ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σημασιολογική περίληψη περιεχομένου στο Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΕΛΙΝΑΣ ΡΑΠΤΑΚΗ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σημασιολογική περίληψη περιεχομένου στο Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΕΛΙΝΑΣ ΡΑΠΤΑΚΗ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από τριμελή εξεταστική επιτροπή την 26η Οκτωβρίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
Ανδρέας-Γεώργιος Σταφυλοπάτης, Καθηγητής Ε.Μ.Π.	Γιώργος Στάμου, Αναπληρωτής Καθηγητής Ε.Μ.Π.	Παναγιώτης Τσανάκας, Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2018

Περίληψη

Τα μέσα κοινωνικής δικτύωσης πλέον αποτελούν βασικό μέσο επικοινωνίας και ενημέρωσης για μεγάλο ποσοστό του πληθυσμού. Στην Ελλάδα, για παράδειγμα, πάνω από το 60% των χρηστών του ίντερνετ υπολογίζεται ότι χρησιμοποιούν υπηρεσίες κοινωνικής δικτύωσης, με την ενημέρωση να είναι ψηλά στους λόγους που οι χρήστες αναφέρουν ότι τα χρησιμοποιούν.

Με την ολοένα και μεγαλύτερη διάδοσή τους, παράγεται τεράστιος όγκος πληροφορίας σε θέματα που απασχολούν τους χρήστες, και μάλιστα με την ιδιαιτερότητα ότι πρόκειται για δεδομένα πραγματικού χρόνου και πολλών μορφών. Έτσι, σε αυτά καταγράφονται γεγονότα, πρόσωπα, οντότητες, που είτε προκύπτουν δυναμικά, είτε εμπλουτίζονται δυναμικά με νέες πληροφορίες.

Ωστόσο, ο τεράστιος αυτός όγκος πληροφορίας καθιστά εξαιρετικά δύσκολη για ένα χρήστη την ολοκληρωμένη παρακολούθηση αυτής της εξέλιξης, με αποτέλεσμα συχνά να «χάνει» σημαντική πληροφορία. Έτσι, είναι πολύ σημαντική μια υπηρεσία περίληψης γεγονότων στη βάση πληροφοριών που εξάγονται από κάποιο μέσο κοινωνικής δικτύωσης.

Σε αυτή την εργασία προτείνουμε μια νέα μέθοδο περίληψης περιεχομένου στο Twitter. Η μέθοδος αυτή χρησιμοποιεί έναν γράφο γνώσης ως βάση για την εξαγωγή της περίληψης, με στόχο τα tweet που θα την απαρτίζουν να μην επαναλαμβάνονται, και να περιγράφουν με πληρότητα το θέμα που εξετάζεται.

Λέξεις-κλειδιά: <<μέσα κοινωνικής δικτύωσης, Twitter, περίληψη, σημασιολογική ανάλυση περιεχομένου, γράφος γνώσης>>

Abstract

Nowadays, social media are a fundamental medium of communication and information for a significant percentage of the population. In Greece for example, more than 60% of Internet users is believed to use social media, with information being among the primary reasons for their use.

Due to their increasing adoption by users, an enormous amount of data is produced in topics that interest users, and importantly real-time data in multiple forms. As a result, events and entities are recorded in them, that either emerge dynamically, or are enriched dynamically with new information.

However, this vast amount of information renders very difficult the complete observation of those developments to the user, who consequently "loses" important information. Thus, a service of content summarization for information extracted by social media is very important.

We are proposing a new method for content summarization in Twitter. This method makes use of a knowledge graph as a basis for the summarization extraction, aiming to choose tweets that aren't repeated and describe with completeness the topic at hand.

Keywords: <<social media, Twitter, summarization, semantic content analysis, knowledge graph>>

Πίνακας Περιεχομένων

1. Εισαγωγή

1.1. Περιεχόμενο του Twitter

Το Twitter είναι μέσο κοινωνικής δικτύωσης και πλατφόρμα micro-blogging, το οποίο επιτρέπει στους χρήστες του να αναρτούν, να αναπαράγουν (retweet) και να διαβάζουν σύντομα μηνύματα (μέχρι 280 χαρακτήρες από το 2017, παλιότερα μέχρι 140 χαρακτήρες), τα λεγόμενα tweets. Τα μηνύματα αυτά μπορούν να αποτελούνται από:

