ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

"Αναγνώριση Προτύπων σε κείμενα ιστοσελίδων Βιοϊατρικού Περιεχομένου με την χρήση της μεθοδολογίας των Hidden Markov Models"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Καραταπάνης Θοδωρής

Επιβλέπων : Κουτσούρης Δημήτριος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός, καθηγητής ΕΜΠ

Αθήνα, Απρίλιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

"Αναγνώριση Προτύπων σε κείμενα ιστοσελίδων Βιοϊατρικού Περιεχομένου με την χρήση της μεθοδολογίας των Hidden Markov Models"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Καραταπάνης Θοδωρής

Επιβλέπων : Κουτσούρης Δημήτριος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός, καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

.....	.....	.....
Κουτσούρης Δημήτριος	Νικήτα Κωνσταντίνα	Γιώργος Ματσόπουλος
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Απρίλιος 2016

ΠΕΡΙΛΗΨΗ

Στην παρούσα διπλωματική εργασία έγινε χρήση τεχνικών μηχανικής μάθησης δίνοντας ιδιαίτερη έμφαση στην ανάπτυξη και μελέτη της τεχνικής των **Hidden Markov Models** για την κατηγοριοποίηση και στην συνέχεια κατάτμηση ιατρικών ημιδομημένων κειμένων που περιγράφουν ιατρικά προϊόντα σε δομημένα έγγραφα ανάλογα με την ιδιότητα στην οποία αναφέρονται.

Πιο συγκεκριμένα το αντικείμενο της διπλωματικής είναι η υλοποίηση εφαρμογής που υποστηρίζει αρχικά τον αυτοματοποιημένο εντοπισμό της περιοχής του κειμένου που εμπίπτει σε μια προκαθορισμένη κατηγορία και εν συνεχεία την περαιτέρω κατάτμηση της περιοχής αυτής σε επιμέρους θεματικές ενότητες που έχουν προκαθοριστεί. Τα κείμενα που χρησιμοποιήθηκαν στα πλαίσια της εργασίας αυτής προέρχονται από βάσεις δεδομένων διεθνών οργανισμών, όπως ο FDA (Food and Drug Administration), και αφορούν ανακλήσεις ιατρικών προϊόντων από τις οποίες επιλέχθηκε να εντοπιστούν ιατρικά προϊόντα που αναφέρονται σε ιατρικές συσκευές από ένα σύνολο κατηγοριών όπως φάρμακα, διατροφικά προϊόντα κ.α και εν συνεχεία προσδιορίστηκε ο εντοπισμός των χωρίων του κειμένου που αναφέρονται στην περιγραφή του προϊόντος (product), τον κωδικό του (code), τον κατασκευαστή (manufacturer), τον λόγο της ανάκλησης (reason), την ποσότητα ανάκλησης (volume) και την κατα τόπους διανομή του προϊόντος (distribution). Ωστόσο σημειώνεται ότι η εφαρμογή που αναπτύχθηκε εύκολα μπορεί να προσαρμοστεί για την δόμηση των κειμένων σε άλλες θεματικές ενότητες ενδιαφέροντος. Τέλος αφού εντοπιστούν τα χωρία μπορούμε να αποθηκεύσουμε τα αρχικά κείμενα σε δομημένη μορφή στην οποία γνωρίζουμε κάθε φορά σε ποιο θέμα σχετικό με το προϊόν κάθε τμήμα αναφέρεται. Επιπλέον, γίνεται δυνατή η περαιτέρω ανάλυση των δεδομένων που βρίσκονται πχ σε μια σχεσιακή βάση δεδομένων από τεχνικές εξόρυξης δεδομένων.

Όλες οι τεχνικές που χρησιμοποιήθηκαν και αναπτύχθηκαν έγιναν σε περιβάλλον **MATLAB** με ιδιαίτερα σημαντική την συμβολή της εργαλειοθήκης **TMG**[3] της matlab για την προεπεξεργασία των κειμένων και την υλοποίηση αλγορίθμων

ταξινόμησης (classification algorithms) καθώς και της εργαλειοθήκης hmm murphy toolbox[8] για την κατασκευή και εκπαίδευση των Hidden Markov Models.

Λέξεις κλειδιά

Μοντέλα Μαρκόφ, αναγνώριση προτύπων, εξόρυξη πληροφορίας από κείμενα ιατρικού περιεχομένου, κατάτμηση κειμένων, κατηγοριοποίηση κειμένων, οργάνωση δομής κειμένου

ABSTRACT

The purpose of this work is to test the application of Hidden Markov Models to the field of text mining. More specifically the project's goal is the automated detection of certain important sections of text (such as sections referring to product description, code, countries of distribution and so on, appearing in reports of medical device recalls) from a set of documents related to the recalls of medical devices, gathered from the web pages of certain agencies such as FDA, TGA and Healthy Canadians. To reach that goal, firstly a binary classifier based on Hidden Markov Models was designed that is able to detect a document that belongs to the category of medical devices with great accuracy.

The method that was used to achieve these good results takes into account besides the word frequencies in each category of the binary classifier(medical devices and other) their relative frequencies of their most frequent words. Furthermore the model was optimized by the removal with a specific method of the least important words.

For the creation of the Hidden Markov Model which can detect the important sections in the documents about the recalls of medical devices, a method was developed to encode the words, given their position in the document and the prevalence of their appearance in each important section, to distinct categories.

These categories can then be used to decode a document in a sequence of observations that can be fed to a Hidden Markov Model which in turn correctly predict the location of the constituent sections of that document.

While the results for documents from individual agencies were excellent, the model can be faulty if we try to combine documents from agencies with very different patterns.

In conclusion while the results for section detection with one and the same model for different agencies weren't good , the methods described in this project could be used to automatically create individual models for different agencies with the ability to discover the location of the important sections (as we describe them) with good accuracy. Finally it is important to note that the methods developed here can be easily

extended to other types of documents (such as newspapers ,general purpose magazines etc).

Keywords

Hidden Markov Models, text mining from medical content, text segmentation, document classification, pattern recognition, text structuring, data mining

Ευχαριστίες

Ευχαριστώ τον καθηγητή Δ.Δ. Κουτσούρη που μου εμπιστεύτηκε αυτήν την πολύ ενδιαφέρουσα διπλωματική καθώς και την Υποψήφια Διδάκτωρα Αγγελική Πανούλια για την βοήθεια και την καθοδήγηση της κατά την διάρκεια της εργασίας μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

Η ΣΗΜΑΣΙΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΑ ΚΕΙΜΕΝΑ (TEXT MINING)

1.1 Η ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΑ ΚΕΙΜΕΝΑ ΚΑΙ ΚΑΠΟΙΕΣ ΠΡΑΚΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ ΤΗΣ

Τα τελευταία χρόνια με την αύξηση των δυνατοτήτων των υπολογιστών, την επέκταση της επιστημονικής γνώσης αλλά και του όγκου πληροφορίας γενικότερα, έχουν προκύψει ανάγκες που δεν υπήρχαν πριν από μερικές δεκαετίες.

Ο όγκος της πληροφορίας είναι τόσο μεγάλος που καθιστά αδύνατη την ολική επίβλεψη του από μικρές ομάδες ανθρώπων με η χωρίς την χρήση υπολογιστών. Ευτυχώς παράλληλα με αυτήν την αύξηση στην ροή της πληροφορίας έχουν αναπτυχθεί και τεχνικές οι οποίες μπορούν να μας βοηθήσουν στην κατασκευή " έξυπνων " φίλτρων για διάφορα είδη πληροφορίας. Η πρακτική αυτών των τεχνικών ονομάζεται Text/Data Mining.

Σκοπός του text mining είναι να αποκτήσει σημαντικές πληροφορίες από κείμενο φυσικής γλώσσας. Ως text mining μπορεί γενικώς να χαρακτηριστεί η διαδικασία ανάλυσης κειμένου προκειμένου να εξαχθούν πληροφορίες που είναι χρήσιμες για συγκεκριμένες εφαρμογές.

Σε σύγκριση με δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων, το κείμενο (text) είναι χωρίς δομή (ή μερικώς δομημένο), άμορφο και δύσκολο να το χειριστούμε αλγοριθμικά. Παρ' όλα αυτά στην εποχή μας ο πιο κοινός τρόπος ανταλλαγής πληροφορίας μεταξύ ανθρώπων, είναι μέσω κειμένου. Το πεδίο του text mining συνήθως αναφέρεται σε κείμενα που προέρχονται από κοινωνικά δίκτυα ή άλλα μέσα ανταλλαγής απόψεων σε γραπτό λόγο και γενικά σε οποιαδήποτε άλλη πληροφορία βρισκεται σε γραπτό λόγο δημοσιευμένη ή μή[5].

Το πεδίο του text mining σήμερα έχει ευρεία εφαρμογή και μπορεί να βοηθήσει στην εκτέλεση των παρακάτω λειτουργιών[6] :

- Φίλτρο spam
- Δημιουργία Recommender όπως το Amazon
- Παρακολούθηση και παρατήρηση προσωπικών απόψεων ατόμων(π.χ. πολιτική ιδεολογία,προτιμήσεις και απόψεις σε διάφορα θέματα) σε κοινωνικά δίκτυα.
- Αυτόματη εξυπηρέτηση πελατών σε e-mails
- Αυτόματη καταχώρηση εγγράφων σε βιβλιοθήκες επιχειρήσεων.
- Μέτρηση προτιμήσεων πελατών μετά από ανάλυση συνεντεύξεων τους.
- Αναγνώριση απάτης.
- Καταπολέμηση cyberbullying ή cybercrime σε chat rooms(im,irc).

  Και άλλα...

1.2 ΣΚΟΠΟΣ ΚΑΙ ΠΕΡΙΛΗΠΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΑΝΤΙΚΕΙΜΕΝΟΥ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ

Σε αυτήν την διπλωματική θα απασχοληθούμε αποκλειστικά με την διαμόρφωση ιατρικών ημιδομημένων κειμένων σε δομημένα κείμενα προκειμένου αυτά να επεξεργαστούν περαιτέρω από ευφυείς αλγορίθμους για την αξιοποίηση του σημασιολογικού τους περιεχομένου. Πιο συγκεκριμένα θα προσπαθήσουμε χρησιμοποιώντας την μεθοδολογία των Hidden Markov Models να εντοπίσουμε συγκεκριμένες περιοχές σε κείμενα που έχουν εξαχθεί από ιστοσελίδες διεθνών οργανισμών όπως ο FDA (http://www.fda.gov/) ο TGA (https://www.tga.gov.au/)

και οι healthy canadians (http://healthycanadians.gc.ca/index-eng.php) και αφορούν ανακλήσεις ιατρικών συσκευών. Θα παρουσιαστεί μια αυτοματοποιημένη μέθοδος που αναπτύχθηκε με την οποία θα μπορούμε να αναγνωρίζουμε τα δεδομένα τα οποία αναφέρονται σε ιατρικές συσκευές από δεδομένα που μπορούν να ανήκουν και σε άλλες θεματικές περιοχές όπως φάρμακα, διατροφικά είδη, βιολογικά προϊόντα και άλλα. Η μέθοδος έχει την δυνατότητα εν συνεχεία να εντοπίζει ενότητες ενδιαφέροντος όπως είναι το τμήμα του κειμένου που αναφέρεται στην περιγραφή του προϊόντος, τον κωδικό του, τον κατασκευαστή, τον λόγο της ανάκλησης, την ποσότητα της ανάκλησης και την διανομή του προϊόντος.

Ένας βασικός λόγος ανάμεσα σε άλλους που κάτι τέτοιο είναι χρήσιμο (ο εντοπισμός αυτών των περιοχών) είναι ότι μπορούν να επιταχυνθούν σε πάρα πολύ μεγάλο βαθμό κάποιες ερωτήσεις σχετικά με αυτά τα προϊόντα από μηχανές αναζήτησης, τον χρήστη ή άλλη εφαρμογή υπολογιστικής ευφυίας.

. Μια αναζήτηση σε κείμενα που δεν έχουν δομή για κάποιες πληροφορίες πρέπει να γίνει στην έκταση όλου του όγκου τους με αποτέλεσμα αρκετά πιο αργές απαντήσεις. Εκτός αυτού η συλλογή κάποιων δεδομένων όπως για παράδειγμα στην περίπτωσή μας, ο κωδικός ή η περιγραφή ιατρικών συσκευών κτλ δεν θα ήταν δυνατή χωρίς την χρήση κάποιων κατηγοριοποιητών οι οποίοι θα μπορούν να αναγνωρίζουν αν τα δεδομένα που επεξεργαζόμαστε έχουν πληροφορία που θέλουμε (στην περίπτωση μας αν θα αναφέρονται σε ιατρικές συσκευές ή όχι). Η πληροφορία που θα συλλεχθεί μπορεί να είναι σημαντική για την εξαγωγή κάποιων συμπερασμάτων, τα οποία μπορούν να χρησιμοποιηθούν από διάφορους φορείς για βελτίωση κάποιων υπηρεσιών ή την λύση κάποιων προβλημάτων.

Στο επόμενο κεφάλαιο θα παρουσιαστούν στοιχεία θεωρίας για τα Hidden Markov Models τα οποία είναι απαραίτητα για την κατανόηση των τεχνικών που θα αναπτυχθούν στα κεφάλαια 4,5 και 6.

ΚΕΦΑΛΑΙΟ 2

HIDDEN MARKOV MODELS

Παρότι τα HMM πρωτοεμφανίσθηκαν στα τέλη του 1960 και στις αρχές του 1970 τα τελευταία χρόνια έχουν γίνει πολύ πιο δημοφιλή στον ερευνητικό χώρο. Υπάρχουν δύο σοβαροί λόγοι που έχει συμβεί αυτό. Καταρχήν τα hmm έχουν καλά ορισμένο μαθηματικό υπόβαθρο που μπορεί αν αποτελέσει την θεωρητική βάση για την χρήση τους σε ένα μεγάλο εύρος εφαρμογών. Ο δεύτερος λόγος είναι ότι όταν τα μοντέλα εφαρμοστούν σωστά μας δίνουν αρκετά καλά αποτελέσματα στην πράξη σε πολλές εφαρμογές[4].

2.1 ΕΙΣΑΓΩΓΗ

Διαδικασίες του πραγματικού κόσμου συνήθως παράγουν παρατηρήσιμες εξόδους που μπορούν να χαρακτηριστούν ως σήματα. Τα σήματα μπορεί να είναι διακριτά στην φύση τους (π.χ. γράμματα του αλφάβητου, κβαντισμένες αποχρώσεις χρωμάτων κ.α.) ή συνεχούς φύσεως (π.χ. θερμοκρασία,δείγματα φωνητικού λόγου, μουσική κλπ). Η πηγή των σημάτων μπορεί να είναι στατική (δηλαδή οι ιδιότητες των σημάτων να μην αλλάζουν με τον χρόνο) είτε μη στατική (δηλαδή οι ιδιότητες μπορεί να αλλάζουν με τον χρόνο). Τα σήματα μπορεί να είναι καθαρά(να έρχονται από μόνο μια πηγή) ή μπορεί να εμπεριέχουν θόρυβο από άλλες πηγές.

Ένα πρόβλημα μείζονος σημασίας είναι να χαρακτηρίσουμε αυτά τα σήματα του πραγματικού κόσμου με όρους μοντέλων για σήματα. Υπάρχουν αρκετοί λόγοι για τους οποίους κάποιος μπορεί να θέλει να χρησιμοποιήσει μοντέλα σημάτων. Πρώτα από όλα, ένα τέτοιο μοντέλο μπορεί να αποτελέσει την βάση για θεωρητική περιγραφή ενός συστήματος επεξεργασίας σημάτων και να χρησιμοποιηθεί μελλοντικά όπου το σύστημα αυτό επεξεργασίας σημάτων χρειάζεται. Για παράδειγμα αν ενδιαφερόμαστε να ενισχύσουμε ένα σήμα ομιλίας το οποίο έχει επηρεαστεί από θόρυβο και έχει αλλοιωθεί λόγο του μέσου μετάδοσης, μπορούμε να

χρησιμοποιήσουμε ένα μοντέλο περιγραφής σημάτων ώστε να φτιάξουμε ένα σύστημα το οποίο θα μπορεί να αφαιρεί τον θόρυβο και να αναιρεί την αλλοίωση, λόγω μέσου μετάδοσης.

Ένας άλλος λόγος πού τέτοια μοντέλα είναι σημαντικά είναι ότι μπορούν να μας δώσουν σημαντικές πληροφορίες για την πηγή των σημάτων (δηλαδή την φυσική διαδικασία που παρήγαγε αυτά τα σήματα) χωρίς να είναι απαραίτητο να έχουμε διαθέσιμη την πηγή. Αυτή η ιδιότητα είναι πολύ σημαντική σε περιπτώσεις που το κόστος δημιουργίας σημάτων από την πραγματική πηγή είναι υψηλό. Σε αυτήν την περίπτωση, με ένα καλό μοντέλο, μπορούμε να προσομοιώσουμε την πηγή και να μάθουμε όσο περισσότερα μπορούμε μέσω των προσομοιώσεων. Τελειώνοντας η σημαντικότερη ίσως συμβολή των μοντέλων περιγραφής σημάτων βρίσκεται στην ανάπτυξη αποδοτικών συστημάτων πρόβλεψης, αναγνώρισης προτύπων, ταυτοποίησης και άλλων ευφυών συστημάτων.

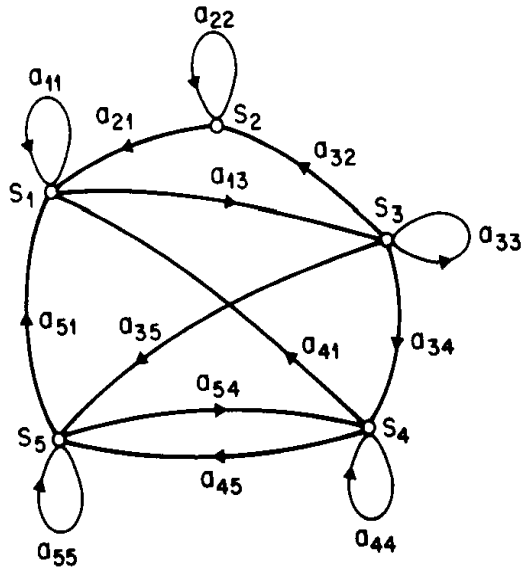
Υπάρχουν αρκετές επιλογές για τον τύπο του μοντέλου που θα χρησιμοποιηθεί για να χαρακτηρίσει τις ιδιότητες ενός σήματος. Ένας γενικός διαχωρισμός αυτών των μοντέλων είναι σε ντετερμινιστικά και στατιστικά μοντέλα. Τα ντετερμινιστικά μοντέλα συνήθως εκμεταλλεύονται κάποιες ειδικές ιδιότητες του σήματος για παράδειγμα ότι το σήμα είναι ημιτονοειδές ή άθροισμα εκθετικών,κλπ. Σε αυτές τις περιπτώσεις η επιλογή του μοντέλου είναι αρκετά απλή, καθώς συνήθως απαιτεί τον υπολογισμό των παραμέτρων του σήματος (π.χ. πλάτος, συχνότητα, φάση, συντελεστές εκθετικών κλπ). Η δεύτερη μεγάλη κλάση μοντέλων για σήματα είναι το σύνολο των στατιστικών μοντέλων στα οποία προσπαθούμε να χαρακτηρίσουμε μόνο τις στατιστικές ιδιότητες του σήματος. Παραδείγματα τέτοιων διαδικασιών είναι οι Γκαουσσιανές, του Πουασόν, Μαρκοβιανές και κρυφές Μαρκοβιανές διαδικασίες. Η υπόθεση που γίνεται για τα στατιστικά μοντέλα είναι ότι το σήμα μπορεί να χαρακτηριστεί εύστοχα σαν μια παραμετρική τυχαία διαδικασία και ότι οι παράμετροι της στοχαστικής διαδικασίας μπορούν να προσδιοριστούν με έναν ακριβή και καλώς ορισμένο τρόπο.

Σε διάφορες εφαρμογές όπως αυτές της επεξεργασίας λόγου (speech) και τα δύο μοντέλα αποδίδουν καλά. Ιδιαίτερα όσον αφορά την δεύτερη κλάση, των

στατιστικών μοντέλων, τα Hidden Markov Models είναι μια πολλά υποσχόμενη τεχνική στην επεξεργασία φωνητικού λόγου και σε πλήθος άλλων εφαρμογών αναγνώρισης προτύπων. Στην συνέχεια θα παρουσιαστεί η απαραίτητη θεωρία για τις αλυσίδες Markov, έπειτα θα επεκταθεί στην εφαρμογή της στα Hidden Markov Models με την βοήθεια κάποιων παραδειγμάτων και τέλος αναλύονται τρία βασικά μαθηματικά προβλήματα που συναντώνται στην πρακτική εφαρμογή της μεθοδολογίας των HMMs στην ανάπτυξη διαφόρων συστημάτων.

2.2 ΔΙΑΚΡΙΤΕΣ ΜΑΡΚΟΒΙΑΝΕΣ ΔΙΑΔΙΚΑΣΙΕΣ

Ας θεωρήσουμε ένα διακριτό σύστημα που μπορεί να βρίσκεται κάθε χρονική στιγμή σε μια κατάσταση από ένα σύνολο N διακριτών καταστάσεων, S_1, S_2,..., S_N. Όπως φαίνεται στην εικόνα 1 οπού το N = 5 στην περίπτωση μας για ευκολία



Εικόνα 1:Μία αλυσίδα Markov με 5 καταστάσεις(S_1 με S_5)

Στον χώρο του διακριτού χρόνου το σύστημα αλλάζει κατάσταση (μπορεί να παραμείνει και στην ίδια κατάσταση) σύμφωνα με κάποιο σύνολο πιθανοτήτων που

αντιστοιχεί σε κάθε κατάσταση. Ορίζουμε τις χρονικές στιγμές όπου μπορεί να προκύψει αλλαγή κατάστασης ως t=1, 2...., και ορίζουμε την κατάσταση που βρισκόμαστε την χρονική στιγμή t ως q_t. Μια πλήρης πιθανοτική περιγραφή του παραπάνω συστήματος, γενικά, απαιτεί την γνώση της κατάστασης την χρονική στιγμή t όπως και όλες τις προηγούμενες καταστάσεις .Για την ειδική περίπτωση της διακριτής Μαρκοβιανής αλυσίδας πρώτης τάξης η περιγραφή περιορίζεται από την γνώση της παρούσας και της προηγούμενης κατάστασης,δηλαδή :

$$P[q_t=S_j|q_{t-1}=S_i,q_{t-2}=S_k,...] = P[q_t = S_j|q_{t-1} = S_i]. \qquad (1)$$

Επιπλέον λαμβάνουμε υπόψιν μόνο τις διαδικασίες στις οποίες το δεξί μέρος της 1 είναι ανεξάρτητο του χρόνου, συνεπώς έχουμε πιθανότητες μετάβασης καταστάσεως της μορφής :

$$a_{ij} = P[q_t = S_j|q_{t-1} = S_i] \qquad 1<=i, j<=N \qquad (2)$$

Οι συντελεστές μετάβασης κατάστασης έχουν τις εξής ιδιότητες :

$$a_{ij}>=0 \qquad (3a)$$

$$\Sigma_{i=1,i<N}\, a_{ij} = 1 \qquad (3b)$$

Η παραπάνω στοχαστική διαδικασία μπορεί να χαρακτηριστεί ώς ένα παρατηρήσιμο Markov Model αφού η έξοδος της διαδικασίας είναι το σύνολο των καταστάσεων κάθε χρονική στιγμή, όπου κάθε κατάσταση αντιστοιχεί σε ένα παρατηρήσιμο γεγονός.

Για να γίνει πιο κατανοητό αυτό ας θεωρήσουμε το απλό markov model 3 καταστάσεων για τον καιρό. Κάνουμε την υπόθεση ότι μια φορά την ημέρα(π.χ το απόγευμα), ο καιρός μπορεί να χαρακτηριστεί ως:

Κατάσταση 1 : Βροχερός-χιονώδης.

Κατάσταση 2 : Συννεφιασμένος.

Κατάσταση 3 : Καθαρός(ηλιοφάνεια).

Υποθέτουμε ότι ο καιρός την t μέρα μπορεί να χαρακτηρίζεται μόνο με τους 3 παραπάνω τρόπους και ότι ο πίνακας Α με τους συντελεστές μετάβασης κατάστασης είναι:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

θεωρώντας ότι ο καιρός την πρώτη μέρα είναι καθαρός (κατάσταση3), μπορούμε να κάνουμε την ερώτηση: Ποια είναι η πιθανότητα (σύμφωνα με το μοντέλο) ο καιρός τις επόμενες 7 μέρες να είναι "καθαρός-καθαρός-βροχερός-βροχερός-καθαρός-συννεφιασμένος-καθαρός" ;

Για να διατυπώσουμε πιο επιστημονικά την ερώτηση ορίζουμε την ακολουθία παρατηρήσεων Ο ως $O=\{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$ που αντιστοιχεί στις χρονικές στιγμές t = 1,2...8. Επιθυμούμε να υπολογίσουμε την πιθανότητα της Ο σύμφωνα με το αυτό το μοντέλο. Αυτή η πιθανότητα μπορεί να εκφραστεί από την παρακάτω σχέση :

$$\begin{aligned} P(O|Model) &= P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|Model] \\ &= P[S_3]*P[S_3|S_3]*P[S_3|S_3]*P[S_1|S_3] \\ &\quad *P[S_1|S_1]*P[S_3|S_1]*P[S_2|S_3]*P[S_3|S_2] \\ &= \pi_3*a_{33}*a_{33}*a_{31}*a_{11}*a_{13}*a_{32}*a_{23} \\ &= 1*(0,8)(0,8)(0,1)(0,4)(0,3)(0,1)(0,2) \\ &= 1,536 * 10^{-4} \end{aligned}$$

Χρησιμοποιούμε τον επόμενο συμβολισμό

$$\pi_i = P(q_1 = S_i), \qquad 1 <= i <= N \qquad (4)$$

Για να ορίσουμε τις πιθανότητες για την πρώτη κατάσταση του μοντέλου

(αρχικοποίηση).

Μια άλλη ενδιαφέρουσα ερώτηση που μπορούμε να κάνουμε είναι : Γνωρίζοντας ότι το μοντέλο βρίσκεται σε μια γνωστή κατάσταση, ποια είναι η πιθανότητα να παραμείνει σε αυτήν την κατάσταση για ακριβώς d μέρες;Αυτή η πιθανότητα μπορεί να υπολογιστεί σαν η πιθανότητα εμφάνισης της ακολουθίας παρατηρήσεων

$O = \{S_i, S_i, \ldots, S_i, S_j\}$ όπως το S_i εμφανίζεται d φορές.

σύμφωνα με το μοντέλο ,το οποίο είναι :

$$P(O \mid \text{Model}, q_1 = S_i) = (a_{ij})^{d-1}(1 - a_{ij}) = p_i(d). \qquad (5)$$

Η ποσότητα $p_i(d)$ είναι η (διακριτή) συνάρτηση πυκνότητας πιθανότητας για διάρκεια παραμονής d στην κατάσταση i. Η εκθετικής μορφής συνάρτηση είναι χαρακτηριστική του χρόνου παραμονής σε μια αλυσίδα Markov. Βασιζόμενοι στην $p_i(d)$ μπορούμε να υπολογίσουμε τις αναμενόμενες παρατηρήσεις (διάρκεια) σε μια κατάσταση με την υπόθεση ότι βρισκόμαστε σε αυτήν την κατάσταση, δηλαδή :

$$\bar{d}_i = \sum_{d=1}^{\infty} d p_i(d) \qquad (6a)$$

$$= \sum_{d=1}^{\infty} d(a_{ii})^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}. \qquad (6b)$$

Συνεπώς περιμένουμε να έχουμε 1/(0,2) = 5 συνεχόμενες μέρες με ήλιο,με σύννεφα 2,5 ημέρες και με βροχή 1,67 ημέρες, σύμφωνα με αυτό το μοντέλο.

2.2.1 ΕΠΕΚΤΕΙΝΟΝΤΑΣ ΤΗΝ ΘΕΩΡΙΑ ΣΤΑ HIDDEN MARKOV MODELS

Μέχρι στιγμής έχουμε χρησιμοποιήσει Markov Models όπου κάθε κατάσταση βρίσκεται σε αντιστοιχία με ένα παρατηρήσιμο γεγονός. Αυτό το μοντέλο είναι πολύ περιοριστικό για να εφαρμοστεί σε αρκετά προβλήματα με ενδιαφέρον. Σε αυτό το τμήμα παρουσιάζεται η επέκταση της ιδέας των Markov Models ώστε οι παρατηρήσεις να είναι στοχαστικές συναρτήσεις των καταστάσεων. Το μοντέλο που προκύπτει ονομάζεται Hidden Markov Model και εκφράζεται στην απλούστερη μορφή του με δύο στοχαστικές διαδικασίες από τις οποίες η μια είναι κρυφή και μπορεί να παρατηρηθεί μόνο μέσω της στοχαστικής διαδικασίας παραγωγής των παρατηρήσεων. Για να γίνει κατανοητή η διαφορά ας θεωρήσουμε το απλό HMM που προσομοιώνει την ρίψη ενός νομίσματος.

Μοντέλα ρίψης νομισμάτων : Ας υποθέσουμε το ακόλουθο σενάριο. Βρίσκεστε σε ένα δωμάτιο και στην μέση του οποίου υπάρχει μια κουρτίνα πίσω από την οποία δεν μπορείτε να ξέρετε τι συμβαίνει. Στην άλλη πλευρά της κουρτίνας υπάρχει ένας άνθρωπος που διεξάγει πείραμα ρίψης ενός (ή πολλών) νομισμάτων. Ο άνθρωπος δεν θα σας πει τίποτα για το ποιο νόμισμα ρίχνει, το μόνο που θα αναφέρει είναι το αποτέλεσμα της ρίψης του νομίσματος. Άρα δημιουργείται μια ακολουθία με τα αποτελέσματα των ρίψεων χωρίς να γνωρίζουμε ποιο νόμισμα χρησιμοποίησε κάθε φορά. Η ακολουθία που δημιουργείται απαρτίζεται από κορώνα (heads) ή γράμματα (tails). Για παράδειγμα το παρακάτω αποτελεί μια ακολουθία παρατηρήσεων για το εν λόγω πείραμα.

$$O = O_1 O_2 O_3 \ldots O_\tau$$
$$= H\ H\ T\ T\ T\ H\ T\ T\ H \ldots H$$

Όπου το Η συμβολίζει την κορώνα και το Τ τα γράμματα.

