά σταθερό και πλήρες σύνολο ετικετών για νέες πηγές με λίγες μόνο ετικέτες. Βάσει αυτού, άλλες ετικέτες που ανήκουν στα συστημένα θέματα μπορούν να συστηθούν. Ο LDA αλγόριθμος συγκρινόμενος με την προσέγγιση του προβλήματος από τους κανόνες συσχέτισης πετυχαίνει μεγαλύτερη ακρίβεια. Ακόμα οι ετικέτες, που συστήνονται, είναι πιο συγκεκριμένες για μία συγκεκριμένη πηγή, και πιο χρήσιμες για αναζήτηση και σύσταση πηγών σε άλλους χρήστες.

1 http://www.flickr.com
2 http: //www.lastfm.com
3 http: //www.delicious.com

4.1.2 Σύσταση με LDA

Η γενική ιδέα του LDA βασίζεται στην υπόθεση ότι ένα άτομο γράφει ένα κείμενο γνωρίζοντας για τι θέματα μιλάει. Το να γράψει ένα θέμα, σημαίνει ότι διαλέγει μία λέξη με μία σίγουρη πιθανότητα από ένα φάσμα λέξεων αυτού του θέματος. Ένα ολόκληρο κείμενο μπορεί να απεικονισθεί ως ένα μείγμα διαφορετικών θεμάτων. Όταν ο συγγραφέας του κειμένου είναι ένα άτομο, αυτά τα θέματα αντανακλούν την άποψη αυτού του ατόμου στο κείμενο καθώς και το προσωπικό του λεξιλόγιο. Στο πεδίο των tagging συστημάτων όπου πολλοί χρήστες σχολιάζουν πηγές, τα θέματα που έχουν εξαχθεί δείχνουν μία πολυδιάστατη άποψη για το κείμενο και οι ετικέτες των θεμάτων απεικονίζουν ένα συνηθισμένο λεξιλόγιο που περιγράφει το κείμενο

Ο LDA εναποθέτει το κάθε κρυφό θέμα ενός κειμένου με μία τιμή πιθανότητας που δείχνει πόσο συμβάλει αυτό το θέμα στην δημιουργία αυτού του κειμένου. Για τα tagging συστήματα, τα κείμενα είναι πηγές $r \in R$ και κάθε πηγή περιγράφεται από ετικέτες $t \in T$ που έχουν εναποθέσει οι χρήστες $u \in U$. Αντί για κείμενα που συντίθενται από όρους έχουμε πηγές που συντίθενται από ετικέτες. Για να φτιάξουμε ένα LDA μοντέλο χρειαζόμαστε πηγές και συσχετισμένες ετικέτες που έχουν εναποθέσει προηγουμένως οι χρήστες. Για κάθε πηγή r χρειαζόμαστε κάποιους σελιδοδείκτες *b(r,ui)* εναποθετημένους από τους χρήστες u_i, $i \in 1 \ldots n$. Στην συνέχεια μπορούμε να απεικονίσουμε κάθε πηγή στο σύστημα όχι με τις πραγματικές του ετικέτες αλλά με τις ετικέτες που έχουν παραχθεί από τον LDA.

Για την νέα πηγή r_{new}, όπου έχουμε μόνο ένα μικρό αριθμό από σελιδοδείκτες $(i \in 1 \ldots n)$, δηλαδή ένας στους πέντε χρήστες έχει σχολιάσει την πηγή, μπορούμε να επεκτείνουμε την κρυφή θεματική απεικόνιση με τις ετικέτες κορυφής για κάθε κρυφό κείμενο. Για να στηρίξουμε το γεγονός ότι μερικές ετικέτες προστίθενται από πολλούς χρήστες όπου άλλες προστίθενται μόνο από ένα ή δύο μπορούμε να χρησιμοποιήσουμε τις πιθανότητες όπου ο LDA εναποθέτει. Οι πιθανότητες εναποτίθενται όχι μόνο στα κρυφά θέματα για μία πηγή, αλλά επίσης για κάθε ετικέτα μέσα σε ένα κρυφό θέμα για να απεικονίσουν την πιθανότητα αυτής της ετικέτας να ανήκει στο συγκεκριμένο θέμα. Απεικονίζουμε κάθε πηγή r_i ως τις πιθανότητες *P(z|ri)* για κάθε κρυφό θέμα $z_j \in Z.$ Κάθε θέμα z_j απεικονίζεται ως οι πιθανότητες *P(t|zj)* για κάθε ετικέτα $t_n \in T.$ Συνδέοντας αυτές τις δύο πιθανότητες για κάθε ετικέτα με την r_{new} , παίρνουμε την τιμή της πιθανότητας για κάθε ετικέτα όπου μπορεί να ερμηνευτεί ως η σχετική συχνότητα της ετικέτας στην πηγή. Θέτοντας μία βάση, μας επιτρέπει να προσαρμόσουμε τον αριθμό των συστημένων ετικετών και να δώσουμε περισσότερο έμφαση στην ακρίβεια.

Tag	Count	Prob.	Tag	Count	Prob.
photography	16452	0.235	howto	23371	0.219
photo	9002	0.129	tutorial	15519	0.145
photos	7739	0.110	reference	14084	0.132
images	6302	0.090	tips	13955	0.131
photoshop	4825	0.069	tutorials	7320	0.069
graphics	2831	0.040	guide	3430	0.032
image	2769	0.040	toread	2948	0.028
art	1910	0.027	article	2376	0.022
stock	1852	0.026	articles	1498	0.014
pictures	1676	0.024	useful	1442	0.013
design	1666	0.024	learning	1147	0.011
gallery	1386	0.020	tricks	1140	0.011
camera	831	0.012	how-to	1081	0.010
digital	802	0.011	help	1054	0.010

Εικόνα 4.1: Οι κορυφαίοι όροι συνθέτοντας τα κρυφά θέματα "photography" και "howto".

Φανταστείτε μία πηγή που ακολουθεί τις ετικέτες : "photo", "photography", και "howto". Η Εικόνα 4.1 δείχνει δύο θέματα με τις εναποθετιμένες λέξεις σε αυτά. Δεδομένων αυτών των θεμάτων μπορούμε πολύ εύκολα να επεκτείνουμε το παρόν σύνολο ετικετών ή να συστήσουμε νέες ετικέτες στους χρήστες κοιτάζοντας τα νέα θέματα. Στο παράδειγμά μας, μπορούμε να συστήσουμε τις λέξεις "photos", "images" "photoshop", "tutorial", "reference", και "tips" αν θέσουμε αρχικά τις πιθανότητες στο 0.045. Ο LDA θα είχε υποθέσει ότι η πηγή μας ανήκει κατά 66% στο θέμα "photo" και κατά 33% στο θέμα "howto". Πολλαπλασιάζοντας αυτές τις πιθανότητες με τις πιθανότητες ετικετών των κρυφών θεμάτων έχει ως αποτέλεσμα μία λίστα κατάταξης των σχετικών ετικετών για την πηγή μας.

4.2 Σύσταση Άρθρων βασισμένη σε θεματικά μοντέλα.

4.2.1 Εισαγωγή

Σήμερα υπάρχουν πάρα πολλές ιστοσελίδες που περιέχουν δωρεάν εκπαιδευτικά κείμενα. Μία από τις πιο αναγνωρισμένες πηγές είναι η εγκυκλοπαίδεια της Wikipedia. Τον Ιούλιο του 2008, υπήρχαν παραπάνω από *2.400.000 άρθρα* διαθέσιμα στα αγγλικά και σε άλλες πολλές γλώσσες. Ο κύριος όγκος των κειμένων της Wikipedia, περιέχει μερικά άρθρα που είναι ακατάλληλα για παιδιά. Τον Μάιο του 2007, τα παιδικά χωριά SOS, ο μεγαλύτερος παγκοσμίως φιλανθρωπικός οργανισμός για ορφανά παιδιά, οργάνωσε μια Συλλογή κειμένων Wikipedia για Σχολεία. Η συλλογή περιείχε *4.625* επιλεγμένα άρθρα βασισμένα στο Εθνικό Πρόγραμμα Σπουδών του Ηνωμένου Βασιλείου και σε άλλα παρόμοια προγράμματα στον κόσμο. Όλα τα άρθρα της συλλογής ελέγχθηκαν για την καταλληλότητά τους στα παιδιά.

Η Εικόνα 4.2 περιέχει μία λίστα με όλες τις κατηγορίες αντικειμένων που παρέχει η συλλογή. Η οργάνωση των άρθρων σε κατηγορίες παρέχει στον χρήστη έναν εύκολο τρόπο να διαβάσει τα άρθρα που αφορούν το ίδιο αντικείμενο. Όμως, άρθρα διαφορετικών κατηγοριών μπορεί να συσχετίζονται. Για παράδειγμα, το άρθρο *Great Wall of China* που τελεί υπό την κατηγορία *Design and Technology: Architecture* θεωρείται ότι σχετίζεται με το άρθρο *Beijing* που ανήκει στην κατηγορία *Geography: Geography of Asia* καθώς επίσης και με το άρθρο *Qui Shi Huang* στην κατηγορία *People: Historical Figures*.

Art	Business Studies	Citizenship	Countries
Design and Technology	Everyday life	Geography	History
IT	Language and literature	Mathematics	Music
People	Religion	Science	

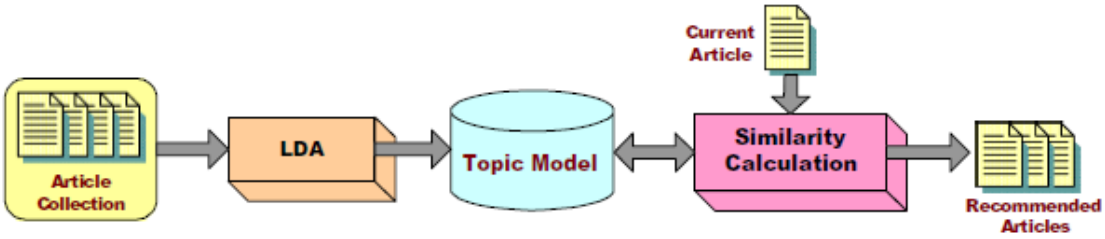
Εικόνα 4.2: Το υποσύνολο των κατηγοριών για την Συλλογή Wikipedia για Σχολεία.

Πέραν της μεθόδου browsing, οι χρήστες μπορούν να ψάξουν άρθρα ακολουθώντας υπερσυνδέσμους που είναι ενσωματωμένοι μέσα στο κείμενο των άρθρων. Όμως, αυτή η μέθοδος είναι εξαρτώμενη από τον συγγραφέα ενός άρθρου. Κάποιοι σύνδεσμοι που συσχετίζονται με άρθρα θα μπορούσαν να αγνοηθούν επανειλημμένα. Επίσης ένα σχετικό άρθρο μπορεί να υπεισέλθει σε έναν υπερσυνδέσμο αν δεν υπάρχει κάποιος όρος που να το περιγράφει μέσα στο παρών άρθρο.

Εξαιτίας των παραπάνω προβλημάτων των μεθόδων ανάκτησης, προτείνεται μία μέθοδος φιλτραρίσματος για σύσταση σχετικών άρθρων βασισμένη σε θεματικά μοντέλα. Για την δημιουργία ενός συνόλου θεμάτων, μπορεί να εφαρμοστεί ο LDA αλγόριθμος στην συλλογή. Ένα θέμα μπορεί να απεικονιστεί ως μία κατανομή ενός συνόλου από όρους. Έτσι, δεδομένου ενός άρθρου, μπορούμε να προτείνουμε σχετικά άρθρα υπολογίζοντας την ομοιότητα τους ως προς την θεματική τους κατανομή. Η σύσταση άρθρων βασισμένη στα θεματικά μοντέλα μπορεί να ανακαλύψει σχετικά άρθρα που μπορεί να προέρχονται από διαφορετικές κατηγορίες και δεν είναι εφικτός ο συσχετισμός τους δια μέσου των υπερσυνδέσμων.

4.2.2 Ο LDA για Σύσταση Άρθρων

Έχουν γίνει πολλές μελέτες για την ανακάλυψη κρυφών θεμάτων για συλλογές κειμένων. Ο LDA είναι το πιο πρόσφατο γενετικό πιθανοτικό μοντέλο για την ανάλυση συνόλων κειμένων. Η βασική ιδέα πίσω από αυτή την προσέγγιση είναι ότι τα κείμενα απεικονίζονται ως τυχαία μείγματα κρυφών θεμάτων. Κάθε θέμα απεικονίζεται ως μία πιθανοτική κατανομή πάνω σε όρους. Κάθε κείμενο απεικονίζεται ως μία πιθανοτική κατανομή πάνω σε θέματα.



Εικόνα 4.3: Η διαδικασία της σύστασης άρθρων για ένα θεματικό μοντέλο.

Η Εικόνα 4.3 δείχνει την διαδικασία της σύστασης άρθρων βασισμένης σε αυτό το θεματικό μοντέλο. Τα δεδομένα που έχουν εισαχθεί για τον LDA αλγόριθμο αποτελούνται από μία συλλογή άρθρων που είναι ένα σύνολο *m* κειμένων δηλωμένα ως $D = \{D_0, \ldots, D_{m-1}\}$. Ο LDA αλγόριθμος δημιουργεί ένα σύνολο από *n* θέματα που δηλώνονται ως $T = \{T_0, \ldots, T_{n-1}\}$. Κάθε θέμα είναι μία κατανομή πιθανοτήτων για *p* λέξεις που δηλώνονται ως $T_i = [w_0^i, \ldots, w_{p-1}^i]$, όπου το w_j^i είναι η πιθανότητα η λέξη *j* να ανατεθεί στο θέμα *i*. Βασισμένοι σε αυτό το μοντέλο, κάθε κείμενο μπορεί να απεικονιστεί ως η πιθανοτική κατανομή σε ένα σύνολο θεμάτων T, για παράδειγμα, $D_i = [t_0^i, \ldots t_{n-1}^i]$, όπου το t_j^i είναι πιθανότητα το θέμα *j* να ανατεθεί στο κείμενο *i*. Για να προτείνουμε σχετικά άρθρα, υπολογίζουμε την ομοιότητα μεταξύ της αναλογίας θεμάτων για ένα δεδομένο άρθρο και όλων των κατανομών των άλλων άρθρων και διαλέγουμε αυτά με την μεγαλύτερη τιμή.

4.2.3 Πείραμα και συζήτηση

Η Συλλογή της Wikipedia για Σχολεία είναι διαθέσιμη από την ιστοσελίδα των παιδικών χωριών SOS[2].

Topic #11

Terms	Prob.
storm	0.017
hurricane	0.011
tropical	0.010
florida	0.010
damage	0.010
wind	0.008
cause	0.007
atlantic	0.007
season	0.006
august	0.006

Topic #29

Terms	Prob.
art	0.015
painting	0.011
italy	0.010
style	0.010
artist	0.008
architecture	0.007
rome	0.007
renaissance	0.005
baroque	0.005
sculpture	0.005

Topic #34

Terms	Prob.
species	0.016
dna	0.010
cell	0.009
plant	0.006
organism	0.006
genetic	0.005
life	0.005
darwin	0.005
protein	0.005
animal	0.004

Topic #44

Terms	Prob.
computer	0.016
windows	0.010
system	0.009
software	0.006
data	0.006
internet	0.005
user	0.005
version	0.005
microsoft	0.005
programming	0.004

Εικόνα 4.4: Παραδείγματα από τα θέματα που δημιουργήθηκαν χρησιμοποιώντας τον LDA.

Η προσέγγιση αυτή της σύστασης άρθρων εφαρμόστηκε όπως περιγράφηκε στην προηγούμενη υποενότητα στην συλλογή άρθρων. Η Εικόνα 4.4 δείχνει μερικά παραδείγματα θεμάτων που δημιουργήθηκαν από τον LDA αλγόριθμο. Κάθε πίνακας περιέχει τους 10 κορυφαίους όρους καταταγμένους ανάλογα με τις πιθανοτικές τους τιμές. Μπορεί να παρατηρηθεί ότι ο LDA μπόρεσε να ομαδοποιήσει τους όρους με υψηλή ομοιότητα των ίδιων θεμάτων.

Η Εικόνα 4.4 δείχνει την σύσταση των 10 πρώτων άρθρων δεδομένων των τίτλων τους. Τα προτεινόμενα άρθρα που δεν εμφανίζονται ως υπερσύνδεσμοι στα δοσμένα άρθρα έχουν σκιαστεί εντονότερα. Μπορεί να παρατηρηθεί ότι τα περισσότερα από τα προτεινόμενα άρθρα δεν περιέχονται σε υπερσυνδέσμους. Παρόλα αυτά η μέθοδος μας μπόρεσε να τα ανακαλύψει.

2 http://www.sos-childrensvillages.org/pages/default.aspx

Article: Gravitation

Related articles	Score
(1) Black hole	0.9920
(2) Redshift	0.9915
(3) Hubble's law	0.9894
(4) Big Bang	0.9837
(5) Metric expansion of space	0.9681
(6) Cosmic microwave background radiation	0.9671
(7) Universe	0.9608
(8) Speed of light	0.9590
(9) Plasma (physics)	0.9509
(10) Cosmic inflation	0.9393

Article: Tsunami

Related articles	Score
(1) Tropical cyclone	0.9638
(2) 2004 Indian Ocean earthquake	0.9461
(3) Eye (cyclone)	0.9431
(4) Tornado	0.9401
(5) Flood	0.9100
(6) 2005 Sumatra earthquake	0.8893
(7) Hurricane Floyd	0.8859
(8) Storm of October 1804	0.8833
(9) Cyclone Rosita	0.8780
(10) Tropical Storm Vamei	0.8775

Article: Isaac Newton

Related articles	Score
(1) Leonhard Euler	0.9429
(2) Georg Cantor	0.9274
(3) Carl Friedrich Gauss	0.9230
(4) Albert Einstein	0.9179
(5) Niels Bohr	0.8813
(6) Paul Dirac	0.8728
(7) David Hilbert	0.8576
(8) Max Planck	0.8549
(9) William Thomson, 1st Baron Kelvin	0.8471
(10) John von Neumann	0.8320

Article: Open source

Related articles	Score
(1) Internet	0.9677
(2) Functional programming	0.9568
(3) Markup language	0.9559
(4) Computer programming	0.9557
(5) World Wide Web	0.9529
(6) Python (programming language)	0.9518
(7) BASIC	0.9474
(8) C++	0.9473
(9) Perl	0.9473
(10) Cryptography	0.9455

Εικόνα 4.5: Παραδείγματα σύστασης άρθρων βασισμένα στην προσέγγιση θεματικών μοντέλων.

4.3 Χαρακτηρίζοντας Microblogs με Μοντέλα Θεμάτων

4.3.1 Εισαγωγή

Εκατομμύρια ανθρώπων στρέφονται προς τις microblogging υπηρεσίες όπως το Twitter για να μάθουν νέα ή απόψεις από άλλα άτομα και ενδιαφέροντα γεγονότα. Τέτοιες υπηρεσίες χρησιμοποιούνται κατά κόρον για κοινωνική διαδικτύωση, για παράδειγμα, να υπάρχει επαφή με φίλους και συναδέλφους. Ακόμα, οι microblogging σελίδες χρησιμοποιούνται ως δημοσιευτικές πλατφόρμες για να δημιουργήσουν και να απορροφήσουν κείμενα από σύνολα χρηστών που έχουν ανόμοια μεταξύ τους ενδιαφέροντα. Θεωρήστε τον υποθετικό χρήστη @jane που ακολουθεί τον χρήστη @frank για τις δημοσιεύσεις του τελευταίου χρήστη με θέμα το ποδόσφαιρο κολεγίου. Όμως ο @frank χρησιμοποιεί το Twitter, επιπλέον, για να ρυθμίζει τις κοινωνικές του επαφές με φίλους και περιστασιακά να δημοσιεύει τις πολιτικές του απόψεις. Προς το παρόν, η @jane έχει λίγα εργαλεία στην διάθεσή της για να φιλτράρει το κείμενο που δεν αφορά το ποδόσφαιρο από τις δημοσιεύσεις του @frank. Με λίγα λόγια, το Twitter υποθέτει πως όλες οι δημοσιεύσεις από τους ανθρώπους που η @jane ακολουθεί, είναι δημοσιεύσεις που θέλει να διαβάσει. Παρόμοια, η @jane έχει ένα περιορισμένο σύνολο από επιλογές για να εντοπίσει νέα άτομα να ακολουθήσει. Μπορεί να κοιτάξει λίστες χρηστών σε ένα κοινωνικό γράφημα (για παράδειγμα όλους τους χρήστες που ακολουθούν τον @frank), ή μπορεί να ψάξει με κάποιες λέξεις-κλειδιά, και να βρει τις σχετικές δημοσιεύσεις. Όμως, παραμένει δύσκολο το πως θα βρει άλλους χρήστες που είναι σαν τον @frank ή -κάτι με μεγαλύτερη πρόκληση – που είναι σαν τον @frank με διαφορετικές πολιτικές απόψεις.

Το παράδειγμα από πάνω αντιπροσωπεύει δύο από τις πολλές πληροφορίες που δεν υπάρχει τρόπος να εξαχθούν από το Twitter μέχρι στιγμής. Για να λυθούν τέτοιες προκλήσεις απαιτείται η εύρεση νέων τεχνικών πέρα από αναλύσεις που βασίζονται στα δίκτυα. Οι τεχνικές αυτές εφαρμόζονται συχνά σε microblogging ιστοσελίδες και κοινωνικά δίκτυα και αναπτύσσουν νέα εργαλεία για την ανάλυση και την κατανόηση κειμένων από το Twitter. Η ανάλυση κειμένου στο Twitter θέτει μοναδικές προκλήσεις: Οι δημοσιεύσεις είναι μικρού μήκους (140 χαρακτήρες ή λιγότεροι) με μία γλώσσα που δεν μοιάζει με τα γραπτά Αγγλικά στην οποία πολλά επιβλεπόμενα μοντέλα μηχανικής εκμάθησης έχουν εκπαιδευτεί και αξιολογηθεί. Για την αποτελεσματική μοντελοποίηση των κειμένων στο Twitter απαιτούνται τεχνικές που χρειάζονται κάποια επίβλεψη.

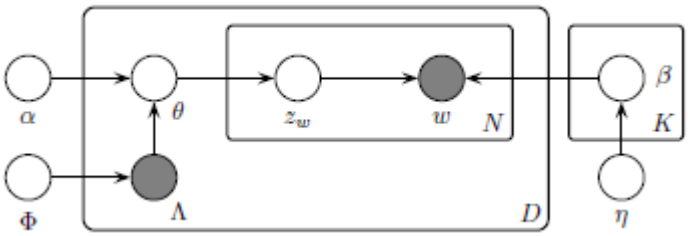
Στο παρόν υποκεφάλαιο θα δανειστούμε τον μηχανισμό των μοντέλων κρυφών μεταβλητών όπως τον LDA. Τα κρυφά θεματικά μοντέλα έχουν εφαρμοστεί ευρέως σε προβλήματα μοντελοποίησης κειμένου, και δεν απαιτούν κατασκευασμένα εκπαιδευτικά δεδομένα. Αυτά τα μοντέλα μετασχηματίζουν συλλογές κειμένων σε κατανομές λέξεων. Όμως ο LDA και τα σχετικά μοντέλα έχουν χρησιμοποιηθεί επιτυχώς σε ανάλυση συλλογών με ειδήσεις ή με επιστημονικά άρθρα, μια αναπάντητη ερώτηση λοιπόν είναι αν μπορούν να χρησιμοποιηθούν σε κείμενα μικρής έκτασης όπως ,για παράδειγμα, οι δημοσιεύσεις στο Twitter με κείμενο που είναι αρκετά διαφορετικό από τις παραδοσιακές συλλογές. Εδώ θα υποθέσουμε ότι η απάντηση είναι θετική και θα χρησιμοποιήσουμε τον Labeled LDA, μια μορφή επιβλεπόμενου LDA, για την ανάλυση των αυτών των κειμένων.

Τι τύπους και τι δομές μπορούμε να εξάγουμε, λοιπόν, με τα θεματικά μοντέλα κρυφών μεταβλητών από τους χρήστες του Twitter; Τα κρυφά θέματα μπορούν να κατηγοριοποιηθούν σε τέσσερις τύπους: τα θέματα ουσίας σχετικά με γεγονότα και ιδέες, τα κοινωνικά θέματα που αναγνωρίζουν την γλώσσα του χρήστη, τα θέματα προσωπικής στάσης του ατόμου που απεικονίζουν την προσωπική θέση του χρήστη, και τα θέματα του προσωπικού στυλ, που ενσωματώνουν την μόδα στην χρήση της γλώσσας.