- Κείμενο
- Hashtag, δηλαδή μια λέξη ή/και έκφραση που ξεκινά με το σύμβολο #, και χρησιμοποιείται για την ομαδοποίηση tweets με βάση ένα συγκεκριμένο θέμα.
- Mention, δηλαδή αναφορά σε κάποιο χρήστη του Twitter, που αποτελείται από το σύμβολο @ και το username του χρήστη.
- URL συνδέσμους
- Εικόνα/Βίντεο
- Poll, δηλαδή διεξαγωγή «δημοσκόπησης» στην οποία ο κάθε χρήστης μπορεί να επιλέξει μία από τις διαθέσιμες απαντήσεις

Σε αντίθεση με τα υπόλοιπα social media, στο Twitter οι σχέσεις μεταξύ των χρηστών δεν είναι αμφίδρομες. Ένας χρήστης μπορεί να «ακολουθήσει» (follow) έναν άλλο χρήστη, χωρίς να πραγματοποιηθεί και το αντίστροφο. Όταν ένας χρήστης «ακολουθεί» κάποιους άλλους, τότε στην αρχική του σελίδα (και συγκεκριμένα στο timeline) θα εμφανίζονται τα tweet αυτών που ακολουθεί, σε αντίστροφη χρονολογική σειρά (τα νεότερα πρώτα). [1]

Επιπλέον, ένας χρήστης μπορεί να αναζητήσει να δει για ένα hashtag, ή οποιαδήποτε λέξη/φράση, όλα τα tweet που έχουν γίνει σε αντίστροφη χρονολογική σειρά (στην ίδια λογική με το timeline), άσχετα με το αν ακολουθεί τους λογαριασμούς που έκαναν αυτά τα tweet. Μάλιστα, στο timeline του ο κάθε χρήστης

μπορεί να δει ποια θέματα είναι πιο δημοφιλή εκείνη τη στιγμή (τα λεγόμενα trends), και απευθείας να δει τη συζήτηση πάνω σε αυτά. Τα trends καθορίζονται από αλγόριθμους του Twitter για κάθε χρήστη με βάση αυτούς που ακολουθεί, τα ενδιαφέροντά του και την τοποθεσία του (αν και ο χρήστης μπορεί να επιλέξει να βλέπει τα trend για μια δεδομένη γεωγραφική περιοχή). Σε κάθε περίπτωση, ο χρήστης βλέπει θέματα (τα οποία εμφανίζονται ως hashtag ή λέξεις/φράσεις) τα οποία είναι δημοφιλή εκείνη τη στιγμή, και όχι αυτά που είναι γενικά δημοφιλή σε ένα μεγάλο χρονικό διάστημα, ή σε καθημερινή βάση. [2]

Με βάση αυτά, προκύπτει ότι το Twitter προσφέρεται για την παρακολούθηση γεγονότων και τον σχολιασμό από τον χρήστη, τη στιγμή που συμβαίνουν. Ουσιαστικά δηλαδή, έχει μεγαλύτερη σημασία το περιεχόμενο, σε σχέση με την «κοινότητα», όπως ισχύει σε άλλα μέσα κοινωνικής δικτύωσης.

1.2. Αντικείμενο της εργασίας

Τα χαρακτηριστικά που περιγράψαμε στην προηγούμενη παράγραφο κάνουν το περιεχόμενο του Twitter ιδιόμορφο, από την άποψη της πληροφορίας που μπορεί να έχει ένα tweet, καθώς μπορεί να περιλαμβάνει πολλές μορφές. Ταυτόχρονα, το μικρό τους μέγεθος, μαζί με την πιο «αυθόρμητη» και γρήγορη παραγωγή τους έχει ως αποτέλεσμα το κάθε tweet, μεμονωμένα, να είναι πολύ φτωχό σε χρήσιμη πληροφορία, ενώ συνολικά να υπάρχουν πάρα πολλά tweet που ουσιαστικά να αναπαράγουν την ίδια πληροφορία.

Ο μεγάλος όγκος tweet που παράγονται σε καθημερινή βάση (περίπου 500 εκατομμύρια τη μέρα [3]), σε συνδυασμό με την χαμηλή «πυκνότητα» σε χρήσιμη πληροφορία, απαιτεί σύνθετες μεθόδους για την εξαγωγή της. Τα εργαλεία που δίνει το Twitter για αυτό το σκοπό, όπως το timeline του χρήστη και η αναζήτηση, δεν επιτρέπουν να βγουν ουσιαστικά συμπεράσματα [4].