Έχοντας το παραπάνω σενάριο έναν ενδιαφέρον πρόβλημα είναι η κατασκευή ενός HMM που να εξηγεί τις παρατηρήσεις της ακολουθίας. Το πρώτο πρόβλημα που

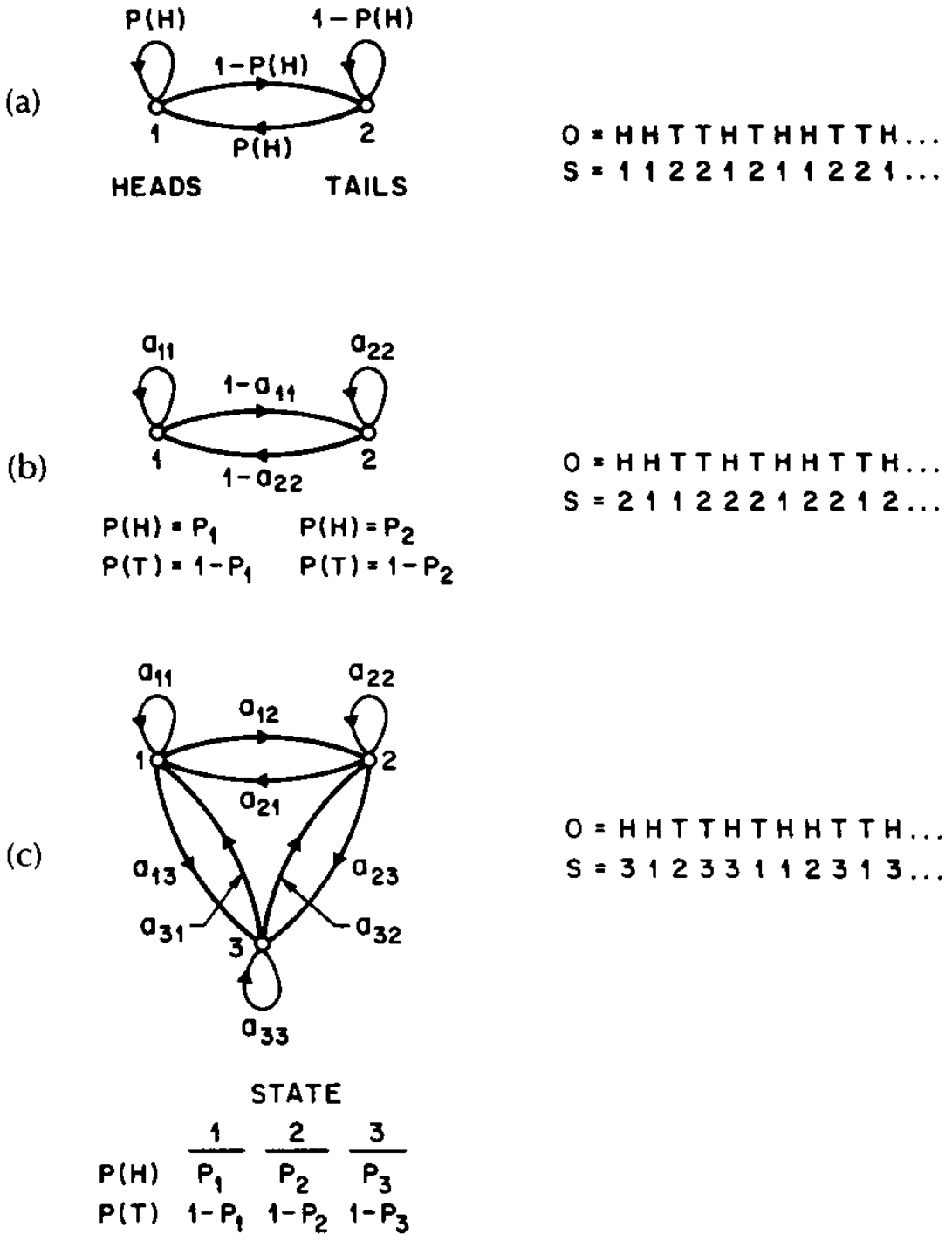
αντιμετωπίζει κάποιος είναι ποια σημασία (φυσική) έχουν οι καταστάσεις του μοντέλου (σε τι αντιστοιχούν) και το δεύτερο είναι πόσες καταστάσεις θα το συγκροτούν. Μια πρώτη υπόθεση θα μπορούσε να είναι ότι οι παρατηρήσεις προκύπτουν από την ρίψη ενός μοναδικού νομίσματος (biased, όχι ισοπίθανο για κάθε πλευρά). Σε αυτήν την περίπτωση, μια δυνατή σχεδίαση για το μοντέλο περιλαμβάνει δύο καταστάσεις, από τις οποίες η μια αντιστοιχεί σε κορώνα και η άλλη σε γράμματα. Αυτό το μοντέλο φαίνεται στην εικόνα 2(a). Με αυτήν την μοντελοποίηση το HMM είναι παρατηρήσιμο και το μόνο που χρειαζόμαστε να βρούμε είναι το πόσο άδικο είναι το νόμισμα (δηλαδή την πιθανότητα για κορώνα). Μπορούμε να μειώσουμε τις καταστάσεις από 2 σε 1, με την κατάσταση να εκφράζει το νόμισμα και την άγνωστη παράμετρο την πιθανότητα ρίψης ενός αποτελέσματος.

Ένας δεύτερος τρόπος μοντελοποίησης ενός HMM φαίνεται στην εικόνα 2(b). Εδώ έχουμε δύο καταστάσεις που αντιπροσωπεύουν δύο διαφορετικά biased νομίσματα. Κάθε κατάσταση χαρακτηρίζεται από πιθανότητες εμφάνισης κορώνας η γραμμάτων, και οι πιθανότητες μεταβάσης από κατάσταση σε κατάσταση βρίσκονται από τον πίνακα μετάβασης κατάστασης (A). Ο μηχανισμός με τον οποίο κάθε φορά επιλέγεται ένα νόμισμα μπορεί να ακολουθεί οποιαδήποτε στοχαστική διαδικασία.

Ένας δεύτερος τρόπος μοντελοποίησης ενός HMM φαίνεται στην εικόνα 2(c). Αυτό το μοντέλο αποτελείται από 3 καταστάσεις που αντιστοιχούν σε 3 biased νομίσματα και η επιλογή νομίσματος κάθε φορά προκύπτει από κάποια πιθανολογική παρατήρηση.

Για ευκολία ανάγνωσης τα τρία μοντέλα παρουσιάζονται στην επόμενη σελίδα

.

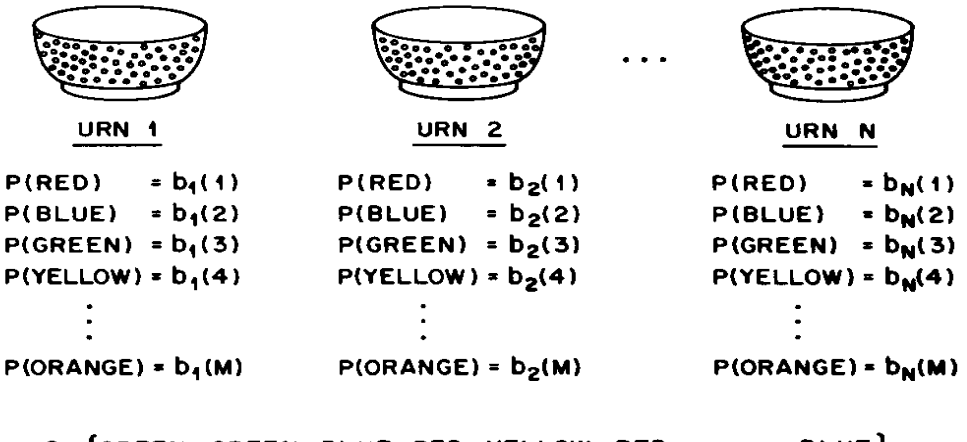


Εικόνα 2. Τρία πιθανά Markov Models τα οποία θα μπορούσαν να ερμηνεύσουν τα αποτελέσματα των κρυφών ρίψεων των νομισμάτων.(a)Μοντέλο με ένα νόμισμα (b)Μοντέλο με δύο νομίσματα. (c) Μοντέλο με τρία νομίσματα.

Το ερώτημα που προκύπτει έχοντας υπόψιν όλα αυτά τα μοντέλα που μπορεί να ερμηνεύσουν το πρόβλημα (και πολλά ακόμα) είναι πως θα επιλέξουμε το καταλληλότερο μοντέλο για το εν λόγω πείραμα. Πρέπει να είναι κατανοητό ότι το μοντέλο με το ένα νόμισμα έχει μόνο μια άγνωστη παράμετρο, το μοντέλο με τα δύο νομίσματα έχει 4 άγνωστες παραμέτρους και το μοντέλο με τα 3 νομίσματα έχει 9 άγνωστες παραμέτρους. Φαίνεται ότι εφόσον τα μεγαλύτερα HMM έχουν περισσότερους βαθμούς ελευθερίας, να είναι πιο ικανά στο να προσομοιώνουν το πείραμα. Παρότι αυτό είναι θεωρητικά σωστό αργότερα θα δούμε ότι υπάρχουν κάποιοι πρακτικοί περιορισμοί που δεν αφήνουν το μοντέλο να είναι πολύ μεγάλο. Εκτός αυτού για παράδειγμα στην περίπτωσή μας το πείραμα μπορεί να γίνεται μόνο με ένα νόμισμα επομένως ένα μοντέλο με τρία νομίσματα δεν θα ανταποκρίνεται στην πραγματικότητα. Ο Rabiner[4] παρουσιάζει καλύτερα την ιδέα των HMMs με το επόμενο πιο πολύπλοκο πείραμα (με σχηματική αναπαράσταση στην εικόνα 3).

Το μοντέλο του δοχείου με τις μπάλες : Υποθέτουμε ότι υπάρχουν N (μεγάλα) γυάλινα δοχεία σε ένα δωμάτιο. Μέσα σε κάθε δοχείο υπάρχει ένας μεγάλος αριθμός από χρωματιστές μπάλες. Κάθε μπάλα μπορεί να έχει ένα χρώμα από M διαφορετικά χρώματα. Η φυσική διαδικασία από την οποία κάνουμε παρατηρήσεις είναι η εξής. Ένα τζίνι υπάρχει στο δωμάτιο και σύμφωνα με κάποια τυχαία διαδικασία επιλέγει ένα αρχικό δοχείο. Από αυτό το δοχείο επιλέγει μια μπάλα κατά τύχη και το χρώμα της είναι η παρατήρησή μας. Η μπάλα στην συνέχεια επανατοποθετείται στο δοχείο από το οποίο προέρχεται. Έπειτα το τζίνι επιλέγει το επόμενο δοχείο από το οποίο θα τραβήξει μια μπάλα σύμφωνα με μία τυχαία διαδικασία που σχετίζεται όμως με το δοχείο της τελευταίας παρατήρησης.

Το πιο απλό HMM που μπορεί να προσομοιώσει αυτήν την διαδικασία περιέχει όσες καταστάσεις όσες και τα δοχεία ενώ οι παρατηρήσεις που μπορεί να γίνουν σε κάθε κατάσταση (χρώματα) αντιστοιχούν σε κάθε χρώμα με μια πιθανότητα. Η διαδικασία επιλογής δοχείου γίνεται μέσο του Πίνακα Μετάβασης Κατάστασης του HMM.



Εικόνα 3.Ένα N-κατάσταστο μοντέλο για το πείραμα με τα δοχεία και τις μπάλες το οποίο απεικονίζει την γενική περίπτωση ενός HMM με διακριτά σύμβολα.

2.2.2 ΣΤΟΧΕΙΑ ΤΩΝ HIDDEN MARKOV MODELS

Τα προηγούμενα παραδείγματα μας δίνουν μια ιδέα για το τι είναι το HMM και πώς μπορεί να εφαρμοστεί σε απλά σενάρια. Στην συνέχεια θα ορίσουμε με σαφήνεια τα στοιχεία του HMM και θα εξηγήσουμε πως το μοντέλο παράγει τις ακολουθίες παρατηρήσεων.

Το HMM χαρακτηρίζεται από :

1) N, τον αριθμό καταστάσεων του μοντέλου. Παρόλο που οι καταστάσεις είναι κρυφές πολύ συχνά οι καταστάσεις ή μια ομάδα από αυτές, αντιπροσωπεύει κάτι στον φυσικό κόσμο. Έτσι στο πείραμα με τα νομίσματα κάθε κατάσταση συμβόλιζε ένα νόμισμα και στο πείραμα με τα δοχεία και τις μπάλες, οι καταστάσεις αντιπροσώπευαν δοχεία. Γενικά οι καταστάσεις είναι συνδεδεμένες με τέτοιο τρόπο ώστε κάθε κατάσταση να μπορεί να φτάσει σε μια άλλη κατάσταση (σε μια χρονική στιγμή). Παρ' όλα αυτά σε κάποιες περιπτώσεις είναι χρήσιμο να εξετάσουμε και άλλες διασυνδέσεις καταστάσεων. Ονομάζουμε τις καταστάσεις (States) $S = \{S_1, S_2, ..., S_N\}$, η κατάσταση την στιγμή t συμβολίζεται με q_t.

2)M, τον αριθμό των ξεχωριστών παρατηρήσεων που μπορούν να γίνουν από κάθε κατάσταση. Για κάθε πείραμα που μπορούμε να διεξάγουμε (με διακριτά στοιχεία) αντιστοιχούμε στα διαφορετικά πιθανά αποτελέσματα (παρατηρήσεις), ένα σύμβολο. Για παράδειγμα στο πείραμα με τα νομίσματα, τα σύμβολα παρατηρήσεων ήταν Η η Τ που αντιστοιχούσαν σε κορώνα και γράμματα. Στο πείραμα με τα δοχεία τα διαφορετικά χρώματα (BLUE ,GREEN κλπ) ήταν τα σύμβολα. Ορίζουμε τα σύμβολα με $V=\{v_1, v_2,..., v_M\}$.

3)Οι πιθανότητες για τις μεταβάσεις καταστάσεων (από κάθε κατάσταση).

$A=\{a_{ij}\}$

όπου $a_{ij}=P[q_{t+1}=S_j|q_t=s_i]$, $1<=i,j<=N.$ (7)

Στην περίπτωση όπου κάθε κατάσταση μπορεί να μεταβεί σε οποιαδήποτε κατάσταση με ένα μόνο βήμα όλα τα a_{ij} είναι μεγαλύτερα του μηδενός. Για άλλα HMM μπορούμε να έχουμε $a_{ij} = 0$ για όσα ζεύγη θέλουμε.

4)Η κατανομή πιθανότητας για την πιθανότητα παρατήρησης ενός συμβόλου για στην κατάσταση j, $B=\{b_j(k)\}$, όπου

$$b_j(k) = P[v_k \text{ στην κατάσταση } t|q_t = S_j], \quad 1 <= j <= N$$
$$1 <= k <= M. \qquad (8)$$

5)Την κατανομή πιθανότητας για την αρχική κατάσταση $\pi=\{\pi_\iota\}$ όπου

$$\pi_\iota = P[q_1 = S_i], \quad 1 <= I <= N. \qquad (9)$$

Γνωρίζοντας τις τιμές των N,M,A,B και π, το HMM μπορεί να χρησιμοποιηθεί για την γέννηση ακολουθιών με παρατηρήσεις.

$$O = O_1 O_2 \ ... \ O_T \qquad (10)$$

(όπου κάθε παρατήρηση O_t είναι ένα από τα σύμβολα του συνόλου V και Τ είναι ο αριθμός συνολικών παρατηρήσεων) με αυτήν την διαδικασία:

1) Γίνεται επιλογή της αρχικής(πρώτης) κατάστασης $q_1 = S_i$ σύμφωνα με την κατανομή πιθανότητας π.

2)Η τιμή του χρόνου t=1.

3)Επιλέγεται το O_t σύμφωνα με την κατανομή πιθανότητας των συμβόλων για την κατάσταση S_i ($b_i(k)$).

4)Γίνεται μετάβαση σε μια νέα κατάσταση μέσω του πίνακα Α που περιέχει τις πιθανότητες μετάβασης καταστάσεων.

5)Επόμενο βήμα $t = t + 1$. Επιστροφή στο βήμα 3 αν $t<T$ αλλιώς η διαδικασία σταματά.

Η παραπάνω διαδικασία μπορεί να χρησιμοποιηθεί σαν γεννήτρια παρατηρήσεων και σαν μοντέλο για τον τρόπο παραγωγής μιας ακολουθίας παρατηρήσεων από ένα κατάλληλο HMM.

Είναι κατανοητό από τα παραπάνω ότι ένας πλήρης προσδιορισμός ενός HMM, χρειάζεται τις τιμές για το πλήθος των καταστάσεων (N) και των διαφορετικών σύμβολα (M), επίσης πρέπει να γνωρίζουμε τις τιμές κατανομής πιθανότητας A, B και π. Χάριν ευκολίας το πλήθος των παραμέτρων του μοντέλου το συμβολίζουμε με λ, δηλαδή :

$$\lambda = (A, B , \pi) \qquad (11)$$

2.2.3 ΤΑ ΤΡΙΑ ΒΑΣΙΚΑ ΠΡΟΒΛΗΜΑΤΑ ΤΩΝ HIDDEN MARKOV MODELS

Σύμφωνα με την μορφή που έχουμε ορίσει προηγουμένως για τα HMM, καλούμαστε να αντιμετωπίσουμε τρία βασικά μαθηματικά προβλήματα που χρειάζεται να επιλύσουμε, προκειμένου η μεθοδολογία των HMMs να έχει πρακτική εφαρμογή. Τα προβλήματα αυτά παρουσιάζονται στην συνέχεια.

Πρόβλημα 1 : Γνωρίζοντας μια σειρά παρατηρήσεων $O = O_1O_2 ...O_T$ και το μοντέλο $\lambda = (A,B,\pi)$ πώς υπολογίζουμε αποδοτικά την πιθανότητα $P(O|\lambda)$ εμφάνισης της ακολουθίας παρατηρήσεων O;

Πρόβλημα 2 : Δεδομένης μιας σειράς παρατηρήσεων $O = O_1O_2 ...O_T$ και με μοντέλο $\lambda = (A,B,\pi)$ πώς επιλέγουμε την αντίστοιχη σειρά εναλλαγής των καταστάσεων η οποία έχει την μέγιστη πιθανότητα από οποιαδήποτε άλλη σειρά διαδοχής καταστάσεων;

Πρόβλημα 3 : Πώς μπορούμε να προσαρμόσουμε τις παραμέτρους $\lambda = (A,B,\pi)$ ώστε να μεγιστοποιήσουμε την πιθανότητα $P(O|\lambda)$ για το μοντέλο μας;

Το πρόβλημα 1 είναι ένα υπολογιστικό πρόβλημα, δηλαδή πρέπει να υπολογίσουμε έχοντας γνωστό το μοντέλο λ και την ακολουθία παρατηρήσεων O, την πιθανότητα παραγωγής αυτής της ακολουθίας από το μοντέλο μας. Αυτό το πρόβλημα εκφράζει την ικανότητα του μοντέλου να προσομοιώνει την φυσική διαδικασία παραγωγής αυτής της εξόδου/παρατηρήσεων. Δηλαδή αν έχουμε πολλαπλά μοντέλα και θέλουμε να επιλέξουμε το καλύτερο για κάποιες σειρές παρατηρήσεων, αυτό που έχει μεγαλύτερες $P(O \mid \lambda)$ προσεγγίζει καλύτερα την λύση.

Το πρόβλημα 2 είναι αυτό το οποίο προσπαθεί να αποκαλύψει τα κρυφά μέρη του HMM, δηλαδή τις καταστάσεις του. **Πρέπει να είναι ξεκάθαρο ότι δεν υπάρχει σωστή σειρά διαδοχής καταστάσεων παρά μόνο για τα εκφυλισμένα μοντέλα.** Άρα για πρακτικές εφαρμογές συνήθως χρησιμοποιούμε ένα κριτήριο

βελτιστοποίησης για να λύσουμε αυτό το πρόβλημα όσο καλύτερα μπορούμε. Δυστυχώς υπάρχουν πολλά τέτοια κριτήρια και η επιλογή του κριτηρίου παίζει μεγάλη σημασία στην "βέλτιστη" σειρά διαδοχής καταστάσεων που θα βρούμε. Κάποιες τυπικές χρησιμότητες αυτού του κριτηρίου είναι για μάθουμε στοιχεία για την δομή του μοντέλου, να βρούμε βέλτιστες ακολουθίες καταστάσεων σε pattern recognition ή να βρούμε στατιστικά για διάφορες καταστάσεις.

Το πρόβλημα 3 προσπαθεί να βελτιστοποιήσει τις παραμέτρους του μοντέλου μας ώστε να εξηγούν (μεγιστοποιούν τις πιθανότητες εμφάνισης των ακολουθιών παρατηρήσεων του μοντέλου) καλύτερα τις παρατηρήσεις O. Οι ακολουθίες παρατηρήσεων που χρησιμοποιούνται για να εκπαιδεύσουμε το μοντέλο ονομάζονται ακολουθίες εκπαίδευσης. Αυτό το πρόβλημα είναι πολύ βασικό για την πρακτική εφαρμογή του μοντέλου, καθώς στις περισσότερες περιπτώσεις θα έχουμε μόνο εκτιμήσεις για το λ (σε κάποιες περιπτώσεις όπως στην τεχνική που ανέπτυξα στο κεφάλαιο 4 πέραν του ορισμού των συμβόλων και του αριθμού καταστάσεων οι παράμετροι υπολογίζονται αυτόματα μέσω εκπαίδευσης) και θα πρέπει να έχουμε ένα τρόπο να εκπαιδεύσουμε το μοντέλο από τις παρατηρήσεις που έχουμε ώστε να βελτιώσει τις προβλέψεις του.

2.3 ΛΥΣΕΙΣ ΣΤΑ 3 ΒΑΣΙΚΑ ΠΡΟΒΛΗΜΑΤΑ ΤΩΝ HIDDEN MARKOV MODELS

2.3.1 ΛΥΣΗ ΣΤΟ ΠΡΩΤΟ ΠΡΟΒΛΗΜΑ

Σε αυτό το πρόβλημα επιθυμούμε να υπολογίσουμε την πιθανότητα εμφάνισης μιας ακολουθίας παρατηρήσεων, $O = O_1 O_2 ... O_T$, δεδομένου μοντέλου λ, δηλαδή το $P(O \mid \lambda)$. Ο πιο απλός τρόπος για να γίνει αυτό, είναι να υπολογίσουμε για όλες τις δυνατές ακολουθίες καταστάσεων μήκους T την πιθανότητα εμφάνισης της O και να βρούμε τον σταθμικό μέσο όρο τους για να έχουμε την $P(O \mid \lambda)$.

Αυτή προκύπτει ότι είναι πολύ χρονοβόρα καθώς για T μήκος ακολουθίας παρατηρήσεων και N αριθμό καταστάσεων του μοντέλου απαιτούνται υπολογισμοί

της τάξης του N^T. Για την ακρίβεια χρειαζόμαστε $(2T - 1)N^T$ πολλαπλασιασμούς και $N^T - 1$ προσθέσεις. Ακόμα και για μικρά νούμερα στα
Ν και Τ οι υπολογισμοί που χρειάζονται για να λυθεί το πρόβλημα με αυτόν τον τρόπο είναι υπερβολικοί (π.χ. για Ν = 5 και Τ = 100 απαιτούνται περίπου 10^{72} πράξεις).Ο αλγόριθμος που χρησιμοποιείται για να λύσει αυτό το πρόβλημα αποδοτικά λέγεται forward-backward algorithm. Η χρονική πολυπλοκότητα αυτού του αλγορίθμου είναι $O(TxN^2)$ η οποία τον κάνει εφαρμόσιμο σε πολλά προβλήματα.

2.3.2 ΛΥΣΗ ΣΤΟ ΔΕΥΤΕΡΟ ΠΡΟΒΛΗΜΑ

Σε αντίθεση με το προηγούμενο πρόβλημα στο οποίο ο forward-backward αλγόριθμος μας δίνει μια ακριβή λύση, υπάρχουν πολλοί πιθανοί τρόποι για να λυθεί, κατά προσέγγιση όμως, το πρόβλημα 2. Αυτό που επιθυμούμε εδώ είναι να βρούμε την βέλτιστη ακολουθία καταστάσεων που σχετίζεται με μια ακολουθία παρατηρήσεων Ο. Το πρόβλημα έγκειται στον προσδιορισμό της βέλτιστης ακολουθίας καταστάσεων χρησιμοποιώντας κάποια κριτήρια. Για παράδειγμα ένα τέτοιο κριτήριο βελτιστοποίησης είναι η επιλογή των καταστάσεων q_t που είναι πιο πιθανές να εμφανιστούν στην ακολουθία παρατηρήσεων Ο (στην θέση t). Με αυτό το κριτήριο μεγιστοποιούμε τον εκτιμώμενο αριθμό των σωστών καταστάσεων στην τελική ακολουθία.

Παρόλο που με τον παραπάνω τρόπο μεγιστοποιούμε τον εκτιμώμενο αριθμό των σωστών καταστάσεων (επιλέγοντας για κάθε χρονική στιγμή t την πιο πιθανή κατάσταση) μπορεί η παραγόμενη ακολουθία καταστάσεων να είναι προβληματική. Για παράδειγμα αν υπάρχουν μηδενικές πιθανότητες μετάβασης καταστάσεων, επειδή με αυτήν την μεθοδολογία υπολογίζουμε την πιο πιθανή κατάσταση κάθε χρονική στιγμή χωρίς να λαμβάνουμε υπόψιν την σειρά διαδοχής τους, μπορεί η λύση να είναι ακόμα και αδύνατο να παραχθεί από το μοντέλο.

Μια πιθανή λύση είναι να αλλάξουμε το κριτήριο βελτιστοποίησης ώστε να ψάχνουμε για την ακολουθία καταστάσεων που μεγιστοποιεί τον σωστό αριθμό δύο διαδοχικών καταστάσεων (q_t , q_{t+1}) ή τρία κλπ , αντί να αναζητάμε μόνο την πιο πιθανή κατάσταση την χρονική στιγμή t . Παρότι αυτά τα κριτήρια μπορεί να είναι

χρήσιμα σε κάποιες εφαρμογές. Το πιο συνηθισμένο κριτήριο που συναντάται είναι η εύρεση της καλύτερης ακολουθίας καταστάσεων έτσι ώστε το P(Q | O, λ) που είναι ισοδύναμο με το P(Q, O | λ) να μεγιστοποιείται. Ο αλγόριθμος που χρησιμοποιείται για την επίλυση αυτού του προβλήματος λέγεται Viterbi (με πολυπλοκότητα $O(TxN^2)$).

2.3.3 ΛΥΣΗ ΣΤΟ ΤΡΙΤΟ ΠΡΟΒΛΗΜΑ

Αυτό το πρόβλημα είναι το πιο δύσκολο των τριών προβλημάτων των Hidden Markov Models. Εδώ αυτό που ψάχνουμε, είναι μια μέθοδο, με την οποία να μπορούμε να ρυθμίσουμε τις παραμέτρους (A, B, π) έτσι ώστε να μεγιστοποιούν την πιθανότητα το μοντέλο να γεννήσει μια ακολουθία παρατηρήσεων O. Δεν υπάρχει κάποιος γνωστός τρόπος με τον οποίο μπορεί να βρεθεί το βέλτιστο μοντέλο λ για πεπερασμένες ακολουθίες παρατηρήσεων. Εντούτοις μπορούμε να επιλέξουμε ένα μοντέλο λ = (A, B, π) τέτοιο ώστε η P(O | λ) να είναι τοπικά μέγιστη, χρησιμοποιώντας μια επαναληπτική μέθοδο όπως η τεχνική του Baum-Welch. Ένα μοντέλο N καταστάσεων μπορεί να εκπαιδευτεί είτε με τυχαίες αρχικές παραμέτρους (A, B, π) είτε με κάποιες αρχικές τιμές που έχουμε επιλέξει εμείς. Επειδή ο Baum-Welch αλγόριθμος βρίσκει τοπικά μέγιστα η επιλογή αρχικών συνθηκών είναι καθοριστικής σημασίας για κάποια προβλήματα. Ο αλγόριθμος χρησιμοποιείται ευρέως στα Hidden Markov Models και έχει βοηθήσει στην κατασκευή Hidden Markov Models που αντιμετωπίζουν προβλήματα σε περιοχές της βιοπληροφορικής, κρυπτανάλυσης, αναγνώρισης λόγου και data mining ανάμεσα σε άλλες.

Στο επόμενο κεφάλαιο θα γίνει παρουσίαση της εργαλειοθήκης TMG η οποία χρησιμοποιήθηκε για την προεπεξεργασία κειμένου και είναι απαραίτητη για την κατανόηση της σχεδίασης των Hidden Markov Models που σχεδιάσθηκαν στα κεφάλαια 4,5 και 6.

ΚΕΦΑΛΑΙΟ 3

ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ ΜΕ ΤΗΝ ΕΡΓΑΛΕΙΟΘΗΚΗ TMG:Text To Document Matrix Generator

3.1 ΕΙΣΑΓΩΓΗ

Η TMG[3] είναι μια εργαλειοθήκη γραμμένη εξολοκλήρου σε Matlab. Το περιβάλλον της Matlab ενδείκνυται σε περιπτώσεις που ένας αλγόριθμος περιλαμβάνει πολλές πράξεις γραμμικής άλγεβρας. Ο σκοπός της εργαλειοθήκης είναι η προεπεξεργασία (με ενσωματωμένους αλγορίθμους) του κειμένου αλλά και η υλοποίηση νέων αλγορίθμων στον χώρο του information retrieval. Στην συνέχεια θα γίνει μια επίδειξη των λειτουργιών της TMG εργαλειοθήκης που χρησιμοποιήθηκαν στην πειραματική διαδικασία για την εκπόνηση της διπλωματικής εργασίας. Πριν την επίδειξη αυτών των λειτουργιών είναι απαραίτητο να αναλυθούν κάποιες ορολογίες και η σημασία τους καθώς θα μας επιτρέψουν να καταλάβουμε καλύτερα τα επόμενα τμήματα της διπλωματικής αλλά θα μας προσφέρουν και εφόδια για την γενικότερη κατανόηση της οικογένειας προβλημάτων που σχετίζονται με το text/data mining.

3.2 VECTOR SPACE MODEL

Το vector space model είναι μια διαδικασία (ή αναπαράσταση ανάλογα με το πως χρησιμοποιείται) για την αναπαράσταση αρχείων με λέξεις (κείμενα) ή και άλλων αντικειμένων όπως queries (ερωτήσεων για ανάκληση συγκεκριμένων κειμένων) με έναν αλγεβρικό τρόπο. Ο λόγος που χρησιμοποιείται το vector space model στην αναπαράσταση κειμένων είναι το γεγονός ότι κάποιες ιδιότητες, χαρακτηριστικές των κειμένων που με άλλο τρόπο δεν θα μπορούσαν να γίνουν διακριτές από έναν υπολογιστή ή ακόμα και από άνθρωπο, πλέον με την χρήση κατάλληλων αλγορίθμων που χρησιμοποιούν γραμμική άλγεβρα πάνω στην αρχική αναπαράσταση του vector space model είναι δυνατόν να εντοπιστούν, όπως ανάλυση περιεχόμενου κειμένων, ομοιότητες μεταξύ κειμένων, σημασιολογικές ομοιότητες μεταξύ κομματιών των κειμένων και άλλα.