4.3.2 Μοντελοποίηση δημοσιεύσεων με lLDA

Η προσέγγιση που χρησιμοποιούμε εδώ βασίζεται στα θεματικά μοντέλα κρυφών μεταβλητών και συγκεκριμένα μια γενικοποίηση του LDA, τον Labeled LDA, εφαρμόζοντας επίβλεψη όπου χρειάζεται. Ο LDA υποθέτει την ύπαρξη ενός συνόλου από ετικέτες Λ, καθεμία από τις οποίες χαρακτηρίζεται από μία πολυωνυμική κατανομή β_k για $k \in 1...|\Lambda|$ πάνω στις λέξεις ενός λεξιλογίου. Το μοντέλο υποθέτει ότι κάθε κείμενο d χρησιμοποιεί μονό ένα υποσύνολο από αυτές τις ετικέτες, που δηλώνονται ως $\Lambda_d \subseteq \Lambda$ και ότι το κείμενο προτιμάει κάποιες ετικέτες από άλλες ετικέτες, κάτι που απεικονίζει η πολυωνυμική κατανομή θ_d πάνω στο Λ_d. Κάθε λέξη w στο κείμενο d επιλέγεται από μία κατανομή λέξεων που συσχετίζεται με τις ετικέτες του κειμένου. Η λέξη επιλέγεται σε αναλογία του κατά πόσο το κείμενο προτιμάει την ετικέτα $\theta_{d,z}$ και του πόσο η ετικέτα προτιμάει την λέξη $\beta_{z,w}$.

Η Εικόνα 4.6 δείχνει το γραφικό μοντέλο δικτύων πιθανοτήτων και την γενετική διαδικασία για τον Labeled LDA. Από αυτή την γενετική διαδικασία, μπορεί να χρησιμοποιηθεί ένας αλγόριθμος συμπερασματολογίας για να ανακατασκευάσει τις ανά κείμενο κατανομές θ ως προς τις ετικέτες και τις ανά ετικέτα κατανομές β ως προς τις λέξεις, αρχίζοντας μόνο από τα κείμενα αυτά κάθε αυτά.



Εικόνα 4.6:*Το γραφικό μοντέλο του Labeled LDA*

Ο Labeled LDA μας επιτρέπει να μοντελοποιήσουμε μία συλλογή από Twitter δημοσιεύσεις σαν ένα μείγμα από κατηγοριοποιημένες διαστάσεις. Το LDA είναι μία ειδική περίπτωση του Labeled LDA. Μπορούμε να μοντελοποιήσουμε K κρυφά θέματα με ετικέτες "Topic 1" έως "Topic K" εναποθετιμένα σε κάθε δημοσίευση της συλλογής. Αν δεν χρησιμοποιούνται άλλες ετικέτες, αυτή η στρατηγική ανάθεσης ετικετών κάνει τον Labeled LDA μαθηματικά ιδανικό με τον παραδοσιακό LDA με K θέματα. Όμως, ο Labeled LDA μας δίνει την ελευθερία να εισάγουμε ετικέτες που ταιριάζουν μόνο σε μερικά υποσύνολα δημοσιεύσεων, έτσι ώστε το μοντέλο να μπορεί να μάθει από σύνολα λέξεων που πηγαίνουν με συγκεκριμένες ετικέτες.

4.3.3 Διαστάσεις με Ετικέτα στο Twitter

Ενώ οι κρυφές διαστάσεις στο Twitter μπορούν να μας βοηθήσουν να ποσοτικοποιήσουμε ευρείες τάσεις, πολλά πρόσθετα μεταδεδομένα είναι διαθέσιμα σε κάθε δημοσίευση και μπορούν να βοηθήσουν στην εύρεση πιο συγκεκριμένων, μικρών τάσεων.

Ένα hashtag είναι μία ανακάλυψη του Twitter που χρησιμοποιείται για να απλοποιήσει την αναζήτηση και την ανακάλυψη τάσεων. Οι χρήστες συμπεριλαμβάνουν ειδικά σχεδιασμένους όρους που αρχίζουν με # μέσα στο κείμενο κάθε δημοσίευσης. Για

παράδειγμα μία δημοσίευση σχετική με μία λίστα εργασιών μπορεί να περιέχει τον όρο #job. Αντιμετωπίζοντας κάθε hashtag ως μία ετικέτα που εφαρμόζεται μόνο σε δημοσιεύσεις που την περιέχουν, ο Labeled LDA ανακαλύπτει ποιες λέξεις είναι καλύτερα συσχετισμένες με κάθε hashtag. Οι τετριμμένες λέξεις που περιγράφονται από κάποια κρυφή διάσταση τείνουν να μην υποστηρίζονται από μία hashtag ετικέτα.

Οι Emoticon ετικέτες εφαρμόστηκαν σε δημοσιεύσεις που χρησιμοποιήθηκαν σε ένα σύνολο από εννιά κανονικά emoticons: χαμόγελο, κατσούφιασμα, κλείσιμο ματιού, πλατύ χαμόγελο, γλώσσα , καρδιά, έκπληξη , παράξενο, και μπερδεμένο. Οι @user ετικέτες εφαρμόστηκαν στις δημοσιεύσεις που απευθύνονται σε έναν χρήστη ως η πρώτη λέξη της δημοσίευσης. Οι ετικέτες απάντησης προστέθηκαν σε όποια δημοσίευση το Twitter API έχει σχεδιαστεί να απαντάει, για παράδειγμα, όταν ένα χρήστης κάνει κλικ σε ένα σύνδεσμο απάντησης σε μία άλλη δημοσίευση. Οι ετικέτες ερώτησης εφαρμόστηκαν σε δημοσιεύσεις που περιέχουν τον χαρακτήρα ερωτηματικό. Επειδή οι emoticons ετικέτες απάντησης, και ερώτησης είναι σχετικά συνήθεις, κάθε μία από αυτές τις ετικέτες βαθμοθετήθηκε με τάξη 10 – για παράδειγμα “:)-0” μέχρι “:)-9”- για να μοντελοποιηθεί η φυσική ποικιλία του πως κάθε ετικέτα χρησιμοποιείται. Ο αριθμός 10 επιλέχθηκε εμπειρικά δεδομένης της σχετικής συχνής χρησιμοποίησης αυτών των συμβόλων σε σχέση με τα hashtags. Οι δημοσιεύσεις περιείχαν 8.8 ετικέτες κατά μέσο όρο μέσα από ένα λεξιλόγιο από 156.223 λέξεις. Η πλειοψηφία των ετικετών ήταν hashtags. Φιλτράραμε τα hashtags που υπήρχαν σε λιγότερες από 30 δημοσιεύσεις, παίρνοντας ως αποτέλεσμα ένα τελικό σύνολο 504 ετικετών.

Emoticons	:)	thanks thank much too hi following love very you're welcome guys awww appreciated ah
		love all guys tweet awesome x nice twitter your goodnight followers later y'all sweet xoxo
	:(	miss sick still feeling ill can't much today already sleep triste him baby her sooo fml
		ah working won't stupid why anymore :'(isn't suck computer isnt ahh yeah nope nothing
Social Signal	Reply	thanks i'm sure ok good will i'll try yeah cool x fine yes definitely hun yep glad xx okay
		lmao yea tho yu wat kno thats nah hell lmfao idk dont doin aint naw already ima gotta we
	@user	haha yeah that's know too oh thats cool its hahaha one funny nice though he pretty yes
	?	did how does anyone know ?? ?! get where ??? really any mean long are ever see
		?! ?? !? who wtf !! huh ??? hahaha wow ?!! ?!? right okay ??!! hahahaha eh oh knew
Hashtags	#travel	travel #traveltuesday #lp hotel #ac ac tip tips #food national air airline #deals countries #tips
	#twilight	#newmoon #twilight twilight watching edward original watch soundtrack Jacob tom_cruise
	#politics	#cnn al_gore hoax climategate fraud #postrank gop inspires policy because why new bill

Πίνακας 4.1: Παράδειγμα κατανομών λέξεων που έχουν παραχθεί από ποικίλες κλάσεις ετικετών.

Ο Πίνακας 4.1 δείχνει μερικά χαρακτηριστικά θέματα που συσχετίζονται με κάθε κλάση ετικετών. Η φυσική ποικιλομορφία της γλώσσας είναι ορατή: Μία από τις ετικέτες χαμόγελου χρησιμοποιείται για να εκφράσει ευγνωμοσύνη και άλλες να εκφράσουν μορφές κοινωνικής γνωριμίας (“xoxo” σημαίνει αγκαλιές και φιλιά). Παρόμοια μια ετικέτα κατσουφιάσματος αποδίδει τη σημασία “άρρωστος” ενώ άλλη την σημασία “προσβεβλημένος”. Ανάλογοι περιορισμοί υπάρχουν και στους άλλους τύπους ετικετών. Εμείς ενδιαφερόμαστε να εξερευνήσουμε εφαρμογές που απομονώνουν κάθε μία από αυτές τις τάσεις, όπως προγράμματα αναζήτησης hashtag ετικετών και μοντελοποίησης συζητήσεων και ερωτήσεων χρησιμοποιώντας κοινωνικές ετικέτες.

4.3.4 Χαρακτηρίζοντας κείμενο στο Twitter

Ο Labeled LDA μπορεί να χρησιμοποιηθεί για να σχεδιάσει δημοσιεύσεις σε εκπαιδευμένες, κρυφές διαστάσεις με ετικέτα, που έχουμε ομαδοποιήσει σε 4 κατηγορίες ουσίας, κατάστασης, στυλ και επικοινωνίας – είτε δια χειρός είτε με κατασκευή. Αυτοί οι σχεδιασμοί μπορούν να δημιουργηθούν από δημοσιεύσεις για να χαρακτηρίσουν τις έντονες τάσεις που επικρατούν στο Twitter.

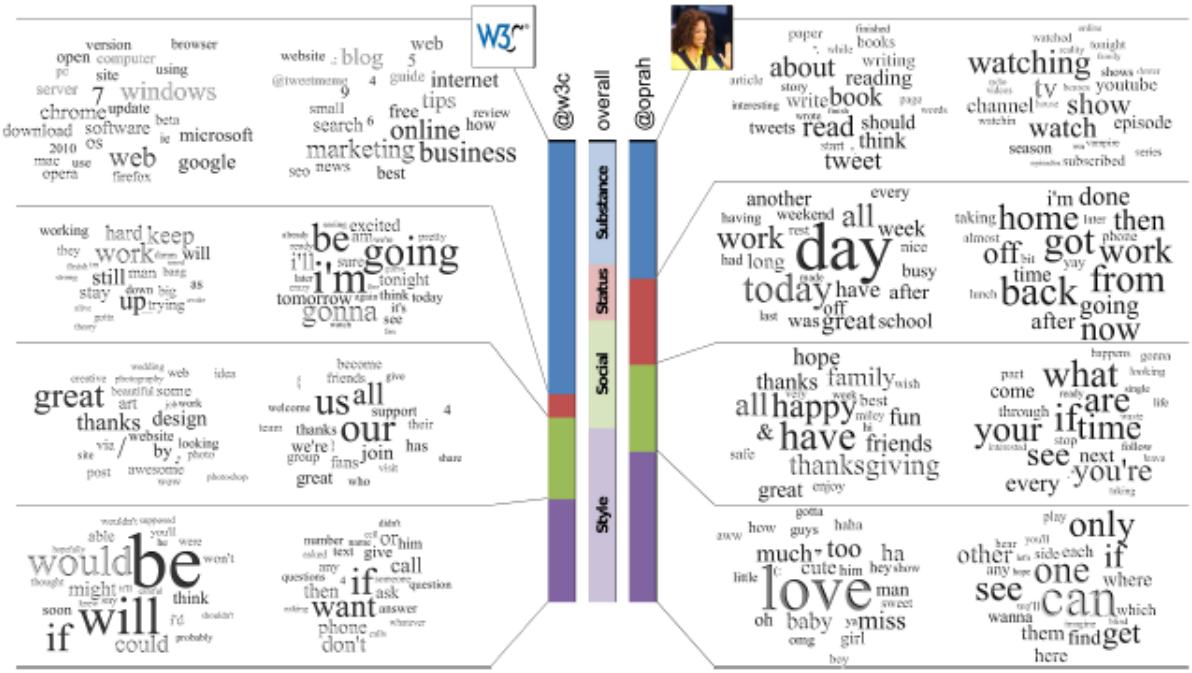
Έχοντας στην διάθεσή μας ένα μεγάλο σύνολο δεδομένων, μπορούμε να παρουσιάσουμε με αρκετά σίγουρη άποψη τι δημοσιεύουν οι χρήστες στο Twitter. Σε επίπεδο λέξεων, το Twitter είναι 11% ουσία, 5% κοινωνική κατάσταση, 16% στυλ, 10% καταστάσεις επικοινωνίας και 56% άλλες καταστάσεις. Πάρα την συνηθισμένη αντίληψη ισχύει το αντίθετο, η χρήση των διαστάσεων ουσίας υπερέχει των διαστάσεων κοινωνικής κατάστασης στο Twitter με αναλογία δύο προς ένα.

Οι άλλες κατηγορίες, διαφορετικές των τεσσάρων, είναι συνηθισμένες γιατί η κατηγοριοποίησή μας σε 4 καταστάσεις αλληλεπιδρά με άλλες κοινές τάσεις στο Twitter. Για παράδειγμα, οι λέξεις που δηλώνουν ώρα ή κάποιον αριθμό περιέχονται σε πολλά θέματα που έχουν ετικέτα “άλλες καταστάσεις”. Η μεγαλύτερη πηγή αυτών, όμως, έρχεται από την αναλογία των γλωσσών στο Twitter. Πιο συγκεκριμένα, περίπου οι μισοί χρήστες έρχονται από χώρες που δεν έχουν την αγγλική γλώσσα ως μητρική. Το μοντέλο διαχωρίζει αποτελεσματικά την χρήση αυτών των γλωσσών στις δικές τους διαστάσεις, οι οποίες ονομάζονται δια χειρός “άλλες καταστάσεις”. Μόνο όταν η γλώσσα έχει αρκετές δημοσιεύσεις το μοντέλο θα έχει αρκετά δεδομένα για να τα διαιρέσει στις 4 κατηγορίες.

Προσθέτοντας τις Labeled LDA διαστάσεις των πρόσφατων δημοσιεύσεων ανάμεσα σε δύο Twitter λογαριασμούς, μπορούμε οπτικά να συγκρίνουμε την χρήση της γλώσσας που χρησιμοποιεί ο καθένας από τους δύο χρήστες. Η Εικόνα 4.7 απεικονίζει την ανάλυση των τεσσάρων κατηγοριών για 200 πρόσφατες δημοσιεύσεις που έχουν γραφτεί από μία διάσημη παρουσιάστρια (@oprah, δεξιά) και από την κοινοπραξία του Παγκόσμιου Διαδικτύου (@w3c, αριστερά). Στο κέντρο, βλέπουμε τις αναλογίες χρήσης των διαστάσεων των δύο χρηστών που απεικονίζονται στις κάθετες γραμμές του γραφήματος. Στην κεντρική κάθετη γραμμή βλέπουμε την γενική αναλογία χρήσης των διαστάσεων, όπου μπορούμε εύκολα να δούμε ότι οι δημοσιεύσεις @w3c είναι στραμμένες προς την κατάσταση ουσίας, και οι δημοσιεύσεις της @oprah αφορούν λίγο περισσότερο από τον μέσο όρο την κοινωνική κατάσταση. Οι πιο συχνές λέξεις για τις επιλεγμένες διαστάσεις για κάθε κατηγορία απεικονίζονται αριστερά και δεξιά. Το μέγεθος κάθε λέξης αντανακλά στο πόσο σημαντική είναι αυτή η λέξη σε κάθε διάσταση από όλους του χρήστες, και η σκίαση εξαρτάται από το κατά πόσο ο χρήστης χρησιμοποιεί την λέξη μέσα στην διάσταση.

Παραδείγματα όπως η Εικόνα 4.7 μπορούν να χρησιμοποιηθούν για να χαρακτηρισμό και σύγκριση μεταξύ χρηστών. Για παράδειγμα μπορούμε να δούμε ότι η @oprah δημοσιεύει για το τηλεοπτικό της show (πάνω δεξιά) κάποια βιβλία. Πιο συγκεκριμένα βλέπουμε ότι η

@oprah χρησιμοποιεί την “book” διάσταση για να μιλήσει για διάβασμα (reading) περισσότερο από τη συγγραφή (writing). Παρόμοια,, ο @w3c δημοσιεύει συχνά για την τεχνολογία (πάνω αριστερά) και το Διαδίκτυο. Μέσα στο θέμα Διαδικτύου, ο @w3c χρησιμοποιεί λέξεις όπως “internet” και “online” αλλά όχι “markteting” ή “seo”. Σε επικοινωνιακό επίπεδο, ο w3c αποδεικνύεται μια ανοιχτή οργάνωση χρησιμοποιώντας λέξεις όπως *join, we, our* και *us*, όπου η @oprah μιλάει στους οπαδούς της (*you, you' re*).



Εικόνα 4.7: Η ανάλυση των 4 κατηγοριών για δύο χρήστες: @w3c (αριστερά) και @oprah (δεξιά). Η χρήση των διαστάσεων για τις κατηγορίες ουσίας(substance), κοινωνική κατάστασης (Status), επικοινωνίας (Social) και στυλ (Style) απεικονίζεται στις ορθογώνιες στήλες, με την μέση χρήση όλων των χρηστών στο Twitter να απεικονίζεται στο κέντρο. Οι συνηθισμένες λέξεις στις επιλεγμένες διαστάσεις για κάθε κατηγορία απεικονίζεται από τις συγκεντρώσεις λέξεων. Το μέγεθος κάθε λέξης είναι αναλογικό στην συχνότητα που την χρησιμοποιούν όλοι οι χρήστες, ενώ το πόσο σκιασμένη είναι μία λέξη είναι ανάλογο ως προς την συχνότητα της λέξης στις πρόσφατες δημοσιεύσεις του χρήστη.

4.4 Εξετάζοντας τις προτιμήσεις πελατών με πιθανοτικά μοντέλα θεμάτων

4.4.1 Εισαγωγή

Το πεδίο της έρευνας ηλεκτρονικής αγοράς έχει από πολύ καιρό πριν εφαρμόσει εξόρυξη δεδομένων (text mining) και τεχνικής εκμάθησης μηχανών σε συναλλαγές λιανικής ηλεκτρονικής αγοράς στις οποίες ένας μεγάλος αριθμός δεδομένων αγοράς έχει αναλυθεί. Η ανάλυση του basket marketing ανακαλύπτει δομές συσχέτισης στις αγορές λιανικής και θέτει τις βάσεις για εφαρμογές όπως product bundling, cross category dependency identification καθώς επίσης και consumer profiling.

Η ανάπτυξη του ηλεκτρονικού εμπορίου έστρωσε τον δρόμο για πολλές τεχνικές και μοντέλα προορισμένα να ενισχύσουν την εμπειρία των πελατών στα ηλεκτρονικά καταστήματα. Ανάμεσα σε αυτές, είναι και οι προσπάθειες εκμάθησης προτιμήσεων για να προσδιοριστούν οι επιθυμίες πελατών με δηλωτικό τρόπο ώστε να υποστηριχθεί η σύσταση νέων προϊόντων. Τυπικά ένα σύστημα σύστασης συγκρίνει το προφίλ ενός χρήστη με μερικά χαρακτηριστικά συμπεριφοράς, και προβλέπει την "κατάταξη" όπου ο πελάτης θα έδινε σε ένα αντικείμενο που δεν έχει υπόψη του ακόμη. Αυτά τα χαρακτηριστικά θα εξαχθούν είτε από το ίδιο το αντικείμενο είτε από το περιβάλλον του χρήστη.

Η εφαρμογή τεχνικών συστημάτων σύστασης σε basket marketing δεδομένα αντιμετωπίζει πολλές προκλήσεις. Πρώτα, η συνηθισμένη προσέγγιση εξαγωγής των κανόνων συσχέτισης παρέχει μία περιορισμένη άποψη για την κρυφή δομή των προτιμήσεων των χρηστών. Ακόμη παρόλο που μπορούμε να χρησιμοποιήσουμε τους κανόνες συσχέτισης για να προβλέψουμε επιτυχώς το υπόλοιπο στο καλάθι του χρήστη, μας λείπει η συνολική άποψη για τις αρέσκειες του χρήστη και τις συσχετίσεις τους. Δεύτερον, υπάρχει ένας αριθμός τεχνικών ζητημάτων που σχετίζονται με τις πιο συνηθισμένες τεχνικές. Οι κανόνες συσχέτισης τείνουν να αγνοούν τα μεγάλα σύνολα αντικειμένων, και η βασισμένη στην μνήμη εκμάθηση εκ συνεργασίας παρουσιάζει έλλειψη ικανότητας κλιμάκωσης. Από την άλλη, τα συστήματα σύστασης που βασίζονται στην ανάλυση κειμένου είναι ακατάλληλα, γιατί οι πληροφορίες για τα προϊόντα λιανικής δεν είναι ούτε διαθέσιμα πάντα ούτε περιγράφονται καταλλήλως.

Σε μία προσπάθεια να διευθετήσουμε τις παραπάνω προκλήσεις των υπάρχοντων τεχνικών στα συστήματα σύστασης που χρησιμοποιούνται στην ανάλυση του basket marketing, εξερευνούμε την χρήση κρυφών θεματικών μοντέλων. Σε αυτή την ενότητα θα εφαρμόσουμε κρυφά θεματικά μοντέλα για να δημιουργήσουμε ένα μοντέλο για τις προτιμήσεις πελατών καθώς επίσης και να κάνουμε μία αποτελεσματική σύσταση προϊόντων στους πελάτες. Πιο συγκεκριμένα, εξερευνούμε κρυφά μοντέλα θεμάτων για να ανακαλύψουμε κρυφά καλάθια και κρυφούς χρήστες από τα δεδομένα αγορών και τέλος προτείνεται ένας μηχανισμός σύστασης βασισμένος σε κρυφά καλάθια και χρήστες.

4.4.2 Εφαρμόζοντας μοντέλα θεμάτων σε καλάθια ηλεκτρονικής αγοράς

Εφαρμόζουμε πιθανοτικά μοντέλα θεμάτων και συγκεκριμένα το LDA μοντέλο στα δεδομένα συναλλαγών λιανικής αγοράς. Το μοντέλο τοποθετεί μια Dirichlet εκ των προτέρων κατανομή στις δύο κατανομές των θεμάτων ως προς τα κείμενα και των λέξεων ως προς τα θέματα. Συνεχίζουμε περιγράφοντας το πως θα εξάγουμε τα θεματικά μοντέλα χρησιμοποιώντας τα σύνολα αντικειμένων που περιέχονται στις συναλλαγές και το άθροισμα των προτιμήσεων των πελατών καθώς ο χρόνος περνάει.

Ορισμός του προβλήματος

Ένα καλάθι ηλεκτρονικής αγοράς συντίθεται από αντικείμενα που αγοράζονται από μία "βόλτα" σε ένα ηλεκτρονικό κατάστημα. Τα πιο σημαντικά χαρακτηριστικά είναι η εξακρίβωση της συναλλαγής και η εξακρίβωση του προϊόντος. Αγνοώντας την ποσότητα που αγοράζεται και την τιμή, κάθε συναλλαγή αντιστοιχεί σε μία αγορά, που συνέβη σε συγκεκριμένη ώρα και μέρος. Αυτή η αγορά μπορεί να συνδεθεί με έναν εξακριβωμένο πελάτη ή με έναν μη εξακριβωμένο πελάτη. Το σύνολο δεδομένων με ποικίλες συναλλαγές μπορεί να ιδωθεί σε έναν πίνακα συσχέτισης. Ο πίνακας είναι ένα σύνολο όλων των συναλλαγών

$$T = \{T_1, T_2, T_3, \ldots, T_n\}. \qquad (1)$$

όπου κάθε συναλλαγή μπορεί να μοντελοποιηθεί σαν ένα δυαδικό διάνυσμα. Το T είναι ένα διάνυσμα όπου t[k] = 1 αν η συναλλαγή που περιέχεται στο αντικείμενο I_k, και t[k] = 0 διαφορετικά.

$$I = \{I_1, I_2, I_3, \ldots, I_K\}. \qquad (2)$$

Βασισμένο σε αυτά τα χαρακτηριστικά (συναλλαγή, αντικείμενο), το καλάθι ηλεκτρονικής αγοράς μπορεί να οριστεί ως N αντικείμενα που αγοράζονται μαζί με μεγάλη συχνότητα. Το επόμενο βήμα είναι να ορίσουμε όλους του πελάτες που έχουν αγοράσει N-m αντικείμενα στο καλάθι και να προτείνουμε την αγορά μερικών m αντικειμένων που λείπουν. Για να πάρουμε αποφάσεις στο marketing, η ανάλυση του basket marketing είναι ένα ισχυρό εργαλείο που ενισχύει την εφαρμογή των cross-selling στρατηγικών.

Εξαγωγή μοντέλου

Για να εφαρμόσουμε ανάλυση θεματικών μοντέλων χρειάζεται να σκεφτούμε ένα σώμα κειμένων όπου αποτελούνται από όρους. Οι όροι στην περίπτωσή μας είναι τα προϊόντα που είναι διαθέσιμα προς πώληση. Θεωρούμε ότι κάθε κείμενο σχηματίζεται είτε από τα προϊόντα που αγοράζονται μαζί σε μία συναλλαγή είτε από τα προϊόντα που αγοράζονται από έναν πελάτη.