Κι ενώ μέχρι το 2015, αυτό το ζήτημα είχε να κάνει μόνο με την αδυναμία κάποιου να επεξεργαστεί για όλο το διάστημα που τον ενδιαφέρει όλα τα tweet που έχουν γίνει (γιατί πρακτικά αυτή τη δυνατότητα έδινε το Twitter), η εισαγωγή του αλγοριθμικού timeline ήρθε να δημιουργήσει ακόμα περισσότερα εμπόδια στην

προσπάθεια για πλήρη αντίληψη ενός γεγονότος στο Twitter. Συγκεκριμένα, το Twitter πλέον δεν εμφανίζει tweet στο χρήστη με αυστηρή χρονολογική σειρά, αλλά προωθεί στην αρχή της σελίδας αυτά που θεωρεί ότι θα ενδιαφέρουν περισσότερο τον χρήστη, παίρνοντας υπ' όψιν τους χρήστες με τους οποίους αλληλεπιδρά πιο συχνά, τα tweet που έχει ασχοληθεί περισσότερο κ.ά. [5]

Αντίστοιχα, στην αναζήτηση ο χρήστης έχει τη δυνατότητα να επιλέξει να δει τα «top results» (μάλιστα αυτά είναι που θα δει πρώτα), τα οποία επιλέγονται από το Twitter βάσει αλγορίθμου, ο οποίος βρίσκει τα πιο σχετικά με τον όρο αναζήτησης tweet για να εμφανίσει στον χρήστη, με κριτήριο ποια tweet έχουν τη μεγαλύτερη αλληλεπίδραση, τι λέξεις-κλειδιά περιλαμβάνουν, και άλλα [6]. Ο ακριβής αλγόριθμος δεν είναι γνωστός, αλλά εκτιμάται ότι παίρνει υπ᾽όψιν και στοιχεία που αφορούν το χρήστη που κάνει την αναζήτηση (θέση, ποιους ακολουθεί κτλ.) [7].

Έτσι, ο χρήστης του Twitter είναι πιθανό να είναι θύμα του «echo chamber effect», δηλαδή το φαινόμενο στο οποίο ένας χρήστης τείνει όχι μόνο να αλληλεπιδρά με περιεχόμενο με το οποίο συμφωνεί περισσότερο, αλλά και να εκτίθεται περισσότερο σε αυτό (ειδικά σε ζητήματα πολιτικής, όπου εμφανίζεται μεγαλύτερη πόλωση) [8].

Με βάση τα παραπάνω, προκύπτει η ανάγκη για εξειδικευμένα εργαλεία και μεθόδους με αντικείμενο την εξαγωγή χρήσιμης πληροφορίας από το περιεχόμενο του Twitter, ώστε ο χρήστης να μη «χάνει» σημαντική πληροφορία που το ίδιο το Twitter, είναι πιθανό να μη του δείξει.

Στην παρούσα εργασία προτείνεται μια μέθοδος με την οποία επιδιώκουμε να εξάγουμε την περίληψη ενός γεγονότος από το Twitter, εντοπίζοντας tweets τα οποία δεν επαναλαμβάνονται στην πληροφορία που έχουν, και συνοψίζουν ολόπλευρα το γεγονός που μας απασχολεί. Η μέθοδος που προτείνουμε βασίζεται στη χρήση ενός γράφου γνώσης για τον εντοπισμό εννοιών στα tweet που διαφοροποιούνται από τη βασική έννοια που θέλουμε να συνοψίσουμε, και άρα περιέχονται σε tweet τα οποία είναι πιθανό να έχουν μεταξύ τους διαφορετικό περιεχόμενο.

Συγκεκριμένα, η μέθοδος αυτή αναγνωρίζει στο κείμενο των tweet λέξεις-κλειδιά, τις οποίες ταξινομεί με τη βοήθεια του γράφου στο κατά πόσο διαφοροποιούνται από τη βασική έννοια της περίληψης, και για αυτές που παρουσιάζουν τη μέγιστη διαφοροποίηση επιλέγει tweet που τις περιλαμβάνουν, και βέβαια είναι δημοφιλή. Έτσι, επιδιώκουμε το τελικό αποτέλεσμα να αποτελείται από tweet τα οποία δεν επαναλαμβάνονται, αφορούν το θέμα της περίληψης και εκφράζουν βασικές πτυχές του θέματος.