ΟΡΙΣΜΟΣ: κάθε κείμενο στο Vector Space Model (VSM) αναπαρίσταται ως πίνακας στον οποίο οι στήλες αναπαριστούν διαφορετικά κείμενα και οι γραμμές αναπαριστούν διαφορετικές λέξεις. Ας ορίσουμε τον document-term πίνακα σαν A με διαστάσεις
MxN όπου N τα διαφορετικά κείμενα που έχουν αναλυθεί και M οι διαφορετικές λέξεις. Κάθε στήλη j με $1<= j <=N$ αποτελεί ένα διάνυσμα
$d_j=\{w_{1j}, w_{2j}, \dots, w_{Nj}\}$ τα w_{ij} είναι συχνότητες εμφάνισης της λέξης i στο κείμενο j.

Η πιο απλή επιλογή για την συχνότητα w_{ij} είναι ο αριθμός εμφανίσεων της λέξης i στο κείμενο j(raw frequency), άλλα είδη συχνότητας που μπορούμε να επιλέξουμε φαίνονται στον παρακάτω πίνακα.

Ονομασία συχνότητας	Συνάρτηση συχνότητας
Δυαδική	0 αν δεν υπάρχει 1 αν υπάρχει
Πραγματική συχνότητα	w_{ij}
Λογαριθμική κανονικοποιημένη	$1 +\log(w_{ij})$ άν $w_{i,j} >0$, αλλιώς 0
Διπλά κανονικοποιημένη με σταθερά K και $0<=K<1$	$K + (1-K)w_{ij}/\max\{i' \text{ ανήκει στο } j\} w_{i'j}$

Οι συχνότητες αυτές ονομάζονται term frequency ή συχνότητες όρου/λέξης

Εκτός από συχνότητες εμφανίσεων λέξεων στα κείμενα υπάρχουν και άλλες συχνότητες που θέλουν να "ζυγίσουν" το πόσο σημαντικές είναι κάποιες λέξεις στο σύνολο των κειμένων(όχι μόνο σε ένα συγκεκριμένο κείμενο) συχνά αναφέρονται στα αγγλικά ως global term weights. Ένας συχνός τρόπος αναπαράστασης αυτής της συχνότητας ονομάζεται inverse document frequency(idf) και υπολογίζεται για έναν όρο i διαιρώντας το πλήθος των κειμένων στα οποία εμφανίζεται ο όρος, με το πλήθος όλων των κειμένων και εν συνεχεία εφαρμόζοντας στον αντίστροφο του λόγου που προκύπτει, την λογαριθμική συνάρτηση.

idf(i,D) = $\log(N/|\{ j \in D : i \in j \}|$ όπου D είναι το σύνολο των κειμένων και N ο αριθμός τους.

Παρακάτω φαίνονται άλλοι τρόποι υπολογισμού των Global Term Weights.

Ονομασία συχνότητας	Συνάρτηση συχνότητας
μοναδική	1
Inverse document frequency	$\log(N/n_i)$, n_i είναι ο αριθμός των κειμένων που έχουν τον όρο i
Inverse document frequency smooth	$\log(1+N/n_i)$
Inverse document frequency max	$\log(1+ \max\{ i' \in j \} n_{i'}/n_i)$
Πιθανοτική inverse document frequency	$\log((N-n_i)/n_i)$

Όπως βλέπουμε τα global term weights είναι συνάρτηση με μοναδικό άγνωστο το i δηλαδή εξαρτώνται μόνο από την λέξη, ας ονομάσουμε g_i το βάρος που αντιστοιχεί στον όρο i.

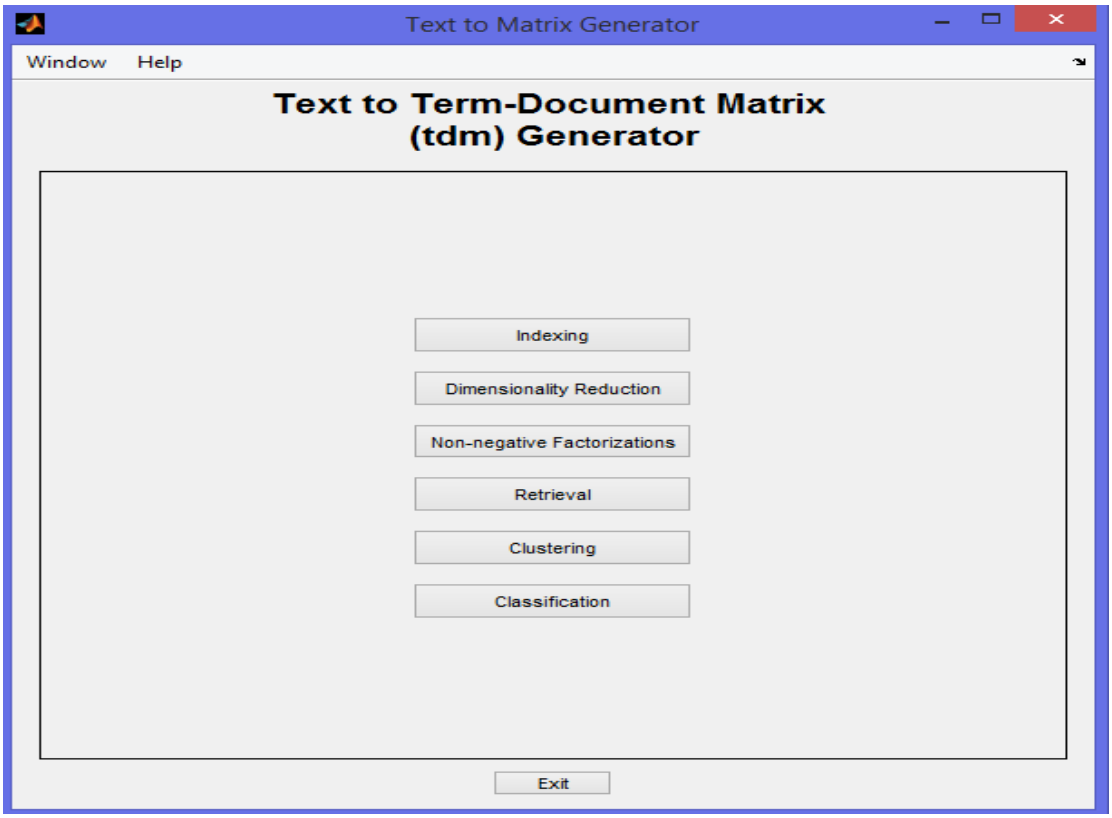
Μετά την αναπαράσταση ενός κειμένου σε vector space model χρησιμοποιώντας το

TMG δημιουργείται ένας πίνακας A διαστάσεων MxN όπου N είναι ο αριθμός των κειμένων και M ο αριθμός των όρων. Κάθε στοιχείο του οποίου ισούται με : $a_{ij} = w_{ij} * g_i * n_{ij}$.

Το μόνο που δεν έχει οριστεί στην παραπάνω εξίσωση είναι το n_{ij} .Αυτός ο όρος ονομάζεται normalization factor και προσπαθεί να περιορίσει την μεγαλύτερη βαρύτητα που παίρνουν οι όροι που εμφανίζονται σε μεγαλύτερα κείμενα.

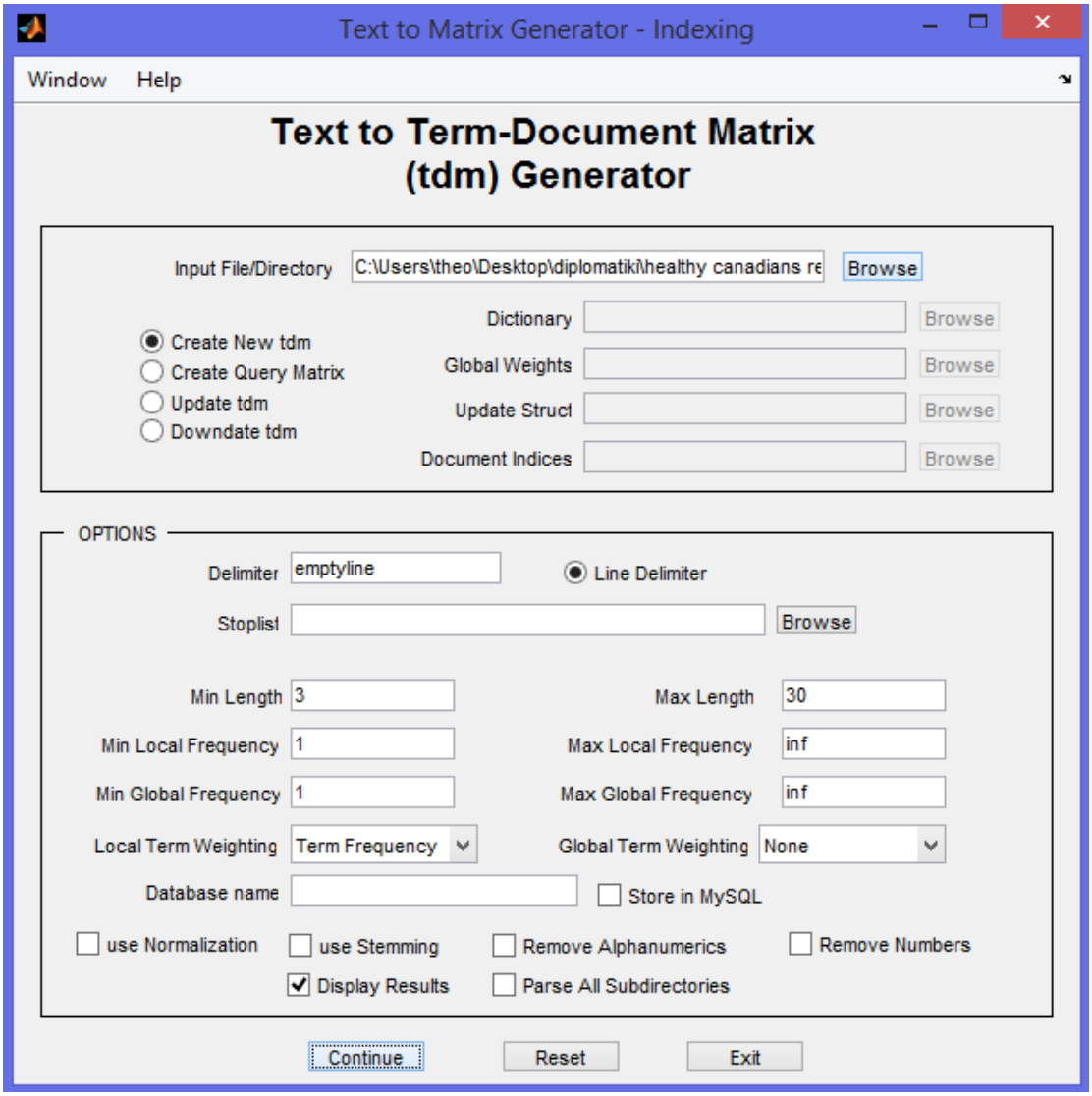
3.3 ΛΕΙΤΟΥΡΓΙΕΣ ΤΗΣ TMG

Έχοντας αναλύσει την βασική ορολογία για το VSM θα προχωρήσω στην ανάλυση των λειτουργιών της TMG που χρησιμοποίησα στις μεθόδους μου για κατηγοριοποίηση και κατάτμηση κειμένου χρησιμοποιώντας Hidden Markov Models. Εγκαθιστώντας την εργαλειοθήκη και τρέχοντας την διεπαφή παρέχεται η δυνατότητα των παρακάτω επιλογών, για την προεπεξεργασία των κειμένων που συνήθως απαιτείται πριν την περεταίρω επεξεργασία τους από ευφυείς αλγορίθμους.



3.3.1 Η ΛΕΙΤΟΥΡΓΙΑ INDEXING

Η επιλογή που χρησιμοποιήθηκε για την ανάλυση των κειμένων που κλήθηκα να ταξινομήσω και να κατατμήσω ήταν η πρώτη, το Indexing. Με αυτήν την επιλογή γίνεται η επεξεργασία κειμένων τα οποία είναι τοποθετημένα σε έναν φάκελο δημιουργώντας έτσι την αναπαράσταση των κειμένων σύμφωνα με το vector space model που περιγράφηκε προηγουμένως. Πατώντας το κουμπί indexing βρισκόμαστε στην σελίδα με όλες τις επιλογές για αυτήν την λειτουργία όπως φαίνεται στην επόμενη εικόνα.



Οι επιλογές Local Term Weighting αφορούν την συχνότητα όρου όπως αναλύθηκε παραπάνω στην παρουσίαση του vector space model. Βλέπουμε ότι υπάρχουν άνω και κάτω όρια για τα global και local term weights ,μια λέξη αν δεν βρίσκεται ανάμεσα σε αυτά τα όρια κόβεται από τον τελικό document-term πίνακα.

Επίσης μπορούμε να επιλέξουμε όρια για τα μήκη των λέξεων. Ένας λόγος να κόψουμε λέξεις λόγω του μήκους τους, είναι ότι πολλές φορές μικρές λέξεις (π.χ. με μήκος < 4) μπορεί να είναι σύνδεσμοι ενώ πολύ μεγάλες λέξεις μπορεί να είναι τυπογραφικά λάθη, συμβολοσειρές η κάτι άλλο που δεν περιέχει σημαντική πληροφορία. Επιπλέον μπορεί να έχουμε παρατηρήσει ότι για συγκεκριμένα κείμενα λαμβάνουμε καλύτερα αποτελέσματα σε κάποια συχνά προβλήματα όπως του classification(κατηγοριοποίηση) αν πάρουμε άλλο όριο λέξεων, πολλές φορές οι βέλτιστες επιλογές αυτών των ορίων βρίσκονται με πειράματα άλλα δεν είναι κάτι που πρέπει να μας απασχολήσει ιδιαίτερα.

Κάποιες άλλες επιλογές που θα ήθελα να αναλύσω περισσότερο είναι οι παρακάτω:

1)Stopword list.
2)Delimeters.
3)Stemming.
4)Αφαίρεση αριθμών και αλφαριθμητικών.

1)Stopword list είναι μια λίστα με λέξεις οι οποίες δεν θα ληφθούν υπόψιν στην κατασκευή του document-term πίνακα. Οι stopword lists συνήθως αποτελούνται από λέξεις οι οποίες εμφανίζονται πολύ συχνά στην γλώσσα γραφής των κειμένων το οποίο συνεπάγεται ότι μπορεί να είναι λέξεις οι οποίες βοηθούν στην κατασκευή κειμένου αλλά δεν περιέχουν πληροφορία όσον αφορά το περιεχόμενου του κειμένου τέτοιες λέξεις μπορεί να είναι στην αγγλική γλώσσα το the , and ,above, whereas και άλλες οι οποίες συνδέουν προτάσεις, είναι προσδιορισμοί κλπ. Stopword lists είναι

πολύ σημαντικό να χρησιμοποιούνται στην φάση του indexing διότι έτσι καθαρίζουμε τα κείμενα από λέξεις χωρίς κάποιο νόημα που εμφανίζονται συχνά με αποτέλεσμα να δημιουργούν φαινομενικές ομοιότητες μεταξύ κειμένων.

2)Delimiters είναι σύμβολα με τα οποία μπορούμε να χωρίσουμε ένα κείμενο σε μικρότερα. Κατά την επεξεργάσια των κειμένων ένα κείμενο θεωρείται διαφορετικό με ένα άλλο είτε όταν ανήκουν σε διαφορετικό αρχείο είτε όταν βρίσκονται στο ίδιο αρχείο αλλά ανάμεσά τους υπάρχει delimeter. Delimeter μπορεί να είναι μια κενή γραμμή η ακόμα και μια λέξη ή μια συμβολοσειρά.

Η συγκεκριμένη επιλογή είναι αρκετά χρήσιμη σε περιπτώσεις που γνωρίζουμε ότι κάποια κείμενα αποθηκεύονται με έναν συγκεκριμένο τρόπο χωρίζοντας κάποιες ενότητες με συγκεκριμένες λέξεις,έτσι ώστε να μπορέσουμε να επεξεργαστούμε κείμενα χρησιμοποιώντας ξεχωριστές ομάδες.

Delimiters επίσης μπορεί να χρησιμοποιηθούν επειδή κάποια κείμενα μπορεί να είναι σειριακά αποθηκευμένα σε ένα αρχείο το οποίο δεν ενδείκνυται όταν θέλουμε να έχουμε καλή ανάλυση με το vector space model.

3)Stemming είναι η διαδικασία με την οποία ορίζουμε μια ρίζα και λέξεις οι οποίες μπορούν να παραχθούν από αυτή την ρίζα. Ουσιαστικά αυτός ο αλγόριθμός κοιτάζει αν με κάποιο κανόνα μια λέξη παράγεται από μία ρίζα η δεν έχει ρίζα. Ο τρόπος με τον οποίον χρησιμοποιείται είναι οι λέξεις που έχουν την ίδια ρίζα να αναγνωρίζονται ως μια (δηλαδή ως την ρίζα) και οι λέξεις που δεν έχουν την ίδια ρίζα ως διαφορετικές. Για παράδειγμα οι λέξεις am, are ,is αναγνωρίζονται ως be και οι λέξεις car, car's ,cars , cars' αναγνωρίζονται ως cars από έναν υποθετικό αλγόριθμο stemming. Έτσι η παρακάτω πρόταση:

the boy's cars are different colors
μετά από χρήση ενός αλγόριθμου stemming μετατρέπεται σε:
the boy car be differ color

Ο λόγος που αυτή η επιλογή είναι πολύ σημαντική στο vector space model είναι ότι

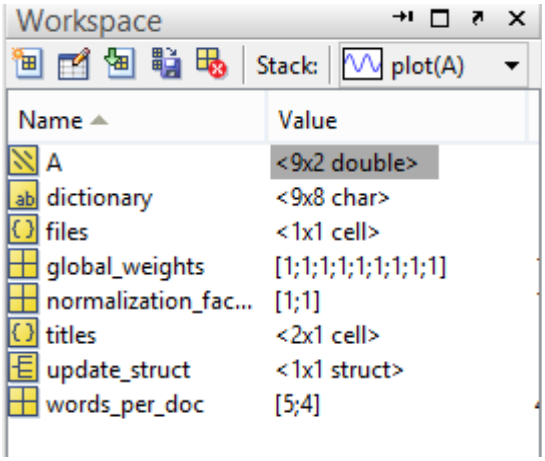
πολλές λέξεις με το ίδιο νόημα(ρίζα) άλλα διαφορετική διατύπωση λαμβάνονται σαν μία λέξη με αποτέλεσμα να αυξάνεται το βάρος που έχει η συγκεκριμένη ρίζα στο σύνολο των κειμένων. Απουσία του stemming μπορεί να διασπούσε το βάρος αυτής της ρίζας/νοήματος σε πολλές λέξεις με αποτέλεσμα να μην ήταν τόσο εμφανές αυτό το χαρακτηριστικό. Εντούτοις το stemming μπορεί να βλάψει κάποιους αλγόριθμους classification ,clustering επομένως θα πρέπει να γίνεται ένας έλεγχος για το άν θα πρέπει να χρησιμοποιηθεί αυτή η επιλογή. Πρέπει να σημειωθεί ότι στην TMG το stemming συμβαίνει μετά την εξάλειψη των stopwords άρα για καλύτερα αποτελέσματα θα πρέπει η λίστα stopwords να περιέχει και όλα τα παράγωγα λέξεων που περιλαμβάνονται σε αυτήν.

4)Αφαίρεση αριθμών και αλφαριθμητικών είναι καλύτερο να γίνεται όταν θέλουμε να διαχωρίσουμε κείμενα τα οποία περιέχουν πολλούς αριθμούς και αλφαριθμητικά που δεν επαναλαμβάνονται, αφενός για αύξηση της απόδοσης (ταχύτητας του αλγόριθμου) και αφετέρου επειδή αυτή η πληροφορία (ένα τυχαίο αλφαριθμητικό όπως ο κωδικός ενός προϊόντος) δεν είναι σημαντική για την κατηγοριοποίηση. Όταν θέλουμε να κάνουμε κατηγοριοποίηση κάποιων κειμένων είτε χρησιμοποιώντας κάποιον γνωστό αλγόριθμο είτε με δική μας μέθοδο είναι συνετό να αφαιρέσουμε ότι θεωρούμε ότι δεν αποτελεί δείκτη ομοιότητας των κειμένων (δηλαδή αν θέλουμε να κάνουμε κατηγοριοποίηση σε κείμενα που θεωρούμε ότι οι αριθμοί δεν παίζουν ρόλο στον προσδιορισμό των κατηγοριών των κείμενων που έχουμε ορίσει, τους αφαιρούμε). Σε αρκετές περιπτώσεις οι αριθμοί και τα αλφαριθμητικά πρέπει να αφαιρούνται άλλα αν πχ θέλουμε να διαχωρίσουμε κείμενα που διαφοροποιούνται ως προς την συχνότητα εμφάνισης αλφαριθμητικών, εξυπηρετεί να διατηρούνται μέσα στα κείμενα.

Γενικά οι παραπάνω παράμετροι μπορούν να επηρεάσουν αρκετά τα αποτελέσματα σε διάφορες εφαρμογές και δεν υπάρχει σωστός και λάθος τρόπος επιλογής τους. Κάθε προγραμματιστής θα πρέπει να αναγνωρίζει τις ιδιαιτερότητες των κειμένων που θέλει να αναλύσει και μετά κάνοντας πειράματα να βρει κάποιες παραμέτρους οι

οποίες του δίνουν λιγότερα λάθη.

Αφότου επιλέξουμε τον φάκελο στον οποίον θέλουμε να εφαρμόσουμε την επιλογή indexing επιλέξουμε τις παραμέτρους και τρέξουμε το πρόγραμμα στο workspace της matlab εμφανίζονται οι εξής μεταβλητές που φαίνονται στην παρακάτω εικόνα:



Από τις παραπάνω μεταβλητές μας ενδιαφέρουν μόνο η μεταβλητή A και η μεταβλητή dictionary. Η μεταβλητή A είναι ο document-term πινακας και η μεταβλητή dictionary βρίσκεται σε αντιστοιχία με τις γραμμές του πίνακα A έτσι ώστε η λέξη που αντιπροσωπεί η γραμμή του πίνακα A να ονοματίζεται από την λέξη που υπάρχει στην ίδια γραμμή του πίνακα dictionary. Τα global weights δεν μας ενδιαφέρουν καθώς στην μέθοδο μου θέλω τα στοιχεία του A να έχουν raw frequencies άρα τα global weights να είναι 1.

3.3.2 ΣΥΝΟΠΤΙΚΗ ΑΝΑΦΟΡΑ ΣΤΙΣ ΥΠΟΛΟΙΠΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΤΗΣ TMG

Καθότι στην εργασία μου η μόνη λειτουργία από την TMG που χρειάστηκε να χρησιμοποιηθεί ήταν το indexing, δεν θα αναλύσω τις υπόλοιπες λειτουργίες αλλά θεωρώ σκόπιμο να αναφέρω συνοπτικά και τις υπόλοιπες δυνατότητες που μας

παρέχει η εργαλειοθήκη TMG..

Η εν λόγω εργαλειοθήκη εκτός από την δυνατότητα indexing παρέχει τα εξής :

Dimensionality reduction /Μείωση διαστάσεων : Μείωση διαστάσεων είναι μια μεθοδολογία που προσπαθεί να προβάλει ένα σύνολο από διανύσματα υψηλής διάστασης σε ένα χώρο μικρότερης διάστασης[1] .

Αυτή η λειτουργία μας επιτρέπει από ένα σύνολο μεταβλητών που αναπαριστούν μια οντότητα (π.χ. κείμενο) να μπορέσουμε να μειώσουμε τις μεταβλητές (π.χ. λέξεις), αφαιρώντας αυτές που παίζουν μικρότερο ρόλο στον χαρακτηρισμό αυτής της οντότητας, δηλαδή τις μεταβλητές με την λιγότερη πληροφορία.
Η TMG παρέχει έξι τεχνικές για μείωση διάστασης, τις εξής :

- Singular Value Decomposition (SVD)
- Principal Component Analysis (PCA)
- Clustered Latent Semantic Indexing (CLSI)
- Centroids Method (CM)
- Semidiscrete Decomposition (SDD)
- SPQR Decomposition

Non-Negative factorizations (NNMF) : Οι αλγόριθμοι που επιτελούν αυτήν την λειτουργία μετατρέπουν έναν πίνακα V, συνήθως σε γινόμενο δύο πινάκων έστω W και H με την ιδιότητα της απουσίας αρνητικών στοιχείων και στους 3 πίνακες. Αυτή η ιδιότητα
καθιστά τους πίνακες W και H (που είναι προφανώς μικρότερης τάξης) πιο εύκολο να εξεταστούν. Οι τεχνικές που υλοποιούν μη αρνητική παραγοντοποίηση βρίσκουν εφαρμογή εκτός άλλων (όπως η όραση υπολογιστών, επεξεργασία ηχητικών σημάτων) και στο πεδίο του text mining (εξόρυξη πληροφορίας κειμένου) [2] .

Εξαιτίας του ότι οι τεχνικές είναι επαναληπτικές το τελικό αποτέλεσμα εξαρτάται από την αρχικοποίηση [3], ένας συνηθισμένος τρόπος προσέγγισης αυτού του προβλήματος είναι με τυχαία αρχικοποίηση, όμως κάποιες πιο σύγχρονες μέθοδοι αρχικοποίησης φαίνεται να εμφανίζουν καλύτερα αποτελέσματα η εργαλειοθήκη

TMG μας παρέχει 4 τεχνικές αρχικοποίησης :

- Non-Negative Double Singular Value Decomposition (NNDSVD)
- Block NNDSVD
- Bisecting NNDSVD
- By clustering

και δύο αλγόριθμους NNMF περαιτέρω βελτίωσης των αποτελεσμάτων :

- Multiplicative Update algorithm by Lee and Seung
- Alternating Non-negativity constrained least squares

Retrieval : Αυτή η λειτουργία αναφέρεται σε text retrieval (ανάκληση κειμένου μετά από κάποια ερώτηση/query). Η TMG μας παρέχει δύο εναλλακτικές για text mining [3] :

- Vector Space Model (VSM)
- Latent Semantic Analysis (LSA)

Clustering : Clustering είναι η αυτόματη κατηγοριοποίηση κειμένων σε κάποιες ομάδες. Η διαφορά του από το classification/κατηγοριοποίηση είναι ότι δεν χρησιμοποιείται κάποιο training data set για τον ορισμό αυτών των ομάδων/κατηγοριών. Η TMG μας παρέχει τους τρεις επόμενους αλγορίθμους για clustering [3] :

- k-means
- Spherical k-means
- Principal Direction Divisive Partitioning (PDDP)

Classification : Η κατηγοριοποίηση των κειμένων μπορεί να γίνει με τους παρακάτω αλγορίθμους που παρέχονται από την TMG [3] :

- k-Nearest Neighbours (KNN)
- Rocchio

- Linear Least Squares Fit (LLSF)

Στο επόμενο κεφάλαιο θα παρουσιασθεί η μέθοδος που ανέπτυξα για την δημιουργία κατηγοριοποιητή με χρήση Hidden Markov Models, ο οποίος θα μας επιτρέπει την αναγνώριση κειμένων που αναφέρονται αποκλειστικά σε ιατρικές συσκευές ανάμεσα σε κείμενα που σχετίζονται με ιατρικά προϊόντα.

ΚΕΦΑΛΑΙΟ 4

ΧΡΗΣΗ HIDDEN MARKOV MODELS ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΩΝ

4.1 ΕΙΣΑΓΩΓΗ

Σε αυτό το μέρος θα εξεταστεί η δυνατότητα χρήσης Hidden Markov Models για την κατηγοριοποίηση κειμένων σε δύο κατηγορίες.

Το dataset που χρησιμοποιήθηκε για την κατασκευή του μοντέλου προέρχεται από ιστοσελίδες του δικτυακού τόπου Healthy Canadians (http://healthycanadians.gc.ca/). Τα δεδομένα που συλλέχθηκαν είναι κείμενα που αφορούν ανακλήσεις προϊόντων από 5 κατηγορίες σχετικές με την υγεία, όπως ιατρικές συσκευές, τρόφιμα, φάρμακα, βιολογικά προϊόντα και άλλα φυσικά προϊόντα. Για το τελικό στάδιο της διπλωματικής ενδιαφερόμαστε μόνο για τα κείμενα που αναφέρονται σε ιατρικές συσκευές άρα είναι χρήσιμο να κατασκευαστεί ένας ταξινομητής ο οποίος θα μπορεί να ξεχωρίζει τις ιατρικές συσκευές από τα υπόλοιπα κείμενα που αναφέρονται στις άλλες κατηγορίες.

Έχοντας κάνει την εισαγωγή για τις γενικές ιδέες των Hidden Markov Models σε αυτό το τμήμα θα γίνει λεπτομερής ανάλυση της διαδικασίας και του τρόπου σκέψης πίσω από την επιλογή αυτής της μεθοδολογίας για την αυτοματοποιημένη κατηγοριοποίηση κειμένου.

Τα Hidden Markov Models είναι στατιστικά μοντέλα και καθώς η ανάλυση των κειμένων βασίζεται σε στατιστικά ανάλογα με εμφανίσεις λέξεων περιμένουμε ότι το μοντέλο θα έχει τουλάχιστον μια καλή επιτυχία ως προς τη ζητούμενη κατηγοριοποίηση επιλέγοντας οι παρατηρήσεις των ακολουθιών εκπαίδευσης και ελέγχου να στηρίζονται στις πιθανότητες εμφάνισης "σημαντικών" λέξεων. Παρακάτω παρουσιάζεται διεξοδικά η σχεδίαση ενός ενδεικτικού HMM.

Αξίζει να σημειωθεί ότι ο παρακάτω τρόπος που θα αναλυθεί είναι ένας από τους δυνατούς τρόπους για κατηγοριοποίηση που μπορούμε να χρησιμοποιήσουμε βασιζόμενοι στα Hidden Markov Models υπάρχουν και άλλοι τρόποι οι οποιοι μπορούν να εκμεταλλευτούν ιδιαίτερα χαρακτηριστικά αυτών των κειμένων (π.χ. το μοντέλο μπορεί να ελέγχει συγκεκριμένη αλληλουχία λέξεων που είναι ξεχωριστή για κάθε κείμενο) αλλά καθαρά για ερευνητικούς σκοπούς θεωρώ ότι ο τρόπος μου μπορεί πολύ εύκολα να γενικευτεί για οποιαδήποτε κείμενα με καθαρά αλγοριθμικό τρόπο σε αντίθεση με τους άλλους τρόπους οι οποίοι απαιτούν την παρατηρητικότητα του προγραμματιστή για τις ιδιαιτερότητες κάθε κειμένου μιας ομάδας κειμένων προς κατηγοριοποίηση.