Κρυφά καλάθια

Πρώτα, θεωρούμε κάθε συναλλαγή ως ένα κείμενο. Τα προϊόντα που αγοράζονται μαζί από έναν πελάτη κατά την διάρκεια μίας "βόλτας" θεωρούνται ως ένα κείμενο που δημιουργήθηκε χρησιμοποιώντας λέξεις από μία λίστα προϊόντων. Αυτό το σύνολο αντικειμένων θεωρείται ότι είναι ένα αποτέλεσμα ενός γενετικού μοντέλου θεμάτων το οποίο προσπαθούμε να υπολογίσουμε. Αυτά τα αποτελέσματα εισέρχονται σε μία συλλογή πιθανοτικών συνόλων δεδομένων και μπορούν να θεωρηθούν ως κρυφά καλάθια.

Κρυφοί χρήστες

Δεύτερον, θεωρούμε μία σειρά από συναλλαγές που έχουν γίνει από έναν πελάτη ως ένα κείμενο. Τα προϊόντα που έχουν αγοραστεί από έναν πελάτη κατά την διάρκεια του χρόνου θεωρούνται ως ένα κείμενο που δημιουργήθηκε χρησιμοποιώντας λέξεις από την λίστα προϊόντων. Αυτό το σύνολο αντικειμένων θεωρείται ως ένα αποτέλεσμα ενός γενετικού μοντέλου θεμάτων όπου προσπαθούμε να περιγράψουμε χρησιμοποιώντας θεματικά μοντέλα. Τα θέματα αυτών των μοντέλων θεωρείται ότι αντανακλούν στις αρεσκείες των καταναλωτών σε μία χρονική περίοδο.

Εξάγοντας συστάσεις

Στη συνέχεια, χρησιμοποιούμε τα εξαγόμενα κρυφά μοντέλα θεμάτων για να δώσουμε κάποιες συστάσεις προϊόντων στους χρήστες. Για κάθε διαθέσιμο αντικείμενο, υπολογίζουμε την ομοιότητά του με τα αντικείμενα που ήδη βρίσκονται στο καλάθι του χρήστη. Τα πιο όμοια αντικείμενα, στην συνέχεια, συστήνονται.

Θέτουμε σε εφαρμογή την διαδικασία Gibbs sampling για να συσχετίσουμε τα προϊόντα μεταξύ τους. Ενισχύουμε, επίσης, την ανάθεση προϊόντων σε πολλά θέματα και συνδυάζουμε τα δύο μοντέλα του καλαθιού και του χρήστη.

Οι παρακάτω τεχνικές έχουν ελεγχθεί για αντικείμενα σύσταση.

1. Κρυφά καλάθια – Gibbs Sampler. Σε αυτή την περίπτωση τα κρυφά καλάθια χρησιμοποιούνται για να προβλέψουν την συμπεριφορά του χρήστη, δεδομένων των αντικειμένων που έχει το καλάθι του αυτή τη στιγμή. Ένας Gibbs sampler χρησιμοποιείται για να εξάγει την πιθανοτική κατανομή των γνωστών αντικειμένων που περιέχονται στο καλάθι. Επομένως τα αντικείμενα με την μεγαλύτερη συσχέτιση με αυτή την κατανομή που δεν βρίσκονται ακόμη μέσα στα γνωστά αντικείμενα, προτείνονται στον χρήστη.
2. Θησαυρός γνώσεων των κρυφών καλαθιών. Όπως και στην προηγούμενη περίπτωση, τα κρυφά καλάθια χρησιμοποιούνται με σκοπό να προβλέψουν την συμπεριφορά του χρήστη. Όμως αντί να χρησιμοποιήσουν στατιστική συμπερασματολογία, ένας *θησαυρός γνώσεων* δημιουργείται από την δημιουργία θεματικών μοντέλων. Για να απομακρύνουμε την εξάρτηση των κειμένων από τους υπολογισμούς μας, εξετάζουμε τις σχέσεις όρο ανά όρο αντί των σχέσεων όρου ανά κείμενο. Η δομή που προκύπτει συνδέει τις λέξεις και τις ομοιότητες μεταξύ τους.

Εδώ θεωρούμε ότι το $S_{LBi,j}$ είναι η υπολογισμένη ομοιότητα μεταξύ δύο αντικειμένων i kai j, κάτι που έχει προκύψει από το θεματικό μοντέλο κρυφών καλαθιών. Χρησιμοποιούμε την εξίσωση 3 για να υπολογίσουμε την ομοιότητα των γνωστών αντικειμένων (KI) του καλαθιού του χρήστη για n διαφορετικά πιθανά αντικείμενα. Στην συνέχεια οι κορυφαίες επιλογές συγκρίνονται με τις αληθινές επιλογές του χρήστη.

$$\{w_1, w_2, w_3, \ldots, w_n\} = \{\sum_{j \in KI} S_{LB1,j}, \sum_{j \in KI} S_{LB2,j}, \sum_{j \in KI} S_{LB3,j}, \ldots, \sum_{j \in KI} S_{LBn,j}\}. \quad (3)$$

3. Κρυφά καλάθια με ενίσχυση συνανάθεσης. Αυτή η περίπτωση είναι παρόμοια με αυτή του θησαυρού, αλλά σε αυτή την περίπτωση τα αντικείμενα που συσχετίζονται με περισσότερα του ενός των γνωστών αντικειμένων από το καλάθι του χρήστη παίρνουν μία μικρή ενίσχυση. Μετράμε τον αριθμό των παρόμοιων αντικειμένων μέσα στο καλάθι του χρήστη και χρησιμοποιούμε αυτόν των αριθμό ως μία δύναμη για τον παράγοντα συνανάθεσης. Στην εξίσωση 4, το M είναι ο αριθμός των αντικειμένων που έχουν βρεθεί παρόμοια στο αντικείμενο 1.

$$w_l = b^{M-1} \sum_{j \in KI} S_{LBl,j} \quad (4)$$

4. Θησαυρός κρυφών χρηστών. Σε αυτή την περίπτωση χρησιμοποιούμε κρυφά θεματικά μοντέλα που έχουν εξαχθεί από τις προτιμήσεις των χρηστών σε κάποιο χρονικό διάστημα. Δημιουργούμε έναν θησαυρό όπου το $S_{LUi,j}$ είναι η υπολογισμένη ομοιότητα μεταξύ των αντικειμένων i και j προερχόμενη από το θεματικό μοντέλο κρυφών χρηστών. Χρησιμοποιούμε αυτές τις ομοιότητες για να προβλέψουμε προτεινόμενα αντικείμενα, δεδομένων των γνωστών αντικειμένων στο καλάθι του χρήστη έχουμε την εξίσωση

$$\{w_1, w_2, w_3, \ldots, w_n\} = \{\sum_{j \in KI} S_{LU1,j}, \sum_{j \in KI} S_{LU2,j}, \sum_{j \in KI} S_{LU3,j}, \ldots, \sum_{j \in KI} S_{LUn,j}\}. \quad (5)$$

5. Κρυφά καλάθια συνδυασμένα με κρυφούς χρήστες. Σε αυτή την περίπτωση η σύσταση των κρυφών θεματικών μοντέλων συμπληρώνεται από ένα μοντέλο κρυφών χρηστών, χρησιμοποιώντας μία παράμετρο ανάμειξης μ, $0<\mu<1$. Για να υπολογίσουμε την ομοιότητα του κειμένου 1 με τα γνωστά αντικείμενα στο καλάθι του χρηστή χρησιμοποιούμε την 6

$$w_l = (1-\mu) \sum_{j \in KI} S_{LBl,j} + \mu \sum_{j \in KI} S_{LUl,j} \quad (6)$$

4. 5 Ανάλυση Οντοτήτων με LDA

4.5.1 Εισαγωγή

Σε πολλές εφαρμογές, υπάρχει μία ποικιλία τρόπων να αναφερθούμε στην ίδια υπονοούμενη οντότητα. Δεδομένης μίας συλλογής από αναφορές οντοτήτων, αναφορές για συντομία, θα θέλαμε α) να καθορίσουμε την συλλογή των "αληθινών" υποκείμενων οντοτήτων και β) να αντιστοιχίσουμε τις αναφορές της συλλογής σε αυτές τις οντότητες. Το πρόβλημα αναδύεται σε πολλές όψεις στην επιστήμη της πληροφορικής. Στα παραδείγματα περιέχονται η μηχανική όραση, όπου χρειάζεται να βρούμε πότε δύο περιοχές δυο διαφορετικών εικόνων αναφέρονται στο ίδιο υποκείμενο αντικείμενο, η επεξεργασία φυσικής γλώσσας όπου θα θέλαμε να καθορίσουμε ποιες φράσεις ουσιαστικών αναφέρονται στην ίδια υποκείμενη οντότητα, και οι βάσεις δεδομένων, όπου όταν ενώνουμε δύο βάσεις ή καθαρίζουμε μία βάση, χρειάζεται να καθορίσουμε δύο στοιχεία αναφέρονται στο ίδιο άτομο.

Ενδιαφερόμαστε στην ανάλυση αναφορών όταν αυτές είναι συνδεδεμένες η μία στη άλλη μέσω συσχετισμένων συνδέσμων, όπως στο πεδίο της βιβλιογραφίας όπου τα ονόματα των συγγραφέων στις εργασίες συνδέονται με συνδέσμους συνεργατών. Η ανάλυση οντοτήτων γίνεται πιο συλλογική καθώς οι αποφάσεις εξαρτώνται από άλλους μέσα από σχετικούς συνδέσμους.

Υπάρχει ένα αρκετά μεγάλο ιστορικό εργασιών στην γενική και την σχετική ανάλυση οντοτήτων. Πρόσφατα, γενετικές και διακριτές πιθανοτικές προσεγγίσεις έχουν προταθεί καθώς επίσης και μη πιθανοτικοί αλγόριθμοι. Το μοντέλο μας διαφέρει από τα περισσότερα παραπάνω στο ότι είναι μη επιβλεπόμενο, δεν υποθέτει ότι οι υποκείμενες οντότητες είναι γωνστές, δεν παίρνει ανά ζεύγος αποφάσεις και μοντελοποιεί με σαφήνεια τις σχέσεις μεταξύ των οντοτήτων χρησιμοποιώντας ομαδοποίηση μελών.

Εισάγουμε το γενετικό πιθανοτικό μοντέλο στην ανάλυση οντοτήτων, που βασίζεται στον LDA. Διαφορετικά από τα υπάρχουσα μοντέλα, δεν εισάγουμε κάποια μεταβλητή απόφασης για κάθε πιθανό ζευγάρι αναφοράς, αλλά αντί αυτού έχουμε μία ετικέτα οντότητας για κάθε αναφορά. Για να μοντελοποιήσουμε σχέσεις συνεργασίας μεταξύ οντοτήτων, εισάγουμε μία ετικέτα ομάδας για κάθε αναφορά, έτσι ώστε οι οντότητες που προέρχονται από την ίδια ομάδα συνεργασίας να είναι πιο πιθανό να παρατηρηθούν σε μία συσχέτιση. Για την ανάλυση συγγραφέων στη βιβλιογραφία, αυτό σημαίνει ότι μοντελοποιούμε τις ομάδες συνεργασίας για να εξηγήσουμε τις σχέσεις συνεργασίας συγγραφής για το ίδιο έργο. Η γενετική διαδικασία στο μοντέλο μας μπορεί να κατανοηθεί ως μία διαδικασία Dirichlet του mixture μοντέλου: οι ετικέτες ομάδας στο μοντέλο επηρεάζουν την επιλογή οντοτήτων για κάθε αναφορά συγγραφέα σε ένα έργο.

4.5.2 Ένα παράδειγμα

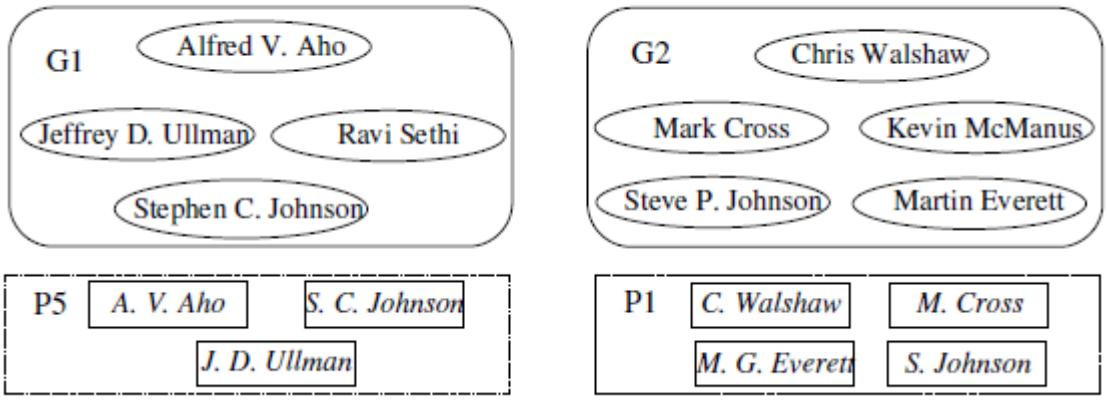
Σε αυτή την ενότητα, δείχνουμε ένα βιβλιογραφικό παράδειγμα για να εξηγήσουμε το πρόβλημα ανάλυσης οντοτήτων για συγγραφείς. Φανταστείτε εάν παράδειγμα έξι βιβλιογραφικών παραπομπών P1 έως P6 από το CiteSeer:

P1: "JOSTLE: Partitioning of Unstructured Meshes for Massively Parallel Machines" C. Walshaw, M. Cross, M. G. Everett, S. Johnson
P2: "Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies", C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus
P3: "Dynamic Mesh Partitioning: A Unified Optimisation and Load-Balancing Algorithm", C. Walshaw, M. Cross, M. G. Everett
P4: "Code Generation for Machines with Multiregister Operations", Alfred V. Aho, Stephen C. Johnson, Jefferey D. Ullman
P5: "Deterministic Parsing of Ambiguous Grammars", A. V. Aho, S. C. Johnson, J. D. Ullman
P6: "Compilers: Principles, Techniques, and Tools", A. Aho, R. Sethi, J. Ullman

Κάθε ένα από τα 6 έργα έχει τις δικές του αναφορές συγγραφέων. Για παράδειγμα, το πρώτο έργο P1 έχει τέσσερις αναφορές "C. Walshaw", "M. Cross", "M. G. Everett" και "S. Johnson". Συνολικά έχουμε 21 αναφορές στα 6 έργα. Ο στόχος είναι να βρούμε σε πόσες διαφορετικές οντότητες συγγραφέων αυτές οι αναφορές ανταποκρίνονται και ποια αναφορά αντιστοιχεί σε κάθε οντότητα. Είναι φανερό ότι όλα τα ονόματα Aho αντιστοιχούν στον ίδιο συγγραφέα, το ίδιο και για τα ονόματα Everret και Ullman. Το ενδιαφέρον όμως συγκεντρώνεται στις δύο οντότητες Johnson: αυτές των έργων P4 και P5 αντιστοιχούν στον Stephen C. Johnson από τα εργαστήρια Bell, ενώ αυτές των έργων P1 και P2 αντιστοιχούν στον Steve P. Johnson από το πανεπιστήμιο του Greenwich στο Λονδίνο. Όμως, κοιτάζοντας μόνο τα ονόματα των αναφορών δεν είναι ξεκάθαρο γιατί ο "Stephen C. Johnson" δεν είναι "S. Johnson", όταν ο "Alfred V. Aho" είναι ο ίδιος με τον "A. Aho". Στόχος μας θα είναι να κάνουμε χρήση των σχέσεων συνεργασίας για να εξάγουμε τα παραπάνω δύο αντικρουόμενα παραδείγματα ταυτόχρονα. Θα θέλαμε να συμπεράνουμε από τις συνεργασίες ότι υπάρχουν δύο διαφορετικές ομάδες συνεργασίας και οι συγγραφείς είναι πιο πιθανό να δημοσιεύσουν κείμενα με τους συγγραφείς της ίδιας ομάδας. Όπως απεικονίζεται στην Εικόνα 4.8 η πρώτη ομάδα G1 έχει τους Aho, Ullman και Sethi ως μέλη. Η άλλη ομάδα, η G2, έχει τους Walshaw, Cross, Everett, και McManus. Ο Steven C. Johnson συσχετίζεται με την πρώτη ομάδα συνεργασίας, ενώ ο S Johnson των εργασιών P1 και P2 είναι διαφορετικό άτομο αφού συσχετίζεται με την δεύτερη ομάδα συνεργασίας.

Για να βγάλουμε αυτά τα συμπεράσματα, το μοντέλο μας εισάγει μια ετικέτα οντοτήτων και μία ετικέτα ομάδας για κάθε αναφορά, και οι δύο είναι κρυφές και πρέπει να βρεθούν. Η διαδικασία συμπερασματολογίας είναι συλλογική, έτσι δεν μπορούμε να θεωρήσουμε κάθε αναφορά ανεξάρτητη – οι συσχετίσεις με τις άλλες αναφορές πρέπει να ληφθούν υπόψη. Επίσης, η ομάδα και οι ετικέτες οντότητας είναι διεξαρτώμενες. Οι ετικέτες οντότητας για τους δύο Johnson εξαρτώνται από τις ετικέτες ομάδας, όπως έχουμε δει. Επίσης, οι ετικέτες ομάδας εξαρτώνται από τις ετικέτες οντοτήτων. Ο Sethia από το έργο P6 και ο Johnson από το P5 ανήκουν στην ίδια ομάδα αφού και οι δύο είναι συνδεδεμένοι με τις ίδιες ετικέτες οντότητας των Aho και Ullman στα δύο αυτά έργα. Αυτές οι δύο κρυμμένες μεταβλητές

είναι οι δύο διαφορές του μοντέλου μας σε σχέση με τα υπόλοιπα. Οι περισσότερες προσεγγίσεις εισαγάγουν μία μεταβλητή απόφασης για κάθε πιθανό ζεύγος για να καταλάβουν αν αντιστοιχούν ή όχι στην ίδια οντότητα, ενώ εμείς εισάγουμε δύο μεταβλητές για κάθε αναφορά δεδομένων. Καθώς το μέγεθος των δεδομένων μεγαλώνει, πιστεύουμε ότι αυτή η διαφορά θα επιφέρει σημαντική επιρροή.



Εικόνα 4.8: Οι Οντότητες συγγραφέων σε δύο διαφορετικές ομάδας συνεργασίας και δύο κείμενα που έχουν δημιουργηθεί. Τα οβάλ είναι οντότητες που ανήκουν σε ομάδες στα κυκλικά ορθογώνια. Τα ορθογώνια με την διακεκομμένη γραμμή απεικονίζουν έργα με αναφορές συγγραφέων που απεικονίζονται ως μικρότερα ορθογώνια. Κάθε έργο έχει δημιουργηθεί από την ομάδα ακριβώς από πάνω του.

Είναι σημαντικό να σημειώσουμε ότι ο ρόλος των έργων P3 και P6 είναι σημαντικός για την εξαγωγή ενός συλλογικού συμπεράσματος για τον Johnson, παρόλο που κανένα από τα δύο δεν περιέχει κάποια αναφορά για αυτόν. Βοηθούν στην ενίσχυση την πεποίθησής μας ότι υπάρχουν δύο ξεχωριστές ομάδες ή κοινότητες όπου οι συγγραφείς-μέλη συνεργάζονται ο ένας με τον άλλο. Παρατηρήστε ότι οι συχνές συνεργασίες μεταξύ Walshaw και Aho, και μεταξύ Everett και Ullman θα είχαν αντίθετο αποτέλεσμα. Έτσι θα σκεφτόμασταν ότι υπάρχει μόνο μία ομάδα και ότι όλοι οι Johnson είναι πιο πιθανό να αναφέρονται στο ίδιο πρόσωπο.

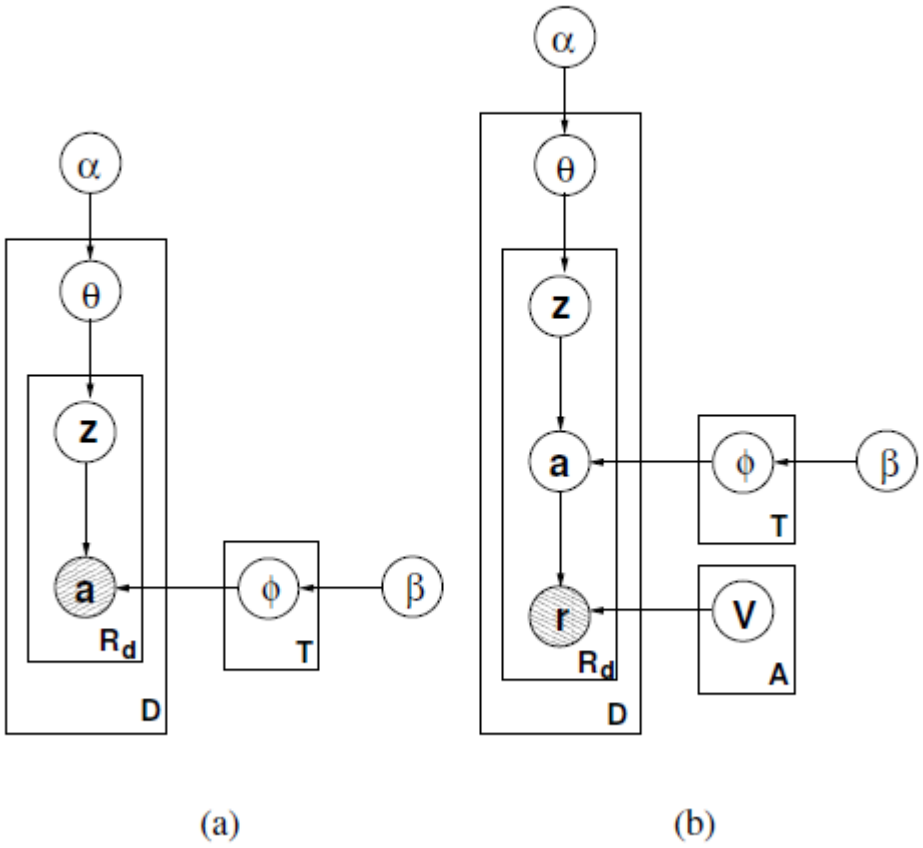
4.5.3 Το LDA Μοντέλο για Συγγραφείς

Σε αυτή την υποενότητα, θα δείξουμε πως το LDA μοντέλο για θέματα και λέξεις μπορεί να προσαρμοστεί σε ένα mixture μοντέλο ομάδας για οντότητες συγγραφέων. Αρχίζουμε με την απλούστερη περίπτωση όπου δεν υπάρχει αμφιβολία για τις αναφορές συγγραφέων. Στην επόμενη ενότητα επεκτείνουμε το μοντέλο για να χειριστούμε αμφίβολες αναφορές συγγραφέων.

Θεωρήστε την συλλογή D κειμένων και ένα σύνολο A συγγραφέων που έχουν γράψει αυτά τα κείμενα. Έχουμε ένα σύνολο από R αναφορές, $\{a_1, \ldots, a_R\}$ σε αυτά τα D κείμενα. Κάθε κείμενο μπορεί να έχει πολλούς συγγραφείς και προ στιγμήν, ας υποθέσουμε ότι οι συγγραφείς για κάθε κείμενο είναι γνωστοί. Για την i-οστή αναφορά, το a_i υποδεικνύει τον συγγραφέα που αντιστοιχεί σε αυτή και το d_i δηλώνει το κείμενο όπου αυτή υπάρχει. Ακόμη εισαγάγουμε την ιδέα των ομάδων συνεργασίας συγγραφέων. Αυτές είναι ομάδες όπου τα

μέλη της καθεμίας τείνουν να γράφουν μαζί. Κάθε αναφορά συγγραφέα α_i, είναι συσχετισμένη με μία ετικέτα ομάδας z_i.

Το πιθανοτικό μοντέλο απεικονίζεται στην Εικόνα 4.9. Η πιθανοτική κατανομή των συγγραφέων για κάθε ομάδα αντιπροσωπεύεται από μία πολυωνυμική μεταβλητή με παραμέτρους φ^j, έτσι η πιθανότητα $P(\alpha = i \mid z = j)$ του i-οστού συγγραφέα στην βάση δεδομένων που έχει επιλεχθεί από την j-οστή ομάδα είναι φ^j_i. Έχουμε T διαφορετικές πολυωνυμικές μεταβλητές, μία για κάθε ομάδα. Κάθε έργο d έχει μοντελοποιηθεί ως ένα μείγμα T ομάδων. Η κατανομή που χρησιμοποιείται, είναι ξανά μία πολυωνυμική με παραμέτρους θ_d, έτσι η πιθανότητα $P_d(z = j)$ της j-οστής ομάδας που επιλέγεται από το κείμενο d είναι θ^d_j. Κάθε θ^d επιλέγεται από μία Dirichlet κατανομή με



Εικόνα 4.9: Απεικόνιση για (α) το LDA μοντέλο για συγγραφείς και (β) το LDA-ER μοντέλο για ανάλυση με αμφίβολες αναφορές. Οι γνωστές μεταβλητές έχουν σκιαστεί.

υπερπαράμετρο α. Παρόμοια κάθε φ^j επιλέγεται από μία Dirichlet κατανομή με υπερπαραμέτρους β.

Για να δείξουμε την γενετική διαδικασία στο μοντέλο, δείχνουμε πως οι συγγραφείς για το έργο P5 επιλέγονται στην Εικόνα 1. Πρώτα, μία κατανομή θ^d ως προς τις ομάδες συνεργασίας επιλέγεται για το έργο. Αυτές είναι πιθανότατα οι ομάδες που θα δώσουν τους συγγραφείς του κειμένου. Κάθε ομάδα έχει μία κατανομή φ_i ως προς τους πιθανούς συγγραφείς. Στο παράδειγμά μας, το φ^{G1} δίνει ίση πιθανότητα στους Aho, Ullman, Sethi, και Stephen C. Johnson και 0 διαφορετικά, ενώ το φ^{G2} επιλέγει μεταξύ των Walshaw, Cross,

Everett, Steve Johnson και McManus με ίση πιθανότητα. Σημειώστε ότι το μοντέλο μας, επιτρέπει σε έναν συγγραφέα να ανήκει σε πολλές ομάδες, παρόλο που δεν απεικονίζεται κάτι τέτοιο εδώ. Η κατανομή θ^d που επιλέγεται για το έργο P1 έχει πιθανότητα 1 για την ομάδα G1 και 0 πιθανότητα για τις άλλες ομάδες. Κάθε συγγραφέας επιλέγεται πρώτα επιλέγοντας μία ομάδα z_i από το θ^d και μετά επιλέγοντας έναν συγγραφέα από την ομάδα z_i. Αφού το θ^d για το P1 δεν έχει μηδενική πιθανότητα μόνο για την ομάδα G1, είναι η ομάδα που επιλέγεται για κάθε συγγραφέα στο P1. Έχοντας επιλέξει την G1 ως την ομάδα για κάθε συγγραφέα, η πρώτη επιλογή από την φ^{G1} δίνει τον Aho ως τον πρώτο συγγραφέα, η δεύτερη δίνει Stephen C. Johnson και η τρίτη δίνει τον Ullman. Οι συγγραφείς των άλλων έργων επιλέγονται παρόμοια. Σημειώστε ότι περισσότερες της μίας ομάδας μπορεί να έχουν μη μηδενική πιθανότητα στην κατανομή θ^d για ένα έργο, έτσι ώστε οι συγγραφείς του ίδιου έργου να μπορούν να προέρχονται από διάφορες ομάδες με μικρότερη πιθανότητα.

4.5.4 LDA-ER μοντέλο για ανάλυση συγγραφέων

Στην προηγούμενη υποενότητα, υποθέσαμε ότι η ταυτότητα του συγγραφέα καθορίζεται χωρίς αμφιβολία από κάθε αναφορά συγγραφέα. Όμως, στην πραγματικότητα δεν συμβαίνει κάτι τέτοιο. Ο ίδιος συγγραφέας μπορεί να αντιπροσωπεύεται με διάφορους τρόπους: “Alfred V. Aho”, “Alfred Aho”, “AV Aho” και ούτω καθεξής. Μπορεί να υπάρχουν τυπογραφικά λάθη. Τέλος, δύο “S. Johnson” μπορεί να μην αναφέρονται στην ίδια οντότητα. Η μία αναφορά μπορεί να αναφέρεται στον “Stephen C. Johnson” και η άλλη στον “Steve P. Johnson”. Το αποτέλεσμα είναι ότι δεν είμαστε πλέον σίγουροι για την αντιστοίχηση της αναφοράς ενός συγγραφέα στην οντότητα του συγγραφέα. Πρέπει να καταφύγουμε στην συμπερασματολογία για να ταυτοποιήσουμε τον σωστό συγγραφέα σε κάθε αναφορά.

Για να το καταφέρουμε αυτό, θα συσχετίσουμε μία ιδιότητα υ_α με έναν συγγραφέα α. Επιπρόσθετα, προσθέτουμε ένα επιπλέον επίπεδο στο μοντέλο που πιθανοτικά επεξεργάζεται τις ιδιότητες V_α για να δημιουργήσει αναφορές $\mathbf{r} = \{r_1, r_2, \ldots, r_R\}$. Κάθε αναφορά δημιουργείται από τη πρώτη επιλογή μίας ομάδας z και μετά από μία οντότητα συγγραφέα όπως και πριν. Στην συνέχεια, η αναφορά συγγραφέα r δημιουργείται από το α καθώς τροποποιείται η ιδιότητα υ_α σύμφωνα με ένα μοντέλο θορύβου N. Η πιθανότητα δημιουργίας μίας αναφοράς r από μία συγκεκριμένη οντότητα συγγραφέα δηλώνεται ως $P(r|\upsilon_\alpha)$. Οι δεσμευμένες κατανομές για κάθε αναφορά κανονικοποιούνται ως προς άθροισμα 1 των οντοτήτων συγγραφέων. Είναι η αναφορά r που είναι γνωστή, καθώς η οντότητα α και η ετικέτα ομάδας z είναι κρυφές μεταβλητές. Το LDA-ER μοντέλο απεικονίζεται στην Εικόνα 4.9(b).

Εφαρμόζοντας το μοντέλο τις Εικόνας 4.9(b) στο παράδειγμά μα, έχουμε ήδη δει πως οι τρεις οντότητες συγγραφέων επιλέγονται για το έργο P1. Οι ιδιότητες υ_α για τους τρεις συγγραφείς είναι “Alfred V. Aho”, “Stephen C. Johnson”, και “Jeffrey D. Ullman”. Όμως τα ολόκληρα ονόματά τους δεν εμφανίζονται πάντα στα έργα ή στις παραπομπές. Σε αυτή την περίπτωση η διαδικασία θορύβου τροποποιεί τις ιδιότητες των τριών επιλεγμένων οντοτήτων για να δημιουργήσει τα “A. V. Aho”, “S. C. Johnson” και “J. D. Ullman” ως τις τρεις αναφορές συγγραφέων στο έργο.

Η πιθανότητα δημιουργίας του συνόλου **r** των αναφορών για ένα σώμα δεδομένων των παραμέτρων *α, β* και *V* είναι

$$
\begin{aligned}
P(\mathbf{r};\alpha,\beta,\mathbf{V}) &= \prod_d P(\mathbf{r}_d;\alpha,\beta,\mathbf{V}) \\
&= \prod_d \sum_{\mathbf{a}_d} P(\mathbf{r}_d \mid \mathbf{a}_d;\mathbf{V}) P(\mathbf{a}_d;\alpha,\beta) \\
&= \int_\phi P(\phi;\beta) \prod_d \sum_{\mathbf{a}_d} P(\mathbf{r}_d \mid \mathbf{a}_d;\mathbf{V}) \\
&\times \int_\theta P(\theta;\alpha) P(\mathbf{a}_d \mid \theta,\phi) d\theta d\phi
\end{aligned}
$$

ΚΕΦΑΛΑΙΟ 5
Εφαρμογή LDA, Ιεραρχικού LDA και Correlated LDA

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο υπήρξε μια αναλυτική μαθηματική περιγραφή του LDA και των υπόλοιπων βασικών παραλλαγών του. Σε αυτό το κεφάλαιο, θα παρουσιάσουμε τρεις από τους αλγόριθμους τους LDA ,τον HLDA και CLDA γραμμένους σε java, μέρος του συστήματος Mallet.

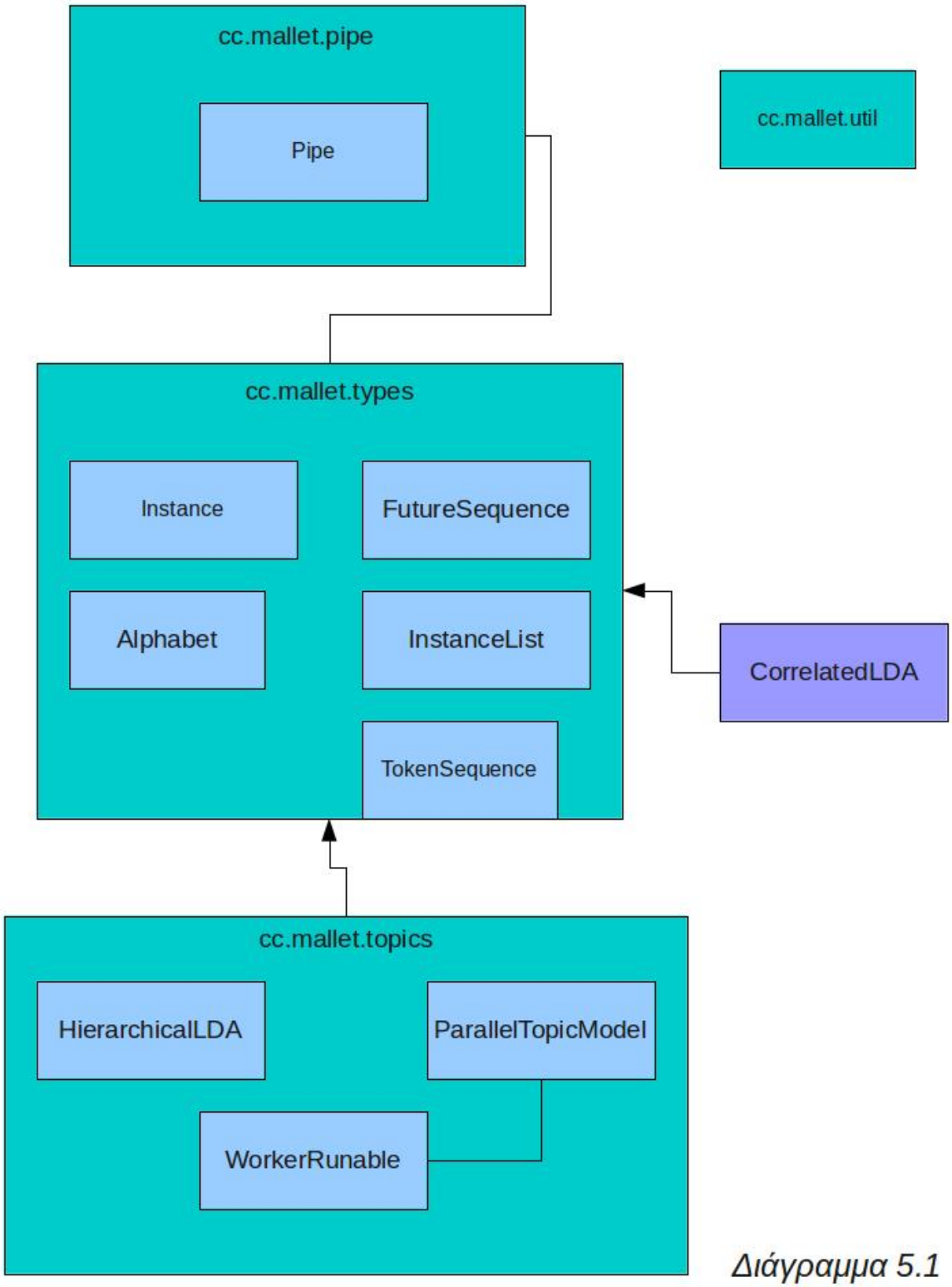
Το Mallet είναι ένα java πακέτο με ποικίλες εφαρμογές, όπως στατιστική επεξεργασία της φυσικής γλώσσας, αρχικοποίηση κειμένων, ομαδοποίηση, *μοντελοποίηση θεμάτων,* εξαγωγή πληροφορίας και εφαρμογές μηχανικής μάθησης. Επιπλέον, εκτός από τις έξυπνες εφαρμογές μηχανικής μάθησης, το Mallet περιέχει κάποιες διαδικασίες ρουτίνας για να μετατραπούν τα κείμενα σε αριθμητικές απεικονίσεις ώστε να επεξεργαστούν αποτελεσματικά. Αυτή η διαδικασία εφαρμόζεται μέσω ενός ευέλικτου συστήματος από "σωλήνες (pipes)", που διαχειρίζονται διάφορες λειτουργίες όπως να τεμαχίζουν τα κείμενα σε strings, να αφαιρούν τις τετριμμένες λέξεις (stop worlds), και να μετατρέπουν σειρές σε αριθμητικά διανύσματα. Για περισσότερες πληροφορίες ανατρέξτε στην ιστοσελίδα "http://mallet.cs.umass.edu/index.php".

Τους αλγόριθμους αυτούς θα τους τρέξουμε για το σώμα *Polarity Dataset Version 2.0* στο επόμενο κεφάλαιο και θα εκθέσουμε τα αποτελέσματά τους. Τέλος θα ακολουθήσει μια παρόμοια ανάλυση ενός αλγορίθμου CTM, που έχει δημιουργηθεί από τον συγγραφέα αυτού του έργου Για την υλοποίηση αυτού του αλγόριθμου χρησιμοποιήθηκαν κλάσεις από το Mallet.

Συνεπώς η δομή αυτού του κεφαλαίου έχει ως εξής: Στην πρώτη ενότητα του κεφαλαίου θα παρουσιάσουμε ένα μέρος του διαγράμματος του Mallet που αφορά την θεματική μοντελοποίηση. Εκεί θα αναλυθούν οι λειτουργίες και οι σχέσεις των πακέτων καθώς και των βασικότερων κλάσεων τους. Στη δεύτερη ενότητα θα ασχοληθούμε με την επιμέρους παρουσίαση των βασικότερων μεθόδων των τριών βασικών κλάσεων που υλοποιούνται οι αλγόριθμοι θεματικών μοντέλων, ParallelTopicModel (Υλοποίηση του LDA), HierarchicalLDA (Υλοποίηση του ιεραρχικού LDA), και CorrelatedLDA (Υλοποίηση του correlated LDA) .

5.2 Παρουσίαση υποδιαγράμματος Mallet και Correlated LDA

Όπως είπαμε και στην εισαγωγή το Mallet είναι ένα σύστημα επεξεργασίας της φυσική γλώσσας. Παρακάτω δίνεται ένα μέρος του διαγράμματος του προγράμματος, αυτό που αφορά την θεματική μοντελοποίηση.



Διάγραμμα 5.1

Ακολουθεί η ανάλυσή των πακέτων που περιέχει.

Τα τέσσερα βασικά πακέτα του Mallet που συντελούν στην θεματική μοντελοποίηση είναι τα cc.mallet.topics, cc.mallet.types, cc.mallet.pipe, και cc.mallet.util. Το πακέτα cc.mallet.pipe και cc.mallet.types επικοινωνούν αμφίδρομα, κλάσεις δηλαδή του ενός γνωρίζουν τις κλάσεις του αλλού. Ενώ οι κλάσεις του cc.mallet.topics εισαγάγουν κλάσεις του cc.mallet.types. Τέλος υπάρχει και το πακέτο cc.mallet.util που υλοποιεί κάποιες βασικές λειτουργίες, όπως διαδικασίες γραμμής εντολών και μαθηματικές πράξεις, τις οποίες χρειάζονται όλες οι κλάσεις το πακέτων. Η κλάση CorrelatedLDA, η οποία δεν ανήκει στο Mallet αλλά αποτελεί προϊόν του συγγραφέα του παρόντος έργου, εισαγάγει κλάσεις από το cc.malle.types και cc.mallet.util. Η απεικόνιση του πακέτου γίνεται στο Διάγραμμα 5.1.

5.2.1 Σημαντικές κλάσεις του Mallet

Το πρώτο πακέτο, cc.mallet.types, περιέχει όλους τους σημαντικούς τύπους του Mallet που πρέπει να μετατραπούν τα δεδομένα για την επεξεργασία τους. Συμπεριλαμβάνονται οι Alphabet, FeutureVector, Instance, Label κτλ. Όταν εισάγουμε δεδομένα στο Mallet αυτά καταχωρούνται σε "σώματα πληροφοριών" τα οποία ορίζονται ως instance. Με λίγα λόγια μία instance είναι η " βασική μονάδα" επεξεργασίας πληροφορία του Mallet. Μία instance που υλοποιείται μέσω της κλάσης Instance. περιέχει τέσσερα γενικά πεδία με προορισμένο όνομα, τα: "data", "targer" , "name", και "source". Στο πεδίο "Data" βρίσκονται όλα τα δεδομένα που υπάρχουν στην instance, το "target' είναι ο δείκτης που συσχετίζεται με την instance, το "name" είναι ένα μικρό όνομα αναγνώρισης, και στο "source" καταγράφεται η πηγή πληροφοριών (όπως για παράδειγμα το αρχικό αρχείο).

Κανένα πεδίο δεν έχει ένα προορισμένο τύπο, και οι τύποι μπορεί να αλλάξουν καθώς η instance επεξεργάζεται. Για παράδειγμα, το κομμάτι δεδομένων μπορεί να αρχίσει ως ένα string ενός ονόματος αρχείου, ύστερα να περάσει από μια Pipe (μια κλάση που θα αναλύσουμε παρακάτω) σε μια κλάση CharSequence που απεικονίζει το περιεχόμενο του αρχείου, και τελικά να μετατραπεί σε ένα διάνυσμα "χαρακτηριστικών" που περιέχει δείκτες που δείχνουν στην κλάση Alphabet. Εξαρτάται από κάθε αντικείμενο της Pipe ποια πεδία θα αλλάξει, το πιο συνηθισμένο είναι να επεξεργαστεί το πεδίο "data".

Η κλάση Alphabet σχηματίζει μία αντίστοιχη μεταξύ ακεραίων και αντικειμένων με όποιο τρόπο κριθεί αυτή αναγκαία. Οι ακέραιοι επιθέτονται διαδοχικά, αρχίζοντας από το μηδέν, καθώς προσθέτονται στην Alphabet. Τα αντικείμενα δεν μπορούν να διαγραφούν από την Alphabet και έτσι οι ακέραιοι δεν ξαναχρησιμοποιούνται.

Η πιο συνηθισμένη χρήση της Alphabet είναι για την δημιουργία ενός λεξικού χαρακτηριστικών ονομάτων που συσχετίζονται με την κλάση FeutureVector σε μία Instance. Σε μία απλή εφαρμογή αρχικοποίησης κειμένου κάθε μοναδική λέξη ενός κειμένου θα έχει μια μοναδική είσοδο στην Alphabet με έναν μοναδικό ακέραιο που συσχετίζεται με αυτή. Οι FeutureVectors βασίζονται στο κομμάτι του ακεραίου για να απεικονίσουν αποτελεσματικά το υποσύνολο της Alphabet.

Το δεύτερο πακέτο, το pipe, περιέχει τις διαδικασίες που μετατρέπουν αφηρημένα δεδομένα σε δομές instance. Περιέχει την αφηρημένη υπερκλάση Pipe μέσω της οποίας μετατρέπονται δεδομένα ενός τύπου σε έναν άλλο. Οι "σωλήνες" (Pipes) χρησιμοποιούνται πολύ συχνά για την εξαγωγή χαρακτηριστικών. Παρόλο που η Pipe δεν έχει "αφηρημένες μεθόδους", για να χρησιμοποιήσουμε μία Pipe υποκλάση πρέπει να κάνουμε override είτε

την pipe μέθοδο είτε την newIteratorFrom μέθοδο. Η πρώτη είναι κατάλληλη όταν η διαδικασία “σωλήνα” μίας Instance είναι αυστηρώς ένα-προς-ένα. Για κάθε Instance που εισέρχεται, υπάρχει ακριβώς μία Instance που εξέρχεται. Η δεύτερη είναι κατάλληλη όταν η διαδικασία “σωλήνα” έχει καταλήξει σε περισσότερα ή λιγότερα δεδομένα τύπου Instance από αυτά που καταφθάνουν στον Iterator.

Ένας σωλήνας (pipe) διαχειρίζεται μία Instance, όπου είναι είναι ένας μεταφορέας, στην ουσία, πληροφοριών. Ο σωλήνας διαβάζει και γράφει κομμάτια της Instance όταν απαιτηθεί να επεξεργαστεί τα δεδομένα της. Εξαρτάται από τον σωλήνα ποια κομμάτια της Instance θα διαβάσει και θα γράψει, αλλά συνήθως θα διαβάσει την είσοδο και θα γράψει την έξοδο στο κομμάτι δεδομένων. Ένας σωλήνας δεν έχει κάποια γνώση της εισόδου ή της εξόδου. Ένα σύνολο βοηθητικών κλάσεων, που εφαρμόζει το interface Iterator, διατρέχει τις συχνά συναντούμενες εισόδους δεδομένων και παραδίδει τα στοιχεία αυτών των δομών δεδομένων σε σωλήνα.

Ένας σωλήνας χρησιμοποιείται συχνά από μια InstanceList. Καθώς εισαγάγονται δεδομένα τύπου Instance στην λίστα, επεξεργάζονται δια μέσου του σωλήνα που συσχετίζεται με την συγκεκριμένη InstanceList και η συγκεκριμένη δομή διατηρείται σε όλη την λίστα.

Ένας FileIterator παίρνει μία λίστα αρχείων για να τα επεξεργαστεί. Ο FileIterator διατρέχει κάθε φάκελο αρχείων, δημιουργώντας μια instance για κάθε αρχείο και βάζοντας τα δεδομένα από το αρχείο στο σώμα δεδομένων αυτής της instance. Ο φάκελος του αρχείου καταγράφεται στο χαρακτηριστικά της instance. Ο FileIterator αποθηκεύει τις δομές instance μιας InstanceList, η οποία τις διατρέχει μέσα από τον “σωλήνα” που έχει συσχετιστεί και κρατάει τα αποτελέσματα.

Οι Pipes μπορούν να συνταχθούν ιεραρχικά. Σε μια τυπική εφαρμογή, δημιουργείται μία SerialPipe, που κρατάει άλλες pipes της λίστας. Όταν βάζουμε μία instance σε μία SerialPipe σημαίνει ότι αυτή θα διατρέξει και από όλα τα παιδιά pipes της αλληλουχίας.

Ένας σωλήνας κρατάει δύο κλάσεις τύπου Alphabet: μία για τα σύμβολα (ονόματα χαρακτηριστικών) που βρίσκονται στα σώματα δεδομένων των instance και διατρέχονται μέσα από τον “σωλήνα”, και μία άλλη για τα σύμβολα που βρίσκονται στα target πεδία.

Τέλος υπάρχει και το πακέτο cc.mallet.util που περιέχει κάποιες βασικές λειτουργίες, όπως η διαδικασίες γραμμής εντολών και οι μαθηματικές πράξεις, τις οποίες χρειάζονται όλες οι κλάσεις το πακέτων.

5.3 Υλοποίηση Αλγορίθμων Θεματικών Μοντέλων

Το πακέτο cc.mallet.topics αποτελεί το επιφανειακό στρώμα της εφαρμογής των αλγορίθμων μοντέλων LDA και ιεραρχικού LDA. Ακολουθεί μια εκτενέστερη ανάλυση των λειτουργιών αυτών των κλάσεων, των μεθόδων τους καθώς και των μεταβλητών τους.

5.3.1 Υλοποίηση απλού LDA

Ο LDA υλοποιείται μέσω της κλάσης ParallelTopicModel.