4.2.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΤΕΧΝΙΚΗ

Πάρα πολύ βασικό στην απόδοση του Hidden Markov Model με σκοπό την κατηγοριοποίηση είναι η σωστή επιλογή του "λεξικού" των παρατηρήσεων (σύνολο των emission symbols) τα οποία πρέπει να επιλεγούν με τέτοιο τρόπο ώστε η συχνότητα εμφάνισης τους να διαφέρει ανάμεσα στις δύο κατηγορίες κειμένων. Όπως γνωρίζουμε ένα Hidden Markov Model όταν εκπαιδευτεί με έναν αλγόριθμο εκπαίδευσης (π.χ. ο Baum Welch) θα προσπαθήσει να μεγιστοποιήσει την πιθανότητα του για να εμφανίσει μια ακολουθία παρατηρήσεων : $O = O_1 O_2 ... O_\tau$ συνεπώς αν στις δύο ομάδες κειμένων υπάρχουν παρατηρήσεις οι οποίες είναι πιο συχνές στην μια από ότι στην άλλη και εκπαιδεύσουμε δύο Hidden Markov Models με ακολουθίες παρατηρήσεων οι οποίες για το ένα θα ανήκουν στην ομάδα των κειμένων που αναφέρονται στις ιατρικές συσκευές και για το άλλο στα κείμενα που ανήκουν στις υπόλοιπες κατηγορίες θα υπολογίζουμε την πιθανότητα εμφάνισης μιας σειράς

παρατηρήσεων (η οποία δημιουργείται από την ανάλυση ενός κειμένου) και αν η πιθανότητα εμφάνισης της ακολουθίας είναι μεγαλύτερη για το Hidden Markov Model που αναφέρεται στις ιατρικές συσκευές, το σύνθετο μοντέλο, θα ταξινομεί το αντίστοιχο κείμενο στην κατηγορία 'ιατρικές συσκευές', αλλιώς θα το εντάσσει στην κατηγορία 'άλλο'.

4.2.2 ΠΑΡΑΓΩΓΗ ΑΚΟΛΟΥΘΙΩΝ ΠΑΡΑΤΗΡΗΣΕΩΝ ΑΠΟ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ

Εφόσον έχω εξηγήσει τον τρόπο με τον οποίο θα γίνεται η κατηγοριοποίηση τώρα θα αναλύσω την διαδικασία με την οποία παράγονται οι ακολουθίες παρατηρήσεων που χρησιμεύουν είτε για την εκπαίδευση είτε ως είσοδος για το μοντέλο, ως αποτέλεσμα της προεπεξεργασίας και ανάλυσης των κειμένων. Πιο συγκεκριμένα παρακάτω προσδιορίζω τα emission symbols που αντιστοιχούν στις καταστάσεις των δύο HMMs.

Για να μπορέσω να προσδιορίσω τα emission symbols χρειάζεται να γίνει ανάλυση πάνω στα δεδομένα. Καταρχήν από ένα σύνολο 3263 κειμένων από τα οποία τα 2000 ανήκουν στην κατηγορία : 'Medical Devices' (ιατρικές συσκευές) και τα υπόλοιπα στην κατηγορία 'other' (άλλο) κράτησα 800 κείμενα της κατηγορίας medical devices και 800 κείμενα της κατηγορίας other για την ανάλυση λέξεων οι οποίες θα χρησιμοποιηθούν για τον προσδιορισμό των emission symbols και στην συνέχεια κάποια από αυτά τα κείμενα αφού διαμορφωθούν σε ακολουθίες παρατηρήσεων με βάση τα emission symbols που προέκυψαν θα χρησιμοποιηθούν για την εκπαίδευση των δύο Hidden Markov Models.

Για την επεξεργασία των κειμένων έγινε χρήση της εργαλειοθήκης TMG για την οποία έχει γίνει αναφορά στο κεφάλαιο της διπλωματικής. Η TMG χρησιμοποιήθηκε για αναπαράσταση των κειμένων σύμφωνα με το Vector Space Model με παραμέτρους επεξεργασίας κειμένου τις εξής :

1)Αφαίρεση λέξεων με μήκος 1 και η αφαίρεση λέξεων με μήκος μεγαλύτερο του 30.
2)Stemming.
3)Αφαίρεση αλφαριθμητικών και αριθμών.

4)Αφαίρεση stopwords.
5)Τα local weights του term-document πίνακα είναι ή συχνότητα εμφάνισης των λέξεων (global weights δεν χρησιμοποιούνται).

Η αφαίρεση stopwords είναι σημαντική καθώς η λίστα με τις stopwords αποτελείται από λέξεις οι οποίες εμφανίζονται πολύ συχνά σε όλα τα κείμενα ανεξαρτήτως του σημασιολογικού τους περιεχομένου (για παράδειγμα αφαιρούνται λέξεις όπως: about, above, after, for, and, between και πολλές άλλες). Οι λέξεις αυτές πρέπει να αφαιρεθούν καθώς η ύπαρξή τους δημιουργεί την ψευδαίσθηση της ομοιότητας των κειμένων, λόγω της μεγάλης συχνότητας με την οποία εμφανίζονται σε όλα τα κείμενα. Η αφαίρεση αριθμών και αλφαριθμητικών έγινε διότι και αυτές οι ομάδες λέξεων εμφανίζονται συχνά και στις δύο κατηγορίες κειμένων. Το stemming χρειάζεται έτσι ώστε λέξεις που έχουν την ίδια λεκτική ρίζα να αναγνωρίζονται ως μια λέξη με αποτέλεσμα να δίνεται μεγαλύτερο βάρος σε αυτές.

Το αποτέλεσμα αυτής της επεξεργασίας με το TMG θα δημιουργήσει τον πίνακα A (document-term) και τον πίνακα dictionary ο οποίος περιέχει όλες τις λέξεις που δεν έχουν κοπεί από τους παραπάνω περιορισμούς και δεν έχουν υποβαθμιστεί στην ρίζα τους από το stemming.

Για τα 800 κείμενα που αναφέρονται σε medical devices κάνω indexing με αυτές τις παραμέτρους. Κάθε γραμμή του πίνακα A αναφέρεται σε μία λέξη η οποία βρίσκεται στην γραμμή με ίδιο δείκτη του πίνακα dictionary (πχ η γραμμή κ του πίνακα A είναι η λέξη που βρίσκεται στην γραμμή κ του πίνακα dictionary). Σε δεύτερο βήμα υπολόγισα τις συνολικές εμφανίσεις κάθε λέξης στο σύνολο των 800 κειμένων και τις αποθήκευσα σε έναν πίνακα έστω F. Στην συνέχεια έχοντας τις συχνότητες εμφάνισης κάθε μοναδικής λέξης έτρεξα έναν αλγόριθμο ταξινόμησης σε αύξουσα σειρά ως προς την συχνότητα εμφάνισης κάθε λέξης το πίνακα F, ο οποίος , επιπλέον, θα μεταβάλει κατα αντιστοιχία τον πίνακα dictionary, ώστε η λέξη που βρίσκεται στην γραμμή ν του πίνακα dictionary να έχει το σύνολο εμφανίσεων της στην γραμμή ν του πίνακα F.

Έπειτα έφτιαξα έναν αλγόριθμο ο οποίος θα αρχίζει από το πρώτο στοιχείο του πίνακα F δηλαδή την λέξη με τις λιγότερες εμφανίσεις στο σύνολο των 800 κειμένων και στο επόμενο βήμα θα προχωράει στην επόμενη λέξη, σε κάθε βήμα θα υπολογίζει το άθροισμα των εμφανίσεων όλων των λέξεων μέχρι το παρόν βήμα προς το σύνολο όλων των εμφανίσεων όλων των λέξεων και των 800 κειμένων. Αυτό θα μας δίνει ένα ποσοστό κάθε φορά. Αυτός ο αλγόριθμος κατασκευάστηκε για να χωρίσουμε τις λέξεις σε 5 ομάδες ανάλογα με το πόσο συχνά εμφανίζονται σε όλα τα κείμενα των medical devices (ο αριθμός των ομάδων δεν είναι απαραίτητο να είναι 5, μπορεί να χρησιμοποιηθούν και άλλοι διαχωρισμοί).

Πιο αναλυτικά ο αλγόριθμος κάνει το εξής:

Έστω ότι σε μια ανάλυση με M πλήθος κειμένων έχουμε U διαφορετικές λέξεις που εμφανίζονται συνολικά S φορές (το άθροισμα λέξεων των M κειμένων).
Ας θεωρήσουμε ότι βρισκόμαστε στο βήμα N του αλγόριθμου (1<=N<=U)
ο αλγόριθμος θα υπολογίσει το ποσοστό επί των συνολικών λέξεων που αντιπροσωπεύουν οι N λιγότερο σημαντικές λέξεις. Δηλαδή αν η πρώτη λέξη εμφανίστηκε E_1 φορές η δεύτερη E_2 και η N-οστή E_N φορές το ποσοστό θα είναι

$$\Pi_N = (E_1 + E_2 + \ldots + E_N)/S$$

Μια από τις 5 ομάδες (και η ομάδα που αποτελείται από τις περισσότερες λέξεις) θα περιέχει τις λέξεις 'σκουπίδια'. Η ομάδα με τις λέξεις σκουπίδια μπορεί να χαρακτηριστεί ως το σύνολο διαφορετικών λέξεων που έχουν τόσο μικρή πιθανότητα εμφάνισης που για να προσδιορισθεί με μικρή απόκλιση θέλει πολύ μεγάλο αριθμό κειμένων για επεξεργασία. Στην πορεία θα παρουσιαστούν πειραματικά αποτελέσματα επομένως θα γίνει πιο κατανοητή αυτή η έννοια.

Μπορεί ο επακριβής προσδιορισμός αυτών των λέξεων ('σκουπίδια') είναι προφανώς αδύνατος διότι ο αριθμός των κειμένων που χρησιμοποιούνται για αυτήν την ανάλυση είναι πεπερασμένος, αλλά όπως θα δούμε στην πορεία μπορούμε να αποκτήσουμε μια καλή προσέγγιση.

Τρέχοντας τον αλγόριθμό που υπολογίζει τα ποσοστά εμφάνισης ομάδων λέξεων και αρχίζοντας από τις λιγότερο συχνές λέξεις υπολογίζουμε κάθε φορά πόσες λέξεις ανήκουν στο top x% (π.χ. top 70% αν Π_N = 30% περιέχει όλες τις λέξεις εκτός από τις N λιγότερο πιθανές) για δύο περιπτώσεις :

Η μία περίπτωση είναι για τα 800 κείμενα που έχουμε επεξεργαστεί και η άλλη είναι για όλα τα κείμενα που έχουμε, δηλαδή τα 2000 που αναφέρονται σε medical devices.

Αυτό που θέλουμε να βρούμε είναι ένα σύνολο λέξεων που να αντιπροσωπεύουν ένα top x% που να μην αλλάζει πολύ από την περίπτωση με τα 800 κείμενα ,στην περίπτωση με τα 2000 κείμενα. Αξίζει να σημειωθεί ότι δεν πρέπει να δοθεί μεγάλη σημασία στον προσδιορισμό του ελάχιστου δυνατού ποσοστού καθώς τα Hidden Markov Models είναι στατιστικά μοντέλα και κάθε ακολουθία παρατηρήσεων απαρτίζεται από πολλές λέξεις επομένως αν για παράδειγμα το top x % αλλάζει κατά 5% περιμένουμε το μοντέλο να μην έχει σημαντική διαφορά με ένα μοντέλο που έχει στηριχθεί σε μια ομάδα λέξεων που αλλάζει πχ κατά 8% ή 1%.

Στατιστικά στοιχεία από την ανάλυση των κειμένων:
τρέχοντας τον παραπάνω αλγόριθμο υπολόγισα τον αριθμό λέξεων που βρίσκονται στο top x% για διάφορα x, τα αποτελέσματα φαίνονται παρακάτω.

διαφορετικές λέξεις για 800 κείμενα = 6006
διαφορετικές λέξεις για 2000 κείμενα = 9279 , Νέες λέξεις=3273

Ποσοστό λέξεων που κρατάμε	800 κείμενα	2000 κείμενα	Ποσοστιαία διαφορά
Top 13/14	2201	2585	17,40%
Top 10/11	1800	2259	14,38%
Top 8/9	1475	1657	12,33%
Top 7/8	1287	1435	11,49%
Top 6/7	1086	1201	10,58%
Top 4/5	636	680	6,91%

Για την τελευταία περίπτωση βλέπουμε ότι από τις 3273 νέες λέξεις μόνο οι 44

(680 – 636) έχουν εισχωρήσει στο top 80%.Οι υπόλοιπες 3229 λέξεις ή το 98,65% των νέων λέξεων έχει προσαρτηθεί στο bottom 20%.

Η ίδια διαδικασία με τις ίδιες ακριβώς παραμέτρους προεπεξεργασίας με το TMG πραγματοποιήθηκε και για κείμενα της κατηγορίας other στην περίπτωση των 800 κειμένων και στην περίπτωση των 1263.

για 800 κείμενα έχω 7.497 διαφορετικές λέξεις.
για 1.263 κείμενα έχω 9.718 διαφορετικές λέξεις και 2.221 νέες λέξεις.

το top 80% για τα 800 κείμενα είναι 546 λέξεις.
το top 80% για τα 1.263 κείμενα είναι 575 λέξεις.
η ποσοστιαία διαφορά τους είναι 5,31% και από τις 2.221 νέες λέξεις οι 2.192 ή το 98,69% των νέων λέξεων έχει εισχωρήσει στο bottom 20%.

Στην κατηγορία other δυστυχώς υπάρχουν λιγότερα κείμενα παρ' όλα αυτά με αυτό το πείραμα θέλουμε να βρούμε στο μέγιστο πλήθος κειμένων που έχουμε αν παρατηρείται παρόμοια συμπεριφορά για την ίδια ομάδα λέξεων. Δεν μας ενδιαφέρει ακριβώς να προσδιορίσουμε το ποσοστό στο οποίο μεγιστοποιείται αυτή η συμπεριφορά αλλά να γνωρίζουμε ότι για αυτήν την ομάδα (“σκουπίδια”) που επιλέξαμε ένα αρκετά μεγάλο ποσοστό αυτών των λέξεων είναι άνευ σημασίας για την κατηγοριοποίηση.

Σύμφωνα με τα παραπάνω αποφάσισα να θεωρήσω την ομάδα 5, δηλαδή την ομάδα με τις λέξεις “σκουπίδια” τις N λέξεις με $\Pi_N = 20\%$. Οι λόγοι που αποφάσισα να δημιουργηθεί μια τέτοια ομάδα είναι :

1)Μειώνει τον χρόνο της μετέπειτα επεξεργασίας των κειμένων κατά πολύ, καθώς χρησιμοποιούμε πολύ λιγότερες λέξεις.

2)Φιλτράρισμα λέξεων με πάρα πολύ μικρές πιθανότητες εμφάνισης, άρα λέξεων που

δεν είναι χαρακτηριστικές της κατηγορίας κειμένου.

Πρέπει να γίνει η παρατήρηση ότι, εφόσον αυτές οι λέξεις εμφανίζονται σπάνια ακόμα και να τις είχα συμπεριλάβει για την εκπαίδευση (σε μετέπειτα βήματα) του μοντέλου θα επηρέαζαν πολύ λίγο τα αποτελέσματα. Παρ' όλα αυτα ο χρόνος επεξεργασίας θα αυξανόταν σημαντικά.

Έχοντας ορίσει την ομάδα 5 οι υπόλοιπες ομάδες χωρίζονται αυθαίρετα με τον παρακάτω τρόπο

	Π_N >=	Π_M <=
Ομάδα 5	0	20
Ομάδα 4	20	40
Ομάδα 3	40	60
Ομάδα 2	60	80
Ομάδα 1	80	100

Κάθε ομάδα απαρτίζεται από τις λέξεις με δείκτες U_I με $N<=I<=M$
(Τα N και M συμβολίζουν τα κάτω και άνω όρια για την αθροιστική συνάρτηση συχνοτήτων των λέξεων και είναι ανεξάρτητα για κάθε κατηγορία.)

Οπού U_t είναι το t στοιχείο του πίνακα dictionary που περιέχει τις μοναδικές λέξεις του συνόλου των κειμένων που επεξεργάστηκε το TMG.

Αφού δημιουργήσω αυτές τις 4 ομάδες λέξεων και για τις δύο κατηγορίες κειμένων το επόμενο βήμα είναι από αυτές τις ομάδες να δημιουργήσω τα emission symbols των καταστάσεων για τα Hidden Markov Models με τον εξής τρόπο. Όπως έχω προαναφέρει θέλω οι ακολουθίες των παρατηρήσεων που θα αντιστοιχιστούν σε κάθε κείμενο να είναι χαρακτηριστικές της κατηγορίας στην οποία ανήκει το κείμενο. Για να κάνω αυτόν τον έλεγχο δηλαδή αν κάποια λέξη εμφανίζεται πιο συχνά στην μια η στην άλλη κατηγορία κάνω το εξής. Για κάθε λέξη που ανήκει σε μια από τις 4 ομάδες της μιας κατηγορίας κοιτάζω αν υπάρχει η ίδια λέξη σε κάποια ομάδα της

άλλης κατηγορίας,δηλαδή αν η λέξη U_1 ανήκει στην ομάδα 1 (top 20%) για τα medical devices ενώ ανήκει στην ομάδα 2(top 40%-top 20%) για τα other θα αντιστοιχιστεί σε ένα emission symbol έστω Em_1 . Αν ή λέξη άνηκε στην ομάδα 2 για τα medical devices και στην ομάδα 3 για τα other θα αντιστοιχιζόταν πάλι στο emission symbol Em_1 . Αν η λέξη άνηκε στην ομάδα 1 για τα medical devices και δεν άνηκε σε καμία ομάδα από τα other θα είχε άλλο emission symbol.

Ο τρόπος υπολογισμού των emission symbols φαίνεται στον επόμενο πίνακα:

Ομάδες για Medical Devices	Ομάδες για other	Emission symbols
ομάδα$_K$	ομάδα$_K$	Η λέξη δεν προκαλεί emission symbol
ομάδα$_K$	ομάδα$_{K-1}$	8
ομάδα$_K$	ομάδα$_{K-2}$	9
ομάδα$_K$	ομάδα$_{K-3}$	10
ομάδα1	Όχι ομάδα	11
ομάδα$_{K-1}$	ομάδα$_K$	5
ομάδα$_{K-2}$	ομάδα$_K$	4
ομάδα$_{K-3}$	ομάδα$_K$	3
Όχι ομάδα	ομάδα1	2
Όχι ομάδα	Όχι ομάδα	Η λέξη δεν προκαλεί emission symbol

Ακολουθώντας αυτήν την συλλογιστική, πετυχαίνω συγκεκριμένα σύμβολα να είναι πιο συχνά σε κάποια κατηγορία δηλαδή τα 8, 9, 10, 11 είναι πιο συχνά στην κλάση medical devices ενώ τα 2 ,3 ,4 και 5 είναι πιο συχνά στην κλάση other. Αν η λέξη ανήκει στην ίδια ομάδα τότε σημαίνει ότι για medical devices και other η λέξη εμφανίζεται με την ίδια συχνότητα άρα δεν μας δίνει σημαντική πληροφορία και άρα δεν μεταφράζεται σε emission symbol.

Το αποτέλεσμα αυτής της διαδικασίας είναι η κατασκευή ενός λεξικού αντιστοίχησης λέξεων σε emission symbols σύμφωνα με τον παραπάνω πίνακα. Αν κάποια λέξη δεν υπάρχει σε αυτό το λεξικό τότε απλά αγνοείται.

4.3 ΔΙΑΔΙΚΑΣΙΑ ΕΚΠΑΙΔΕΥΣΗΣ ΤΩΝ HIDDEN MARKOV MODELS

Έχοντας καθορίσει τώρα ποιες λέξεις θα κρατάμε και σε τι σύμβολο θα αντιστοιχούν μπορούμε να προχωρήσουμε στην εκπαίδευση του Hidden Markov Model. Για την εκπαίδευση του μοντέλου χρησιμοποίησα την πρώτης φόρα 40 κείμενα της κατηγορίας medical devices και 40 κείμενα της κατηγορίας other για να εκπαιδεύσω 2 Hidden Markov Models όπου το ένα θα αντιστοιχεί στα medical devices και το άλλο στα other. Η διαδικασία για την εκπαίδευση είναι ίδια και για τις δύο κατηγορίες οπότε αρκεί να αναλύσω μόνο για την κατηγορία των medical devices.

Αρχικά πήρα 40 κείμενα από την κατηγορία των medical devices, τα οποία δεν είχαν χρησιμοποιηθεί σε προγενέστερο βήμα και έτρεξα το TMG με τις ίδιες ρυθμίσεις που είχαν χρησιμοποιηθεί για τα 800 κείμενα. Στην συνέχεια διέσχισα κάθετα το term-document πίνακα για να δω ποιες λέξεις εμφανίζονται σε κάθε κείμενο και να δημιουργήσω μια ακολουθία για κάθε ένα από τα κείμενα σύμφωνα με το αν η λέξη αντιστοιχίζεται σε κάποιο emission symbol και με τον αριθμό εμφανίσεων της στο κείμενο. Το αποτέλεσμα αυτής της διαδικασίας ήταν να δημιουργηθεί ένας πίνακας με 40 γραμμές και μεγάλο αριθμό στηλών.

Σε αυτό το σημείο πρέπει να αναφερθεί κάτι σχετικά με την εκπαίδευση των Hidden Markov Model. Αν έχω μια σειρά από παρατηρήσεις $O_1O_2 \dots O_T$ και άλλη μια σειρά $O_1O_2 \dots O_K$ με $K>>T$ που αντιστοιχούν σε δύο κείμενα οπού το δεύτερο κείμενο είναι αρκετά μεγαλύτερο από το πρώτο, τότε το δεύτερο κείμενο θα επηρεάσει αρκετά περισσότερο τις τελικές παραμέτρους του μοντέλου.

Για να αντιμετωπίσω αυτό το πρόβλημα προσπάθησα να επιλέξω κείμενα με παρόμοιο μέγεθος έτσι ώστε να αποφύγω την ύπαρξη κειμένων τα οποία θα μονοπωλούν την εκπαίδευση του Hidden Markov Model με ενδεχόμενο να μειώσουν την απόδοση του μοντέλου. Ένας λόγος παραπάνω για την επιλογή αυτή είναι, ότι η

συνάρτηση εκπαίδευσης του μοντέλου που παρέχεται από την εργαλειοθήκη HMM MURPHY TOOLBOX, η οποία και χρησιμοποιήθηκε για την ανάπτυξη του μοντέλου, δέχεται αναγκαστικά ως όρισμα ακολουθίες ιδίου μήκους. Η προσαρμογή των κειμένων στο ίδιο μήκος έγινε με δύο τρόπους :

Α)Αφαίρεσα παρατηρήσεις από τις ακολουθίες με μεγαλύτερο μήκος προσθέτοντας αυτές τις παρατηρήσεις σε ακολουθίες με μικρότερο μήκος έτσι ώστε όλες οι ακολουθίες να έχουν το ίδιο μήκος. Με αυτόν τον τρόπο δεν χάνονται παρατηρήσεις αλλά μπορεί να αλλάζει η μορφή των ακολουθιών κάποιων κειμένων. Έχοντας όμως ως δεδομένο ότι την μεγαλύτερη σημασία στην κατηγοριοποίηση των κειμένων παίζουν οι συχνότητες εμφάνισης παρατηρήσεων δεν θεωρώ ότι αυτός ο τρόπος προκαλεί ζημιά στο μοντέλο.

Β)Σύμφωνα με την δεύτερη προσέγγιση βρήκα την μικρότερη ακολουθία και έκοψα από όλες τις υπόλοιπες τα παραπάνω στοιχεία. Τα κείμενα έχουν επιλεγεί με τέτοιο τρόπο ώστε να έχουν περίπου ίδιο αριθμό λέξεων το οποίο όμως δεν συνεπάγεται ότι θα δημιουργήσουν και περίπου ίδιο αριθμό παρατηρήσεων. Με αυτόν τον τρόπο δεν αλλάζει η δομή της ακολουθίας κάθε κειμένου αλλά γίνεται πιο μικρή. Καθώς η συγκεκριμένη μοντελοποίηση βασίζεται καθαρά σε στατιστικά στοιχεία περιμένω αυτός ο τρόπος να έχει χειρότερα αποτελέσματα. Το πείραμα έγινε με 100 κείμενα της κατηγορίας medical devices και 100 κείμενα της κατηγορίας other.

Επειδή οι πίνακες με τις πιθανότητες εκπομπής παρατηρήσεων προέκυψε να έχουν μηδενικές πιθανότητες εμφάνισης για τα λιγότερο πιθανά σύμβολα τόσο για τα κείμενα της κατηγορίας medical devices όσο και για τα 'other' κείμενα, πρόσθεσα μια μικρή σταθερά στα στοιχεία του πίνακα. Στην συνέχεια τον κανονικοποίησα ώστε το άθροισμα όλων των πιθανοτήτων των παρατηρήσεων κάθε κατάστασης να κάνει 1. Ο λόγος που είναι σκόπιμο να γίνει κάτι τέτοιο έγκειται στο γεγονός ότι κείμενα που μπορεί να τύχει να εμφανίσουν αυτό το σύμβολο και να είναι της άλλης κατηγορίας δεν θα μπορέσουν να κατηγοριοποιηθούν σωστά. Η υπόθεση που γίνεται είναι ότι προσθέτοντας μια πολύ μικρή σταθερά στα στοιχεία του πίνακα των πιθανοτήτων για

τις παρατηρήσεις κάθε κατάστασης, δεν θα επηρεαστεί πολύ το log-likehood του μοντέλου για μια ακολουθία παρατηρήσεων με αποτέλεσμα να βελτιωθεί η δυνατότητα αναγνώρισης κειμένων. Αυτή η υπόθεση όμως θα εξεταστεί στην συνέχεια. Η σταθερά που προστίθεται αν δεν λέγεται κάτι άλλο είναι 10^{-7} πριν την κανονικοποίηση κατά γραμμή του πίνακα (δηλαδή στην matlab B=B + 10^{-7} , και μετά κανονικοποιώ κατά γραμμή έτσι ώστε το άθροισμα κάθε γραμμής να ισούται με 1, όπου B = Πίνακας με τις πιθανότητες εμφάνισης παρατηρήσεων).

4.4 ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΓΙΑ ΤΗΝ ΕΡΜΗΝΕΙΑ ΤΩΝ ΠΕΙΡΑΜΑΤΙΚΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Πριν παρουσιάσω τα αποτελέσματα πρέπει να εισάγω κάποιες ορολογίες οι οποίες χρησιμεύουν στην αξιολόγηση των αποτελεσμάτων του μοντέλου ταξινόμησης και εμφανίζονται στην συνέχεια. Τα στοιχεία που μας ενδιαφέρουν εκτός από τα κείμενα σε απόλυτο αριθμό που κατηγοριοποιήθηκαν σωστά είναι το precision και το recall.

Το recall για κάθε κατηγορία ορίζεται ως N_R/N_O όπου το N_O είναι το σύνολο των κειμένων αυτής της κατηγορίας και N_R είναι τα κείμενα αυτής της κατηγορίας που αναγνωρίστηκαν να ανήκουν σε αυτήν. Το recall είναι ανεξάρτητο του αριθμού κειμένων που χρησιμοποιήθηκαν σε κάθε κατηγορία δηλαδή παραμένει το ίδιο αν ο λόγος του πλήθους των κειμένων της μιας κατηγορίας προς το πλήθος των κειμένων της άλλης κατηγορίας είναι 1:1 είτε M:1 όπου M ένας θετικός πραγματικός αριθμός.

Το precision για κάθε κατηγορία ορίζεται ως N^+/N όπου N είναι όλα τα κείμενα που το μοντέλο αναγνώρισε να ανήκουν στην κατηγορία και N^+ είναι όσα από αυτά πραγματικά ανήκουν. Πρέπει να δοθεί μεγάλη προσοχή σε αυτό το σημείο διότι το precision εξαρτάται από τους λόγους του αριθμού κειμένων σε κάθε κατηγορία. Διαφορετικοί λόγοι μας δίνουν και διαφορετικά αποτελέσματα στο precision κάτι που δεν θέλουμε.

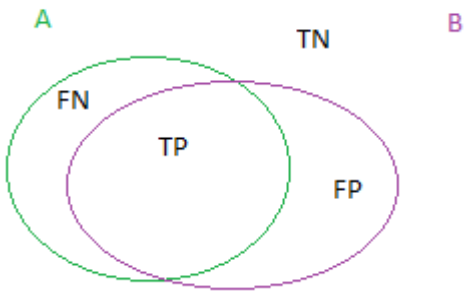
Στο πείραμα για την αξιολόγηση του ταξινομητή έχω χρησιμοποιήσει 658 κείμενα από την κατηγορία medical devices και 304 κείμενα από την κατηγορία other. Στο

σύνολο δεδομένων που διαθέτω έχω 2000 κείμενα που υπάγονται στην κατηγορία των medical devices και 1263 της κατηγορίας other, συνεπώς ο λόγος τους είναι 61,3 : 38,7. Άρα για τον υπολογισμό του precision θα πρέπει να χρησιμοποιηθεί ο λόγος 61,3 : 38,7 και όχι ο λόγος 658:304 . Η αιτία που έγινε αυτό (να επιλέξω στο πείραμα 658 και 304 κείμενα) είναι για να χρησιμοποιήσω όσο περισσότερα κείμενα μπορώ από κάθε κατηγορία ώστε να έχω μεγαλύτερο δείγμα για τα πειράματα. Αυτό όμως δεν είναι πρόβλημα καθώς το precision ανηγμένο στον σωστό λόγο υπολογίζεται από τον νόμο του Bayes. Σε όλα τα πειράματα positive θεωρείται το medical device και negative το other.

Ο νόμος τους Bayes είναι ο εξής:
αν έχουμε ένα γεγονός A και ένα γεγονός B η πιθανότητα του A δεδομένου του B υπολογίζεται ως :

$$P(A \mid B) = P(B \mid A)P(A)/[P(B \mid A)P(A) + P(B \mid -A)P(-A)]$$

Έτσι αν θεωρήσουμε ότι το γεγονός B είναι να έχουμε positive (δηλαδή το κείμενο να καταχωρείται ως medical device) και το γεγονός A το κείμενο να ανήκει στην κατηγορία των medical devices. Το $P(A) = Prevalence^{+}$ και το $P(B| A) = Recall^{+}$ όπου $Recall^{+}$ είναι το recall για τα κείμενα της κατηγορίας των medical devices και $Prevalence^{+}$ είναι η πιθανότητα εμφάνισης κειμένου κατηγορίας medical device στα συνολικά κείμενα μας. Επειδή ο αναμενόμενος λόγος από το σύνολο όλων των κειμένων είναι 61,3 : 38,7 η πιθανότητα εμφάνισης ($prevalence^{+}$) είναι 61,3%. Το $P(B \mid -A)$ είναι $1-Recall^{-}$ όπου $Recall^{-}$ είναι το recall για την κατηγορία other, το P(-A) είναι προφανώς 38,7%. Παρακάτω παραθέτω και ένα ενδεικτικό διάγραμμα Venn για τα ενδεχόμενα A και B.



Στους επόμενους πίνακες αποτελεσμάτων οι συμβολισμοί σημαίνουν τα εξής :

N = Αριθμός καταστάσεων για το Hidden Markov Model.

TP = Σωστά κατανεμημένα medical devices (true positives).

FP = Λάθος κατανεμημένα medical devices (false positives).