ParallelTopicModel

Συγγραφείς: David Mimno, Andrew McCallum

Σύνοψη Πεδίων	
protected double[]	**alpha** Είναι η σταθερά άλφα του μοντέλου, ένα διάνυσμα θετικών τιμών, όπου μέσω Dirichlet κατανομής, καθορίζει τις μεταβλητές θ, τις κατανομής θεμάτων ανά κείμενο δηλαδή. Είναι προκαθορισμένη από το πρόγραμμα ως ένα μοναδιαίο διάνυσμα, ο χρήστης μπορεί να το αλλάξει δίνοντας μια τιμή εισόδου, κάθε στοιχείο του διανύσματος άλφα θα είναι ίσο με την τιμή της εισόδου δια τον συνολικό αριθμό των θεμάτων.
protected Alphabet	**alphabet** Ένα διάνυσμα αντιστοίχησης των λέξεων που έχουν βρεθεί στο σώμα με ακεραίους αριθμούς ώστε να μπορέσουν να επεξεργαστούν.
protected double	**beta** Η εκ των προτέρων αρχική κατανομή που δίνει την αναλογία λέξεων και θεμάτων, την πιθανότητα δηλαδή κάθε λέξης να αντιστοιχεί σε ένα θέμα. Το μοντέλο με μία δεύτερη Dirichlet με βάση αυτή την σταθερά σε συνδυασμό με την αναλογία θεμάτων και ενός κειμένου που έχει βρει, θα επιλέξει ένα θέμα για αυτή την μια λέξη του κειμένου. Η σταθερά αυτή είναι προκαθορισμένη ίση με ένα, ωστόσο ο χρήστης μπορεί άμεσα να την αλλάξει.
int	**burninPeriod** Ο αριθμός των επαναλήψεων της διαδικασίας Gibbs sampling όπου τα αποτέλεσμα δεν απόθηκεύονται αφού ο αλγόριθμος θεωρούμε ότι δεν έχει εκπαιδευτεί καλά μέχρι στιγμής. Είναι προκαθορισμένη ίση με 200. Και ο χρήστης μπορεί να την αλλάξει.
protected java.util.ArrayList <TopicAssignment>	**data** Είναι μια ArrayList, που περιέχει τις λέξη σε μορφή Instance καταχωρημένες σε αντιστοιχία μίας ετικέτας, όπου είναι ο αντίστοιχος αριθμός της λέξης στην Alphabet.
protected int[]	**docLengthCounts**

	Διάνυσμα όπου κάθε στοιχείο του περιέχει τον συνολικό αριθμό λέξεων κάθε κειμένου του σώματος εκπαίδευσης.
int	**numIterations** Ο αριθμός των επαναλήψεων της διαδικασίας Gibbs. Είναι ίσος με 1000 αλλά ο χρήστης έχει την δυνατότητα να τον αλλάξει.
protected int	**numTopics** Ο αριθμός των κειμένων που επιθυμεί ο χρηστής , ο αλγόριθμος να αναλύσει τα κείμενα. Είναι βασικό να μην είναι ούτε πολύ μεγάλος γιατί θα υπάρξουν θέματα πολύ εξειδικευμένα ούτε πολύ μικρός γιατί θα τα θέματα θα είναι πολύ γενικά χωρίς να παρέχουν κάποια χρήσιμη πληροφορία. Η σταθερά έχει τεθεί ίση με 20.
protected int	**numTypes** Ο συνολικός αριθμός των λέξεων της Alphabet τους σώματος κειμένων
protected int[]	**tokensPerTopic** Ο αριθμός των λέξεων-δειγμάτων που έχουν εναποτεθεί σε κάθε θέμα.
protected int[][]	**topicDocCounts** Ιστόγραμμα αναλογίας θεμάτων/κείμενων

Σύνοψη Μεθόδων	
void	**addInstances**(InstanceList training) Προσθέτει όλες τις λέξεις δομημένες σε μορφή Instance σε μία λίστα.
void	**estimate**() Η μέθοδος αυτή καλεί ένα αντικείμενο της κλάσης WorkerRunnable όπου με σε αυτή γίνονται κάποιες βασικές διεργασίες και βοηθά στην υλοποίηση του Gibbs sampling.
TopicInferencer	**getInferencer**() Υπολογίζει τις κατανομές θεμάτων για νέα κείμενα.
java.lang.Object[][]	**getTopWords**(int numWords) Επιστρέφει ένα διάνυσμα (ένα στοιχείο από κάθε θέμα) από διανύσματα λέξεων, που είναι οι πιο πιθανές λέξεις για το θέμα σε καθοδική σειρά.
static ParallelTopicModel	**read**(java.io.File f) Διαβάζει τα δεδομένα από ένα αρχείο
void	**setBurninPeriod**(int burninPeriod) Θέτει την περίοδο οπού τα αποτελέσματα του Gibbs sampling δεν αποθηκεύονται
void	**setNumIterations**(int numIterations) Θέτει τον αριθμό των επαναλήψεων του αλγορίθμου.
void	**setTopicDisplay**(int interval, int n) Θέτει τον αριθμό των λέξεων που θα εκθέτονται σε κάθε θέμα.
void	**write**(java.io.File serializedModelFile) Μεταφέρει τα αποτελέσματα σε ένα αρχείο.

5.3.2 Υλοποίηση Ιεραρχικού LDA

Η κλάση που υλοποιεί τον ιεραρχικό LDA λέγεται HierarchicalLDA. Αξίζει να σημειώσουμε πως υπάρχει και η HierarchicalLDATUI όπου είναι πιο φιλική και μπορεί ο χρήστης να την εκτελέσει μέσω γραμμής εντολών. Ακολουθεί η σύνοψη της HierachicalLDA.

HierachicalLDA

Σύνοψη Πεδίων	
int	**numLevels** Ο αριθμός των επιπέδων του δέντρου.
int	**numofDocuments** Ο συνολικός αριθμός των κειμένων του σώματος που δίνουμε στον αλγόριθμο για την εκπαίδευσή του.
int	**numTypes** Ο αριθμός των διαφορετικών λέξεων του σώματος κειμένου.
double	**alpha** Μια σταθερά για την ομαλοποίηση των αναλογιών θεμάτων ίση με 10.
double	**gamma** Ο αριθμός των "φανταστικών πελατών" στο επόμενο, μέχρι στιγμής μη χρησιμοποιημένο, τραπέζι. Ο αριθμός αυτός τίθεται από τον κατασκευαστή ίσος με 1.
double	**eta** Σταθερά ομαλοποίησης των κατανομών λέξεων/θεμάτων προκαθορισμένη ίση με 0.1.
NCRPNode[]	**DocumentLeaves** Αποθηκεύει το παρόν μονοπάτι του δέντρου.

Σύνοψη Μεθόδων	
void	**estimate**(int iterations) Η μέθοδος εφαρμόζει την διαδικασία Gibbs για των αριθμό των επαναλήψεων που έχουν οριστεί.
void	**initialize**(InstanceList instances, InstanceList testing, int numLevels, Randoms random) Αρχικοποίηση της διαδικασίας.
TopicInferencer	**samplePath**(int doc, int iteration) Επιλογή του μονοπατιού ενός κειμένου. Η μέθοδος αυτής χρησιμοποιείται από την estimate για να υπολογίσει τα μονοπάτια κειμένων για κάθε επανάληψη.
void	**sampleTopics**(int doc) Επιλέγεται θέμα για κάθε λέξη ενός κειμένου.
void	**setAlpha**(double alpha) Θέτει την σταθερά alpha.
void	**setEta**(double eta) Θέτει την σταθερά eta.
void	**setGamma**(double gamma) Θέτει την σταθερά gamma.
void	**setProgressDisplay**(boolean showProgress) Αυτή η παράμετρος καθορίζει αν θα παίρνουμε τα αποτελέσματα από κάθε επανάληψη Gibbs sampling ή όχι.
void	**setTopicDisplay**(int interval, int words) Καθορίζει τις διαστάσεις του πίνακα κατανομής θεμάτων ανά λέξη.

Υλοποίηση Correlated LDA

Ο αλγόριθμος του Correlated LDA δεν υπάρχει στο Mallet. Έγινε λοιπόν μια προσπάθεια από τον συγγραφέα της παρόντος έργου για υλοποίηση του συγκεκριμένου αλγορίθμου. Η επίτευξη αυτού του στόχου δεν ήταν εύκολη δεδομένου ότι δεν υπάρχουν πολλές πηγές που μπορεί να ανατρέξει κάποιος για να υλοποιήσει τον αλγόριθμο. Η υλοποίηση του έγινε με την βοήθεια της εργασίας Variational EM Algorithms for Correlated Topic Models (Mohammad Emtiyaz Khan και Guillaume Bouchard, 14 September 2009). Αξίζει να τονίσουμε ότι ο αλγόριθμος δεν χρησιμοποιεί για την εύρεση των κατανομών τον Gibbs sampling αλγόριθμο αλλά τον EM.

Ο αλγόριθμος θέτει μια βοηθητική κατανομή q_d για κάθε κείμενο d και προσπαθεί να ελαχιστοποιήσει την Kullback-Leibler διαφορά μεταξύ βοηθητικών κατανομών και πραγματικής κατανομής p.

$$-KL\left(q_d||p\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\log|\mathbf{V}_d| - \frac{1}{2}\left\{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{V}_d\right) + \left(\mathbf{m}_d - \boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{m}_d - \boldsymbol{\mu}\right)\right\} + \mathrm{cnst}$$

Σε γενικές γραμμές οι δομή του αλγόριθμου σύμφωνα με την εργασία των Emitiyaz και Boychard έχει ως εξής:

1. Αρχικοποίηση των $\boldsymbol{\mu}^{(0)} = \mathbf{m}_d = 0, \boldsymbol{\Sigma}^{(0)} = \mathbf{I}_T, \mathbf{V}_d = \mathbf{I}_T.$

όπου $\boldsymbol{\mu}^{(0)}$ και $\boldsymbol{\Sigma}^{(0)}$ το διάνυσμα των πραγματικών μέσων κατανομών των θεμάτων και ο πίνακας πραγματικής διασποράς θεμάτων που συσχετίζει τα θέματα μεταξύ τους και $\mathbf{m}_d$ και $\mathbf{V}_d$ το διάνυσμα μέσων κατανομών θεμάτων και ο πίνακας διασποράς θεμάτων του κειμένου d για την βοηθητικής κατανομή που ορίσαμε.

2. Επανάληψη μεταξύ E και M βημάτων μέχρι την σύγκλιση.

3. E-βήμα: Για d=1, 2, …, D, λύση των παρακάτω εξισώσεων,

$$\mathbf{S}_d\mathbf{c}_d - \boldsymbol{\Sigma}^{-1}(\mathbf{m}_d - \boldsymbol{\mu}) - \frac{W_d}{\xi_d}e^{\mathrm{diag}\left(\frac{1}{2}V_{tt,d}+m_{t,d}\right)_{1:T}} = 0$$

$$-\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\mathbf{V}_d^{-1} - \frac{W_d}{2\xi_d}e^{\mathrm{diag}\left(\frac{1}{2}V_{tt,d}+m_{t,d}\right)_{1:T}} = 0$$

όπου $\xi_d = \sum_t e^{\frac{1}{2}V_{tt,d}+m_{t,d}}$, $\mathbf{S}_d = [\mathbf{s}_1, \ldots, \mathbf{s}_V]$ με $\mathbf{s_u}$ ένα διάνυσμα $s_{u,t}\log\beta_{u,t}$ για κάθε θέμα t όπου $s_{u,t}$ και $\beta_{u,t}$ την πιθανότητα της λέξης u στο θέμα t της βοηθητικής κατανομής και της πραγματικής αντίστοιχα.

4. M-βήμα:

$$\boldsymbol{\mu} = \frac{1}{D}\sum_d \mathbf{m}_d$$

$$\boldsymbol{\Sigma} = \frac{1}{D}\sum_d \mathbf{V}_d + (\mathbf{m}_d - \boldsymbol{\mu})(\mathbf{m}_d - \boldsymbol{\mu})^T$$

$$\beta_{v,t} \propto \sum_d c_{v,d} s_{v,t,d}$$

$$s_{v,t,d} \propto \beta_{v,t} e^{m_{t,d}}$$

με κανονικοποίηση ως προ υ για την πρώτη ποσότητα και κανονικοποίηση t για την δεύτερη.

Το πρόγραμμα κατασκευάστηκε σε Java και χρησιμοποιεί τα πακέτα c.mallet.types, cc.mallet.pipe, και cc.mallet.util κυρίως για την τμηματοποίηση του κειμένου και την κατάλληλη επεξεργασία του. Ακολουθεί η σύνοψη.

CorrelatedLDA

Σύνοψη Πεδίων	
protected Alphabet	**alphabet** Το διάνυσμα αντιστοίχησης των λέξεων που έχουν βρεθεί με ακεραίους αριθμούς ώστε να μπορέσουν να επεξεργαστούν.
protected double[][]	**beta** Το διάνυσμα δύο διαστάσεων λέξεων/θεμάτων που δίνει την αναλογία λέξεων και θεμάτων, την πιθανότητα δηλαδή κάθε λέξης να αντιστοιχεί σε ένα θέμα. Αρχικά στο διάνυσμα μπαίνουν κάποιες τυχαίες τιμές .
protected java.util.ArrayList<TopicAssignment>	**data** Είναι μια ArrayList, που περιέχει τις λέξη σε μορφή Instance καταχωρημένες σε αντιστοιχία μίας ετικέτας, όπου είναι ο αντίστοιχος αριθμός της λέξης στην Alphabet.
protected double[][]	**sigma** Διάνυσμα τριών διαστάσεων κειμένων/θεμάτων/λέξεων. Είναι η βοηθητική κατανομή που εισαγάγουμε για να υλοποιήσουμε την ανίσωση Jensen. Στην αρχή δίνονται τυχαίες τιμές στον πίνακα μικρότερες του 1 και στην συνέχεια το η sigma κατασκευάζει την beta.
protected int	**numTopics** Ο αριθμός των κειμένων που επιθυμεί ο χρηστής , ο αλγόριθμος να αναλύσει τα κείμενα. Είναι βασικό να μην είναι ούτε πολύ μεγάλος γιατί θα υπάρξουν θέματα πολύ εξειδικευμένα ούτε πολύ μικρός γιατί θα τα θέματα θα είναι

	πολύ γενικά χωρίς να παρέχουν κάποια χρήσιμη πληροφορία. Η σταθερά έχει τεθεί ίση με 20.
protected static double[]	**mean** Η κρυμμένη πραγματική μέση τιμή των θεμάτων των κειμένων. Αρχικά το διάνυσμα παίρνει μηδενικές τιμές.
protected static double[][]	**Cov** Ο κρυμμένος πίνακας διασποράς διαστάσεων. θεμάτων/θεμάτων. Ο πίνακας αυτός μας δείχνει κατά πόσο τα θέματα συσχετίζονται μεταξύ τους. Αρχικά ο πίνακας
protected int[][]	**topicDocCounts** Ιστόγραμμα αναλογίας θεμάτων/κείμενων
protected int	**numTypes** Ο συνολικός αριθμός των λέξεων της Alphabet τους σώματος κειμένων
protected int[]	**tokensPerTopic** Ο αριθμός των λέξεων-δειγμάτων που έχουν εναποτεθεί σε κάθε θέμα.
protected static double[]	**x** Είναι η μεταβλητή ξ που υπολογίζεται σε κάθε βήμα Ε σύμφωνα με τον τύπο $$\xi_d = \sum_t e^{\frac{1}{2}V_{tt,d}+m_{t,d}}$$ όπου V και m που είναι ο πίνακας διασποράς θεμάτων και το διάνυσμα μέσης τιμής θεμάτων για δεδομένο κείμενο d.
protected static double[][]	**m** Ο πίνακας των διανυσμάτων μέσης τιμής της βοηθητικής κατανομής που έχουμε ορίσει για κάθε κείμενο.
protected int[]	**V** Πίνακας τριών διαστάσεων κείμενων/θεμάτων/θεμάτων όπου περιέχει τους πίνακες διασποράς της βοηθητικής κατανομής για κάθε κείμενο.

Σύνοψη Μεθόδων	
void	**InitializeEM**(InstanceList training) Η μέθοδος που αρχικοποιεί την διαδικασία. Δέχεται ως είσοδο το σώμα κειμένων σε μορφή InstanceList, θέτει αρχικές τιμές στις μεταβλητές ή καλεί τις μεθόδους που είναι υπεύθυνες για ανάθεση τιμών μεταβλητών. Και στην συνέχεια καλεί την estimate όπου αρχίζει να εφαρμόζεται ο Αλγόριθμος EM. Τέλος τυπώνει και το τελικό αποτέλεσμα κατανομής θεμάτων/λέξεων.
void	**estimate**(int alphabet,int numberOfTopics,int numberOfDocuments,int tokensperDocument[] ,double ini_m[][], double ini_V[][][],double c[][]) Η μέθοδος αυτή αποτελεί στην ουσία ττην υλοποίηση του αλγορίθμου EM. Σε πρώτη φάση ελέγχει την διαφορά μεταξύ καινούριου και παλιού ορίου Kullback-Leibler. Και εφόσον δεν είναι μικρότερη από έναν συγκεκριμένο κλάμα (εδώ έχει τεθεί ίσο με 0.0001) καλείς τις μεθόδους Estep και Mstep.
private void	**eStep**(int documentTokensNumber,double previous_m[], double previous_V[][],int numberOfTopics,double c[],double Sd[][],int d) Το Ε βήμα το αλγόριθμου όπου ο αλγόριθμος λύνει της εξισώσεις $\mathbf{S}_d\mathbf{c}_d - \boldsymbol{\Sigma}^{-1}(\mathbf{m}_d - \boldsymbol{\mu}) - \frac{W_d}{\xi_d} e^{\mathrm{diag}\left(\frac{1}{2}V_{tt,d}+m_{t,d}\right)_{1:T}} = 0$ $-\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\mathbf{V}_d^{-1} - \frac{W_d}{2\xi_d} e^{\mathrm{diag}\left(\frac{1}{2}V_{tt,d}+m_{t,d}\right)_{1:T}} = 0$ για κάθε κείμενο d.
private void	**mStep**(double[][] md, double[][][] V,int numberOfDocuments, int numberOfTopics, int alphabet,double[][] c) Το Μ βήμα του αλγορίθμου. Από τις εξισώσεις του Μ βήματος βρίσκουμε το καινούριο πραγματικό διάνυσμα μέσων τιμών, τον πίνακα διασποράς θεμάτων, την καινούρια πραγματική κατανομή θεμάτων/λέξεων καθώς και την καινούριες τιμές της βοηθητικής σύμφωνα με τους αντίστοιχους τύπους

	$$\boldsymbol{\mu} = \frac{1}{D}\sum_d \mathbf{m}_d$$ $$\boldsymbol{\Sigma} = \frac{1}{D}\sum_d \mathbf{V}_d + (\mathbf{m}_d - \boldsymbol{\mu})(\mathbf{m}_d - \boldsymbol{\mu})^T$$ $$\beta_{v,t} \propto \sum_d c_{v,d} s_{v,t,d}$$ $$s_{v,t,d} \propto \beta_{v,t} e^{m_{t,d}}$$
`private void`	**`initializeRandomSigma`**(int tokenssize[],int numberOfDocs) Θέτει τυχαίες μεταβλητές στην κατανομή sigma.
`private void`	**`estimateBetaSigma`**(int numberOfDocs,double c[][],double md[][],int alphabet) Υπολογίζει τις καινούριες κατανομές των κατανομών beta και sigma σε κάθε M βήμα. $$\beta_{v,t} \propto \sum_d c_{v,d} s_{v,t,d}$$ $$s_{v,t,d} \propto \beta_{v,t} e^{m_{t,d}}$$
`private double[][]`	**`normalizeMatrix`**(double vector[][],int alphabet,boolean isSecond) Κανονικοποιεί sigma και beta ανάλογα την boolean. Καλείται στο τέλος του eStep.
`private double`	**`FindLikelihood`**(double c[][], int DocNumber,int alphabet) Βρίσκει την καινούρια πιθανοφάνεια της εξίσωσης Kullback_Leibler.

ΚΕΦΑΛΑΙΟ 6
Αξιολόγηση των αλγορίθμων LDA, HLDA και CLDA

Στο παρών κεφάλαιο θα κάνουμε μία περιγραφή των σημαντικότερων μεθόδων αξιολόγησης θεματικών μοντέλων. Στην συνέχεια θα τρέξουμε του τρεις αλγορίθμους LDA, hLDA και cLDA που παρουσιάστηκαν στο κεφάλαιο 5 εφαρμόζοντάς τους στο σώμα *Sentiment Polarity Dataset Version 2.0.* Επόμενο βήμα μας θα είναι να αξιολογήσουμε τους δύο πρώτους αλγορίθμους (LDA και hLDA).

6.1 Αξιολόγηση Μοντέλων Θεμάτων

6.1.1 Εισαγωγή

Η στατιστική μοντελοποίηση θεμάτων όπως προαναφέραμε στο προηγούμενο κεφάλαιο είναι ένα χρήσιμο εργαλείο ανάλυσης μεγάλων μη δομημένων συλλογών κειμένων. Υπάρχει ένας σημαντικός όγκος εργασίας πάνω στην ανάπτυξη έξυπνων μοντέλων θεμάτων και εφαρμογών τους. Η μη επιβλεπόμενη φύση, όμως, αυτών των μοντέλων καθιστά δύσκολη την επιλογή ενός μοντέλου σε κάποιο πρόβλημα. Για μερικές εφαρμογές ίσως υπάρχουν κάποιοι ειδικοί στόχοι, όπως ανάκτηση πληροφοριών ή αρχικοποίηση κειμένων, των οποίων η απόδοση μπορεί να αξιολογηθεί. Όμως, υπάρχει η ανάγκη για μια καθολική μέθοδο που θα μετράει την ικανότητα γενίκευσης ενός θεματικού μοντέλου με έναν τρόπο που να είναι ακριβής, υπολογιστικά δυνατός, και ανεξάρτητος από κάθε συγκεκριμένη εφαρμογή.

Σε αυτή την ενότητα θα θεωρήσουμε μόνο το απλούστερο θεματικό μοντέλο, τον LDA, και θα συγκρίνουμε ένα αριθμό από μεθόδους για τον υπολογισμό της πιθανότητας held-out κειμένων δεδομένου ενός εκπαιδευμένου μοντέλου. Οι περισσότερες από τις μεθόδους που παρουσιάζονται, όμως, είναι εφαρμόσιμες και στα πιο πολύπλοκα θεματικά μοντέλα. Επιπλέον, για να συγκρίνουμε μεθόδους αξιολόγησης που χρησιμοποιούνται σήμερα στην βιβλιογραφία της θεματικής μοντελοποίησης, παρουσιάζουμε και άλλες εναλλακτικές μεθόδους.

6.1.2 Αξιολογώντας τον LDA

Ο LDA αξιολογείται συνήθως είτε μετρώντας την απόδοσή του σχετικά με την επίτευξη δευτερευόντων στόχων, όπως είναι η αρχικοποίηση κειμένων ή η ανάκτηση πληροφορίας, είτε υπολογίζοντας την πιθανότητα των held-out κειμένων που έχουν κρατηθεί από τα κείμενα εκπαίδευσης του αλγορίθμου. Ένα καλύτερο μοντέλο θα δώσει μια αυξημένη πιθανότητα στα held-out κείμενα, σε γενικές γραμμές.

Η πιθανότητα του συνόλου των held-out κειμένων W δεδομένου ενός συνόλου εκπαιδευτικών κειμένων W', μπορεί να γραφτεί ως εξής

$$P(W \mid W') = \int \mathrm{d}\Phi \, \mathrm{d}\alpha \, \mathrm{d}\boldsymbol{m} \, P(W \mid \Phi, \alpha\boldsymbol{m}) \, P(\Phi, \alpha\boldsymbol{m} \mid W').$$

Αυτό το ολοκλήρωμα μπορεί να υπολογιστεί βρίσκοντας την μέση τιμή P($W|\Phi,\alpha\boldsymbol{m}$) με δείγματα από το P($\Phi,\alpha\boldsymbol{m}|W'$), ή αξιολογώντας έναν υπολογισμό σημείου. Θα πάρουμε την τελευταία προσέγγιση. Οι στατιστικές μέθοδοι και οι MCMC μέθοδοι είναι αποτελεσματικές στο να περιθωριοποιούν εναποθέσεις θεμάτων Z που συσχετίζονται με τα εκπαιδευτικά δεδομένα που βρίσκουν το Φ και το $\alpha\boldsymbol{m}$.

Σε αυτό το κεφάλαιο ενδιαφερόμαστε να αξιολογήσουμε το

$$P(\mathcal{W} \mid \Phi, \alpha\boldsymbol{m}) = \prod_d P(\boldsymbol{w}^{(d)} \mid \Phi, \alpha\boldsymbol{m}). \qquad (6)$$

Επειδή οι εναποθέσεις θεμάτων για ένα κείμενο είναι ανεξάρτητες από της εναποθέσεις θεμάτων σε άλλα θέματα, κάθε held-out κείμενο μπορεί να αξιολογηθεί ξεχωριστά. Για το υπόλοιπο της ενότητας, θα αναφερόμαστε στο παρών κείμενο σαν **w**, στις κρυφές θεματικές εναποθέσεις ως **z**, και στην θεματική κατανομή που αφορά ένα συγκεκριμένο κείμενο **ως θ.**

Πολλές από τις μεθόδους αξιολόγησης απαιτούν την ικανότητα λήψης ενός συνόλου εναποθέσεων θεμάτων **z** για ένα κείμενο όπως το **w** χρησιμοποιώντας την διαδικασία Gibbs sampling. Η Gibbs sampling περιέχει την επανασύλλεξη κάθε z_n από την δεσμευμένη εκ των υστέρων κατανομή δεδομένων των $\boldsymbol{w}$, Φ, $\alpha\boldsymbol{m}$, $\mathbf{z}_{\backslash n}$:

$$\begin{aligned} P(z_n = t \mid \boldsymbol{w}, \boldsymbol{z}_{\backslash n}, \Phi, \alpha\boldsymbol{m}) & \\ & \propto P(w_n \mid z_n = t, \Phi)\, P(z_n = t \mid \boldsymbol{z}_{\backslash n}, \alpha\boldsymbol{m}) \\ & \propto \phi_{w_n|t} \frac{\{N_t\}_{\backslash n} + \alpha m_t}{N - 1 + \alpha}, \end{aligned} \qquad (7)$$

όπου το $\{N_t\}_{\backslash n}$ είναι ο αριθμός των φορών που το θέμα t που βρέθηκε στο κείμενο, εξαιρώντας την θέση n, και το N ο ολικός αριθμός των δειγμάτων στο κείμενο.