TN = Σωστά κατανεμημένα other (true negatives).

FN = Λάθος κατανεμημένα other (false negatives).

PrecD = Το ποσοστό των true positives προς το σύνολο όλων των positives ή αλλιώς precision του μοντέλου ως προς τα κείμενα που αναφέρονται σε medical devices για την σωστή κατανομή (61,3:38,7).

PrecO = Το ποσοστό των true negatives προς το σύνολο όλων των negatives ή αλλιώς precision του μοντέλου ως προς τα κείμενα που αναφέρονται σε other για την σωστή κατανομή.

RecO = Recall για τα other κείμενα.

RecD = Recall για τα medical devices κείμενα.

4.5 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.5.1 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ Α ΤΡΟΠΟ

Ο επόμενος πίνακας δείχνει τα πειραματικά αποτελέσματα για 658 κείμενα από medical devices και 304 κείμενα από την κατηγορία other, για Hidden Markov Models με διαφορετικό αριθμό καταστάσεων, για 40 κείμενα εκπαίδευσης.

N	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
1	644	14	1	303	99,79 %	98,67%	97,72%	99,67 %
3	645	13	1	303	99,79%	98,76%	98,02%	99,67 %
6	647	11	1	303	99,79 %	98,95%	98,32%	99,67 %
8	644	14	1	303	99,79 %	98,67%	97,72%	99,67 %
10	646	12	1	303	99,79%	98,85%	98,17%	99,67 %
50	645	13	1	303	99,79 %	98,76%	98,02%	99,67 %
100	642	16	1	303	99,79 %	98,48%	97,56%	99,67 %

Όπως παρατηρούμε στο συγκεκριμένο πείραμα η επιλογή καταστάσεων δεν παίζει σημαντικό ρόλο στην τελική ευστοχία του μοντέλου. Παρατηρείται μια μικρή πτώση στο recall για τα medical devices στο πείραμα με τις 100 καταστάσεις. Δυστυχώς η υπολογιστική ισχύς που απαιτείται για την εκπαίδευση μοντέλων με περισσότερες καταστάσεις με εμπόδισε από το να κάνω περισσότερα πειράματα.

Βλέπουμε επίσης ότι το precision για τα medical devices (PrecD) είναι αρκετά υψηλό και είναι ο δείκτης που μας ενδιαφέρει περισσότερο, καθώς το ο σκοπός του κατηγοριοποιητή είναι να εντοπίζει τα κείμενα που αναφέρονται σε medical device για να τα επεξεργαστούμε κάνοντας τους text segmentation το οποίο θα αναπτυχθεί

στο επόμενο κεφάλαιο της διπλωματικής.

Ο παρακάτω πίνακας δείχνει τα πειραματικά αποτελέσματα για 658 κείμενα από medical devices και 304 κείμενα από την κατηγορία other, για Hidden Markov Models με διαφορετικό αριθμό καταστάσεων, για 100 κείμενα εκπαίδευσης.

N	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
1	645	13	1	303	99,79 %	98,76%	98,02%	99,67 %
3	645	13	1	303	99,79%	98,76%	98,02%	99,67 %
6	646	12	1	303	99,79 %	98,85%	98,17%	99,67 %
8	646	12	1	303	99,79 %	98,85%	98,17%	99,67 %
10	648	12	1	303	99,79%	99,05%	98,48%	99,67 %
50	651	7	1	303	99,79 %	99,26%	98,93%	99,67 %
100	651	7	1	303	99,79 %	99,26%	98,93%	99,67 %

Χωρίς πρόσθεση σταθεράς

N	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
1	658	0	2	302	99,59 %	100%	100%	99,34 %
3	658	0	2	302	99,59 %	100%	100%	99,34 %
6	658	0	2	302	99,59 %	100%	100%	99,34 %
8	658	0	2	302	99,59 %	100%	100%	99,34 %
10	658	0	2	302	99,59 %	100%	100%	99,34 %
50	658	0	2	302	99,59 %	100%	100.00%	99,34 %
100	658	0	2	302	99,59 %	100%	100.00%	99,34 %

Παρατηρώ ότι χωρίς την πρόσθεση σταθεράς στον πίνακα πιθανοτήτων των παρατηρήσεων όλων των καταστάσεων, στα 658 κείμενα για medical devices δεν υπήρχε κάποιο κείμενο που δεν κατατάχθηκε σωστά, όμως δύο από τα κείμενα της κατηγορίας other αναγνωρίστηκαν σαν medical devices. Παρ' όλα αυτά ο δείκτης που μας ενδιαφέρει δηλαδή το precision για τα medical devices έχει μειωθεί έτσι κρίνω σκόπιμο να πραγματοποιήσω ένα πείραμα προσθέτοντας διαφορετικές σταθερές στον αρχικό πίνακα.

Για Hidden Markov Model μιας κατάστασης εξετάστηκαν κανονικοποιήσεις του πίνακα πιθανοτήτων παρατηρήσεων των καταστάσεων για τις τιμές που φαίνονται στην πρώτη στήλη του επόμενου πίνακα και παρουσιάζονται τα πειραματικά αποτελέσματα.

S	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
10^{-7}	645	13	1	303	99,79 %	98,76%	98,02%	99,67 %
10^{-14}	653	5	1	303	99,79 %	99,5%	99,2%	99,67 %
10^{-20}	657	1	1	303	99,79 %	99,9%	99,84%	99,67 %
10^{-25}	657	1	2	302	99,58 %	99,9%	99,84%	99,34 %

Βλέπουμε ότι για σταθερά 10^{-20} έχουμε τους καλύτερους δείκτες δηλαδή υψηλό recall που σημαίνει ότι θα χάσουμε λίγα κείμενα και το ίδιο καλό precision που σημαίνει ότι θα έχουμε πολύ λίγο 'θόρυβο' στα δεδομένα που θέλουμε να επεξεργαστούμε (κείμενα που αναφέρονται αποκλειστικά σε medical devices).

4.5.2 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ Β ΤΡΟΠΟ

Ο επόμενος πίνακας δείχνει τα πειραματικά αποτελέσματα για 658 κείμενα από medical devices και 304 κείμενα από την κατηγορία other, για Hidden Markov Models με διαφορετικό αριθμό καταστάσεων.

N	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
1	643	15	1	303	99,79 %	98,58%	97,72%	99,67 %
3	644	14	1	303	99,79 %	98,67%	97,87%	99,67 %
6	643	15	1	303	99,79 %	98,58%	97,72%	99,67 %
8	643	15	1	303	99,79 %	98,58%	97,72%	99,67 %
10	642	16	1	303	99,79 %	98,48%	97,57%	99,67 %
50	646	12	1	303	99,79 %	98,85%	98,17%	99,67 %
100	643	15	1	303	99,79 %	98,58%	97,72%	99,67 %

Παρατηρούμε πάλι ότι δεν παίζει τόσο μεγάλο ρόλο η επιλογή αριθμού καταστάσεων για το συγκεκριμένο πρόβλημα. Τα αποτελέσματα δεν είναι πολύ χειρότερα από τον τρόπο Α. Όμως πρέπει να σημειωθεί ότι τόσο στην περίπτωση των 40 κειμένων όσο και στην περίπτωση των 100 κειμένων, ο τρόπος Α δίνει καλύτερα αποτελέσματα.

Για μια κατάσταση και διαφορετικές σταθερές κανονικοποίησης:

S	TP	FP	FN	TN	PrecD	PrecO	RecD	RecO
10^{-7}	643	15	1	303	99,79 %	98,58%	97,72%	99,67 %
10^{-14}	644	14	1	303	99,79 %	98,67%	97,87%	99,67 %
10^{-20}	646	12	1	302	99,58 %	98,85%	98,17%	99,34 %
10^{-25}	646	12	2	302	99,58 %	98,85%	98,17%	99,34 %
0	649	9	8	296	98,34 %	99,12%	98,63%	97,37 %

Παρατηρείται η ίδια συμπεριφορά για την σταθερά κανονικοποίησης όπως και στο πείραμα Α με την διαφορά ότι το μοντέλο κάνει περισσότερα λάθη χωρίς σταθερά κανονικοποίησης.

4.6 ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΣΚΕΨΕΙΣ ΓΙΑ ΒΕΛΤΙΩΣΗ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ

Λόγω του ότι ο τρόπος προσέγγισης του προβλήματος είναι με στατιστικούς δείκτες (αυτή την συμπεριφορά έχουν οι παρατηρήσεις/emission symbols) η καλύτερη συμπεριφορά του τρόπου Α αποδεικνύει ότι είναι προτιμότερο να κρατάμε περισσότερες παρατηρήσεις για την εκπαίδευση ακόμα και αν αυτές χαλάσουν την σειρά κάποιων ακολουθιών, παρά να επιλέξουμε να διατηρήσουμε την σειρά των ακολουθιών χωρίς να μολυνθεί από τις παρατηρήσεις άλλων κείμενων εκπαίδευσης αλλά να χάσουμε σε πλήθος παρατηρήσεων.

Πρέπει να σημειωθεί ότι εφόσον τα δεδομένα λαμβάνονται από τον term-document πίνακα του TMG και όχι από την ανάγνωση λέξεων με την σειρά από τα κείμενα. Κάποια patterns που μπορεί να υπάρχουν στο αρχικό κείμενο χάνονται. Αυτό όμως δεν σημαίνει ότι δεν υπάρχουν patterns διότι ο τρόπος που το TMG βάζει τις λέξεις είναι συγκεκριμένος (όχι τυχαίος) και άρα θα μπορούσε να εξαχθεί κάποιος κανόνας από την σειρά εμφάνισης κάποιων λέξεων.

Σε περίπτωση που μετά την εκπαίδευση υπάρχουν κάποιες παρατηρήσεις με μηδενική πιθανότητα εμφάνισης σε όλο το μοντέλο καλό είναι να εξετάζεται η πρόσθεση μιας πολύ μικρής σταθεράς. Εκτός του ότι μπορούμε να αποφύγουμε περιπτώσεις κείμενων που δεν μπορούν να κατηγοριοποιηθούν αν και τα δύο μοντέλα έχουν μηδενική πιθανότητα να εμφανίσουν την ακολουθία παρατηρήσεων (αυτό βέβαια είναι πάρα πολύ σπάνιο) η προσθήκη μιας μικρής σταθεράς στις πιθανότητες εμφάνισης των παρατηρήσεων μπορεί επιπλέον να βελτιώσει το μοντέλο.

Επιπλέον ενδέχεται να πρέπει να εξεταστούν κατηγοριοποιητές οι οποίοι υλοποιούν την κατηγοριοποίηση σε 2 (μπορεί και παραπάνω στάδια). Με τα επόμενα στάδια να εκπαιδεύονται περισσότερο με κείμενα τα οποία τα προηγούμενα στάδια δεν μπόρεσαν ταξινομήσουν σωστά.

Έχοντας πλέον έναν αυτοματοποιημένο τρόπο αναγνώρισης κειμένων που αναφέρονται σε ανακλήσεις ιατρικών συσκευών στα επόμενα δύο κεφάλαια θα

αναπτύξουμε μεθόδους χρησιμοποιώντας HMM, με τις οποίες θα προσπαθήσουμε να αναγνωρίσουμε τμήματα αυτών των κειμένων, που μας ενδιαφέρουν (όπως οι χώρες διανομής της συσκευής, ο λόγος ανάκλησης κ.α.).

ΚΕΦΑΛΑΙΟ 5

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΣΕ ΙΑΤΡΙΚΑ ΚΕΙΜΕΝΑ ΜΕ ΧΡΗΣΗ HIDDEN MARKOV MODELS ΜΕ ΣΤΟΧΟ ΤΗΝ ΚΑΤΑΤΜΗΣΗ ΤΟΥΣ.

5.1.1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ.

Το κυρίως αντικείμενο αυτής της διπλωματικής ήταν η δοκιμή των Hidden Markov Models στην κατάτμηση κειμένου σε συγκεκριμένα τμήματα που έχουν κάποια νοηματική συνοχή και περιεχόμενο πληροφορίας που μας ενδιαφέρει για περαιτέρω επεξεργασία. Πιο συγκεκριμένα παίρνοντας ένα κείμενο το οποίο μπορεί να περιέχει πολλαπλά τέτοια τμήματα ο σκοπός μου είναι να κατασκευάσω ένα Hidden Markov Model με το οποίο να μπορώ να τα εντοπίσω. Τα κείμενα αυτά θα αναφέρονται αποκλειστικά σε ανακλήσεις ιατρικών συσκευών και τα οποία θα λαμβάνονται από ιστοσελίδες γνωστών οργανισμών όπως ο FDA , οι Healthy Canadians και ο TGA αφου περάσουν από τον έλεγχο ότι πράγματι ανήκουν στην κατηγορία αυτή με την βοήθεια του ταξινομητή που υλοποιήθηκε και περιγράφηκε στο 4.
. Επειδή ένα τέτοιο εγχείρημα (text segmentation) γίνεται αρκετά πιο πολύπλοκο για το πλήθος των διαφορετικών κειμένων (διαφορετικά formats των οργανισμών) στα οποία θέλουμε να κάνουμε κατάτμηση αρχικά, θα εξεταστεί αυτός το εντοπισμός τμημάτων στα κείμενα του FDA. Στο επόμενο μέρος της διπλωματικής θα προσπαθήσω να χρησιμοποιήσω τις τεχνικές αυτές για να γενικευτεί το μοντέλο ώστε να μπορεί να εντοπίσει αυτά τα τμήματα και σε άλλους οργανισμούς.

Στην συνέχεια θα δούμε ποια είναι αυτά τα τμήματα και θα γίνει λεπτομερής ανάλυση της διαδικασίας που ακολουθήθηκε για την επίτευξη κατάτμησης κειμένου για κείμενα του πάρθηκαν από το site του FDA (http://www.fda.gov/Safety/Recalls/) και αφορούν ανακλήσεις ιατρικών συσκευών.

Ο FDA για την καταγραφή των Medical Devices χρησιμοποιεί 6 τμήματα, τα τμήματα αυτά είναι :

PRODUCT
CODE
MANUFACTURER
REASON
VOLUME OF PRODUCT IN COMMERCE
DISTRIBUTION

Ο σκοπός είναι λαμβάνοντας ένα αρχείο κείμενου το οποίο μπορεί να αναφέρεται σε ένα ή σε πολλά medical devices να εντοπίσω χρησιμοποιώντας τις δυνατότητες των Hidden Markov Models αυτές τις περιοχές. Ένα τέτοιο μοντέλο μπορεί να χρησιμοποιηθεί έτσι ώστε να μπορέσουμε να αποθηκεύσουμε τα αρχεία βάση αυτών των τμημάτων. Δηλαδή αρχίζοντας από έναν σύνολο κειμένων τα οποία αναφέρονται σε medical devices να καταλήξουμε σε 6 σύνολα κειμένων που το κάθε σύνολο περιέχει τα τμήματα των κειμένων που αναφέρονται σε αυτό (δηλαδή ένα σύνολο για το product ένα για το code κλπ). Έχοντας αυτά τα 6 σύνολα μπορούμε πολύ εύκολα και γρήγορα να αποκτήσουμε πληροφορίες που μας ενδιαφέρουν. Για παράδειγμα μπορεί κάποιος να θέλει να μάθει για κάποιο προϊόν, τις χώρες που διανέμεται, έτσι αρκεί να κάνει μια αναζήτηση στο σύνολο που αναφέρεται στο product για να εντοπίσει το προϊόν και αφού το εντοπίσει από το σύνολο με τις χώρες διανομής να βρει γρήγορα την απάντηση. Αν δεν έχουμε αποθηκεύσει τα κείμενα με τέτοιο τρόπο κάτι τέτοιο είναι πολύ πιο χρονοβόρο καθώς η αναζήτηση θα γινόταν στο σύνολο των κειμένων και όχι μόνο στις περιοχές που θα μπορούσαν να είναι προϊόντα. Εκτός αυτού ο εντοπισμός των χωρών διανομής απαιτεί να ξέρουμε ότι βρισκόμαστε σε αυτό το τμήμα διαφορετικά θα έπρεπε να ανακληθεί όλο το αρχείο για να δει αυτός που έκανε την ερώτηση τις χώρες διανομής.

5.1.2 ΟΙ ΔΥΣΚΟΛΙΕΣ ΤΟΥ TEXT SEGMENTATION.

Καταρχήν όταν υπάρχει ένα τέτοιο πρόβλημα, δηλαδή εντοπισμού συγκεκριμένων χωρίων μέσα σε ένα κείμενο, πρέπει να γίνει ανάλυση για τις ιδιαιτερότητες κάθε τμήματος. Δεν υπάρχει γενικός τρόπος η κάποιος αλγόριθμος που θα μας βοηθήσει ώστε να εντοπίζουμε αυτά τα τμήματα. Το ζήτημα του text segmentation είναι ένα ανοιχτό πρόβλημα και πολλές φορές είναι αδύνατο να χωρίσουμε ένα κείμενο σε ενότητες που έχουν κάποιο νόημα για εμάς αν αυτές δεν πληρούν κάποιες προϋποθέσεις. Για παράδειγμα ας θεωρήσουμε το εξής πρόβλημα :

Ένα υποθετικό κείμενο πρέπει να χωριστεί σε δύο ενότητες Α και Β, κάθε λέξη του κειμένου ανήκει είτε στην μια ενότητα είτε στην άλλη. Ας θεωρήσουμε την απλούστερη περίπτωση όπου μιλάμε για ένα κείμενο το οποίο έχει αριθμό λέξεων Μ και ξέρουμε ότι αν αρχίσουμε από την ενότητα Α όταν συναντήσουμε την ενότητα Β τότε δεν θα ξανασυναντήσουμε την ενότητα Α (εξετάζω την πιο απλή περίπτωση για να μπορέσω να αναδείξω την δυσκολία του προβλήματος). Τότε αυτό που μας ζητείται είναι να βρούμε την τελευταία λέξη της πρώτης ενότητας και αν αυτή η ενότητα είναι η Α ή η Β. Θα κάνουμε άλλες δύο υποθέσεις,οι ενότητες δεν έχουν κάποια λέξη με την οποία μπορούν να τελειώσουν αλλά ένα σύνολο λέξεων Τ, επίσης ξέρουμε όλες τις συχνότητες εμφάνισης των λέξεων σε όλη την έκταση του κειμένου. Ας θεωρήσουμε το σύνολο των πρώτων λέξεων για κάθε κείμενο ως Α και το σύνολο των λέξεων στο κυρίως κείμενο ως Κ. Έτσι για αυτές τις δύο ενότητες έχουμε 6 σύνολα

A_A και A_B για τις λέξεις που εμφανίζονται στην αρχή

K_A και K_B για τις λέξεις που εμφανίζονται στο κύριο μέρος

T_A και T_B για τις λέξεις που εμφανίζονται στο τέλος.

Για κάθε λέξη που ανήκει σε κάθε σύνολο αντιστοιχούμε μια πιθανότητα για παράδειγμα αν υπάρχει η λέξη w1 στο T_A τότε θα ξέρουμε και την πιθανότητα αυτή να είναι η τελευταία λέξη αν η λέξη w_2 υπάρχει στο K_A ξέρουμε και τη πιθανότητα της εμφάνισης αυτής της λέξης στο κύριο μέρος. Γνωρίζοντας όλες αυτές τις

πιθανότητες και την αλληλουχία λέξεων L_1 , L_2 ,.....,L_M αυτό που πρέπει να κάνουμε είναι να ελέγξουμε όλες τις πιθανότητες κατανομής των λέξεων στα παραπάνω σύνολα με το δεδομένο ότι διατηρείται η συνοχή των ενοτήτων (όταν τελειώνει μια ενότητα δεν ξαναρχίζει). Η αλληλουχία που μεγιστοποιεί αυτήν την πιθανότητα θεωρείται ότι είναι η κατάτμηση του κειμένου.

Αλληλουχία μπορεί να είναι της μορφής

$$A_A(L_1),K_A(L_2),.....K_A(L_N),T_A(L_{N+1}),A_B(L_{N+2}),K_B(L_{N+3}),....,T_B(L_M).$$

Έτσι γνωρίζοντας αυτές τις πιθανότητες δηλαδή την πιθανότητα η L_1 να ανήκει στο σύνολο A_A την πιθανότητα η L_N να ανήκει στο σύνολο K_A κλπ. Βρίσκοντας την ακολουθία με την μεγαλύτερη πιθανότητα τότε έχουμε την καλύτερη προσέγγιση για τις ενότητες του κειμένου.

Τα προβλήματα που υπάρχουν είναι τα εξής αν τα σύνολα Τ και Κ έχουν λέξεις που εμφανίζονται με παρόμοια πιθανότητα τότε είναι πολύ δύσκολο να κάνουμε διαχωρισμό αν μια λέξη ανήκει στο κύριο μέρος η αν ανήκει στο τέλος επίσης αν τα κύρια μέρη και των δύο ενοτήτων έχουν παρόμοιες πιθανότητες τότε υπάρχει μεγαλύτερο πρόβλημα στον διαχωρισμό. Επίσης επειδή αναφερόμαστε σε πιθανότητες είναι σπάνιο οι λέξεις που υπάρχουν πχ ακριβώς στα κύρια μέρη των ενοτήτων Α και Β να έχουν λόγους εμφάνισης όσους οι πιθανότητες εμφάνισης που είναι γνωστές από τα σύνολα Κ. Συνεπώς κείμενα με μεγάλες ομοιότητες στα παραπάνω σύνολα είναι εκ φύσης δύσκολο να διαχωριστούν μόνο και μόνο στο ότι αναφερόμαστε σε πιθανότητες. Εκτός αυτού το γεγονός ότι αναφερόμαστε σε πιθανότητες, σημαίνει ότι οι τυπικές αποκλίσεις πιθανοτήτων εμφάνισης λέξεων που βρίσκονται κοντά στα όρια μπορεί να εισάγει επιπλέον αστοχία στο μοντέλο ακόμα και αν δεν υπάρχει μεγάλη ομοιότητα. Ένα άλλο πρόβλημα που υπάρχει είναι ο ακριβής προσδιορισμός αυτών των πιθανοτήτων, στην πράξη είναι αδύνατο να ξέρουμε τις ακριβείς πιθανότητες εμφάνισης όλων των λέξεων (απαιτείται πολύ μεγάλος όγκος κειμένων εκπαίδευσης) και ακόμα αν γνωρίζαμε όλες τις πιθανότητες αν θεωρούσαμε κάθε λέξη ως ξεχωριστή παρατήρηση οι καταστάσεις που θα απαιτούνταν για να εκφράσουν τις πιθανές

αλληλουχίες σωστά σε ένα Hidden Markov Model θα αύξαναν πολύ την πολυπλοκότητα (O viterbi αλγόριθμος που υπολογίζει το πιο πιθανό μονοπάτι έχει πολυπλοκότητα
$O = (T *|S|^2)$ όπου S ο αριθμός των καταστάσεων του μοντέλου και T το μήκος της ακολουθίας παρατηρήσεων) σε βαθμό που μπορεί να μην ήταν αξιοποιήσιμο το μοντέλο.

Το παραπάνω πρόβλημα δεν θέλει να αποδείξει με μαθηματικό τρόπο τις αδυναμίες γενικών στατιστικών μοντέλων στην κατάτμηση κειμένου, παρά μόνο να αναδείξει κάποια προβλήματα που πρέπει με κάποιον τρόπο να αντιμετωπιστούν για να έχουμε καλύτερα αποτελέσματα.

5.2 ΣΧΕΔΙΑΣΜΟΣ HIDDEN MARKOV MODEL ΓΙΑ ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΣΕ ΙΑΤΡΙΚΑ ΚΕΙΜΕΝΑ

Για αυτό το πρόβλημα χρησιμοποιήθηκαν αρχεία κειμένου που αναφέρονται σε ιατρικές συσκευές από τον FDA για τα έτη 2005 και 2006.Τα αρχεία είναι χωρισμένα σε δύο κατηγορίες. Στην μια κατηγορία κάθε αρχείο περιέχει τα πλήρη κείμενα για μία η περισσότερες ιατρικές συσκευές ,ενώ στην δεύτερη κατηγορία τα αρχεία είναι χωρισμένα σε έξι φακέλους με ονομασίες product , code κλπ και κάθε αρχείο μέσα σε κάθε φάκελο περιέχει το τμήμα του κειμένου που προσδιορίζεται από το όνομα και αναφέρεται σε μία ιατρική συσκευή.

Τα κείμενα της πρώτης κατηγορίας τα χρειαζόμαστε για να ελέγξουμε πόσο καλά θα λειτουργήσει το Hidden Markov Model. Τα κείμενα της δεύτερης κατηγορίας θα τα χρησιμοποιήσουμε για να αναλύσουμε τις ιδιαιτερότητες κάθε ενότητας ώστε να μπορέσουμε να κατασκευάσουμε ένα μοντέλο το οποίο θα μπορεί να διακρίνει αυτές τις ενότητες.

Καταρχήν, πριν ξεκινήσει η ανάλυση της διαδικασίας πρέπει να επισημάνω ότι

κατασκευάζοντας ένα Hidden Markov Model θα πρέπει να επιλέξω πόσες διαφορετικές παρατηρήσεις/emission symbols θα μπορεί να παράγει το μοντέλο.

Γενικά ένα Hidden Markov Model δεν περιορίζεται χρονικά (η πολυπλοκότητα του) από την παρουσία παραπάνω διαφορετικών παρατηρήσεων, αλλά για να αποτυπωθούν οι διαφορετικές σχέσεις που μπορεί να συνδέουν αυτές τις παρατηρήσεις χρειάζονται αρκετά μεγάλος αριθμός καταστάσεων , του οποίου το τετράγωνο είναι ανάλογο με την χρονική πολυπλοκότητα του Viterbi αλγορίθμου (όπως φάνηκε στο προηγούμενο πείραμα 100 καταστάσεις έκαναν το μοντέλο αρκετά αργό στον υπολογιστή μου). Παρ' όλα αυτά ένα μικρότερο πλήθος από παρατηρήσεις αλλά με συγκεκριμένο νοηματικό περιεχόμενο είναι δυνατόν να βοηθήσει το μοντέλο. Για παράδειγμα από το να έχουμε μια ομάδα Ν λέξεων που εμφανίζουν μια κοινή ιδιότητα ως διαφορετικές είναι καλύτερο να της ομαδοποιήσουμε σε μία παρατήρηση η οποία θα εκφράζει και θα παρουσιάζει αυτήν την ιδιότητα.

Τέλος ένας άλλος παράγοντας, καθοριστικής σημασίας για την απόδοση του μοντέλου, είναι η αρχιτεκτονική του, δηλαδή η επιλογή των καταστάσεων και του τρόπου με τον οποίο αυτές διασυνδέονται.

5.2.1 ΜΕΤΑΤΡΟΠΗ ΣΥΜΒΟΛΟΣΕΙΡΩΝ ΚΕΙΜΕΝΩΝ ΣΕ ΑΚΟΛΟΥΘΙΕΣ ΠΑΡΑΤΗΡΗΣΕΩΝ

Διαδικασία:

Μελετώντας τα κείμενα αποφάσισα να χωρίσω τις παρατηρήσεις σε 8 ομάδες οι οποίες δεν έχουν κοινά στοιχεία,τις εξής :

λέξη
αριθμός
αλφαριθμητικό (εξαιρούνται τα παρακάτω)
λέξεις που περιέχουν τον χαρακτήρα -
αριθμούς που περιέχουν τον χαρακτήρα -
αλφαριθμητικά που περιέχουν τον χαρακτήρα -

μέρος δηλαδή όταν κάποια συμβολοσειρά τελειώνει με)
αλφαριθμητικά που περιέχουν τον χαρακτήρα – και /

Εκτός των λέξεων οι υπόλοιπες ομάδες επιλέχθηκαν με τέτοιο τρόπο ώστε οι συμβολοσειρές που ανήκουν σε αυτές να εμφανίζονται με διαφορετική συχνότητα σε διαφορετικές ενότητες των κειμένων των medical devices (για παράδειγμα η ενότητα code περιέχει περισσότερους αριθμούς από τα υπόλοιπα μέρη) στη συνέχεια θα εξηγήσω ξεχωριστά για κάθε ομάδα την ιδέα πίσω από την επιλογή της αλλά πριν από αυτό θα ασχοληθώ με την πρώτη ομάδα τις λέξεις.

Η ομάδα με τις λέξεις χωρίζεται στις 7 παρακάτω ομάδες:
1)Λέξεις που εμφανίζονται περισσότερο στο product.
2)Λέξεις που εμφανίζονται περισσότερο στο code.
3)Λέξεις που εμφανίζονται περισσότερο στο reason.
4)Λέξεις που εμφανίζονται περισσότερο στο manufacturer.
5)Λέξεις που εμφανίζονται περισσότερο στο volume.
6)Λέξεις που εμφανίζονται περισσότερο στο distribution.
7)Λέξεις που δεν έχουμε συναντήσει στα κείμενα εκπαίδευσης ή εμφανίζονται συχνά σε περισσότερες από μια ενότητες.

Για τον εντοπισμό αυτών των λέξεων χρησιμοποιήθηκαν 1000 κείμενα από κάθε ενότητα στα οποία χρησιμοποίησα την λειτουργία του indexing της TMG χωρίς την χρήση stemming αλλά με την αφαίρεση αριθμών και αλφαριθμητικών κόβοντας τις λέξεις που έχουν μήκος ένα ή μεγαλύτερο του 30.Stemming δεν χρησιμοποιήθηκε διότι κατασκεύασα μόνος μου αλγόριθμο ο οποίος θα διαβάζει σειριακά όλες τις λέξεις κάποιων αρχείων και θα ελέγχει σε ποια κατηγορία ανήκουν,το πρόβλημα είναι ότι η TMG δεν έχει συνάρτηση η οποία να κάνει stemming σε μεμονωμένες λέξεις και έτσι λέξεις στις οποίες έχει εφαρμοστεί stemming και έχουν αντικατασταθεί από την ρίζα τους δεν θα κατατάσσονταν στην σωστή κατηγορία

λέξεων. Για κάθε ενότητα έτσι κατασκευάζεται ένας term-document πίνακας με όλες τις λέξεις που συναντήθηκαν σε αυτά τα 1000 κείμενα. Οι πίνακες για την ενότητα product και reason είχαν 2589 και 2112 λέξεις αντίστοιχα. Σε αυτά τα τμήματα που έχουν μεγαλύτερη χρήση λέξεων λόγω της περιγραφικής τους φύσης, αποφάσισα να κόψω τις λέξεις με πολύ μικρή συχνότητα εμφάνισης (την ομάδα που απαρτίζει το 20% από τις μικρότερες εμφανίσεις λέξεων), το οποίο γίνεται κυρίως για λόγους απόδοσης. Οι λέξεις αν θέλουμε θα μπορούσαν κρατηθούν και όλες και να μην αλλάξουν σημαντικά τα αποτελέσματα, αλλά επέλεξα αυτήν την αντιμετώπιση επειδή σε μεγαλύτερο training data set θα έπρεπε να γίνει τέτοια διαδικασία για να διαχωρίσουμε τις λέξεις που εμφανίζονται τυχαία (δηλαδή με πολύ μικρές πιθανότητες εμφάνισης) στην ενότητα. Με αυτόν τον τρόπο δημιούργησα 6 dictionaries με τις λέξεις που εμφανίζονται στα αντίστοιχα χωρία. Το επόμενο βήμα ήταν να ενοποιήσω αυτά τα λεξικά αφαιρώντας όλες τις λέξεις οι οποίες εμφανίζονταν και σε άλλη ενότητα.