6.1.3 Υπολογίζοντας το P(w| Φ,αm)

Ο υπολογισμός της πιθανότητας P($\mathbf{w}|\Phi,\alpha\boldsymbol{m}$) για το held-out κείμενο **w** μπορεί να θεωρηθεί σαν την κανονικοποίηση της σταθεράς που σχετίζει την εκ των υστέρων κατανομή της **z** με την απο κοινού κατανομή ως προς τις **w** και **z** με τον κανόνα του Bayes:

$$P(\boldsymbol{z} \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{z}, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}. \qquad (8)$$

Υπάρχουν πολλές υπάρχουν μέθοδοι για την εκτίμηση κανονικοποιημένων σταθερών. Σε αυτή την ενότητα, κάνουμε μια επισκόπηση σε ορισμένες από αυτές τις μεθόδους, και επίσης δείχνουμε και δύο εναλλακτικές: έναν Chib-style αλγόριθμο και έναν “left-to-right” αλγόριθμο αξιολόγησης.

Importance sampling μέθοδοι

Γενικά, δεδομένου ενός μοντέλου παρατηρούμενες μεταβλητές **w** και με άγνωστες μεταβλητές **z**, μπορεί να χρησιμοποιηθεί η διαδικασία importance sampling για να υπολογιστεί η πιθανότητα $P(\boldsymbol{w}) = \sum_{\boldsymbol{h}} P(\boldsymbol{w}, \boldsymbol{h})$ ων με $\int \mathrm{d}\boldsymbol{h}\, P(\boldsymbol{w}, \boldsymbol{h})$. είτε είτε Αν η Q(h) είναι κάποια απλή, ευάγωγη κατανομή ως προς το **h**,- "η προτεινόμενη κατανομή" - τότε το

$$P(\boldsymbol{w}) \simeq \frac{1}{S} \sum_{s} \frac{P(\boldsymbol{w}, \boldsymbol{h}^{(s)})}{Q(\boldsymbol{h}^{(s)})}, \quad \boldsymbol{h}^{(s)} \sim Q(\boldsymbol{h}), \qquad (9)$$

είναι ένας αμερόληπτος εκτιμητής. Για να ασφαλίσουμε μια χαμηλή διασπορά, η Q(**h**) πρέπει να είναι ίδια με την "κατανομή-στόχο" P(**h,w**) και πρέπει να μην είναι μηδενική εκεί που η P(**w,h**) δεν είναι μηδενική.

Σε αυτή την υποενότητα θα εξηγήσουμε πως η P(**w**|Φ,$\alpha\boldsymbol{m}$) μπορεί να υπολογιστεί χρησιμοποιώντας importance sampling είτε (α) ολοκληρώνοντας ως προς **θ** και χρησιμοποιώντας την εκ των προτέρων κατανομή πάνω στην **h=z** ως την προτεινόμενη κατανομή, είτε (β) χρησιμοποιώντας την εκ των προτέρων κατανομή πάνω στην **h=θ** ως την προτεινόμενη κατανομή, με αποτέλεσμα να αφήνουμε τις θεματικές εναποθέσεις **z** να περιθωριοποιούνται άμεσα.

Αν η προτεινόμενη κατανομή είναι η εκ των προτέρων κατανομή πάνω στην **z,**

$$\begin{aligned} P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) &= \sum_{\boldsymbol{z}} P(\boldsymbol{w} \mid \boldsymbol{z}, \Phi)\, P(\boldsymbol{z} \mid \alpha\boldsymbol{m}) \\ &\simeq \frac{1}{S} \sum_{s} P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)}, \Phi), \qquad (10) \end{aligned}$$

όπου $\boldsymbol{z}^{(s)} \sim P(\boldsymbol{z} \mid \alpha\boldsymbol{m})$. Δυστυχώς οι θεματικές εναποθέσεις που επιλέγονται από την εκ των προτέρων, χωρίς να λαμβάνονται υπόψη τα αντίστοιχα δείγματα, είναι απίθανο να παράγουν μια σωστή επεξήγηση του **w**. Η εκ των προτέρων κατανμή συνήθως δεν είναι κοντά στην κατανομή-στόχο εκτός και αν το **w** είναι πολύ μικρό.

Καλύτερες προτεινόμενες κατανομές μπορούν να κατασκευαστούν για το $\mathbf{z}^{(s)}$ λαμβάνοντας υπόψη και το **w**. Ο απλούστερος τρόπος είναι να σχηματίσουμε μια κατανομή πάνω στα θέματα του κάθε δείγματος w_n, αγνοώντας τις εξαρτήσεις $Q(z_n) \propto \alpha m_{z_n} \phi_{w_n|z_n}$. Μια πιο έξυπνη μέθοδος γνωστή και ως "επαναληπτικές ψευδο-μετρήσεις", περιέχει μια διαρκή ενημέρωση του $Q(z_n)^{(0)}$ για κάθε sampling επανάληψη $Q(z_n)^{(0)} \propto \alpha m_{z_n} \phi_{w_n|z_n}$, ο ο κανόνας ενημέρωσης είναι

$$Q(z_n)^{(s)} \propto (\alpha m_{z_n} + \sum_{n' \neq n} Q(z_{n'})^{(s-1)})\, \phi_{w_n|z_n}. \qquad (11)$$

Εναλλακτικά, η P(**w**|Φ,$\alpha\boldsymbol{m}$) μπορεί να γραφτεί ως ένα ολοκλήρωμα πανώ στην θεματική

κατανομή **θ**:

$$P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) = \int \mathrm{d}\boldsymbol{\theta}\, P(\boldsymbol{w} \mid \boldsymbol{\theta}, \Phi)\, P(\boldsymbol{\theta} \mid \alpha\boldsymbol{m}) \simeq \frac{1}{S} \sum_s P(\boldsymbol{w} \mid \boldsymbol{\theta}^{(s)}, \Phi), \qquad (12)$$

όπου το $\boldsymbol{\theta}^{(s)}$ σχεδιάζεται από την $P(\boldsymbol{\theta}|\alpha\mathbf{m}) = \mathrm{Dir}(\boldsymbol{\theta};\alpha\mathbf{m})$. Ο εκτιμητής της (12) υπολογίζεται εύκολα επειδή οι τοπικές εναποθέσεις είναι ανεξάρτητες του **θ**:

$$P(\boldsymbol{w} \mid \boldsymbol{\theta}^{(s)}, \Phi) = \prod_n P(w_n \mid \boldsymbol{\theta}^{(s)}, \Phi) = \prod_n \sum_{z_n} P(w_n, z_n \mid \boldsymbol{\theta}^{(s)}, \Phi). \qquad (13)$$

Αν οι πιθανότητες $P(w|\boldsymbol{\theta}^{(s)}, \Phi)$ υπολογιστούν από ένα συνθετικό κείμενο, τυχαία δημιουργημένο χρησιμοποιώντας την θ(s), ο εκτιμητής αντιστοιχεί στην εμπειρική πιθανοφάνεια που περιγράφεται από του Li και McCallum (2006). Αν χρησιμοποιηθούν άμεσα, όμως, θα δώσουν το ίδιο αποτέλεσμα με τον να χρησιμοποιήσουμε απείρως μεγάλα συνθετικά κείμενα.

Η διαδικασία importance sampling δεν λειτουργεί σωστά όταν επιτελούμε sampling σε μεγάλων διαστάσεων κατανομές. Εκτός και αν η προτεινόμενη κατανομή έχει υπολογιστεί σχεδόν τέλεια κοντά στην κατανομή-στόχο, η διασπορά της λύσης είναι πάρα πολύ μεγάλη. Αν κάνουμε sampling σε συνεχόμενες τιμές όπως το **θ**, η λύση μπορεί να έχει και άπειρη διασπορά.

Μέθοδος αρμονικού μέσου

Η μέθοδός αρμονικού μέσου (Newton & Raftery, 1994) βασίζεται στην παρακάτω αμερόλιπτη λύση:

$$\frac{1}{P(\boldsymbol{w})} = \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{z} \mid \boldsymbol{w})}{P(\boldsymbol{w} \mid \boldsymbol{z})} \simeq \frac{1}{S} \sum_s \frac{1}{P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)})}, \qquad (14)$$

όπου το $\mathbf{z}^{(s)}$ σχεδιάζεται από την $P(\mathbf{z}|\mathbf{w})$. Δεσμεύοντας το Φ και το $\alpha\mathbf{m}$ δίνει έναν εκτιμητή για την $P(\mathbf{w}|\Phi,\alpha\boldsymbol{m})$:

$$P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \simeq \frac{1}{\frac{1}{S}\sum_s \frac{1}{P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)}, \Phi)}} = \mathrm{HM}\left(\{P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)}, \Phi)\}_{s=1}^{S}\right), \qquad (15)$$

όπου $\mathbf{z}^{(s)} \sim P(\mathbf{w}|\Phi,\alpha\boldsymbol{m})$ και HM(.) δηλώνει την αρμονική μέση τιμή. Πρακτικά τα $\{\mathbf{z}^{(s)}\}_{s=1}^{S}$ είναι S δείγματα επιλεγμένα από έναν Gibbs sampler μετά από μια burn-in περίοδο B επαναλήψεων. Αφού τα δείγματα χρησιμοποιούνται για αν υπολογίσουν μια προσδοκώμενη τιμή, δεν χρειάζεται να είναι ανεξάρτητα. Συνεπώς, το κόστος της λύσης είναι αυτό των S+B Gibbs επαναλήψεων.

Οι Newton και Raftery (1994) εξέφρασαν κάποιες επιφυλάξεις σχετικά με την μέθοδος αρμονικού μέσου ο Neal προσέθεσε περισσότερη κριτική. Παρόλη όμως την κριτική που δέχτηκε, η μέθοδος αυτή έχει χρησιμοποιηθεί σε πολλές διατριβές θεματικής μοντελοποίησης.

Annealed importance sampling

Η annealed importance sampling (AIS) μπορεί να θεωρηθεί ως μια απλή importance sampling διαδικασία ορισμένη σε ένα πολυδιάστατο χώρο καταστάσεων (Neal 2001). Πολλές από τις βοηθητικές μεταβλητές εισάγονται για να φτιάξουν μια προτεινόμενη κατανομή ως πιο κοντά στην κατανομή-στόχο. Όταν η AIS χρησιμοποιείται για να υπολογίσει την $P(\mathbf{w}|\Phi,\alpha\boldsymbol{m})$, εφαρμόζει την παρακάτω σειρά πιθανοτικών κατανομών:

$$P_s(\boldsymbol{z}) \propto P(\boldsymbol{w}\,|\,\boldsymbol{z},\Phi)^{\tau_s}\,P(\boldsymbol{z}\,|\,\alpha\boldsymbol{m}),$$

που ορίζεται από ενά σύνολο "αντίστροφων θερμοκρασιών" $0=\tau_0<\tau_1< \ldots <\tau_S = 1$. Όταν $s = 0$, $\tau_s = 0$, η $P_0(z)$ είναι η εκ των προτέρων κατανομή $P(\mathbf{w}|\alpha\boldsymbol{m})$. Παρόμοια, όταν $s = S$, η $P_S(z)$ είναι η εκ των υστέρων κατανομή $P(\mathbf{w}|\Phi,\alpha\boldsymbol{m})$. Οι ενδιάμεσες τιμές του s παρεμβάλλονται μεταξύ των εκ των προτέρων και των εκ των υστέρων κατανομών. Για κάθε $s = 1, \ldots, S\text{-}1$, ένας διαχειριστής μετάβασης μιας Μαρκοβιανής αλυσίδας $T_s(z' \leftarrow z)$ που αφήνει την PS(z) απαράλλακτη πρέπει επίσης να οριστεί. Όταν υπολογίζεται η $P(\mathbf{w}|\Phi,\alpha\boldsymbol{m})$, $T_s(z' \leftarrow z)$ είναι ο Gibbs sampling διαχειριστής που επιλέγει σειριακά από την

$$P_s(z_n\,|\,\boldsymbol{z}_{\setminus n}) \propto P(w_n\,|\,z_n,\Phi)^{\tau_s}\,P(z_n\,|\,\boldsymbol{z}_{\setminus n},\alpha\boldsymbol{m}). \quad (16)$$

Η διαδικασία sampling της (16) είναι τόσο εύκολη όσο και αυτή της (7).

Η AIS κατασκευάζει μια προτεινόμενη κατανομή Q(Z) πάνω σε έναν εκτεταμένο χώρο καταστάσεων $Z = \{z^{(1)}, \ldots, z^{(S)}\}$ επιλέγοντας πρώτα την ευάγωγη εκ των προτέρων $P_0(\mathbf{z})$ και μετά εφαρμόζοντας μια σειρά από διαχειριστές μετάβασης $T_1, T_2, \ldots, T_{S-1}$ όπου "μετακινούν" το δείγμα μέσα από τις ενδιάμεσες κατανομές $P_s(\mathbf{z})$ προς την εκ των υστέρων $P_S(\mathbf{z})$. Η πιθανότητα της σειράς καταστάσεων Z δίνεται από την

$$Q(Z) = P_0(\boldsymbol{z}^{(1)}) \prod_{s=1}^{S-1} T_s(\boldsymbol{z}^{(s+1)} \leftarrow \boldsymbol{z}^{(s)}). \quad (17)$$

Η κατανομή-στόχος για την προτεινόμενη Q(Z) είναι

$$P(Z) = P_S(\boldsymbol{z}^{(S)}) \prod_{s=1}^{S-1} \tilde{T}_s(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s+1)}), \quad (18)$$

1: initialize $0 = \tau_0 < \tau_1 < \ldots < \tau_S = 1$
2: sample $\boldsymbol{z}^{(1)}$ from the prior $P_0(\boldsymbol{z}) = P(\boldsymbol{z} \mid \alpha \boldsymbol{m})$.
3: **for** $s = 2 : S$ **do**
4: sample $\boldsymbol{z}^{(s)} \sim T_{s-1}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s-1)})$
5: **end for**
6: $P(\boldsymbol{w} \mid \Phi, \alpha \boldsymbol{m}) \simeq \prod_{s=1}^{S} P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)}, \Phi)^{\tau_s - \tau_{s-1}}$

Αλγόριθμος 6.1: Annealed importance sampling

όπου το $\tilde{T}_s$ είναι ο αντίστροφος διαχειρίστης μετάβασης, δοσμένος από

$$\tilde{T}_s(\boldsymbol{z}' \leftarrow \boldsymbol{z}) = T_s(\boldsymbol{z} \leftarrow \boldsymbol{z}') \frac{P_s(\boldsymbol{z}')}{P_s(\boldsymbol{z})}. \qquad (19)$$

Έχοντας επιλέξει μια σειρά από θεματικές εναποθέσεις από την Q(Z), ένα βάρος βαθμωτής σημασίας κατασκευάζεται:

$$\begin{aligned} w_{\mathrm{AIS}} &= \frac{P(\boldsymbol{w} \mid \Phi, \alpha \boldsymbol{m})\, P(Z)}{Q(Z)} \\ &= \frac{P(\boldsymbol{w}, \boldsymbol{z}^{(S)} \mid \Phi, \alpha \boldsymbol{m}) \prod_{s=1}^{S-1} \tilde{T}_s(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s+1)})}{P_0(\boldsymbol{z}^{(1)}) \prod_{s=1}^{S-1} T_s(\boldsymbol{z}^{(s+1)} \leftarrow \boldsymbol{z}^{(s)})} \\ &= \prod_{s=1}^{S} P(\boldsymbol{w} \mid \boldsymbol{z}^{(s)}, \Phi)^{\tau_s - \tau_{s-1}}. \end{aligned}$$

Δεδομένου ενός συνόλου από την Q(Z), τα αντίστοιχα βάρη σημαντικότητας μπορούν να χρησιμοποιηθούν για τον υπολογισμό της P(**w**|*Φ,α**m***) λόγω της παρακάτω εξίσωσης:

$$\begin{aligned} P(\boldsymbol{w} \mid \Phi, \alpha \boldsymbol{m}) &= P(\boldsymbol{w} \mid \Phi, \alpha \boldsymbol{m}) \sum_{Z} P(Z) \\ &= \mathbb{E}_{Q(Z)}\left[w_{\mathrm{AIS}}\right]. \qquad (20) \end{aligned}$$

Οι διαχειριστές μετάβασης δεν χρειάζεται αναγκαστικά να είναι εργοδικοί. Ο απλός importance sampling υπολογισμός στην (10), με τον οποίο η προτεινόμενη κατανομή είναι P(**w**|*α**m***), ανακτάται χρησιμοποιώντας διαχειριστές μετάβασης που δεν κάνουν τίποτα: $T_s(z' \leftarrow z) = \delta$ (**z'** – **z**) για όλα τα s.

Ο AIS αλγόριθμος περιγράφεται περιληπτικά στον αλγόριθμο 6.1.

Chib-style υπολογισμός

Για κάθε "ειδικό" σύνολο κρυφών θεματικών εναποθέσεων $\boldsymbol{z}^\star$ ο Bayes κανόνας δίνει τη παρακάτω ισότητα:

$$P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{z}^\star \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})}. \qquad (21)$$

Ο Chib (1995) εισήγαγε μια οικογένεια αλγορίθμων που πρώτα επιλέγουν $\boldsymbol{z}^\star$ και μετά υπολογίζουν τον π$P(\boldsymbol{z}^\star \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})\, P(\boldsymbol{z}^\star, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) = \; P(\boldsymbol{w} \mid \boldsymbol{z}^\star, \Phi)\, P(\boldsymbol{z}^\star \mid \alpha\boldsymbol{m})$

Ο αριθμητής είναι γνωστό από τις (4) και (5).

Κάθε διαχειριστής Τ Μαρκοβιανής αλυσίδας για την διαδικασία sampling της εκ των υστέρων κατανομής, ικανοποιεί

$$P(\boldsymbol{z}^\star \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}) = \sum_{\boldsymbol{z}} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z})\, P(\boldsymbol{z} \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}). \qquad (22)$$

1: initialize $\boldsymbol{z}^*$ to a high posterior probability state
2: sample s uniformly from $\{1, \ldots, S\}$
3: sample $\boldsymbol{z}^{(s)} \sim \widetilde{T}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^\star)$
4: **for** $s' = (s+1) : S$ **do**
5: sample $\boldsymbol{z}^{(s')} \sim T(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'-1)})$
6: **end for**
7: **for** $s' = (s-1) : -1 : 1$ **do**
8: sample $\boldsymbol{z}^{(s')} \sim \widetilde{T}(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'+1)})$
9: **end for**
10: $P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \simeq P(\boldsymbol{w}, \boldsymbol{z}^\star \mid \Phi, \alpha\boldsymbol{m}) \Big/ \frac{1}{S} \sum_{s'} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}^{(s')})$

Αλγόριθμος 6.2: Ένας Chib-style αλγόριθμος

Η (22) μπορεί να αντικατασταθεί από την (21) για να δώσει

$$P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}{\sum_{\boldsymbol{z}} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z})\, P(\boldsymbol{z} \mid \boldsymbol{w}, \Phi, \alpha\boldsymbol{m})} \simeq \frac{P(\boldsymbol{z}^\star, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}{\frac{1}{S} \sum_{s=1}^{S} T(\boldsymbol{z}^\star \leftarrow \boldsymbol{z}^{(s)})},$$

όπου Z = $\{z^{(1)}, \ldots, z^{(S)}\}$ μπορεί να βρεθεί με Gibbs sampling από την P(**z**|**w**,Φ,α**m**). Οι Murray και Salakhutdinov (2009) έδειξαν ότι αυτός ο εκτιμητής μπορεί να υπερεκτιμήσει την επιθυμητή προσδοκώμενη πιθανότητα. Αντί αυτού κατασκεύασαν την παρακάτω

προτεινόμενη κατανομή:

$$Q(Z) = \frac{1}{S}\sum_{s=1}^{S} \tilde{T}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{\star}) \prod_{s'=s+1}^{S} T(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'-1)}) \cdot \prod_{s'=1}^{s-1} \tilde{T}(\boldsymbol{z}^{(s')} \leftarrow \boldsymbol{z}^{(s'+1)}).$$

Επειδή η μπροστινή μετάβαση του διαχειριστή T αποτελείται από την εφαρμογή της (7) στις θέσεις από 1 έως N, ο αντίστροφος διαχειριστής μετάβασης $\tilde{T}$ μπορεί να κατασκευαστεί απλά εφαρμόζοντας την (7) με αντίστροφη σειρά.

Χρησιμοποιώντας τον ορισμό του $\tilde{T}$ στην (19) μπορεί να δειχθεί ότι

$$P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \simeq \frac{P(\boldsymbol{z}^{\star}, \boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m})}{\frac{1}{S}\sum_{s=1}^{S} T(\boldsymbol{z}^{\star} \leftarrow \boldsymbol{z}^{(s)})}, \qquad (23)$$

με δείγματα από την Q(Z). Σε αυτή την εφαρμογή, με Gibbs επιλογείς που πηγαίνουν εμπρός και πίσω, ο εκτιμητής είναι επίσημα αμερόληπτος, ακόμα και σε περιορισμένα τρεξίματα της αλυσίδας.

Η πιθανότητα μετακίνησης του $\boldsymbol{z}^{\star}$ δίνεται από την

$$T(\boldsymbol{z}^{\star} \leftarrow \boldsymbol{z}) = \prod_{n} P(z_n^{\star} \mid \boldsymbol{z}_{<n}^{\star}, \boldsymbol{z}_{>n}, \boldsymbol{w}, \Phi, \alpha\boldsymbol{m}). \qquad (24)$$

Ο Chib-style εκτιμητής είναι έγκυρος για κάθε επιλογή "ειδική κατάστασης" $\boldsymbol{z}^{\star}$. Θέτουμε το $\boldsymbol{z}^{\star}$ μεγιστοποιώντας επαναληπτικώς την (7) για τις θέσεις 1, … , N και εκτελώντας για μερικές επαναλήψεις την διαδικασία Gibbs sa$\boldsymbol{z}^{\star}$)ling. Σε όλα τα πειράματα που έχουν γίνει λιγότερο από το 1% του υπολογιστικού χρόνου χρησιμοποιήθηκε για να βρεθεί το

Η Chib-style μέθοδος δείχνεται στον Αλγόριθμο 6.2.

1: initialize $l := 0$
2: **for** each position n in $\boldsymbol{w}$ **do**
3: initialize $p_n := 0$
4: **for** each particle $r = 1$ to R **do**
5: **for** $n' < n$ **do**
6: sample $z_{n'}^{(r)} \sim P(z_{n'}^{(r)} \mid w_{n'}, \{\boldsymbol{z}_{<n}^{(r)}\}_{\setminus n'}, \Phi, \alpha\boldsymbol{m})$
7: **end for**
8: $p_n := p_n + \sum_t P(w_n, z_n^{(r)} = t \mid \boldsymbol{z}_{<n}^{(r)}, \Phi, \alpha\boldsymbol{m})$
9: sample $z_n^{(r)} \sim P(z_n^{(r)} \mid w_n, \boldsymbol{z}_{<n}^{(r)}, \Phi, \alpha\boldsymbol{m})$
10: **end for**
11: $p_n := p_n \,/\, R$
12: $l := l + \log p_n$
13: **end for**
14: $\log P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \simeq l$

Αλγόριθμος 6.3: Ο "left-to-right" αλγόριθμος αξιολόγησης.

Ο "left-to-right" αλγόριθμος αξιολόγησης

Μια άλλη προσέγγιση για τον υπολογισμό της P(**w**|*Φ,αm*) προτάθηκε πρόσφατα από τον Wallach (2008). Αυτή η μέθοδος, που λειτουργεί με έναν αυξητικό, ¨αριστερά προς τα δεξιά" τρόπο, αναλύει την P(w|Φ,αm) ως

$$\begin{aligned} & P(\boldsymbol{w} \mid \Phi, \alpha\boldsymbol{m}) \\ & \quad = \textstyle\prod_n P(w_n \mid \boldsymbol{w}_{<n}, \Phi, \alpha\boldsymbol{m}) \\ & \quad = \textstyle\prod_n \sum_{\boldsymbol{z}_{\leq n}} P(w_n, \boldsymbol{z}_{\leq n} \mid \boldsymbol{w}_{<n}, \Phi, \alpha\boldsymbol{m}). \end{aligned} \quad (25)$$

Κάθε άθροισμα μεγαλύτερο του $\boldsymbol{z}_{\leq n}$ μπορεί να υπολογιστεί χρησιμοποιώντας μια προσέγγιση εμπνευσμένη από τις ακολουθητικές Monte Carlo μεθόδους., όπως στον αλγόριθμο 6.3. Αυτή η μέθοδος είναι πιο κατάλληλη για μια ευρεία περιοχή εφαρμογών – συμπεριλαμβανομένων των συστημάτων αναγνώρισης ομιλίας – σε σύγκριση με τις άλλες μεθόδους σε αυτήν ενότητα, λόγω της "αριστερά προς τα δεξιά" λειτουργίας του. Αυτός ο αλγόριθμος χρησιμοποιείται πλέον στο Mallet και αυτόν θα χρησιμοποιήσουμε για αν αξιολογήσουμε τον LDA.