Ιδιαίτερη περίπτωση αποτελεί η ενότητα distribution καθώς σε αυτήν περιέχονται μόνο λέξεις που αντιστοιχούν σε χώρες η αρκτικόλεξα (digraphs) χωρών. Άρα στο λεξικό που δημιουργήθηκε από την επεξεργασία με το TMG προσετέθησαν όλες οι χώρες και όλες οι συντομεύσεις των χωρών με δύο γράμματα ref123.

Αφότου δημιουργηθεί αυτό το λεξικό σε κάθε λέξη αντιστοιχείται ένας ακέραιος με τον παρακάτω τρόπο :

Αν η λέξη προέρχεται από το product αντιστοιχείται το 1
Αν η λέξη προέρχεται από το code αντιστοιχείται το 2
Αν η λέξη προέρχεται από το reason αντιστοιχείται το 3
Αν η λέξη προέρχεται από το manufacturer αντιστοιχείται το 4
Αν η λέξη προέρχεται από το volume of product in commerce αντιστοιχείται το 5
Αν η λέξη προέρχεται από το distribution αντιστοιχείται το 6

Τέλος είναι πολύ σημαντικό να λάβουμε υπόψιν τις λέξεις που δεν έχουμε συναντήσει στην εκπαίδευση, καθώς αυτές θα απαρτίζουν ένα μεγάλο αριθμό των λέξεων που θα διαβάσουμε, έτσι ώστε να τις μετατρέψουμε σε παρατηρήσεις που θα βοηθήσουν το Hidden Markov Model να έχει καλύτερη ακρίβεια. Αυτές οι λέξεις

εμφανίζονται με διαφορετικές συχνότητες σε διαφορετικές ενότητες και μας δίνουν σημαντική πληροφορία για την δομή κάθε ενότητας (π.χ η ενότητα distribution έχει ελάχιστες άγνωστες λέξεις καθώς σε αντίθεση με το product ή το reason) .

Όλες οι λέξεις της Αγγλικής γλώσσας οι οποίες δεν κατατάσσονται σε κάποια από τις παραπάνω 6 ομάδες μεταφράζονται σε παρατήρηση με τον αριθμό 7.

Αυτές οι αντιστοιχίσεις γίνονται διότι τα Hidden Markov Model αντιλαμβάνονται ως παρατηρήσεις διαφορετικούς θετικούς ακεραίους. Τα αποτελέσματα της παραπάνω διαδικασίας αντιστοιχίσεων είναι ο χωρισμός των λέξεων της Αγγλικής γλώσσας σε 7 ομάδες οι οποίες έχουν επιλεχθεί με τέτοιο τρόπο ώστε να εμφανίζονται με διαφορετική συχνότητα σε κάθε ενότητα.

Η ανάλυση γενικά πρέπει να γίνει σε τρία μέρη της ενότητας κάθε κειμένου, στην αρχή, στο τέλος και στις υπόλοιπες λέξεις που συγκροτούν τον κορμό.

Μέρος Α, λέξεις εισαγωγής σε ενότητα (αρχή) :

Σε αυτό το μέρος δεν χρειάστηκε να γίνει ανάλυση καθώς κάθε ενότητα αρχίζει από την αντίστοιχη λέξη με κεφαλαία. Συνεπώς για κάθε ενότητα οι λέξεις με τις οποίες μπαίνουμε σε αυτήν είναι οι λέξεις PRODUCT,CODE κλπ.

Μέρος Β, λέξεις για το κύριο μέρος της ενότητας :

Η ανάλυση των λέξεων για το κύριο μέρος δεν γίνεται με το TMG αλλά με δικό μου αλγόριθμο ο οποίος διαβάζει σειριακά όλες τις λέξεις του μέρους Β , αντικαθιστώντας την κάθε μία σε μια ομάδα (λέξη,αλφαριθμητικό κλπ), στην συνέχεια υπολογίζεται μια πιθανότητα εμφάνισης αυτών των ομάδων για την συγκεκριμένη ενότητα, αυτό είναι σημαντικό διότι από αυτές τις πιθανότητες θα κατασκευάσουμε το Hidden Markov Model.

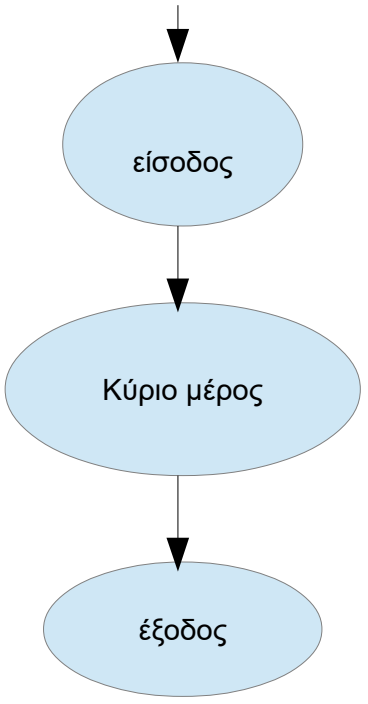
Μέρος Γ, λέξεις οι οποίες βρίσκονται στο τέλος κάθε ενότητας :

Η ίδια διαδικασία γίνεται και για την τελευταία λέξη κάθε κειμένου εκπαίδευσης που αντιστοιχεί σε αυτήν την ενότητα. Από τον αλγόριθμο δημιουργείται ένας

πίνακας που περιέχει τις πιθανότητες εμφάνισης κάθε ομάδας συμβολοσειράς.

Παρατηρήσεις/Emission symbols :

Με την παραπάνω διαδικασία δηλαδή χωρίζοντας κάθε ενότητα σε 3 μέρη έχει νόημα οι παρατηρήσεις να μπορούν να αποτυπώσουν αυτήν την δομή. Ως εκτούτου, μια γενική ιδέα, όσον αφορα στην κατασκευή της αρχιτεκτονικής του μοντέλου, είναι να υπάρχει μια κατάσταση που να αντιπροσωπεύει την αρχή δηλαδή την είσοδο στην ενότητα, μια άλλη κατάσταση/ή καταστάσεις να αντιπροσωπεύουν το κύριο μέρος και τέλος μια κατάσταση να αντιπροσωπεύει την έξοδο/τελευταία λέξη από την ενότητα. Σχηματικά παρακάτω φαίνεται η αναπαράσταση μίας ενότητας με τρείς καταστάσεις. Στο παρακάτω μοντέλο το κύριο μέρος της ενότητας αντιστοιχεί σε μία κατάσταση. Τα βελάκια υποδηλώνουν τις δυνατές μεταβάσεις κατάστασης, ενώ το πρώτο βελάκι για την είσοδο υποδηλώνει απλά ότι μια ενότητα θα καταλήγει στην είσοδο της επόμενης.



Κατάσταση εισόδου στην ενότητα, περιέχει σαν παρατηρήσεις μόνο τις πρώτες λέξεις που μπορεί να έχει μια ενότητα(σε αυτήν την περίπτωση οι λέξεις είναι μοναδικές για κάθε χωρίο).

Οι παρατηρήσεις αυτής κατάστασης είναι οι ομάδες συμβολοσειρών που μπορεί να εμφανιστούν στο κυρίως κείμενο.

Οι παρατηρήσεις αυτής της κατάστασης είναι οι ομάδες συμβολοσειρών που μπορεί να βρίσκονται στο τέλος της ενότητας(τελευταία λέξη).

Έχοντας το παραπάνω σχέδιο στο μυαλό μας, αντιλαμβανόμαστε ότι εφόσον

μεγάλο ρόλο στην αναγνώριση της ενότητας παίζει η κατάσταση εισόδου λογικό είναι οι παρατηρήσεις που εμφανίζονται σε αυτήν να είναι διαφοροποιημένες από ενότητα σε ενότητα, δηλαδή αν είχαμε 6 σύνολα από λέξεις (συμβολοσειρές) που αντιστοιχούν στις λέξεις με τις οποίες μπορεί να αρχίσει μια ενότητα τότε καθένα από αυτά τα σύνολα θα πρέπει να αντιστοιχεί σε διαφορετική παρατήρηση (η να περιέχει διαφορετικές παρατηρήσεις) ώστε να φαίνεται η σημασία που έχουν αυτές οι λέξεις στην οριοθέτηση των ενοτήτων. Στην περίπτωση μας, στο dataset του FDA αυτά τα σύνολα αποτελούνται από μία λέξη δηλαδή product, code κλπ.

Το κύριο μέρος του κειμένου απαρτίζεται από πολλές λέξεις και εδώ εισάγονται παρατηρήσεις/emission symbols που θα μπορέσουν να αποτυπώσουν την διαφορά των ενοτήτων στην σύσταση συμβολοσειρών, οι 8 βασικές ομάδες έχουν συζητηθεί νωρίτερα και έχουν επιλεχθεί με αυτόν τον βασικό γνώμονα. Στην συνέχεια θα γίνει αναλυτικότερη αναφορά σε αυτές.

Οι τελευταίες λέξεις του κειμένου παίζουν και αυτές ρόλο στην αναγνώριση ενότητας αλλά δεν είναι τόσο σημαντικές όσο οι λέξεις εισαγωγής σε μια ενότητα διότι είναι περισσότερες σε πλήθος. Παρ' όλα αυτά ανεξαρτήτως του πόσο μπορούν να βοηθήσουν οι τελευταίες λέξεις στην αναγνώριση χωρίων κειμένου στην περίπτωση μας, πρέπει να συμπεριληφθούν καθώς αποτελούν μια εννοιολογική κλάση λέξεων, η εισαγωγή της οποίας στην μοντελοποίηση, μπορεί να βελτιώσει τα αποτελέσματα. Οι ομάδες λέξεων που βρίσκονται συχνά στο τέλος (τελευταία λέξη) μιας ενότητας με αυτό τον τρόπο πρέπει να αποτυπωθούν σαν διαφορετική παρατήρηση.

Έχοντας εξηγήσει την ιδέα πίσω από τον προσδιορισμό των παρατηρήσεων στην συνέχεια θα εξηγήσω γιατί επέλεξα τις ομάδες συμβολοσειρών που προανέφερα, για την δημιουργία των ακολουθιών παρατηρήσεων από τα κείμενα του dataset.Έχοντας υπόψιν μας ένα ενδεικτικό κείμενο ανάκλησης ιατρικών συσκευών από τον FDA, όπως φαίνεται στην συνέχεια, ο λόγος για τον οποίο επιλέχθηκαν οι παρακάτω ομάδες συμβολοσειρών είναι :

Recall FDA 2004-01-07

_____ PRODUCT Lorad M-IV and Lorad M-IV Platinum mammography systems. Mammographic X-Ray systems to conduct mammography. Model Number: M-IV and M-IV Platinum. Recall # Z-0148/0149-04. CODE None. RECALLING FIRM/MANUFACTURER Lorad, Danbury, Connecticut, by CPA letter December 23, 2003. Firm initiated recall is ongoing. REASON System failed to meet the mAs accuracy specifications at low mAs values. VOLUME OF PRODUCT IN COMMERCE 4,206. DISTRIBUTION Nationwide. \n

Αριθμός : οι αριθμοί έχουν επιλεγεί σαν ξεχωριστή παρατήρηση (δηλαδή κάθε αριθμός εισάγεται στο Hidden Markov Model με το ίδιο σύμβολο παρατήρησης/emission symbol) διότι έχουν διαφορετική συχνότητα εμφάνισης στις ενότητες, με μεγαλύτερη συχνότητα στις ενότητες code και volume.

Αλφαριθμητικό : τα αλφαριθμητικά αποτελούν ξεχωριστή παρατήρηση καθώς μπορούν να εμφανιστούν με αυξημένη συχνότητα σε κάποιες ενότητες όπως στο code και στο product. Είναι σημαντικό να σημειωθεί ότι η εισαγωγή μιας παρατήρησης όπως τα αλφαριθμητικά μπορεί να γίνει σε ένα Hidden Markov Model χωρίς να γνωρίζουμε εκ των προτέρων αν εμφανίζεται με διαφορετικές συχνότητες σε κάποιες ενότητες. Η εκπαίδευση του μοντέλου θα το δείξει αυτό. Αν φανεί ότι η παρατήρηση είναι περιττή μπορεί να αφαιρεθεί στην συνέχεια.

Λέξεις που περιέχουν τον χαρακτήρα - : Όπως φαίνεται και από το κείμενο του FDA που παρέθεσα οι λέξεις που περιέχουν τον χαρακτήρα ' - ' εμφανίζονται περισσότερο στο product αρά είναι σημαντικό να προστεθεί μια παρατήρηση που να εκφράζει αυτήν την συμπεριφορά.

Αριθμούς που περιέχουν τον χαρακτήρα - : Αν και δεν φαίνεται από το συγκεκριμένο κείμενο του FDA, αριθμοί που περιέχουν τον χαρακτήρα ' - ' εμφανίζονται πιο συχνά στην ενότητα του code.

Αλφαριθμητικά που περιέχουν τον χαρακτήρα -: Αν παρατηρήσουμε το product σε αυτό το κείμενο τελειώνει με την λέξη Z-0148/0149-04 όμως ένα πολύ μεγάλο ποσοστό των product (θα φανεί στην εκπαίδευση αυτό) τελειώνουν με συμβολοσειρές τύπου Z-0516-04 οι οποίες ανήκουν σε αυτήν την παρατήρηση. Αυτή η παρατήρηση είναι στατιστικά τόσο συχνή ώστε πρέπει να έχει ξεχωριστή κατηγορία.

"Μέρος" δηλαδή όταν κάποια συμβολοσειρά τελειώνει με) : Πολλές φορές το product ή το code έχουν την μορφή a).....

b).....

Αυτές οι 'παρατηρήσεις' εμφανίζονται μόνο σε αυτά τα 2 μέρη άρα είναι σημαντικό να τις συμπεριλάβουμε.

Α**λφαριθμητικά που περιέχουν τον χαρακτήρα – και / :** Κάποιες φόρες το product τελειώνει με συμβολοσειρά που ανήκει σε αυτήν την κατηγορία (όπως στην περίπτωση αυτού του κειμένου).

Οι παραπάνω παρατηρήσεις αναφέρονται σε συμβολοσειρές που εμφανίζονται στο κύριο μέρος μιας ενότητας ή στο τέλος τής ,στην συνέχεια θα προστεθούν παρατηρήσεις οι οποίες θα ' αιχμαλωτίζουν ' την είσοδο του μοντέλου σε μια νέα ενότητα. Επειδή στην περίπτωση των κειμένων του FDA οι λέξεις που εμφανίζονται στην αρχή κάθε ενότητας είναι πάντα ίδιες έχουν προστεθεί 6 παρατηρήσεις που ταυτίζονται με αυτές τις λέξεις , δηλαδή:

Παρατηρήσεις για την είσοδο σε ενότητα:

Παρατήρηση PRODUCT : Εμφανίζεται πάντα στην αρχή της ενότητας product.
Παρατήρηση CODE : Εμφανίζεται πάντα στην αρχή της ενότητας code.
Παρατήρηση REASON : Εμφανίζεται πάντα στην αρχή της ενότητας reason.
Παρατήρηση RECALLING : Εμφανίζεται πάντα στην αρχή της ενότητας manufacturer.
Παρατήρηση VOLUME : Εμφανίζεται πάντα στην αρχή της ενότητας volume.
Παρατήρηση DISTRIBUTION : Εμφανίζεται πάντα στην αρχή της ενότητας distribution.

Πρέπει να γίνει σαφές ότι η μέθοδος που έχει αναλυθεί είναι γενική,το γεγονός ότι κάθε ενότητα αρχίζει με μια λέξη βοηθάει το μοντέλο πολύ, αλλά είναι συγκυριακό, αν κάποια ενότητα άρχιζε με παραπάνω λέξεις (δηλαδή όχι πάντα με τον ίδιο τρόπο) τότε αυτές οι λέξεις θα αποτελούσαν μια παρατήρηση.

Έχοντας πλέον αναφερθεί στο σύνολο των δυνατών συμβόλων παρατηρήσεων αυτού του μοντέλου μπορούμε να προχωρήσουμε στην περαιτέρω ανάλυση των κειμένων για να εκπαιδεύσουμε καλύτερα το μοντέλο. Πριν γίνει αυτό θα πρέπει να αντιστοιχίσουμε στο Hidden Markov Model για κάθε παρατήρηση έναν θετικό ακέραιο (είναι σύμβαση αυτό). Οι παρατηρήσεις στο Hidden Markov model αναπαρίστανται από τους αριθμούς που φαίνονται στον επόμενο πίνακα.

Παρατήρηση	Emission Symbol
Λέξη που εμφανίζεται περισσότερο στο product.	1
Λέξη που εμφανίζεται περισσότερο στο code.	2
Λέξη που εμφανίζεται περισσότερο στο reason.	3
Λέξη που εμφανίζεται περισσότερο στο manufacturer.	4
Λέξη που εμφανίζεται περισσότερο στο volume.	5
Λέξη που εμφανίζεται περισσότερο στο distribution.	6
Άγνωστη λέξη ή λέξη που εμφανίζεται συχνά σε παραπάνω από μια ενότητες	7
Αριθμός	8
Λέξη με -	9
Αλφαριθμητικό με -	10
Αλφαριθμητικό	11
Αριθμός με -	12
Μέρος δηλαδή όταν κάποια συμβολοσειρά τελειώνει με)	13
Αλφαριθμητικό με – και /	14
PRODUCT	15
CODE	16
RECALLING	17
REASON	18
VOLUME	19
DISTRIBUTION	20

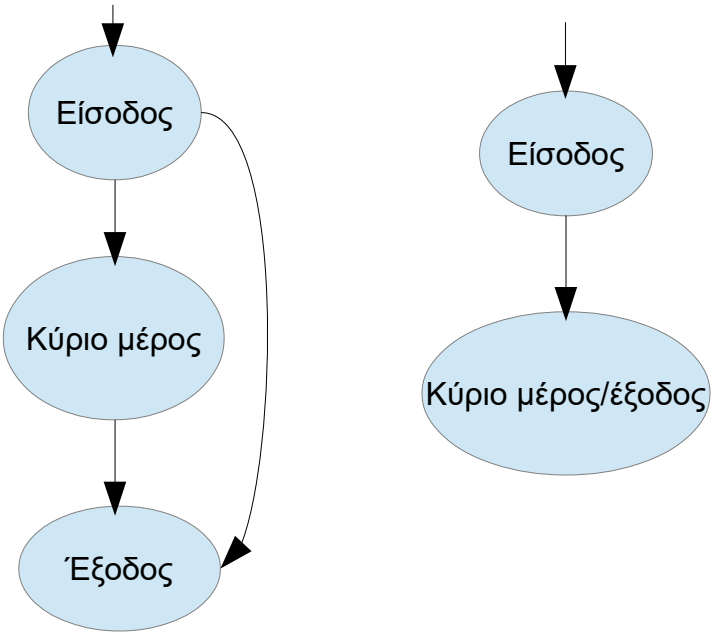
5.2.2 ΤΡΟΠΟΙ ΥΠΟΛΟΓΙΣΜΟΥ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ HIDDEN MARKOV MODEL

Έχοντας ορίσει όλες τις παρατηρήσεις μπορούμε τώρα να προχωρήσουμε στην εκπαίδευση του μοντέλου. Η εκπαίδευση Hidden Markov Models μπορεί να γίνει με δύο τρόπους:

A)Unsupervised Learning: Δηλαδή να γίνεται ο υπολογισμός παραμέτρων του μοντέλου με τον αλγόριθμο Baum Welch. Ο αλγόριθμος χρησιμοποιείται όταν είναι γνωστή μόνο η ακολουθία παρατηρήσεων (όχι η ακολουθία καταστάσεων).

B)Supervised Learning: Δηλαδή γνωρίζοντας τις ακολουθίες κατάστασης (όχι μόνο παρατηρήσεων) να υπολογίζουμε τις παραμέτρους του μοντέλου αλγεβρικά.

Επειδή στο συγκεκριμένο μοντέλο έχουμε προσδιορίσει ότι κάθε ενότητα θα εκφράζεται με τους παρακάτω τρόπους :



Έχει προστεθεί η περίπτωση που το κύριο μέρος έχει πιθανότητα να έχει μία λέξη (μετάβαση από την είσοδο (πρώτη λέξη) κατευθείαν στην έξοδο (τελευταία λέξη)

ή να μην έχει σημαντική διαφορά το τέλος του με το κύριο μέρος οπότε έχουν συγχωνευθεί οι καταστάσεις κύριο μέρος και έξοδος.

Γνωρίζουμε την ακολουθία των καταστάσεων επομένως η εκπαίδευση μπορεί να γίνει με επιβλεπόμενη μάθηση (supervised learning). Γενικώς μπορούμε να χρησιμοποιούμε συνδυασμούς των δύο μεθόδων για να εκπαιδεύουμε οποιαδήποτε Hidden Markov Models καθώς μπορεί να ξέρουμε μερικώς την αλληλουχία καταστάσεων, οπότε για τις καταστάσεις στις οποίες έχουμε πλήρη γνώση χρησιμοποιούμε supervised learning ενώ για μέρη του μοντέλου που έχουμε γνώση μόνο των παρατηρήσεων χρησιμοποιούμε unsupervised learning. Ένα παράδειγμα θα ήταν να χρησιμοποιούμε περισσότερες καταστάσεις για να αποτυπώσουμε το κύριο μέρος, οπότε το κύριο μέρος θα αποτελούσε ένα ξεχωριστό Hidden Markov Model στο οποίο οι παράμετροι του θα γίνονταν γνωστές μέσω των παρατηρήσεων εκπαίδευσης για αυτό το κομμάτι, αλλά η σύνδεση του με το υπόλοιπο Hidden Markov Model θα γινόταν με επιβλεπόμενη μάθηση.

Το γεγονός ότι το μοντέλο με μια κατάσταση για το κύριο μέρος είχε πολύ καλά αποτελέσματα (και πιθανά προβλήματα δεν οφείλονταν στον αριθμό καταστάσεων του κύριου μέρους) δεν εξετάστηκαν μοντέλα με παραπάνω καταστάσεις για το κύριο μέρος.

Με την επιβλεπόμενη μάθηση μπορούμε να υπολογίσουμε τις παραμέτρους του μοντέλου με την παρακάτω διαδικασία:

Γνωρίζοντας την ακολουθία καταστάσεων οι όροι του πίνακα(A) με τις πιθανότητες μετάβασης από κατάσταση σε κατάσταση υπολογίζονται ως εξής :

$A(i,j) = A_{ij}/A_i$

ο A_{ij} αριθμός των μεταβάσεων από την κατάσταση i στην κατάσταση j ,

και A_i οι συνολικές μεταβάσεις από την i κατάσταση προς οποιαδήποτε κατάσταση.

Οι όροι του πίνακα E με διαστάσεις MxN όπου M ο αριθμός καταστάσεων και N ο αριθμός των παρατηρήσεων. E(1,1) είναι η πιθανότητα η κατάσταση 1 να

δημιουργήσει την παρατήρηση 1 (στο μοντέλο μας η παρατήρηση 1 αντιστοιχεί σε λέξη που εμφανίζεται πιο συχνά στην ενότητα product). Υπενθυμίζω ότι το άθροισμα κατά γραμμή (δηλαδή το άθροισμα όλων των πιθανών παρατηρήσεων μιας κατάστασης) πρέπει να ισούται με 1. Τα στοιχεία του Ε επομένως υπολογίζονται με τον εξής τρόπο :

$E(i,j) = E_{ij}/E_i$

όπου E_{ij} ο αριθμός όλων των j παρατηρήσεων που παράγονται από την κατάσταση i και E_i οι συνολικές παρατηρήσεις που συλλέχθηκαν στην συγκεκριμένη κατάσταση.

5.2.3 ΜΙΑ ΕΝΑΛΛΑΚΤΙΚΗ ΜΕΘΟΔΟΣ ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΠΑΡΑΜΕΤΡΩΝ

Όπως ανέφερα παραπάνω εκτός από τον γενικό τρόπο που χρησιμοποιούμε για να υπολογίσουμε τους συντελεστές του πίνακα με τις πιθανότητες μετάβασης κατάστασης, μπορούμε να προσδιορίσουμε τους συντελεστές με έναν εναλλακτικό τρόπο. Για να εξηγήσω την ιδέα πίσω από αυτόν τον τρόπο θα αναφέρω ένα υποθετικό πρόβλημα.

Έστω ότι έχουμε 2 πιθανές διαφορετικές παρατηρήσεις. Με αυτές θέλουμε να διαφοροποιήσουμε δύο διαφορετικά συστήματα. Το ένα σύστημα εμφανίζει την πρώτη παρατήρηση με ποσοστό 80% και την δεύτερη με 20%. Το άλλο σύστημα εμφανίζει την πρώτη παρατήρηση με ποσοστό 90% και την δεύτερη με 10%. Ας υποθέσουμε επίσης ότι κάθε σύστημα παράγει κατά μέσο όρο 10 παρατηρήσεις πριν το διαδεχθεί το άλλο σύστημα . Θεωρούμε μια ακολουθία από 20 παρατηρήσεις και την περίπτωση που οι 10 πρώτες παρατηρήσεις δημιουργούνται από το πρώτο σύστημα και οι υπόλοιπες από το άλλο σύστημα έχοντας τις αντίστοιχες αναλογίες εμφάνισης παρατηρήσεων. Δηλαδή ας θεωρήσουμε ότι στις πρώτες 10 παρατηρήσεις οι 2 ανήκουν στην δεύτερη παρατήρηση και από τις επόμενες 10 μια ανήκει στην δεύτερη παρατήρηση.

Σύμφωνα με το supervised learning για ένα HMM 2 καταστάσεων, θα έπρεπε να

δημιουργηθεί ο επόμενος πίνακας με πιθανότητες μετάβασης κατάστασης (η κατάσταση 1 αντιπροσωπεύει το πρώτο σύστημα και η 2 το δεύτερο).

A	1	2
1	0.9	0.1
2	0.1	0.9

Παρ' όλα αυτά χρησιμοποιώντας τον viterbi algorithm στην ακολουθία παρατηρήσεων:

1 1 2 1 2 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1

θεωρώντας π=[0,5 0,5] δηλαδή είναι ισοπίθανο να αρχίσουμε από την μία η την άλλη κατάσταση και πίνακα με πιθανότητες εμφάνισης παρατηρήσεων για κάθε κατάσταση/σύστημα :

E	1	2
1	0.8	0.2
2	0.9	0.1

Το πιο πιθανό μονοπάτι θα είναι να παραμείνουμε στην κατάσταση 1 για όλες τις παρατηρήσεις. Ο λόγος που συμβαίνει αυτό είναι ότι το πιο πιθανό μονοπάτι δεν καθορίζεται μόνο από τις πιθανότητες εμφάνισης κάποιας παρατήρησης από κάποια κατάσταση αλλά και από τις πιθανότητες μετάβασης. Το γεγονός ότι οι 2 καταστάσεις δεν διαφέρουν πάρα πολύ στην συμπεριφορά τους εμποδίζει το μοντέλο, με αυτές τις παραμέτρους, να αναγνωρίσει έστω και με απόκλιση την ύπαρξη 2 συστημάτων (καταστάσεων) για την παραγωγή αυτών των παρατηρήσεων. Με πράξεις αυτό γίνεται αντιληπτό διότι η πιθανότητα να παραμείνουμε στην κατάσταση 1, για 20 παρατηρήσεις και να παράγουμε αυτήν την ακολουθία είναι $P_1 = 0,5*0,9^{19} *0,8^{17} *0,2^3$

(0,5 είναι η πιθανότητα να αρχίσουμε στην κατάσταση 1, $0,9^{19}$ είναι η πιθανότητα να παραμείνουμε στην κατάσταση 1 19 φορές, το $0,8^{17}$ είναι η πιθανότητα να έχουμε 17 πρώτες παρατηρήσεις και $0,2^3$ είναι η πιθανότητα να έχουμε 3 δεύτερες παρατηρήσεις). Η πιθανότητα να εμφανιζόταν η αλλαγή στην 11 παρατήρηση είναι $P_2 = 0,5*0,9^9*0,8^8*0,2^2*0,1*0,9^9*0,9^9*0,1$ υπολογίζοντας τον λόγο τους βλέπουμε ότι $P_1/P_2 = 0,9^{-8}*0,8^9*0,2*10^2 = 6,23$ δηλαδή το μοντέλο θεωρεί πολύ πιο πιθανό σε ακολουθίες μήκους 20 με αυτήν την μορφή να μην αλλάζει η κατάσταση.

Για να ρυθμίσουμε το μοντέλο ώστε να μπορεί να αναγνωρίζει καλύτερα το τι αντιπροσωπεύει η ακολουθία, θα πρέπει να μειώσουμε την πιθανότητα παραμονής σε κατάσταση ώστε να είναι πιο εύκολη η μετάβαση από την μια κατάσταση στην άλλη. Ένας τρόπος για να γίνει αυτό είναι αν θεωρήσουμε ότι οι 10 πρώτες παρατηρήσεις ανήκουν στο σύστημα 1, να υπολογίσουμε τις πιθανότητες παραγωγής των ακολουθιών για τις επόμενες 10 παρατηρήσεις για κάθε σύστημα (κατάσταση) με άγνωστη την πιθανότητα παραμονής στην κατάσταση και στην συνέχεια να τις εξισώσουμε (για την ακρίβεια η πιθανότητα εμφάνισης της 2ης ακολουθίας από την κατάσταση 1 να είναι ελάχιστα μικρότερη ώστε να συμβαίνει η μετάβαση). Με τα συγκεκριμένα νούμερα η νέα πιθανότητα μετάβασης (σε αυτήν την περίπτωση δεν λαμβάνω καν υπόψιν την πιθανότητα αρχικής κατανομής διότι βλέπω το σύστημα στο άπειρο και με ενδιαφέρουν μόνο οι διαδοχές όχι η αρχική συνθήκη) υπολογίζεται σε 0,69217 για παραμονή και 0,30783 για αλλαγή κατάστασης (το ίδιο ισχύει και για τα δύο συστήματα). Με τα νέα νούμερα το path αλλάζει από την 9 παρατήρηση που είναι πολύ καλύτερη προσέγγιση (δηλαδή με τις νέες πιθανότητες μετάβασης το μοντέλο θεωρεί ότι το πρώτο σύστημα παράγει τις πρώτες 8 παρατηρήσεις και οι υπόλοιπες παράγονται από το δεύτερο σύστημα).

Αυτό που θέλω να κρατήσουμε από το παραπάνω παράδειγμα είναι ότι για να προσδιορίσουμε τους συντελεστές του πίνακα Α μπορεί να είναι καλύτερο σε κάποιες περιπτώσεις να μην λάβουμε υπόψιν μόνο τις αλλαγές κατάστασης και να υπολογίσουμε ένα πηλίκο, αλλά να σκεφτούμε πως θέλουμε να λειτουργεί το μοντέλο και να μειώσουμε η να αυξήσουμε αυτούς τους συντελεστές ώστε να έχουμε καλύτερα αποτελέσματα.