Τα σχετικά κόστη των μεθόδων

Η πλειοψηφία των μεθόδων όπως περιγράφτηκε προηγουμένων είναι βασισμένη στο Gibbs sampling, που κυριαρχεί και στα κόστη: υπολογίζοντας την $P(z_n|w_n,\mathbf{z}_{\backslash n},\Phi,\alpha\boldsymbol{m})$ έχει σημαντικά περισσότερο κόστος σε σχέση με τον υπολογισμό της $P(w_n|z_n, \Phi)$ – την ποσότητα δηλαδή, που χρησιμοποιείται για να κατασκευάσει τους εκτιμητές δεδομένων των δειγμάτων. Η Chib-style μέθοδοι είναι μια εξαίρεση: κατασκευάζοντας τον αλγόριθμο έχει ένα κόστος το πολύ ίσο με αυτούς του Gibbs sampling.

Το importance sampling, που χρησιμοποιεί την εκ των προτέρων κατανομή πάνω στο $\boldsymbol{\theta}$ ως την κατανομή επιλογής, δεν εμπεριέχει Gibbs sampling. Το $\sum_{z_n} P(z_n, w_n|\boldsymbol{\theta}^{(s)}, \Phi)$ μπορεί να υπολογιστεί από κάθε held-out δείγμα w_n. Το κόστος του απλού imortance sampling χρησιμοποιώντας μια κατανομή ως προς το $\mathbf{z}$ είναι δυσκολότερο να το εκφράσουμε, και εφαρμογή θα είναι εξαρτημένη. Λίγο άδικο για αυτές τις μεθόδους υποθέτουμε ότι το κόστος των δημιουργημένων δειγμάτων είναι άμεσα συγκρίσιμο με το Gibbs sampling. Το κόστος μπορεί να εξεταστεί πιο εξονυχιστικά αν μια τέτοια μέθοδος φέρει καλά αποτελέσματα.

1: initialize $0 = \tau_0 < \tau_1 < \ldots < \tau_S = 1$
2: sample $\boldsymbol{z}^{(1)}$ from $P_0(\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m})$.
3: **for** $s = 2 : S$ **do**
4: sample $\boldsymbol{z}^{(s)} \sim T_{s-1}(\boldsymbol{z}^{(s)} \leftarrow \boldsymbol{z}^{(s-1)})$
5: **end for**
6: $P(\boldsymbol{w}^{(2)}|\boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) \simeq \prod_{s=1}^{S} P(\boldsymbol{w}^{(2)}|\boldsymbol{z}^{(s)}, \boldsymbol{w}^{(1)}, \Phi)^{\tau_s - \tau_{s-1}}$

Αλγόριθμος 6.4: Ο AIS για την μέθοδο ολοκλήρωσης κειμένου

Ολοκλήρωση Κειμένου

Ένας άλλος τρόπος αξιολόγησης θεματικών μοντέλων είναι να συγκρίνουμε την απόδοση που έχουμε προβλέψει υπολογίζοντας την πιθανότητα του δεύτερου μισού ενός κειμένου, δεδομένου του πρώτου. Αυτό τυπικά επιτυγχάνεται προσθέτοντας το πρώτο μισό κάθε held-out κειμένου στο σύνολο εκπαίδευσης, και κρατώντας το δεύτερο μισό για αξιολόγηση. Θέτοντας το $\mathbf{w}^{(1)}$ ως το πρώτο μισό και το $\mathbf{w}^{(2)}$ ως το δεύτερο μισό ο στόχος είναι να υπολογίσουμε

$$P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) = \frac{P(\boldsymbol{w}^{(2)}, \boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})}{P(\boldsymbol{w}^{(1)} \mid \Phi, \alpha\boldsymbol{m})}, \quad (26)$$

που είναι μια αναλογία κανονικοποιημένων σταθερών. Κάθε μία από τις μέθοδος για τον υπολογισμό της $P(\mathbf{w}|\Phi,\alpha\boldsymbol{m}) \equiv P(\mathbf{w}^{(1)},\mathbf{w}^{(2)}|\Phi,\alpha\boldsymbol{m})$ περιγράφτηκε στην προηγούμενη ενότητα και μπορεί να τρέξει μόνο για το $\mathbf{w}^{(1)}$ για να υπολογίσουμε την $P(\mathbf{w}^{(1)},\mathbf{w}^{(2)}|\Phi,\alpha\boldsymbol{m})$. Όμως ειδικές τεχνικές μπορεί να είναι πιο αποτελεσματικές.

Estimated θ

Η estimated θ μέθοδος περιλαμβάνει την επιλογή δειγμάτων $\mathbf{z}^{(1,s)} \sim P(\mathbf{z}^{(1)}|\mathbf{w}^{(1)},\Phi,\alpha\boldsymbol{m})$ και μετά τον σχηματισμό της

$$\hat{\theta}_t^{(s)} = P(t \mid \boldsymbol{z}^{(1,s)}, \alpha\boldsymbol{m}) = \frac{N_t^{(1,s)} + \alpha m_t}{N^{(1)} + \alpha}, \qquad (27)$$

όπου το $N^{(1)}$ είναι ο αριθμός των δειγμάτων $\mathbf{w}^{(1)}$. Τότε για το $\mathbf{w}^{(2)}$ ισχύει

$$P(\boldsymbol{w}^{(2)} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}) \simeq \frac{1}{S} \sum_s \prod_n \sum_t \phi_{w_n^{(2)}|t}\, \hat{\theta}_t^{(s)}.$$

Importance sampling και AIS

Οι importance sampling αλγόριθμοι που περιγράψαμε παραπάνω μπορούν να προσαρμοστούν ώστε να υπολογίσουν την (26) άμεσα με την χρησιμοποίηση δειγμάτων δεσμευμένων ως προς $\mathbf{w}^{(1)}$. Για τον AIS, χρησιμοποιούμε την παρακάτω ακολουθία κατανομών

$$P_s(\boldsymbol{z}) \propto P(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)} \mid \boldsymbol{z}, \Phi)^{\tau_s}\, P(\boldsymbol{z} \mid \boldsymbol{w}^{(1)}, \Phi, \alpha\boldsymbol{m}).$$

Η (26) μπορεί να υπολογιστεί από τον αλγόριθμο 4.

Ο "left-to-right" αλγόριθμος αξιολόγησης

Ο αλγόριθμος "left-to-right" μπορεί να υπολογίσει την $P(\mathbf{w}^{(2)}|\mathbf{w}^{(1)},\Phi,\alpha\boldsymbol{m})$ απευθείας. Αν οι λέξεις στο ς έχουν ταξινομηθεί ετσί ώστε το $\mathbf{w}^{(1)}$ να είναι πλήρως παρατηρούμενο πριν κάποια από τις λέξεις του w(2) παρατηρηθεί, τότε ένας δεύτερος εκτιμητής μπορεί να συσσωρευτεί όπως την σειρά 12 του αλγορίθμου 3, για τις θέσεις που περιέχουν δείγματα από το $\mathbf{w}^{(2)}$.

6.2 Υλοποίηση LDA, Ιεραρχικού LDA και Correlated LDA

6.2.1 Εισαγωγή

Οι τρεις αλγόριθμοι που όπου αναλύσαμε την υλοποίηση τους στο κεφάλαιο 5 (Latent Dirichlet Allocation, Ιεραρχικό LDA και Correlated LDA) εφαρμόστηκαν στο σώμα *Sentiment Polarity Dataset Version 2.0*, μία συλλογή κειμένων αποτελούμενη συνολικά από 2000 κείμενα με κριτικές ταινιών, 1000 με θετική κριτική και 1000 με αρνητική.

6.2.2 Εφαρμογή LDA

Εφαρμόσαμε τον αλγόριθμο ParallelTopicModel σε 1900 κείμενα από το σώμα Polarity Dataset θέτοντας την διάσταση των θεμάτων στο 20, και τις σταθερές άλφα και βήτα όπως αυτές έχουν προσαρμοστεί από το ίδιο το Mallet,η άλφα δηλαδή ως ένα μοναδιαίο διάνυσμα και η βήτα ίση με 1. Τα υπόλοιπα 100 κείμενα του σώματος τα κρατήσαμε ως held-out κείμενα για την μετέπειτα αξιολόγηση του αλγορίθμου πάνω στο συγκεκριμένο πίνακα. Πήραμε τον παρακάτω πίνακα:

0		1		2		3		4	
horror	281	lawyer	82	truman	39	movie	5,419	eyes	59
scream	242	flynt	80	tarantino	115	film	2,198	bobby	52
killer	178	case	75	jackie	108	good	1,762	cruise	52
original	106	oscar	64	max	105	time	1,353	tom	50
series	97	bulworth	59	shakespeare	90	bad	1,342	snake	49
julie	88	courtroom	55	carrey	89	don	1,015	vincent	47
slasher	83	larry	51	fiction	82	plot	937	jolie	45
summer	78	cole	50	pulp	79	movies	900	green	43
dvd	77	ghost	50	patch	77	big	896	hunt	41
5		6		7		8		9	
funny	269	batman	193	music	91	ryan	155	life	851
comedy	264	vampire	122	rock	87	harry	130	love	656
jokes	196	robin	114	girls	84	war	128	family	484
smith	192	effects	95	nights	82	west	104	father	417
humor	140	arnold	91	band	77	private	92	man	414
ben	109	spawn	78	spice	5	troopers	84	mother	398
bob	109	blade	8	angels	73	starship	76	young	328
laughs	98	vampires	76	musical	68	simon	76	wife	318
gags	93	horror	76	boogie	4	wild	71	son	296

Πίνακας 6.1: Ο πίνακας δείχνει τα 10 πρώτα θέματα, που δημιουργήθηκαν από τον ParallelTopicModel στο σώμα Polarity Dataset, μαζί με τις 9 πιο συχνές λέξεις. Δίπλα αναγράφεται το μη κανονικοποιημένο βάρος τους δηλαδή οι φορές που έχει ανατεθεί η λέξη στο συγκεκριμένο θέμα.

Ο Πίνακας 6.1 απεικονίζει τα 10 πρώτα θέματα που δημιούργησε ο αλγόριθμος εφαρμοσμένος στο σώμα *Polarity Dataset,* μαζί με τις 9 λέξεις που έχουν την μεγαλύτερη κατανομή ως προς το κάθε θέμα. Δίπλα σε κάθε λέξη αναγράφεται ο αριθμός των φορών

που η λέξη έχει εναποτεθεί στο συγκεκριμένο θέμα. Παρατηρήστε το συσχετισμό των λέξεων μεταξύ τους.

Αξιολόγηση LDA

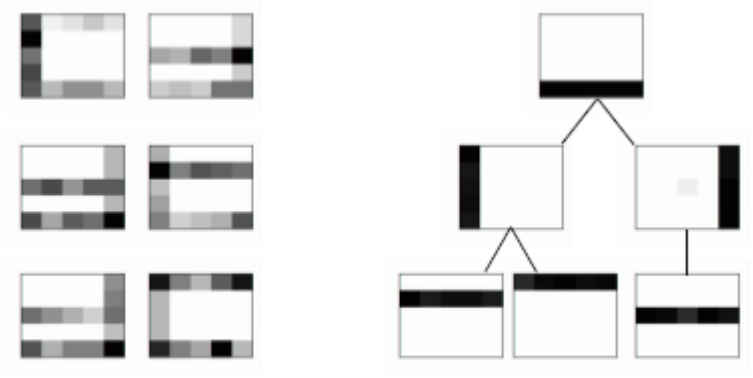
Στην συνέχεια αξιολογήσαμε τον αλγόριθμο με την μέθοδο Left-to-Right που έχει υλοποιηθεί από το Mallet. Η αξιολόγηση έγινε με τα 100 τελευταία κείμενα του σώματος Polarity Dataset που τα είχαμε κρατήσει ως held-out. Με την μέθοδο αυτή υπολογίσαμε την ποσότητα :

$$\sum_d \log P(\boldsymbol{w}^{(d)} \mid \Phi, \alpha \boldsymbol{m})$$

Η ποσότητα αυτή βρέθηκε ίση με -299809.38929934683.

6.2.3 Εφαρμογή ιεραρχικού LDA

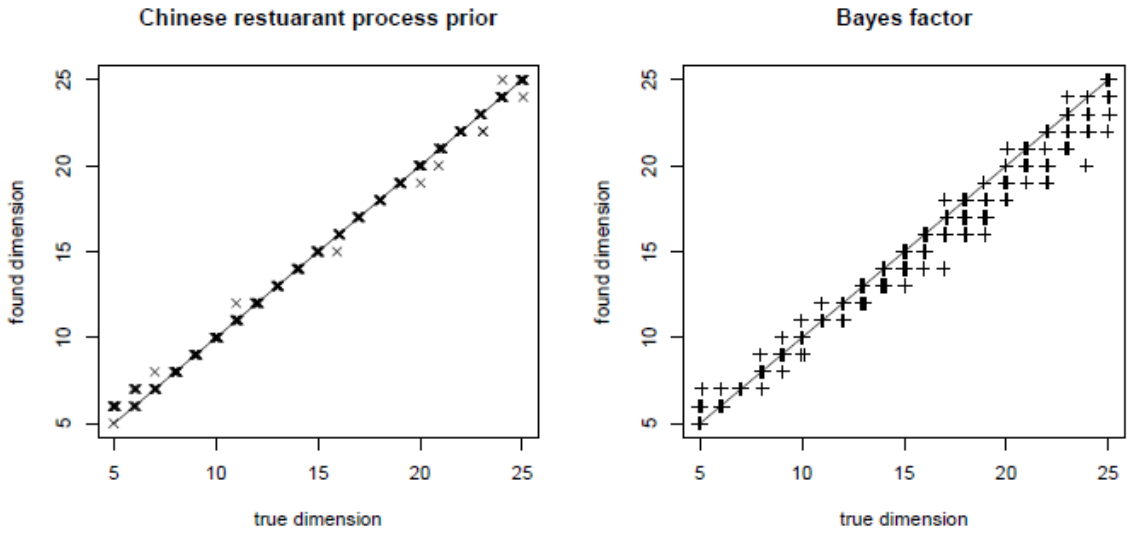
Πριν αρχίσουμε την ανάλυση της εφαρμογής και αξιολόγηση του ιεραρχικού LDA, είναι καλό να παραθέσουμε μερικά παραδείγματα και εμπειρικά αποτελέσματα από την εργασία "Hierarchichal Topic Models and the Nested Chinese Restaurant Process (Blei, Griffiths, Jordan και Tenebaum)". Στην εργασία τους απέδειξαν ότι η nested CRP διαδικασία είναι ικανή να μάθει από ιεραρχίες κειμένου μέσω του ιεραρχικού LDA χρησιμοποιώντας ένα σώμα με ένα μικρό λεξιλόγιο. Δημιούργησαν ένα σώμα 100 κειμένων με 1000 λέξεις το καθένα από μία ιεραρχία τριών επιπέδων με ένα λεξιλόγιο 25 όρων. Σε αυτό το σώμα, τα θέματα σε ένα λεξιλόγιο μπορούν να αναπαρασταθούν ως μπάρες σε ένα πλαίσιο 5 επί 5. Το θέμα-ρίζα τοποθετεί την μάζα πιθανότητας στην κάτω μπάρα. Στο δεύτερο επίπεδο, το ένα θέμα ταυτοποιείται από την πιο αριστερή μπάρα ενώ η ταυτοποίηση του δεύτερου γίνεται με την μπάρα που είναι πιο δεξιά. Το πιο αριστερό θέμα έχει δύο υποθέματα ενώ το δεξιότερο έχει ένα υποθέμα. Η Εικόνα 6.1(a) αναπαριστά έξι κείμενα που έχουν δημιουργηθεί από αυτό το μοντέλο. Η Εικόνα 6.1(b) αναπαριστά την ιεραρχία χρησιμοποιώντας την διαδικασία Gibbs sampling που περιγράφεται στο υποκεφάλαιο 3.2.



Εικόνα 6.1:(a) Έξι κείμενα που έχουν δημιουργηθεί από ένα σώμα 100 κειμένων χρησιμοποιώντας την ιεραρχία 3 επιπέδων. Κάθε κείμενο έχει 1000 λέξεις από ένα λεξιλόγιο 25 όρων. (b) Η σωστή ιεραρχία που βρέθηκε από των Gibbs sampler σε αυτό το σώμα.

Για να συγκρίνουν την CRP μέθοδο με το LDA μοντέλο δημιούργησαν 210 σώματα των

100 κειμένων 1000 λέξεων καθένα από ένα LDA μοντέλο με $K \in$ ¿ {5, ..., 25}, L=5, μέγεθος λεξιλογίου ίσο με 100, και οι τυχαίοι παράγοντες δημιουργήθηκαν από μία κατανομή Dirichlet (η = 0.1). Για σύγκριση με την CRP εκ των προτέρων κατανομή, χρησιμοποίησαν την μέθοδο προσεγγιστικών Bayes παραγόντων για την επιλογή του μοντέλου, όπου επιλέγεται το μοντέλο που μεγιστοποιεί την πιθανότητα $p(\text{δεδομένα} \mid K)p(K)$ για διάφορα K και μία προσεγγιστική εκ των υστέρων κατανομή. Με το LDA μοντέλο, η μέθοδος των παραγόντων Bayes είναι πολύ πιο αργή σε σχέση με την CRP καθώς περιέχει πολλές επαναλήψεις του Gibbs sampler με ταχύτητα συγκρίσιμη μίας επανάληψης του CRP sampler. Επιπλέον, με την μέθοδο των παραγόντων Bayes πρέπει να επιλεχθεί μία κατάλληλη κλίμακα του K. Με την CRP εκ των προτέρων κατανομή, η μόνη ελεύθερη παράμετρος είναι η γ (εδώ χρησιμοποιούν $\gamma = 0.1$). Όπως φαίνεται στην Εικόνα 6.2, η CRP εκ των προτέρων κατανομή είναι πιο αποτελεσματική σε σχέση με τους παράγοντες Bayes πάνω σε αυτά τα δεδομένα. Σημειώστε ότι και η CRP μέθοδος αλλά και οι παράγοντες Bayes είναι ευαίσθητοι στην επιλογή του η, την υπερπαράμετρο της εκ των προτέρων κατανομής στα θέματα. Όμως, εδώ αυτή η παράμετρος ήταν γνωστή και έτσι μπορεί να γίνει μία δίκαια σύγκριση.



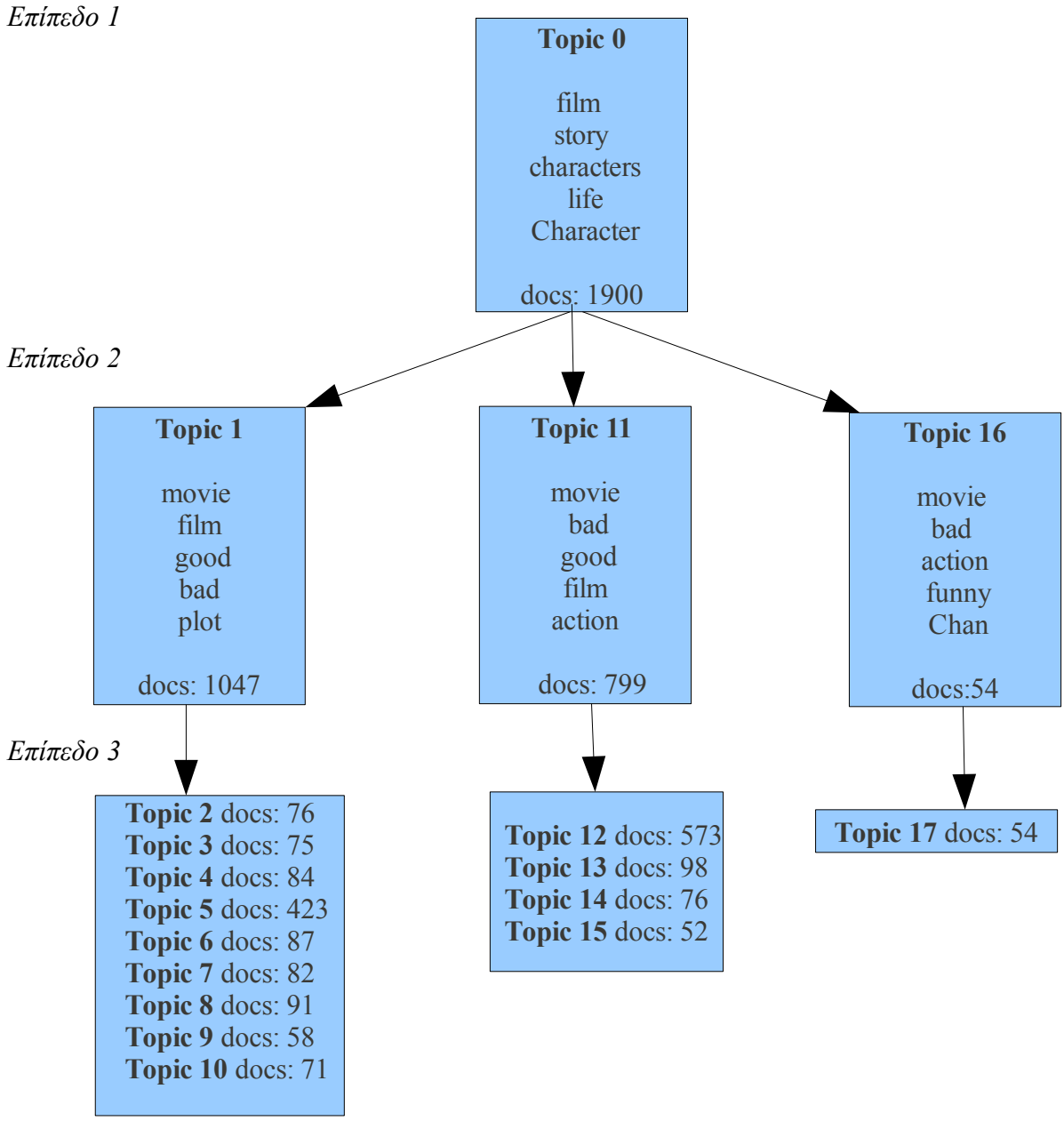
Εικόνα 6.2: (Αριστερά) Η μέση διάσταση που βρέθηκε από μία CRP εκ των προτέρων κατανομή σχεδιασμένη σε σχέση με την αληθινή διάσταση των δεδομένων που προσομειώθηκαν. Για κάθε διάσταση δημιουργήθηκαν δέκα σώματα με μέγεθος λεξιλογίου ίσο με 100. Κάθε σώμα περιέχει 100 κείμενα 1000 λέξεων. (Δεξιά) Αποτελέσματα της επιλογής μοντέλου με τους παράγοντες Bayes.

Structure	Leaf error			Other
	0	1	2	
3 (7 6 5)	70%	14%	4%	12%
4 (6 6 5 5)	48%	30%	2%	20%
4 (6 6 6 4)	52%	36%	0%	12%
5 (7 6 5 5 4)	30%	40%	16%	14%
5 (6 5 5 5 4)	50%	22%	16%	12%

Εικόνα 6.3: Αποτελέσματα των υπολογισμένων ιεραρχιών στα δεδομένα που προσομειώθηκαν.

Σε ένα παρόμοιο πείραμα, δημιούργησαν 50 σώματα ένα για κάθε πέντε διαφορετικές ιεραρχίες χρησιμοποιώντας ένα hLDA μοντέλο και την ίδια συμμετρική Dirichlet εκ των προτέρων κατανομή θεμάτων. Κάθε θέμα έχει 100 κείμενα 1000 λέξεων από ένα λεξιλόγιο 100 όρων. Η Εικόνα 6.3 δείχνει τα αποτελέσματα του sampling από την προκύπτουσα εκ των υστέρων κατανομή των δέντρων με τον Gibbs Sampler από την υποενότητα 3.2.4. Σε όλες τις περιπτώσεις, προκύπτει σωστή δομή περισσότερο από οποιαδήποτε άλλη. Σε όλα τα πειράματα, καμία προκύπτουσα δομή δεν έχει μεγαλύτερη απόκλιση περισσότερο από τρεις κόμβους από την σωστή

Εμείς εφαρμόσαμε τον αλγόριθμο HierarchicalLDA του Mallet στα ίδια 1900 κείμενα του Polarity Dataset σώματος. Για την εφαρμογή του θέσαμε ένα δέντρο τριών επιπέδων. Η παράμετρος άλφα τέθηκε ίση με 10, η παράμετρος γάμμα ίση με 1 και η ήτα ίση με 0.1.