5.2.4 ΥΠΟΛΙΣΜΟΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΜΟΝΤΕΛΟΥ

Για να υπολογίσω τις παραμέτρους με επιβλεπόμενη μάθηση δημιούργησα κάποιες συναρτήσεις στην matlab οι οποίες θα κάνουν αυτούς τους υπολογισμούς με τον εξής τρόπο. Τα κείμενα εκπαίδευσης αποτελούν 1000 κείμενα από κάθε ενότητα.

Για τους προσδιορισμούς των πιθανοτήτων των παρατηρήσεων για την κατάσταση που εκπροσωπεί το κύριο μέρος, έφτιαξα έναν αλγόριθμο ο οποίος διαβάζει λέξη προς λέξη κάθε κείμενο από την αρχή μέχρι την προτελευταία λέξη (στα κείμενα εκπαίδευσης δεν περιέχονται οι λέξεις PRODUCT,CODE κλπ για αυτό τα διαβάζω από την αρχή τους). Για την ανάγνωση του κειμένου έχουν χρησιμοποιηθεί delimeters (για των διαχωρισμό λέξεων) τα σύμβολα ' , ' , ' . ' , ' : ' το οποίο σημαίνει ότι αυτοί οι χαρακτήρες δεν θα υπάρχουν ποτέ σε μία λέξη (μία λέξη που ακολουθείτε από κόμμα όπως π.χ. το and, αν δεν είχα χρησιμοποιήσει το ' , ' ως delimiter θα καταγραφόταν ως and, και όχι and με αποτέλεσμα να μην αντιστοιχηθεί στην παρατήρηση που πρέπει). Στην συνέχεια δημιουργείται ένας πίνακας που έχει τις λέξεις με την σειρά εμφάνισης τους από πάνω προς τα κάτω. Αυτός ο πίνακας εισάγεται σε συνάρτηση η οποία αναλαμβάνει την μετάφραση των λέξεων/συμβολοσειρών σε παρατηρήσεις, υπολογίζει τις εμφανίσεις κάθε παρατήρησης. Όταν τελειώσει η διαδικασία ανάγνωσης όλων των αρχείων , είναι γνωστός ο συνολικός αριθμός παρατηρήσεων για αυτά τα 1000 κείμενα όπως και οι ξεχωριστές εμφανίσεις κάθε παρατήρησης άρα μπορούν εύκολα να υπολογιστούν οι συντελεστές E(i,j) του πίνακα με τις πιθανότητες εμφάνισης παρατηρήσεων.

Για τους προσδιορισμούς των πιθανοτήτων των παρατηρήσεων για την κατάσταση που εκπροσωπεί την τελευταία λέξη (Έξοδος). Ακολουθείται η ίδια διαδικασία με την διαφορά ότι αυτήν την φορά κρατάω μόνο την τελευταία λέξη από κάθε κείμενο.

Στην συνέχεια παραθέτω τα αποτελέσματα της matlab για τον πίνακα E. Επειδή ο πίνακας είναι 16x20 τον παρουσιάζω με δύο υποπίνακες. Ο ένας για τις πρώτες 10 παρατηρήσεις και ο άλλος για τις υπόλοιπες 10. Οι καταστάσεις που αντιστοιχούν σε κύριος μέρος είναι οι καταστάσεις 2,5,8,11,14 και 16. Οι καταστάσεις που

αντιστοιχούν στην έξοδο/τελευταία λέξη, είναι οι 3,6,9,12. Οι υπόλοιπες καταστάσεις αναφέρονται στην κατάσταση είσοδος (πρώτη λέξη της ενότητας).

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0.0145	0.0901	0.0168	0.1139	0.0025	0.1179	0.4000	0.0829	0.0194	0.0351
3	0	0	0	0	0	0	1.0000e-03	0.0070	0	0.9710
4	0	0	0	0	0	0	0	0	0	0
5	6.5684e-04	0.0012	7.4641e-04	0.1661	2.9857e-05	0.0139	0.0876	0.3377	0.0174	0.0839
6	0.0020	0	0.0020	0.1477	0	0.0010	0.0724	0.4593	0.0010	0.0382
7	0	0	0	0	0	0	0	0	0	0
8	0.0065	5.5785e-04	0.2408	0.1227	0	0.1839	0.3091	0.1288	0.0065	0
9	0	0	0.8520	0.1400	0	1.0000e-03	1.0000e-03	0.0060	0	0
10	0	0	0	0	0	0	0	0	0	0
11	0.1547	0.0181	0.0128	0.1372	0.0022	0.1456	0.4984	0.0157	0.0065	0.0011
12	0.2010	0.0270	0.0250	0.1480	0.0020	0.0130	0.5390	0.0180	0.0050	0.0020
13	0	0	0	0	0	0	0	0	0	0
14	0	0.0123	0.0084	0.0302	0.0134	0.0905	0.0966	0.7285	0.0017	0.0039
15	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0.9500	0.0500	0	0	0

	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	1	0	0	0	0	0
2	0.0408	0.0150	0.0280	8.0495e-04	2.2342e+13	0	0	0	0	0
3	1.0000e-03	0.0150	0	0.0050	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0
5	0.2501	0.0244	0.0160	2.5378e-04	0	0	0	0	0	0
6	0.2161	0.0573	0	0.0030	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0
8	7.2520e-04	4.4628e-04	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0
11	0.0059	9.8706e-04	5.7146e-04	1.5585e-04	0	0	0	0	0	0
12	0.0170	0.0020	0	1.0000e-03	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0
14	0.0056	0.0017	0.0034	0.0039	0.1000	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0

Από τους παραπάνω πίνακες φαίνεται ότι δεν έχω χρησιμοποιήσει κατάσταση εξόδου για 2 περιπτώσεις. Οι περιπτώσεις που θεώρησα ότι δεν χρειάζεται κατάσταση για την τελευταία λέξη είναι το distribution και το volume.

Η ενότητα distribution περιέχει αποκλειστικά λέξεις που αναφέρονται σε χώρες είτε είναι συνδετικές λέξεις. Είναι βέβαιο ότι θα τελειώσει η ενότητα με την παρατήρηση 6 , αλλά εφόσον είναι τόσο μικρή η διαφορά (95% για το κύριο μέρος και 100% για την έξοδο) και ο διαχωρισμός της ενότητας ουσιαστικά γίνεται από την διαφορετική συμπεριφορά με τις επόμενες καταστάσεις δεν υπάρχει λόγος να προστεθεί κατάσταση.

Στην ενότητα volume θα μπορούσε να προστεθεί μια extra κατάσταση για την

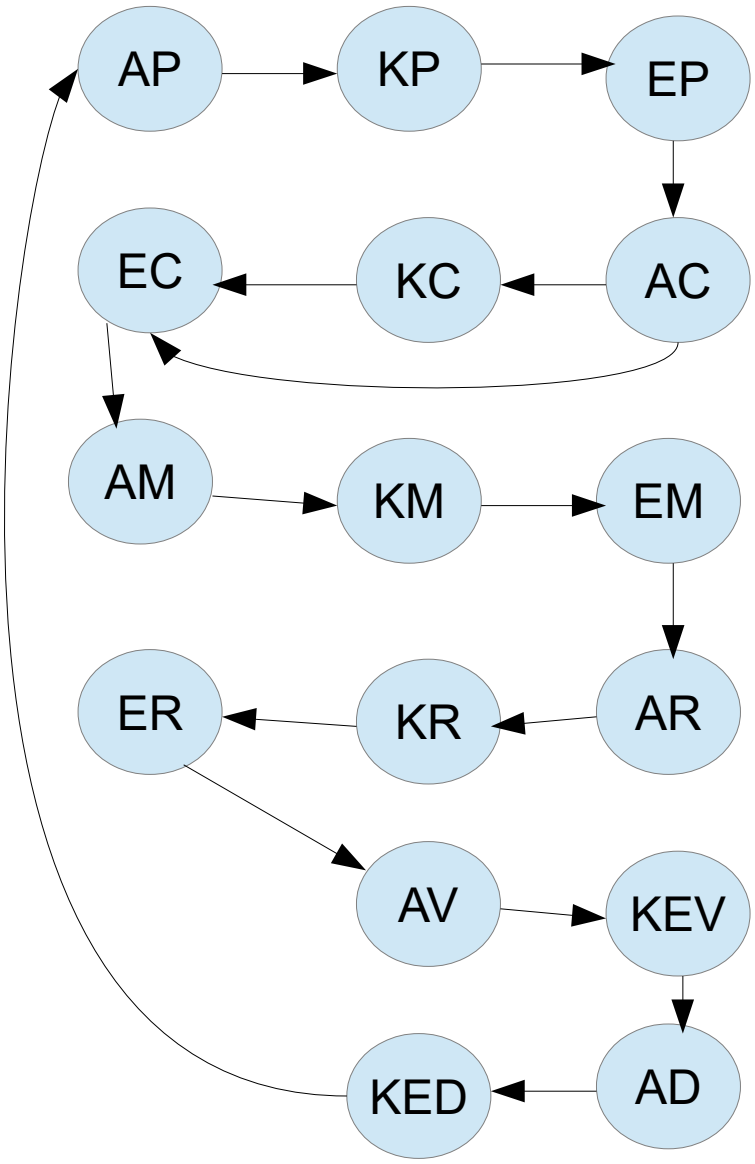
έξοδο (αν και είναι πολύ διαφορετική από τις υπόλοιπες καταστάσεις έχοντας πιθανότητα εμφάνισης αριθμού 72,85%). Λόγω των εξαιρετικών αποτελεσμάτων του πειράματος δεν έκανα όμως αλλαγή.

Στην συνέχεια θα μιλήσουμε για τον υπολογισμό του πίνακα με τις πιθανότητες μετάβασης των καταστάσεων (πίνακας Α). Οι καταστάσεις έτσι όπως εμφανίζονται στις δύο εικόνες αντιστοιχούν στις εξής ενότητες όπως φαίνεται στον παρακάτω πίνακα.

	Είσοδος	Κύριο μέρος	Έξοδος
product	1	2	3
code	4	5	6
reason	7	8	9
manufacturer	10	11	12
distribution	13	14	-
volume	15	16	-

Πριν γίνει ο υπολογισμός του πίνακα Α πρέπει να δοθεί το σχήμα με την αρχιτεκτονική του Hidden Markov Model που επιλέχθηκε για την υλοποίηση της τμηματικής κατάτμησης των κειμένων. Ο γραφικός σχεδιασμός του μοντέλου είναι σημαντικός στην κατανόηση του προβλήματος, επίσης μπορεί να μας βοηθήσει στην ρύθμιση κάποιων παραμέτρων με μη κλασσικό τρόπο (δηλαδή όχι αυτό που ανέφερα στο supervised learning) αλλά με τρόπο που προσπαθεί να προσομοιώσει καλύτερα την σχηματική αναπαράσταση του μοντέλου.

Στο επόμενο διάγραμμα φαίνονται οι πιθανές μεταβάσεις καταστάσεων. Κάθε κατάσταση έχει όνομα στο οποίο το πρώτο γράμμα αναφέρεται στο αν η κατάσταση είναι για την είσοδο (Α) για το κύριο μέρος (Κ) ή την έξοδο (Ε), το επόμενο γράμμα είναι το πρώτο γράμμα της ενότητας την οποία εκφράζει η κατάσταση δηλαδή για το product to P για το code to C κλπ. Οι καταστάσεις που έχουν τρία γράμματα είναι οι καταστάσεις στις οποίες δεν υπάρχει διαχωρισμός τελευταίας λέξης από κύριο μέρος.



Όπως φαίνεται από το διάγραμμα η κατάσταση που εκφράζει το κύριο μέρος της distribution συνδέεται με την κατάσταση εισόδου στην ενότητα product. Αυτό συμβαίνει διότι μπορεί να υπάρχουν παραπάνω recalls για medical devices σε ένα αρχείο επομένως θα πρέπει να μπορεί το μοντέλο να επανέλθει στις καταστάσεις του product για να μπορέσει να συνεχιστεί η κατάτμηση κατά μήκος όλου του κειμένου .

Στην συνέχεια φαίνονται οι συντελεστές του πίνακα με τις πιθανότητες μετάβασης κατάστασης, όπως έχουν υπολογιστεί με την μέθοδο του supervised learning.

	1	2	3	4	5	6	7	8
1	0	1	0	0	0	0	0	0
2	0	0.9804	0.0196	0	0	0	0	0
3	0	0	0	1	0	0	0	0
4	0	0	0	0	1	0	0	0
5	0	0	0	0	0.9855	0.0045	0.0100	0
6	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0.9443
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0
16	0.3066	0	0	0	0	0	0	0

	9	10	11	12	13	14	15	16
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0.0557	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0
11	0	0	0.9481	0.0519	0	0	0	0
12	0	0	0	0	1	0	0	0
13	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0.8241	0.1759	0
15	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	0.6934

Όπως φαίνεται από τους παραπάνω πίνακες οι καταστάσεις που αναφέρονται σε είσοδο ή έξοδο επειδή έχουν μόνο μία παρατήρηση (κατασκευαστικά) έχουν πιθανότητα μετάβασης στην επόμενη κατάσταση 100%. Ο αλγόριθμός ο οποίος

υπολογίζει τις πιθανότητες μετάβασης των καταστάσεων που αναφέρονται στο κύριο μέρος ουσιαστικά μετράει το μέσο μήκος αυτού του μέρους σε παρατηρήσεις και αφαιρεί 1 για να βρει τον μέσο όρο. Το αντίστροφο είναι η πιθανότητα να μεταβούμε στην επόμενη κατάσταση. Είναι πολύ σημαντικό να σημειωθεί ότι η διαδοχή των ενοτήτων με σταθερό τρόπο δηλαδή product->code->manufacturer->reason->volume->distribution->product->.....->distribution είναι καθοριστική στην κατασκευή του μοντέλου και στην ευστοχία του στα αποτελέσματα, αν δεν υπήρχε αυτή η διαδοχή θα έπρεπε να γίνουν αλλαγές στην κατασκευή και να χρησιμοποιηθεί ο Baum Welch αλγόριθμός για εντοπίσει κάποιες συμπεριφορές στην αλληλουχία ενοτήτων.

Επειδή αρχίζουμε πάντα από την ενότητα product ο πίνακας π για την κατανομή πιθανοτήτων αρχικής κατάσταση έχει 1 στην κατάσταση 1.

5.3 ΕΦΑΡΜΟΓΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

5.3.1 ΣΗΜΕΙΩΣΕΙΣ ΠΡΙΝ ΤΗΝ ΧΡΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

Η κατάσταση που περιγράφει το κύριο μέρους του distribution έχει δυνατότητα να εμφανίσει μόνο δύο παρατηρήσεις. Καλό είναι επειδή στα δεδομένα μπορεί να υπάρχει κάποιο σφάλμα (κάποιο σύμβολο) ή κάποια παρατήρηση που να μην υπήρχε στα κείμενα εκπαίδευσης, να προστεθεί μια πολύ μικρή σταθερά στις πιθανότητες εμφάνισης των υπόλοιπων παρατηρήσεων αυτής της κατάστασης έτσι ώστε να μην κολλήσει το μοντέλο (στην συνέχεια φυσικά γίνεται κανονικοποίηση της γραμμής ώστε οι πιθανότητες εμφάνισης όλων των παρατηρήσεων της κατάστασης να έχουν άθροισμα την μονάδα) σε περίπτωση κάποιας άγνωστης παρατήρησης. Επίσης παρατηρήθηκε ότι κάποια κείμενα μπορεί να εμφανίζουν για ένα medical device μόνο product και code για αυτό προστέθηκε η δυνατότητα μετάβασης από την έξοδο του code στην είσοδο του product, η πιθανότητα αυτή είναι αρκετά μικρή και πρέπει να υπάρχει για να μην κολλήσει το μοντέλο σε περίπτωση που ένα recall έχει μόνο product και code.

5.3.2 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Για να ελέγξω την ευστοχία του μοντέλου χρησιμοποίησα κείμενα τα οποία είχαν κυμαινόμενο μήκος και μπορούν να αναφέρονται σε πληθώρα medical devices. Λόγω του ότι στην δομή του κειμένου εμφανίζονται κάποια patterns τα αποτελέσματα είναι αρκετά καλά. Παρ' όλα αυτά επειδή μπορεί να εμφανιστούν κάποια λάθη κατά την εξαγωγή δεδομένων από ένα site, όταν αυτά εμφανίζονται, το μοντέλο αποτυγχάνει. Παρακάτω παραθέτω έναν πίνακα με τα αποτελέσματα του μοντέλου σε 10 αρχεία κυμαινόμενου μήκους. Η πρώτη στήλη αναφέρεται στην ημερομηνία της εμφάνισης της ανάκλησης (recall) της ιατρικής συσκευής στο site του FDA. Η τελευταία στήλη αναφέρεται στον αριθμό των συσκευών που υπάρχουν στο αρχείο.

Αρχείο	Ευστοχία σε αναγνώριση segment	Μήκος αρχείου σε χαρακτήρες	Διαφορετικές συσκευές
02/15/06	100.00%	7292	9
01/25/06	3.50%	4443	28
11/23/05	100.00%	18410	22
01/21/04	100.00%	7794	59
03/24/04	100.00%	5739	20
07/28/04	100.00%	20504	20
02/16/05	100.00%	6947	31
05/18/05	100.00%	4927	8
08/04/04	100.00%	10369	8
12/22/04	100.00%	5863	18

Στο δεύτερο κείμενο το μοντέλο απέτυχε διότι μετά το πρώτο recall ακολουθήθηκε ένα ιδιότυπο pattern που δεν ήταν γνωστό και το μοντέλο δεν μπόρεσε να το ερμηνέυσει σωστά.

5.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα αποτελέσματα αυτού του μοντέλου ήταν σχεδόν άψογα, τα μόνα σφάλματα που εμφανίστηκαν ήταν όταν τα κείμενα που ελέγχθηκαν εμφάνισαν ένα διαφορετικό pattern το οποίο δεν είχε συναντηθεί κατά την εκπαίδευση. Το γεγονός ότι μετά την εμφάνιση κάποιου σφάλματος το μοντέλο θα κάνει λάθη και στην κατάτμηση των επόμενων χωρίων του κειμένου που αναφέρονται σε ιατρικές συσκευές, θα μπορούσε να αντιμετωπιστεί με κατάτμηση των εγγράφων ώστε να περιέχουν μόνο μια ανάκληση (να αναφέρονται μόνο σε μια ιατρική συσκευή), χρησιμοποιώντας τον κατάλληλο delimeter στην TMG (στην προκειμένη περίπτωση το σύμβολο _). Με αυτόν τον τρόπο το κείμενο που εμφάνισε σφάλμα θα είχε χωριστεί σε 28 κείμενα και θα αποτύγχανε μόνο σε 1 κείμενο από τα 28 αντί των 27.

Ο λόγος που το μοντέλο είχε τόσο καλά αποτελέσματα ήταν ότι η δομή των αναφορών ανακλήσεων που δημοσιεύει ο FDA για τις ιατρικές συσκευές μας επέτρεψε να φτιάξουμε ένα εξειδικευμένο μοντέλο που θα εκμεταλλεύεται αυτήν την δομή έτσι ώστε να αναγνωρίζει τις επιμέρους ενότητες που μας ενδιαφέρουν.

Δυστυχώς η εξειδίκευση ενός μοντέλου στην αναγνώριση ενός προτύπου έρχεται με το τίμημα της αδυναμίας γενίκευσης του σε άλλα πρότυπα. Για αυτόν τον λόγο στο επόμενο κεφάλαιο θα αναπτυχθεί ένα λιγότερο εξειδικευμένο μοντέλο, ως μια προσπάθεια ανάπτυξης μιας πιο γενικευμένης τεχνικής, το οποίο θα δοκιμαστεί σε ανακλήσεις διαφορετικού προτύπου που προέρχονται από δύο διαφορετικούς οργανισμούς.

ΚΕΦΑΛΑΙΟ 6

ΓΕΝΙΚΕΥΜΕΝΟΣ ΤΡΟΠΟΣ ΚΑΤΑΤΜΗΣΗΣ ΚΕΙΜΕΝΟΥ ΜΕ ΧΡΗΣΗ HIDDEN MARKOV MODELS

6.1 ΕΙΣΑΓΩΓΗ

Σε αυτό το μέρος διερευνάται η χρήση της μεθοδολογίας των HMMs στο πρόβλημα της κατάτμησης κειμένων, τα οποία υπακούουν σε διαφορετικό πρότυπο, ήτοι στην αναγνώριση του προτύπου των κειμένων ανάκλησης. Πιο συγκεκριμένα, επιχειρείται η ανάπτυξη μοντέλου HMM με την δυνατότητα να αναγνωρίζει το πρότυπο των ανακλήσεων για κείμενα που προέρχονται από τον διαδικτυακό τόπο του FDA και του TGA. Λόγω του ότι το μοντέλο θα πρέπει να συλλαμβάνει τις ιδιομορφίες και άλλων προτύπων καταγραφής κειμένων ιατρικών ανακλήσεων, έχει χρησιμοποιηθεί ένας νέος τρόπος που δεν περιορίζεται από την μορφή προτυποποίησης μιας συγκεκριμένης ιστοσελίδας.

Η διαδικασία περιλαμβάνει δύο βήματα. Το πρώτο βήμα είναι να μελετήσουμε τις ιδιομορφίες των κειμένων του TGA και να φτιάξουμε ένα κάπως γενικότερο μοντέλο που θα μπορεί να πετύχει καλά αποτελέσματα στα κείμενα του. Το δεύτερο βήμα είναι να προσθέσουμε τις καταστάσεις που χρειαζόμαστε για να μπορεί ποιοτικά να κάνει κατάτμηση και σε κείμενα του FDA, να εκπαιδεύσουμε το μοντέλο με ίδιο αριθμό κειμένων και από τους δύο οργανισμούς(ώστε να μην “μεροληπτεί” το μοντέλο προς έναν οργανισμό) και τέλος να αξιολογήσουμε τα αποτελέσματα.

Ο σκοπός αυτού του τμήματος της διπλωματικής δεν είναι να λύσει το πρόβλημα της γενίκευσης του text segmentation άλλα να παρουσιάσει κάποιους τρόπους προσέγγισης του συγκεκριμένου προβλήματος και στην συνέχεια να γίνει ερμηνεία των αποτελεσμάτων.

6.2 ΣΧΕΔΙΑΣΜΟΣ ΓΕΝΙΚΕΥΜΕΝΟΥ ΜΟΝΤΕΛΟΥ ΓΙΑ ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΠΟΥ ΕΦΑΡΜΟΖΕΤΑΙ ΣΤΟΝ TGA

Παρακάτω παραθέτω ένα κείμενο του TGA για ανάκληση ιατρικής συσκευής (το οποίο ακολουθεί ένα γενικό πρότυπο), που θα μας βοηθήσει να καταλάβουμε καλύτερα το πρόβλημα (τις επικεφαλίδες κάθε ενότητας τις έχω γράψει με έντονη γραφή για ευκολία ανάγνωσης).

TGA Recall Reference RC-2012-RN-00920-1 **Product Name/Description** Eclipse Treatment Planning System, (radiation therapy treatment planning system). Versions: 8.9 and 10 Product Code: H48 Multiple serial numbers affected ARTG Number: 119983 Recall Action Level Hospital Recall Action Classification Class II Recall Action Commencement Date 7/09/2012 **Responsible Entity** Varian Medical Systems Australasia Pty Ltd **Reason/Issue** When the clock on the Eclipse client workstation is running behind the clock on the Database/System server, it is possible that a change to the assigned HU value for a structure dose not invalidate the density image stored in the cache, and for the subsequent dose and Monitor Unit calculation to be based on the outdated HU assignment. The resultants MUs may be higher or lower than expected. treatment of the patient using these values can lead to under or overdosing. Recall Action Recall for Product Correction Recall Action Instructions Varian is providing users with work around instructions and will be providing a technical correction to permanently correct the issue when available. Contact Information 1800 657 036 - Varian Oncology Helpdesk \n

Τα κείμενα του TGA παρουσιάζουν μεγάλες διαφορές από τα κείμενα του FDA καθώς τα μέρη που θέλουμε να αναγνωρίσουμε είναι 4. Οι ενότητες που θέλουμε εντοπίσουμε είναι η περιοχή που αναφέρεται στο product , στο reason , στο manufacturer και τέλος σε περιοχή που δεν έχει σχέση με κάποια ενότητα. Σε αντίθεση στα κείμενα του FDA είχαμε 6 περιοχές που η μια διαδεχόταν την άλλη και δεν είχαμε ενδιάμεσες περιοχές κειμένου που δεν ανήκουν σε κάποια περιοχή.

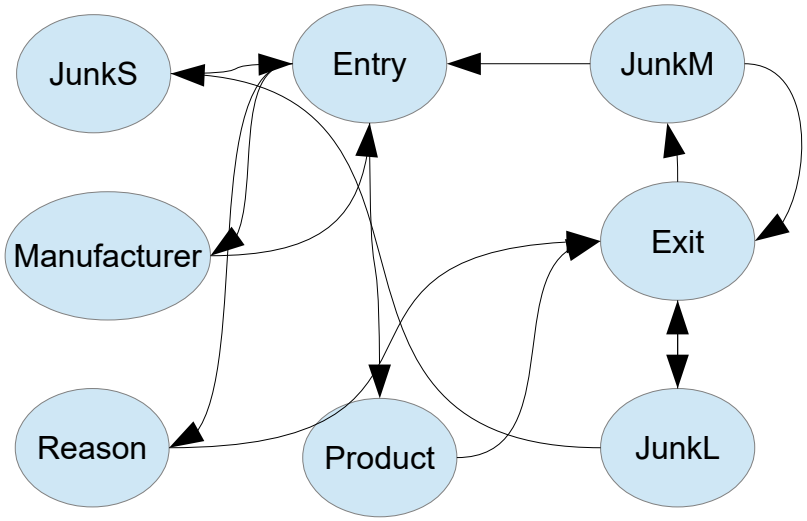
Επιπλέον κάθε ενότητα του TGA έχει διαφορετικές επικεφαλίδες από τις αντίστοιχες του FDA. Η ύπαρξη της επιπρόσθετης περιοχής που αναπαριστά πληροφορίες που δεν μας ενδιαφέρουν καθώς και η έλλειψη των περιοχών code,volume και distribution καθιστούν την αλληλουχία ενοτήτων που είχαμε θεωρήσει στο Hidden Markov Model για την περίπτωση του FDA μη λειτουργική για την παρούσα περίπτωση. Τέλος η σύσταση λέξεων κάθε περιοχής αλλάζει από τις αντίστοιχες του FDA από μικρό έως πολύ μεγάλο βαθμό αλλά αυτό θα φανεί πιο καθαρά στην ανάλυση που ακολουθεί.

Οι προαναφερθείσες διαφορές με ανάγκασαν να ακολουθήσω έναν αρκετά διαφορετικό τρόπο ανάπτυξης του μοντέλου, το οποίο δεν θα επηρεάζεται τόσο πολύ από την αλληλουχία των ενοτήτων όπως και από συγκεκριμένες λέξεις κλειδιά που σημαίνουν την έναρξη κάποιας ενότητας. Επίσης η ύπαρξη κειμένου που πρέπει να αφαιρεθεί από το μοντέλο (ότι κείμενο δεν ανήκει σε καμία ενότητα ή αλλιώς junk) εισάγει στο μοντέλο μια νέα παράμετρο όπου δεν υπήρχε νωρίτερα. Παρ' όλα αυτά η γενική ιδέα αντιμετώπισης αυτών των προβλημάτων, για δημιουργία συνόλων με λέξεις που εμφανίζονται πιο συχνά σε συγκεκριμένα μέρη του κειμένου ισχύει και θα χρησιμοποιηθεί και εδώ.

Καταρχήν εκμεταλλεύτηκα το γεγονός ότι οι λέξεις που εμφανίζονται στις επικεφαλίδες κάθε ενότητας δεν εμφανίζονται σε άλλα μέρη του κειμένου (η εμφανίζονται πολύ σπάνια). Συνεπώς η εμφάνιση μιας λέξης που ανήκει σε επικεφαλίδα σημαίνει την έναρξη μιας ενότητας. Επίσης όπως παρατηρείται στο κείμενο του TGA που παρέθεσα, κάθε περιοχή junk κειμένου αρχίζει με τις λέξεις Recall Action . Δημιούργησα δύο πίνακες ο ένας περιέχει όλες τις λέξεις πριν από κάθε ενότητα(Product , Name/Description , Responsible , Entity και reason/issue) και ο άλλος τις λέξεις πριν από “μή-ενότητα” ή το junk (Recall και Action). Ο πρώτος πίνακας έτσι περιέχει τις “entry words” και ο δεύτερος τις “exit words”. Οι λέξεις που ανήκουν στους δύο πίνακες θα αναπαριστούν δύο διαφορετικές παρατηρήσεις (σε αντίθεση με το μοντέλο του FDA που είχαμε 6 διαφορετικές παρατηρήσεις , που σηματοδοτούσαν την έναρξη κάθε ενότητας).

Αν λοιπόν αυτές οι παρατηρήσεις ακολουθούνται από αλληλουχία παρατηρήσεων που διαφέρουν από ενότητα σε ενότητα μπορούμε να πετύχουμε ένα καλό text

segmentation. Στην επόμενη σελίδα θα παραθέσω το σχήμα του Hidden Markov Model για τον TGA και θα εξηγήσω πως κατέληξα σε αυτήν την σχεδίαση.



Καταρχήν θα εξηγήσω τι αναπαριστά η κάθε κατάσταση :

Κατάσταση Entry : Περιέχει παρατηρήσεις που μπορεί να είναι από λέξεις ανήκουν στον πίνακα entry words.

Κατάσταση Exit : Περιέχει παρατηρήσεις που μπορεί να είναι από λέξεις που ανήκουν στον πίνακα exit words.

Κατάσταση Reason : Περιέχει τις παρατηρήσεις που έχουν προκύψει από την ανάλυση 200 κειμένων που αναφέρονται στο reason.

Κατάσταση Product : Περιέχει τις παρατηρήσεις που έχουν προκύψει από την ανάλυση 200 κειμένων που αναφέρονται στο product.

Κατάσταση manufacturer : Περιέχει τις παρατηρήσεις που έχουν προκύψει από την ανάλυση 200 κειμένων που αναφέρονται στο manufacturer.