Σχεδιάγραμμα 6.1

Στο Σχεδιάγραμμα 6.1 μπορούμε να δούμε το δέντρο των τριών επίπεδων που σχηματίστηκε από τον HierarchicalLDA. Τα docs δηλώνουν τον αριθμό των κειμένων που έχουν εναποτεθεί σε κάθε θέμα. Παρατηρείστε ότι ο το άθροισμα των docs των παιδιών κάθε κόμβου γονέα ισούται τον αριθμό των docs που έχουν εναποτεθεί σε αυτόν. Έτσι για παράδειγμα, Υπάρχουν 76 κείμενα που έχουν ως θέμα το Topic 2. Σε αυτά ανήκουν λέξεις που έχουν εναποτεθεί και στα Θέματα 1 και 0. Που στην ουσία είναι το μονοπάτι δέντρου από το αρχικό θέμα 0 στο τελικό θέμα 2. Για να γίνει πιο κατανοητό παραθέτουμε και έναν πίνακα με κάποια θέματα μαζί με τις 10 πιο συχνά εναποτεθειμένες λέξεις σε αυτά.

0 /267325	1 /84574	2 /5442	5 /33317	8 /5957
film story characters life character time man people love director	movie film good bad plot big time action don funny	truman carrey toy show shrek woody gibson princess eddie mel	effects star film special alien ship earth action planet wars	wedding husband town leila roberts sisters egoyan party romantic marriage
11 /67779	12 /36917	14 /4919	16 /2437	17 /3097
movie bad good film action plot time big movies funny	film horror movie scream killer harry alien effects mission original	nbsp war series japanese television files col malick chris chucky	movie bad action funny chan big guy tarzan schumacher flubber	flynt movie larry movies rocky spoon fight tibbs chocolat stretch

Πίνακας 6.2: Ο Πίνακας απεικονίζει 10 από τα 18 θέματα μαζί με τις 10 πιο συχνές λέξεις εναποτεθείμενες σε αυτά. Ο αριθμός δίπλα από την πλάγια γραμμή υποδηλώνει τον συνολικό αριθμό των λέξεων-δειγμάτων που έχουν ανατεθεί στο συγκεκριμένο θέμα.

Αξιολόγηση ιεραρχικού LDA

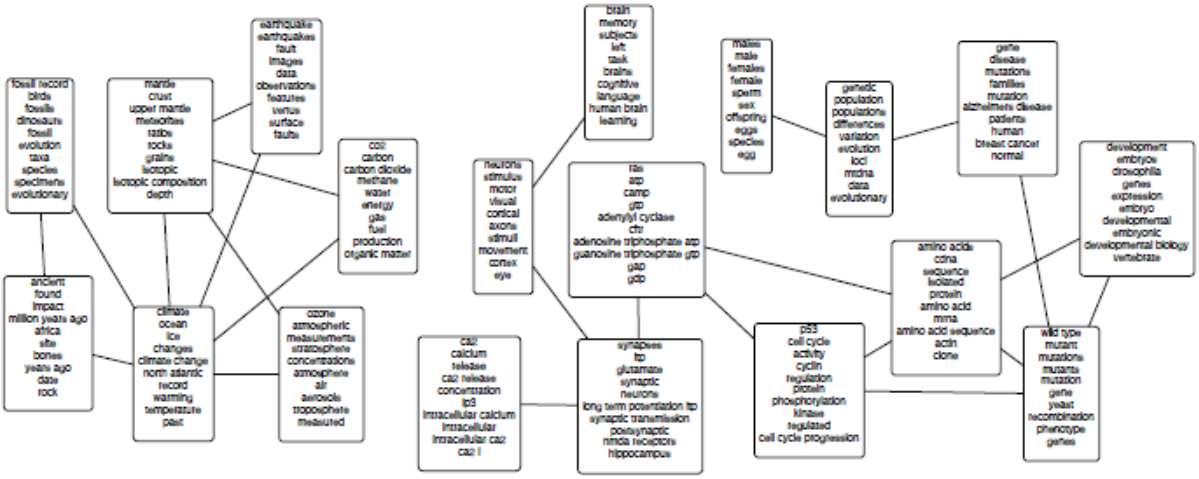
Ο αλγόριθμος αξιολογήθηκε μέσω της μεθόδου empiricalLikelihood που περιέχεται στο Mallet.
Η μέθοδος σχηματίζει ένα μονοπάτι στο δέντρο, συλλέγει μία πολυωνυμική ως προς τα θέματα σε αυτό το μονοπάτι, και στο τέλος επιστρέφει το άθροισμα του βάρους των λέξεων. Όπως και στον ParallelTopicModel κρατήσαμε ως held-out τα 100 τελευταία κείμενα του σώματος Polarity Dataset. Η εμπειρική λογαριθμική πιθανοφάνεια βρέθηκε:

-300206.7253642578

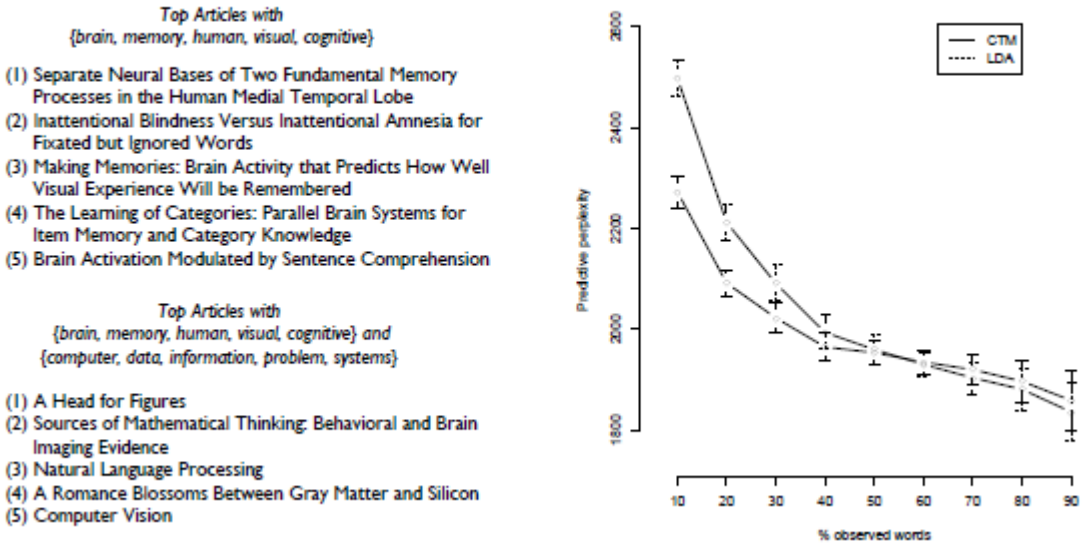
6.2.4 Εφαρμογή Correlated LDA

Πριν αναλύσουμε το δικός πείραμα θα παρουσιάσουμε τα εμπειρικά αποτελέσματα της εργασίας “Correlated Topic Models (Blei και Lafferty)”. Οι παραπάνω για να παρουσιάσουν τον Correlated LDA, χρησιμοποίησαν 16,351 επιστημονικά άρθρα από το 1990 μέχρι το 1999. Κατασκεύασαν ένα γράφημα των κρυφών θεμάτων και των συσχετίσεων μεταξύ τους εξετάζοντας τις πιο πιθανές λέξεις από κάθε θέμα και τις σχέσεις ανάμεσα στα θέματα. Ένα μέρος αυτού του γραφήματος στην Εικόνα 6.4. Σε αυτό το υπογράφημα, υπάρχουν

συνδεδεμένες συλλογές από κείμενα: επιστήμη υλικών, γεωλογία, και βιολογία κυττάρων. Επιπλέον, ένας correlated LDA μπορεί να χρησιμοποιηθεί για να εξερευνήσει μη δομημένα γνωστά κείμενα. Στην Εικόνα 6.5, παρουσιάζεται μία λίστα από άρθρα που έχουν εναποτεθεί στο θέμα "cognitive science" και άρθρα που έχουν εναποτεθεί δύο θέματα "cognitive science" και "visual neuroscience". Ο ενδιαφερόμενος μπορεί να επισκεφτεί την ιστοσελίδα *"http://www.cs.cmu.edu/˜lemur/science/"*.

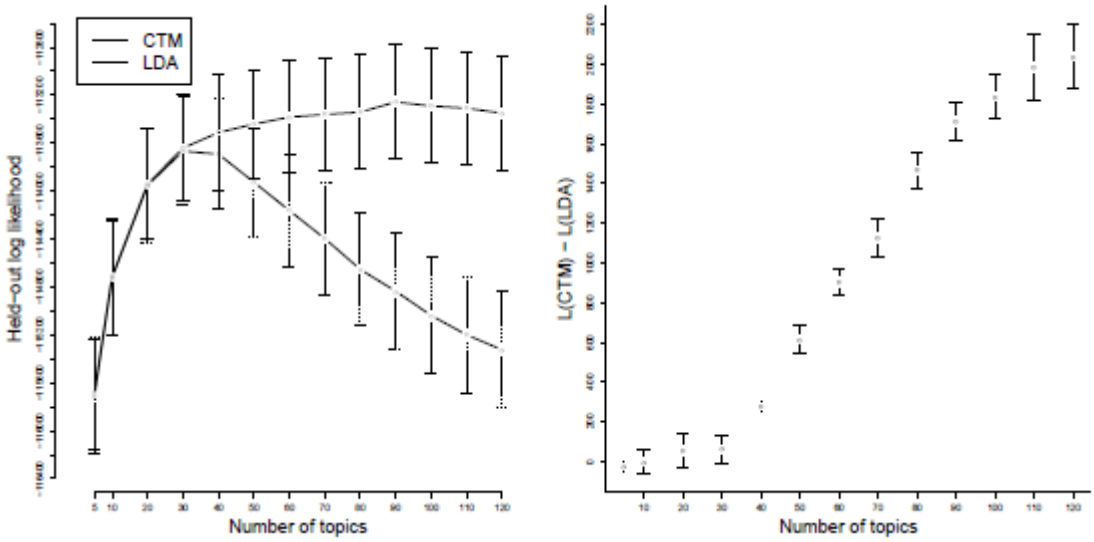


Εικόνα 6.4: Ένα κομμάτι του γραφήματος θεμάτων που δημιουργήθηκε μαθαίνοντας από 16,351 άρθρα από την Science. Κάθε κόμβος αντιπροσωπεύει ένα θέμα.



Εικόνα 6.5: (Αριστερά) Εξερευνώντας μία συλλογή μέσα από τα θέματά της.((Δεξιά) Προσδοκώμενη πολυπλοκότητα για τα μερικώς γνωστά held-out κείμενα από το σώμα 1960 Science.

Ο correlated LDA συγκρίθηκε με τον απλό LDA εισάγοντας μία μικρότερη συλλογή από άρθρα στα μοντέλα θεμάτων. Αυτή η συλλογή περιέχει τα 1,452 κείμενα από το 1960. Χρησιμοποιώντας την μέθοδο “ten-fold cross validation”, υπολογίστηκε η λογαριθμική πιθανοφάνεια των held-out δεδομένων . Ένα καλύτερο μοντέλο θα δώσει υψηλότερη πιθανοφάνεια στα held-out δεδομένα.



Εικόνα 6.6: (Αριστερά) Η μέση held-out πιθανοφάνεια: Ο correlated LDA υποστηρίζει περισσότερα θέματα από τον LDA. (Δεξιά) Τα σφάλματα της held-out πιθανοφάνειας. Θετικοί αριθμοί δείχνουν ότι ο correlated LDA αποκρίνεται καλύτερα.

Η Εικόνα 6.6 δείχνει την μέση held-out λογαριθμική πιθανότητα για κάθε μοντέλο και την μέση διαφορά μεταξύ τους. Ο correlated LDA υποστηρίζει περισσότερα θέματα από τον LDA: Η πιθανοφάνεια για τον LDA κορυφώνεται στα 30 θέματα ενώ η πιθανοφάνεια για τον correlated LDA κορυφώνεται κοντα στα 90 θέματα. Οι μέσοι και σφάλματα της διαφοράς στην λογαριθμική επιφάνεια των μοντέλων παρουσιάζονται στην αριστερή πλευρά της Εικόνας 6.6: και δείχνουν ότι ο correlated LDA αποκρίνεται καλύτερα

Το Mallet δεν διαθέτει κάποιον αλγόριθμο που να υλοποιεί το correlated LDA. Για αυτό τον λόγο δημιουργήθηκε ένα πρότυπο του αλγορίθμου από τον συγγραφέα αυτής της εργασίας από το έργο “Variational EM Algorithms for Correlated Topic Models” των Mohammad Emtiyaz Khan και Guillaume Bouchard, το οποίο αναλύθηκε στο Κεφάλαιο 5. Ο αλγόριθμος έτρεξε με είσοδο 100 κειμένων από το σώμα Polarity Dataset. Δυστυχώς δεν καταφέραμε να δώσουμε περισσότερα κείμενα ως είσοδο διότι η επεξεργασία τους απαιτούσε μία αρκετά μεγάλη ποσότητα υπολογιστικής μνήμης που δεν ήταν διαθέσιμη. Ο κύριος λόγος απαίτησης μνήμης ήταν η δημιουργία ενός πίνακα τριών διαστάσεων, μία διάστασης μεγέθους όσο και πλήθος των θεμάτων που θέλαμε ο αλγόριθμος να επιστρέψει, μία μεγέθους λεξιλογίου των δειγμάτων λέξεων και μία μεγέθους. Ο μόνος τρόπος για να λειτουργήσει ο αλγόριθμος ήταν να μειωθεί δραματικά η είσοδος των κειμένων έτσι ώστε οι δύο τελευταίες διαστάσεις του πίνακα να μειωθούν σε αρκετό βαθμό. Η μείωση των θεμάτων δεν φάνηκε να λύνει σημαντικά το πρόβλημα αφού η μείωσή τους φέρει ως αποτέλεσμα την μείωση της μίας διάστασης μόνο. Έτσι από τα 1900 κείμενα που έτρεξαν οι δύο παραπάνω αλγόριθμοι (ParallelTopicModel και HierachicalLDA) περιοριστήκαμε στα 100. Ο αριθμός θεμάτων που θα αναλυθούν οι λέξεις παρέμεινε 20. Τα αποτελέσματα στα

10 από τα 20 συνολικά θέματα δίνονται στον παρακάτω πίνακα.

0		1		2		3		4	
love	0.1524	character	0.0598	find	0.1401	make	0.1884	film	0.0897
movie	0.0710	giving	0.0588	movie	0.0808	film	0.0936	smart	0.0626
big	0.0643	films	0.0369	action	0.0758	didn	0.0578	people	0.0408
film	0.0628	movie	0.0313	making	0.0452	actors	0.0422	story	0.0365
characters	0.0378	film	0.0279	high	0.0406	movie	0.0298	work	0.0363
good	0.0260	kids	0.0267	story	0.0368	characters	0.0287	comedy	0.0360
time	0.0235	involved	0.0260	comedy	0.0355	scene	0.0242	year	0.0338
long	0.0212	great	0.0246	thinking	0.0242	time	0.0201	great	0.0283
mess	0.0187	life	0.0216	movies	0.0209	screenplay	0.019	mystery	0.0274
made	0.0183	town	0.0211	talent	0.0182	director	0.0186	wife	0.0236
5		6		7		8		9	
funny	0.1215	star	0.0896	movie	0.0721	film	0.1389	big	0.0971
worst	0.0514	kids	0.0630	cast	0.0645	movie	0.0679	scenario	0.0619
movie	0.0415	character	0.0413	obvious	0.0552	give	0.0450	film	0.0536
wife	0.0376	time	0.0412	film	0.0544	good	0.0380	give	0.0492
film	0.0332	story	0.0409	lot	0.0341	problems	0.0284	head	0.0469
mark	0.0303	movie	0.0339	version	0.0263	back	0.0279	wrote	0.0356
immediately	0.02	don	0.0336	place	0.0262	movies	0.0274	original	0.0299
character	0.0244	life	0.0262	lots	0.0257	action	0.0258	movies	0.0297
plays	0.0242	films	0.0257	character	0.0249	car	0.0237	year	0.0245
movies	0.0241	plot	0.0234	interesting	0.0225	production	0.023	mind	0.0236

Πίνακας 6.3: Ο πίνακας κατανομής 10 θεμάτων ως προς τις 10 πιο πιθανές λέξεις του καθενός.

Ο παραπάνω Πίνακας απεικονίζει την κατανομή θεμάτων λέξεων για 10 θέματα μαζί με τις 10 πιο πιθανές λέξεις για το κάθε κείμενο. Οι αριθμοί στα δεξιά των λέξεων δείχνουν την κανονικοποιημένη κατανομή των λέξεων ως προς το συγκεκριμένο θέμα. Όπως παρατηρείτε οι λέξεις “film” και “movie” παρουσιάζονται σε όλα σχεδόν τα θέματα. Αυτό οφείλεται στο ότι αυτές οι λέξεις απατώνται πολύ συχνά σε κάθε κείμενο γεγονός που δημιουργεί ισχυρές συσχετίσεις με όλα σχεδόν με όλες τις υπόλοιπες λέξεις-δείγματα.

Αξιολόγηση Correlated LDA

Στο υποκεφάλαιο 6.1 περιγράψαμε κάποιους αλγόριθμους αξιολόγησης της μεθόδου LDA. Αυτές οι μέθοδοι είχαν ως προϋπόθεση την εφαρμογή της εκάστοτε εκδοχής του LDA με την μέθοδο Gibbs Sampling. Ο CorrelatedLDA όμως πραγματοποιήθηκε με την μέθοδο Expectation Maximamization (EM) συνεπώς δεν είναι δυνατή η αξιολόγησή του με μία από τις παραπάνω μεθόδους.

6.2.5 Σύγκριση Αλγορίθμων

Παρατηρώντας την πιθανοφάνεια στους δύο πρώτους αλγορίθμους ParallelTopicModel και HierachicalLDA (299809.38929934683 και -300206.7253642578 αντίστοιχα) βλέπουμε μία ελάχιστη υπεροχή του ParallelTopicModel. Έτσι με μία ελάχιστη απόκλιση μπορούμε να θεωρήσουμε τους δύο αλγορίθμους ισοδύναμους. Η σύγκριση με τον CorrelatedLDA δεν ήταν δυνατή αφού ο αλγόριθμος δεν έχει αξιολογηθεί.

7 Συμπεράσματα

Περιγράψαμε και αναλύσαμε τον αλγόριθμο Latent Dirichlet Allocation καθώς και τις εφαρμογές του στην εξερεύνηση κειμένων μεγάλων συλλογών. Περιγράψαμε ακόμα 4 επεκτάσεις του αλγορίθμου: Μία εκδοχή που εισαγάγει μία ιεραρχία στα θέματα μέσω ενός δέντρου (ιεραρχικός LDA), μία που υποθέτει συσχέτιση μεταξύ θεμάτων (correlated LDA), δύο που επιτρέπουν στα θέματα να εξελίσσονται σε διακριτό και σε συνεχή χρόνο (Δυναμικό Μοντέλο Θεμάτων και Δυναμικό Θεματικό Μοντέλου Συνεχούς Χρόνου αντίστοιχα), και τέλος μία εκδοχή η οποία λαμβάνει υπόψη μία μεταβλητή απόκρισης (επιβλεπόμενος LDA). Είδαμε πως η μοντελοποίηση θεμάτων παρέχει μία χρήσιμη γνωσιολογία μίας μεγάλης συλλογής κειμένων.

Υπάρχουν πάρα πολλά πλεονεκτήματα με την γενετική πιθανοτική προσέγγιση στην θεματική μοντελοποίηση, σε σύγκριση με μία μη πιθανοτική μέθοδο όπως το LSA (Deerwester et al. 1990) ή με την μη αρνητική παραγοντοποίηση πινάκων (Lee και Seung, 1999). Πρώτον τα γενετικά μοντέλα εφαρμόζονται πιο εύκολα σε νέα δεδομένα. Αυτό είναι απαραίτητο για πεδία όπως αυτά της ανάκτησης δεδομένων και της αρχικοποίησης. Δεύτερον, τα γενετικά μοντέλα μπορούν να προσαρμοστούν και να εξελιχθούν σε πιο πολύπλοκα θεματικά μοντέλα. Για παράδειγμα ο LDA έχει χρησιμοποιηθεί σε μοντέλα ανάλυσης βιβλιογραφίας (Bhattacharya και Getoor, 2006), για σύσταση χρηστών στα κοινωνικά δίκτυα (Frankhauser, Wolfgang, Nejdl. 2009), για ανάλυση προσωπικότητας και συμπεριφορά χρηστών στα κοινωνικά δίκτυα (Ramage, Dumais, Liebling, 2010), για σύσταση προϊόντων στην ηλεκτρονική αγορά (Christidis, Apostolou, Mentzas, ?), για σύσταση άρθρων (Haruechaiyasak, Damrongrat, 2008). Τέλος, τα γενετικά μοντέλα είναι γενικεύμενα με την έννοια ότι εκτός από την ανάλυση κειμένων μπορούν να χρησιμοποιηθούν και στην ανάλυση οντοτήτων που δεν περιέχουν λέξεις ως όρους . Για παράδειγμα έχουν χρησιμοποιηθεί LDA μοντέλα για ανάλυση εικόνων(Fei-Fei και Perona, 2005), σε δεδομένα γενετικής (Pritchard et al., 2000), και σε δεδομένα κοινωνικών δικτύων (Airoldi et al., 2007).

Αναλύσαμε ένα μέρος του προγράμματος Mallet: είδαμε τις δύο κλάσεις υλοποίησης LDA και ιεραρχικού LDA (ParallelTopicModel και HierchicalLDA). Καθώς και υλοποιήσαμε τον correlated LDA (CorrelatedLDA), ένας αλγόριθμος ο οποία κατασκευάστηκε με την βοήθεια της εργασίας των "Variational EM Algorithms for Correlated Topic Models (Mohammad Emtiyaz Khan και Guillaume Bouchard, 14 Σεπτεμβρίου 2009)" και χρησιμοποίησε πακέτα από το Mallet, κυρίως για την διευκόλυνση επεξεργασίας των λέξεων (cc.mallet.pipe, cc.malle.types).

Τελειώνοντας, αξίζει να σημειώσουμε το εξής. Τα θέματα που ανακαλύπτονται από τον LDA και τα άλλα μοντέλα δεν είναι "ορισμένα". Η προσαρμογή ενός θεματικού μοντέλου σε μία συλλογή θα φέρει δομές που μπορεί να έχουν ή και να μην έχουν "φυσική υπόσταση" σε ένα σώμα.

Τα θεματικά μοντέλα είναι ένα χρήσιμο εργαλείο εξερεύνησης. Τα θέματα παρέχουν μία περίληψη ενός σώματος κειμένων που είναι αδύνατο να γίνει με το χέρι. Η θεματική ανάλυση μπορεί να ανακαλύψει συνδέσεις ανάμεσα και μέσα στα κείμενα που δεν είναι φανερά με το γυμνό μάτι, και να βρει συσχετίσεις όρων που κάποιος δεν τις θεωρεί δεδομένες.

8 Βιβλιογραφία

[1] D. M. Blei, A. Ng, and M. I. Jordan(2003). *Latent Dirichlet allocation.* In Journal Machine Learning Research MLR, 3:993-1022.

[2] M. Steyvers, T. Griffiths (2006). *Probabilistic Topic Models.* In Latent Semantic Analysis: A Road to Meaning.

[3] D. M. Blei, J. D. Lafferty (2009) *Topic Models.* In Classification, Clustering and Applications.

[4] D. M. Blei and J. D. Lafferty (2006). *Dynamic topic models.* In ICML.

[5] M. J. Wainwright and M. I. Jordan (2005). *A Variational Principle for Graphical Models.* In New Directions in Statistical Signal Processing.

[6] K. Christidis, D. Apostolou, G. Mentzas (2010). E*xploring Customer Preferences with Probabilistic Topic Models.*

[7] D. M. Blei, J. D. Lafferty (2008). *Correlated Topic Models.*

[8] D. M. Blei, J. D. Lafferty (2007). *A Correlated Topic Model of Science*.

[9] D. M. Blei, J. D. McAuliffe (2010). *Supervised topic models.*

[10] T. K. Landauer, P. W. Foltz (1998). *An Introduction to Latent Semantic Analysis*. In Discourse Processes.

[11] C. Wang, D. M. Blei, D. Heckerman (2007). *Continuous Time Dynamic Topic Models.*

[12] I. Bhattacharya, L. Getoor (2006). *A Latent Dirichlet Model for Unsupervised Entity Resolution*.

[13] R. Krestel, P. Frankhauser, Wolfgang Nejdl (2009). Latent Dirichlet Allocation for Tag Recommendation.

[14] C. M. Bishop (2006). *Graphical Models.* In Pattern Recognition and Machine Learning, 8:359-370.

[15] D. Ramage, S. Dumais, D. Liebling (2010). C*haracterizing Microblogs with Topic Models.*

[16] C. Haruechaiyasak, C. Damrongrat (2008). *Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools.*

[17] D. M. Blei, T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum (2004). Hie*rarchical Topic Models and the Nested Chinese Restaurant Process. In Advances in Neural Information Prossecing Systems 16.*

[18] H. M. Wallach, I. Murray, R. Salakhutdinov, D. Mimmo (2009). *Evaluation Methods for Topic Models.*

[19] McCallum, A. Kachiters (2002). *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.