Καταστάσεις Junk : Αυτές οι καταστάσεις είναι τρεις και όχι μια , διότι τα μέρη του κειμένου που πρέπει να πετάξουμε δεν έχουν πάντα το ίδιο μήκος, αυτές οι 3 καταστάσεις ουσιαστικά είναι τρεις τρόποι που μπορεί να εμφανίζεται το junk. Αν παρατηρήσουμε το κείμενο του TGA θα δούμε ότι πρέπει να αφαιρέσουμε κείμενο στην αρχή (πριν το product), αμέσως μετά το product και πριν το manufacturer και τέλος μετά το reason όπως παρατηρούμε καθένα από αυτά τα τμήματα έχει διαφορετικό μήκος και συνδέεται με διαφορετικές καταστάσεις. Για λόγους γενικότητας αυτές οι καταστάσεις έχουν θεωρηθεί να έχουν παρόμοια σύσταση και δεν έχουν χρησιμοποιηθεί λέξεις που μπορεί να τις χαρακτηρίζουν (π.χ. τόσο η πρώτη όσο και η δεύτερη σχεδόν σε κάθε περίπτωση περιέχουν τις ίδιες λέξεις). Για να βρω τις συχνότητες εκπομπής παρατηρήσεων ανέλυσα από 20 κείμενα τις συχνότητες εμφάνισης τους και προσάρμοσα τις πιθανότητες μεταβάσεις στο μήκος κάθε τμήματος. Το τελευταίο γράμμα κάθε junk κατάστασης συμβολίζει το μήκος της (Long, Medium και Short).

Όπως ανέφερα νωρίτερα δεν ήθελα να δώσω με αυτό το μοντέλο μεγάλη βαρύτητα στην αλληλουχία των ενοτήτων για αυτό δημιούργησα μια γενική κατάσταση η οποία σηματοδοτεί την έναρξη κάποιας ενότητας. Όπως βλέπουμε όλες οι ενότητες που θέλουμε να εντοπίσουμε συνδέονται με την κατάσταση entry. Η κατάσταση exit επίσης σηματοδοτεί την έναρξη ενότητας αλλά επειδή χαρακτηρίζει τις ενότητες που δεν μας ενδιαφέρουν ονομάζεται exit. Όπως φαίνεται από το σχήμα όλες οι ενότητες junk συνδέονται με αυτήν την κατάσταση. Η κατάσταση JunkL επειδή βρίσκεται στο τέλος κάθε κειμένου νοηματικά πρέπει να συνδεθεί με την κατάσταση JunkS που βρίσκεται στην αρχή (για την περίπτωση που εφαρμόζουμε το μοντέλο σε κείμενο που αναφέρεται σε 2 ή παραπάνω ανακλήσεις ιατρικών συσκευών). Επειδή υπάρχει η περίπτωση αυτές οι καταστάσεις να εμφανίσουν exit παρατήρηση έχει προστεθεί η δυνατότητα μετάβασης στην κατάσταση exit.

6.2.1 ΥΠΟΛΟΓΙΣΜΟΙ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΜΟΝΤΕΛΟΥ

Το σύνολο των συμβόλων παρατηρήσεων που χρησιμοποιήθηκε και σε αυτήν την περίπτωση είναι το σύνολο του συμβόλου παρατηρήσεων του μοντέλου που είχε παρουσιασθεί στο προηγούμενο κεφάλαιο , η διαφορά είναι ότι οι παρατηρήσεις που αντιστοιχούν στους αριθμούς 15-20 δεν χρησιμοποιούνται ενώ η παρατήρηση 21 αναφέρεται σε exit words και η 22 σε entry words. Όπως θα φανεί στην συνέχεια στους πίνακες που παρουσιάζονται οι πιθανότητες εμφάνισης παρατηρήσεων , οι παρατηρήσεις 15-20 παρόλο που δεν χρησιμοποιούνται, έχουν τιμές αλλά αυτές είναι πολύ μικρές (10^{-15}) διότι έχει προστεθεί μια πολύ μικρή σταθερά στον πίνακα παρατηρήσεων. Η προσθήκη της σταθεράς είναι απαραίτητη, προκειμένου να μην υπάρχουν 0 σε κάποιες περιπτώσεις που οφείλονται καθαρά σε ελλειπή εκπαίδευση.

Για την ανάλυση των ενοτήτων product, reason , manufacturer και junk. Χρησιμοποιήθηκαν 200 κείμενα για κάθε ενότητα από εκεί κατά τα γνωστά βγήκαν οι πιθανότητες εμφάνισης των παρατηρήσεων. Για τις ενότητες entry και exit επειδή αυτές καλούνται να χωρίζουν τμήματα του κειμένου κατασκευαστικά έχουν φτιαχτεί ώστε να έχουν πιθανότητα 100% εμφάνισης της αντίστοιχης παρατήρησης. Οι πιθανότητες μετάβασης κατάστασης έχουν υπολογιστεί από τον μέσο όρο του μήκους κάθε ενότητας όπως και στην περίπτωση του FDA. Επειδή τα κείμενα του TGA αρχίζουν πάντα από την κατάσταση JunkS ο πίνακας π(της αρχικής κατανομής) αντιστοιχεί πιθανότητα 100% σε αυτήν την κατάσταση.

Παρακάτω φαίνεται ο πίνακας με τις πιθανότητες εμφάνισης παρατηρήσεων για όλες τις καταστάσεις οι οποίες αντιστοιχούν στις καταστάσεις του γράφου ως εξής.

Κατάσταση 1 -> product

Κατάσταση 2 -> manufacturer

Κατάσταση 3 -> reason

Κατάσταση 4 -> Entry

Κατάσταση 5 -> Exit

Κατάσταση 6 -> JunkS

Κατάσταση 7 -> JunkM

Κατάσταση 8 -> JunkL

	1	2	3	4	5	6	7	8	9	10	11
1	0.0103	0.0208	0.0264	0.1805	9.3738e-04	0.0750	0.4408	0.1650	0.0126	0.0118	0.0486
2	0.0117	0	0.3431	0.0105	0	0.1645	0.4609	0	0.0070	0	0.0023
3	0.1169	0.0153	0.0192	0.1287	0.0015	0.1388	0.5469	0.0155	0.0084	0.0021	0.0055
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	9.9998e-07	9.9998e-07	9.9998e-07
7	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	9.9998e-07	9.9998e-07	9.9998e-07
8	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	9.9998e-07	9.9998e-07	9.9998e-07

	12	13	14	15	16	17	18	19	20	21	22
1	0.0069	3.7495e-04	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	7.0264e-04	3.5132e-04	7.0264e-05	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	1	0
6	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	0	0
7	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	0	0
8	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	9.9998e-07	0	0

Συνεχίζουμε με τον πίνακα πιθανότητας μετάβασης κατάστασης:

	1	2	3	4	5	6	7	8
1	0.9600	0	0	0	0.0400	0	0	0
2	0	0.3330	0	0.6670	0	0	0	0
3	0	0	0.9860	0	0.0140	0	0	0
4	0.1110	0.1110	0.1110	0.6670	0	0	0	0
5	0	0	0	0	0.6670	0.3033	0.0303	0.0030
6	0	0	0	0.1000	0.1000	0	0	0.8000
7	0	0	0	0.0800	0.0100	0	0	0.9100
8	0	0	0	0	1.0000e-03	0.0399	0	0.9600

Επειδή υπάρχει μια πολύ μικρή πιθανότητα να εμφανιστεί στην JunkL παρατήρηση που ανήκει σε Exit έχει προστεθεί δυνατότητα μετάβασης στην κατάσταση Exit της τάξης του 10^{-3} .

Όπως φαίνεται από τους δύο πίνακες το μόνο που αλλάζει στις καταστάσεις junk είναι οι δυνατές μεταβάσεις ,και η πιθανότητα παραμονής σε κάθε τέτοια κατάσταση (δηλαδή το μήκος της).

Πριν την χρησιμοποίηση του μοντέλου είναι καλό να προστεθεί μια πολύ μικρή σταθερά (πρόσθεσα σταθερά 10^{-15}) διότι κάποιες καταστάσεις μπορεί να εμφανίσουν με πολύ μικρή συχνότητα παρατηρήσεις που δεν έχουν συναντηθεί στα κείμενα εκπαίδευσης. Σε αυτήν την περίπτωση το μοντέλο κολλάει αν δεν γίνει η πρόσθεση της σταθεράς. Η σταθερά πρέπει να είναι πολύ μικρή ώστε να μην αλλάζει η δομή του μοντέλου και με αυτόν τον τρόπο να ακολουθούνται σπάνιες και μη εγκυρες διαδρομές μέσα στις καταστάσεις του μοντέλου, όταν αυτό συναντήσει για πρώτη φορά μια παρατήρηση σε μια ακολουθία παρατηρήσεων.

6.2.2 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Το μοντέλο αυτό δοκιμάστηκε σε 20 κείμενα του TGA(τα κείμενα αυτά έχουν ονομασίες στα δεδομένα που παραθέτονται με την διπλωματική και έχουν κατάληξη από 000A9 ως 000CF). Σε κάθε κείμενο το μοντέλο κατάφερε να εντοπίσει τις ενότητες.

Πρέπει να σημειωθεί ότι επειδή entry words από την ενότητα product εμφανίζονται και μετά αυτής σε κάποια κείμενα έχει προστεθεί μια extra λέξη για αυτήν την ενότητα παρ' όλα αυτά οι υπόλοιπες λέξεις είναι σωστές και θεωρώ ότι η εισαγωγή μιας extra λέξεις σε ένα χωρίο που έχει κατά μέσο όρο 25 λέξεις είναι ασήμαντη και θα μπορούσε εύκολα να αφαιρεθεί με κάποιο πρόγραμμα.

6.3 ΕΠΕΚΕΤΑΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΓΙΑ ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΣΕ FDA - TGA

6.3.1 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ

Μετά την δημιουργία ενός γενικότερου Hidden Markov Model που πέτυχε καλά αποτελέσματα σε κείμενα για ανακλήσεις ιατρικών συσκευών από τον διαδικτυακό TGA τώρα επιχειρώ να δοκιμάσω την σχεδίαση αυτού του μοντέλου και σε κείμενα κάποιου άλλου οργανισμού που ακολουθούν ένα διαφορετικό πρότυπο μορφοποίησης. Σε αυτό το μέρος θα χρησιμοποιήσω την προηγούμενη τεχνική, στα κείμενα του FDA που είχαμε δει στην προηγούμενη ενότητα άλλα το μοντέλο θα εκπαιδευτεί εξίσου με κείμενα και των δύο οργανισμών.

Είναι σαφές ότι επειδή αναζητούμε την γενίκευση θα πρέπει ο κορμός του μοντέλου να παραμείνει σταθερός (έτσι έχει κατασκευαστεί άλλωστε ώστε να μην λαμβάνει ιδιαίτερα υπόψιν αλληλουχίες ενοτήτων όπως στην περίπτωση του πρώτου Hidden Markov Model).

Παρατηρώντας τα κείμενα του FDA, διαπιστώνουμε ότι υπάρχουν άλλες 3 ενότητες που δεν υπάρχουν στα κείμενα του TGA αυτές είναι οι ενότητες code,volume και distribution. Επίσης υπάρχουν και οι ενότητες product , reason και manufacturer άλλα έχουν κάπως διαφορετική σύσταση λέξεων από αυτές του TGA. Συνεπώς το προηγούμενο μοντέλο θα πρέπει να προσαρμοστεί ώστε να μπορεί να εντοπίσει αυτές τις 3 νέες ενότητες άλλα και να προσαρμόσει τις ήδη υπάρχουσες καταστάσεις του ώστε να εντοπίζουν και κείμενα του FDA.

Επειδή και στην περίπτωση του μοντέλου για τον FDA είχε γίνει στατιστική ανάλυση των παρατηρήσεων σε κάθε ενότητα (οι καταστάσεις που αφορούν το κύριο μέρος του κειμένου κάθε ενότητας) οι καταστάσεις product , reason και manufacturer έχουν προσαρμοστεί ώστε να λαμβάνουν τις παρατηρήσεις κάθε οργανισμού με βάρος 50%. Δηλαδή για κάθε παρατήρηση του reason η πιθανότητα εμφάνισης της είναι $\Pi_T = 0.5x\Pi_{FDA} + 0,5x\Pi_{TGA}$. Εξαίρεση αποτελεί η κατάσταση του product του FDA διότι εμφανίζει την λέξη Recall (έχει σημασία ο κεφαλαίος χαρακτήρας) η οποία κατατάσσεται στην νέα παρατήρηση 21, για αυτόν τον λόγο ξανά

υπολογίστηκαν οι πιθανότητες παρατηρήσεων της product. Τα νέα αποτελέσματα φαίνονται στην επόμενη σελίδα :

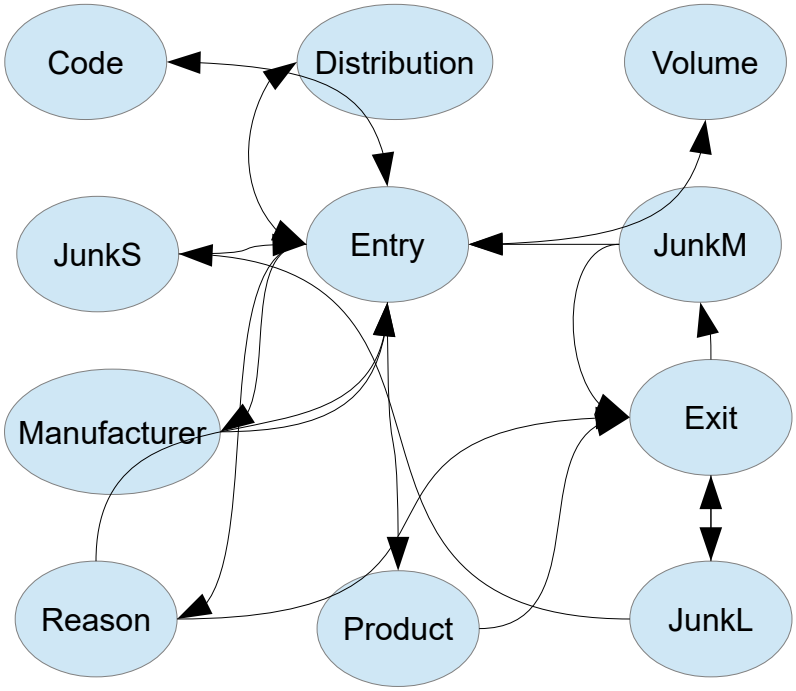
	1	2	3	4	5	6	7	8	9	10	11
1	0.0172	0.0506	0.0229	0.1112	0.0030	0.1178	0.4034	0.0717	0.0181	0.0519	0.0544

	12	13	14	15	16	17	18	19	20	21	22
1	0.0086	0.0263	4.6512e-04	0	0	0	0	0	0	0.0422	0

Όπως φαίνεται το 4.22% των παρατηρήσεων της product του FDA έχει αλλάξει καθώς από την παρατήρηση 7 έχουν μεταφερθεί στην παρατήρηση 21, στην συνέχεια οι πιθανότητες εκπομπής της μικτής κατάστασης TGA -FDA για product υπολογίζεται κατά τα γνωστά.

Για τις νέες ενότητες οι πιθανότητες εμφάνισης παρατηρήσεων είναι οι ίδιες με αυτές του FDA εφόσον δεν εμφανίζονται στον TGA.
Ο πίνακας Entry words έχει αλλάξει και συμπεριλαμβάνει τις λέξεις ετικέτας των κειμένων του FDA (PRODUCT , CODE , REASON , RECALLING , FIRM/MANUFACTURER , VOLUME , OF , IN , COMMERCE , DISTRIBUTION).
Ο πίνακας με τα Exit words παραμένει ο ίδιος.
Κάθε νέα ενότητα θα πρέπει να συνδέεται με την κατάσταση Entry και θα πρέπει να καταλήγει πάλι εκεί έτσι ώστε να μπορεί να αναγνωριστεί η επόμενη ενότητα.

Το σχήμα του νέου μοντέλου φαίνεται παρακάτω



Όπως φαίνεται η μόνη σχηματική αλλαγή που κάναμε από το μοντέλο του TGA είναι η πρόσθεση πιθανότητας μετάβασης από το Reason στο Entry. Επίσης επειδή στο μοντέλο του FDA η πρώτη κατάσταση είναι πάντα το product, ο νέος πίνακας της αρχικής κατανομής θα έχει 50% στην κατάσταση JunkS και 50% στην κατάσταση Entry. Κατά τα άλλα οι επιπρόσθετες καταστάσεις συνδέονται μόνο με την κατάσταση Entry.

Στην συνέχεια φαίνεται ο πίνακας με τις πιθανότητες εμφάνισης παρατηρήσεων σε κάθε κατάσταση (έχει προστεθεί μικρή σταθερά για τους γνωστούς λόγους)
Κατάσταση 1 -> product
Κατάσταση 2 -> manufacturer
Κατάσταση 3 -> reason
Κατάσταση 4 -> Entry
Κατάσταση 5 -> Exit

Κατάσταση 6 -> JunkS

Κατάσταση 7 -> JunkM

Κατάσταση 8 -> JunkL

Κατάσταση 9 -> code

Κατάσταση 10 -> volume

Κατάσταση 11-> distribution

	1	2	3	4	5	6	7	8	9	10	11
1	0.0124	0.0555	0.0216	0.1472	0.0017	0.0964	0.4104	0.1240	0.0160	0.0235	0.0447
2	0.0091	2.7892e-04	0.2919	0.0666	5.0000e-15	0.1742	0.3850	0.0644	0.0067	5.0000e-15	0.0015
3	0.1358	0.0167	0.0160	0.1330	0.0018	0.1422	0.5227	0.0156	0.0074	0.0016	0.0057
4	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
5	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
6	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	1.0000e-14	1.0000e-14	1.0000e-14
7	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	1.0000e-14	1.0000e-14	1.0000e-14
8	0.0822	9.9998e-07	0.0274	0.2329	9.9998e-07	0.1096	0.4520	0.0959	1.0000e-14	1.0000e-14	1.0000e-14
9	6.5684e-04	0.0012	7.4641e-04	0.1661	2.9857e-05	0.0139	0.0876	0.3377	0.0174	0.0839	0.2501
10	1.0000e-14	0.0123	0.0084	0.0302	0.0134	0.0905	0.0966	0.7285	0.0017	0.0039	0.0056
11	2.8785e-04	8.1558e-04	0.0041	0.0021	1.0000e-14	0.9557	0.0344	8.6356e-04	2.8785e-04	6.2368e-04	3.3583e-04

	12	13	14	15	16	17	18	19	20	21	22
1	0.0109	0.0142	4.0247e-04	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	0.0211	1.0000e-14
2	2.2314e-04	5.0000e-15	5.0000e-15	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
3	8.4485e-04	4.6139e-04	1.1306e-04	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
4	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000
5	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000	1.0000e-14
6	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
7	1.0000e-14	9.9998e-07	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
8	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
9	0.0244	0.0160	2.5378e-04	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
10	0.0017	0.0034	0.0039	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14
11	9.5951e-05	3.8380e-04	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14	1.0000e-14

Ο πίνακας για τις πιθανότητες μετάβασης κατάστασης

	1	2	3	4	5	6	7	8	9	10	11
1	0.9740	0	0	0.0130	0.0130	0	0	0	0	0	0
2	0	0.9141	0	0.0430	0.0430	0	0	0	0	0	0
3	0	0	0.9790	0.1050	0.0105	0	0	0	0	0	0
4	0.0555	0.0555	0.0555	0.6670	0	0	0	0	0.0555	0.0555	0.0555
5	0	0	0	0	0.6670	0.3033	0.0303	0.0030	0	0	0
6	0	0	0	0.1000	0.1000	0	0	0.8000	0	0	0
7	0	0	0	0.0800	0.0100	0	0	0.9100	0	0	0
8	0	0	0	0	1.0000e-03	0.0399	0	0.9600	0	0	0
9	0	0	0	0.0145	0	0	0	0	0.9855	0	0
10	0	0	0	0.4064	0	0	0	0	0.0594	0	0
11	0	0	0	0.3066	0	0	0	0	0.6934	0	0

6.3.2 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

TGA:

Για έλεγχο χρησιμοποίησα τα ίδια κείμενα που είχα χρησιμοποιήσει στο πρώτο μοντέλο. Σε κείμενα που αποτελούνται από πολλά recalls η ενότητα product με αυτό το μοντέλο ανιχνεύει και άλλες περιοχές, κυρίως όμως junk ως product λόγω της εισαγωγής της παρατήρησης 21 στην κατάσταση product. Η περιοχή manufacturer έχει καλύτερη ευστοχία από ότι η product αλλά κάποιες λέξεις που πριν ανιχνεύονταν ως manufacturer τώρα μπορεί να εντοπιστούν στην reason ή στην product. Η κατάσταση reason χάνει κάποια κομμάτια κειμένου επίσης από την product. Σε κείμενα που αποτελούνται από μόνο μια ανάκληση ιατρικής συσκευής τα αποτελέσματα είναι παρόμοια με του πρώτου μοντέλου(το οποίο όμως δεν είχε πρόβλημα σε κείμενα με πολλές ανακλήσεις)

FDA:

Στον FDA τα κείμενα που εξετάστηκαν ήταν λιγότερα και αφορούσαν μόνο μια ανάκληση. Για τον FDA οι ενότητες distribution και volume σχεδόν δεν ανιχνεύονταν καθόλου, η ενότητα product μπορούσε να ανιχνεύσει την περιοχή της αλλά σε αυτήν εισχωρούσαν και μέρη του κειμένου που άνηκαν σε άλλες περιοχές. Οι περιοχές reason και manufacturer μπορούσαν να εντοπιστούν σε κάποιο βαθμό αλλά ποτέ σε όλη την έκταση της ενότητας.

6.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η κατάτμηση κειμένου για πολλαπλά πρότυπα δεν είναι εύκολο να αναλυθεί με έναν γενικό τρόπο για αυτόν τον λόγο τα αποτελέσματα δεν ήταν καλά στην γενική μέθοδο κατάτμησης κειμένου. Το γεγονός ότι η γενική μέθοδος στον TGA έδωσε σχεδόν άριστα αποτελέσματα (το πρώτο μοντέλο) για κείμενα που αναφέρονταν ακόμα και σε 20 συσκευές, είναι ενθαρρυντικό για πιθανές τροποποιήσεις στο μοντέλο ώστε να μπορεί να συμπεριλαμβάνει και πρότυπα άλλων οργανισμών.

Πιθανώς θα πρέπει να εξεταστούν άλλες διατάξεις καταστάσεων για να εκφράσουν ενότητες (όχι μόνο μια κατάσταση για κάθε ενότητα) αλλά οι συνδυασμοί που πρέπει να ελεγχθούν είναι πάρα πολλοί.

Καλό θα ήταν να γίνει έλεγχος για παράλληλες καταστάσεις που μπορούν να εκφράζουν μια ενότητα οι οποίες θα προσπαθούν να ανιχνεύσουν καλύτερα ενότητες που διαφέρουν πολύ από το μέσο μήκος μιας ενότητας. Για παράδειγμα ενώ η ενότητα code στον FDA έχει μέσο μήκος περίπου 69 υπάρχουν αρκετά κείμενα στα οποία το code έχει μόνο μια λέξη (το none) ενώ υπάρχουν και κείμενα τα οποία έχουν τόσο μεγάλο μήκος τα οποία αυξάνουν τον μέσο όρο λέξεων το οποίο όμως δεν ανταποκρίνεται στα περισσότερα κείμενα. Μπορεί να πρέπει να γίνουν στατιστικές αναλύσεις για τα μήκη των κειμένων και οι συντελεστές μετάβασης κατάστασης να μην υπολογίζονται από τους τύπους του supervised learning για Hidden Markov Models, αλλά να φιλτράρονται αρχεία με πολύ μεγάλο μήκος (η πολύ μικρό σε άλλες περιπτώσεις) εκτός εκπαίδευσης για να μην χαλάνε το μοντέλο.

Επίσης η κατάσταση entry μπορεί να χωριστεί σε περισσότερες καταστάσεις ώστε να υπάρχει πιο σαφής διαχωρισμός μεταξύ ενοτήτων. Η αντίθετη περίπτωση για τον FDA με 6 καταστάσεις “Entry” δούλευε άριστα.

Ένα άλλο πρόβλημα είναι ότι οι ενότητες που μπορεί να αναφέρονται για παράδειγμα στο product μπορεί να έχουν ίδιο όνομα άλλα έχουν διαφορετικό περιεχόμενο λέξεων. Αυτή η αλλαγή σύστασης λέξεων σε ενότητες με ίδιο όνομα που βρίσκονται σε διαφορετικούς οργανισμούς μπορεί να απαιτεί ενδεχομένως την ύπαρξη παράλληλων καταστάσεων (τόσες καταστάσεις όσες και οι οργανισμοί) για να εκφραστεί καλύτερα η διαφορά σύστασης των ενοτήτων από οργανισμό σε

οργανισμό. Αυτό βέβαια μπορεί να εισάγει άλλα προβλήματα καθώς αν μια ενότητα με διαφορετικό όνομα από τον οργανισμό Α πχ μοιάζει με διαφορετική ενότητα από τον οργανισμό Β τότε μπορεί να γίνεται σύγχυση αυτών των ενοτήτων.

Πρέπει να είναι σαφές ότι παρότι αυτό το γενικό μοντέλο δεν μπόρεσε να κάνει καλή κατάτμηση κειμένου, ένα μοντέλο το οποίο είναι σχεδιασμένο να κάνει κατάτμηση στον FDA και στον TGA μπορεί να κατασκευαστεί, αλλά φυσικά θα έχει πρόβλημα με την εισαγωγή νέων κειμένων από άλλους οργανισμούς. Το πρόβλημα θα υπήρχε αν χρειαζόμασταν να σχεδιάσουμε μοντέλο που κάνει text segmentation σε 10 οργανισμούς επειδή εκεί λόγο του πλήθους διαφορετικών προτύπων είναι δύσκολη ακόμα και η ειδική λύση.

Μια ενδεχομένως καλή λύση του προβλήματος του text segmentation που ακολουθούν ένα γενικό πρότυπο (δηλαδή αναφέρονται σε ιατρικές συσκευές) μπορεί να ήταν η δημιουργία ενός classifier ο οποίος μπορεί να κατατάξει το κείμενο σε έναν οργανισμό και μετά να χρησιμοποιηθεί ένα μοντέλο κατάτμησης κειμένου για αυτόν τον οργανισμό. Η δημιουργία Hidden Markov Model για να κάνει κατάτμηση κειμένου σε έναν οργανισμό είναι αρκετά απλή, επομένως η μόνη δυσκολία θα ήταν στην ευστοχία του classifier. Παρ' όλα αυτά ακόμα και σε περίπτωση που ο classifier έχει πρόβλημα στο να ξεχωρίσει αν κάποιο κείμενο ανήκει ανάμεσα σε λίγους οργανισμούς. Μπορεί να είναι δυνατή η δημιουργία ενός εξειδικευμένου μοντέλου που θα κάνει κατάτμηση κειμένου σε αυτούς τους οργανισμούς.

Το πρόβλημα της κατάτμησης κειμένου για την περίπτωση των ιατρικών συσκευών κατά την γνώμη μου δεν είναι δύσκολο να λυθεί με πολλά εξειδικευμένα μοντέλα. Αλλά ένα γενικό μοντέλο το οποίο εύκολα θα μπορεί να ανιχνεύσει διαφορετικές ενότητες ανάμεσα στα διαφορετικά πρότυπα των οργανισμών είναι πολύ δύσκολο να δημιουργηθεί.

ΚΕΦΑΛΑΙΟ 7

ΣΥΝΟΨΗ

Αυτό που επιχειρήθηκε κατά την φάση πραγματοποίησης της διπλωματικής εργασίας είναι, έχοντας υλοποιήσει μια πολύ αποδοτική, βάσει των αποτελεσμάτων, τεχνική αναγνώρισης προτύπου σε κείμενα για τους επιμέρους οργανισμούς, να δοκιμαστεί η κατασκευή ενός μοντέλου αναγνώρισης προτύπου για κείμενα οποιουδήποτε προτύπου και να αναδειχθεί το συγκεκριμένο πρόβλημα. Το εγχείρημα αυτό, που υπάγεται στην γενικότερη κατασκευή ενός μοντέλου ικανού να αναπαριστά ένα οποιοδήποτε σύστημα παρουσιάζει μια παρεμφερή κανόνικότητα με τα άλλα συστήματα, ξεπερνώντας τις μορφολογικές ιδιαιτερότητες που καθένα διατηρεί, είναι ομολογουμένως ένα από τα δύσκολοτερα ερευνητικά θέματα, αλλά και από τα πιο ενδιαφέροντα για να μελετήσει κανείς.

Συνοψίζοντας τα αποτελέσματα των πειραμάτων αυτής της διπλωματικής είναι εμφανές ότι τα Hidden Markov Models μπορούν να χρησιμοποιηθούν για την κατασκευή ενός δυαδικού ταξινομητή όπως αυτού του κεφαλαίου 4 με πολύ καλά αποτελέσματα. Το γεγονός ότι οι δυαδικοί ταξινομητές μπορούν να εφαρμοστούν σε σε πληθώρα προβλημάτων καθιστά την μέθοδο που αναπτύχθηκε χρήσιμη και για την ανάπτυξη άλλων εφαρμογών πέραν της συγκεκριμένης.

Η διαδικασία κατασκευής μοντέλου που να μπορεί να εντοπίζει ενότητες σε κείμενα οργανισμών με διαφορετικά πρότυπα παρότι δεν είχε καλά αποτελέσματα, αν περιορισθεί σε έναν οργανισμό έχει εξαιρετικά αποτελέσματα. Το γεγονός ότι αυτή η μέθοδος μπορεί πολύ εύκολα να αυτοματοποιηθεί και να παράγει “εύστοχα” μοντέλα για κάθε οργανισμό ξεχωριστά σημαίνει ότι μπορεί να αξιοποιηθεί αν και δεν είναι η βέλτιστη λύση. Όπως με την περίπτωση του δυαδικού ταξινομητή, αυτή η μέθοδος, με μικρές αλλαγές, μπορεί να εφαρμοστεί στην αναγνώριση ενοτήτων και άλλων ειδών κειμένων πέραν αυτών που αναφέρονται σε ανακλήσεις ιατρικών συσκευών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1].Τεχνικές μείωσης διαστάσεων Σ. Φωτόπουλος – Α. Μακεδόνας (2008)

[2]. Exploring Non-negative Matrix Factorization Holly Jin, LinkedIn Corp, Workshop on Algorithms for Modern Massive Data Sets Stanford University, Ιούνιος 25–28, 2008

[3]. Ο διαδικτυακός χώρος της εργαλειοθήκης TMG http://scgroup20.ceid.upatras.gr:8000/tmg/

[4]. A tutorial on hidden markov models and selected applications in speech recogntion. LAWRENCE R. RABINER, proceedings of the IEEE, vol 77, no. 2 Φεβρουάριος 1989

[5]. Text mining Ian H. Witten (2005), *Practical handbook of internet computing*, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida.

[6]. http://text-analysis.sourceforge.net/practical-applications

[7]. Automatic segmentation of text into structured records. Sunita sarawagi, Vinayak Borkar, Kaustubh Deshmukh. SIGMOD conference 2001

[8]. Ο διαδικτυακός χώρος της εργαλειοθήκης HMM murphy´s toolbox : https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[9]. Η ιστοσελίδα του TGA **:** https://www.tga.gov.au/

[10]. Η ιστοσελίδα του FDA **:** http://www.fda.gov/

[11]. Η ιστοσελίδα των healthy canadians **:** http://healthycanadians.gc.ca/index-eng.php