2. Τεχνολογικό Υπόβαθρο και Προηγούμενες Εργασίες

Στο κεφάλαιο αυτό θα παραθέσουμε συνοπτικά την ερευνητική δουλειά που έχει γίνει σε πεδία συναφή με αυτή την εργασία. Καθώς δεν βρέθηκε κάποια εργασία που να έχει συνδυάσει το γράφο γνώσης με την περίληψη περιεχομένου στο Twitter, δεν μπορούμε να εξετάσουμε τις προηγούμενες εργασίες πάνω στο συγκεκριμένο θέμα. Έτσι, επιλέγουμε να παραθέσουμε την προηγούμενη ερευνητική δουλειά που έχει γίνει συνολικά πάνω στο πρόβλημα της περίληψης περιεχομένου στο Twitter.

Ταυτόχρονα, χρειάζεται να γίνει ειδική μνεία για τον γράφο γνώσης, τόσο στο γιατί επιλέξαμε να τον χρησιμοποιήσουμε στα πλαίσια αυτής της εργασίας, όσο και στο πώς τον κατασκευάσαμε. Τέλος, αναφερόμαστε ξεχωριστά στην επιλογή να βασιστούμε στη Wikipedia τόσο για την κατασκευή του γράφου γνώσης, όσο και για την επιλογή των λέξεων-κλειδιών από τα tweets.

2.1. Γράφοι γνώσης

Ως γράφος γνώσης ορίζεται ένας συνεκτικός γράφος, του οποίου οι κόμβοι αναπαριστούν θέματα (concepts) και οι ακμές αναπαριστούν διάφορες σχέσεις μεταξύ τους. Με τη χρήση του θέλουμε να εκμεταλλευτούμε τις σημασιολογικές σχέσεις των διάφορων εννοιών, ώστε να εντοπίσουμε αυτές που διαφοροποιούνται σημαντικά μεταξύ τους.

Η χρήση γράφων γνώσης, έχει ήδη δοκιμαστεί για την αξιολόγηση περιεχομένου στα μέσα κοινωνικής δικτύωσης [9]. Στην παρούσα εργασία θα δοκιμάσουμε το κατά πόσο ο γράφος γνώσης μπορεί να αξιοποιηθεί για να αξιολογήσει και να ξεχωρίσει tweet που καλύπτουν διαφορετικές πτυχές του θέματος που εξετάζεται.

2.2. Wikipedia και αναγνώριση λημμάτων της σε κείμενο

Το ζήτημα του πώς θα κατασκευάσουμε το γράφο γνώσης συνδέεται με το ποιες λέξεις-κλειδιά μας ενδιαφέρει να αντιμετωπίσουμε στα tweets ώστε να τα ταξινομήσουμε. Ιδανικά, θέλουμε οι έννοιες που θα βρεθούν στα tweet να μπορούν αυτόματα να συγκριθούν με τις έννοιες που αποτελούν το γράφο γνώσης, ώστε να αποφευχθεί η εισαγωγή παραπάνω βημάτων και ενδεχομένως να χάνεται χρήσιμη πληροφορία.

Για τους παραπάνω λόγους, επιλέξαμε να βασιστούμε στη Wikipedia. Η Wikipedia προσφέρεται για την κατασκευή του γράφου γνώσης, όπως περιγράφηκε στην προηγούμενη παράγραφο, για πολλούς λόγους. Καταρχάς, το μέγεθός της (η αγγλική έκδοσή της έχει πάνω από 5,5 εκατομμύρια άρθρα το 2018) προσφέρει πολύ μεγάλη κάλυψη σε έννοιες, και ιδιαίτερα σε έννοιες πολύ εξειδικευμένες ή επώνυμες, που είναι πιθανό να μην εμφανίζονται σε άλλες πηγές, όπως το WordNet, ενώ η πληθώρα εσωτερικών συνδέσμων (τα λεγόμενα wikilinks) εξασφαλίζει την καλή δομή όλης αυτής της γνώσης, κάτι που δεν θα μπορούσε να επιτευχθεί μέσω μιας μηχανής αναζήτησης. Επιπλέον, το περιεχόμενο της Wikipedia αναπτύσσεται «κοινωνικά», καθώς συντηρείται από την κοινότητά των χρηστών της, αναπαριστά τη γνώμη της πλειοψηφίας και είναι εύκολα κατανοητό, επομένως μπορούμε να το εμπιστευτούμε για να βρούμε σχέσεις μεταξύ των εννοιών του γράφου οι οποίες θα είναι έγκυρες και όχι πολύ εξειδικευμένες. Τέλος, τα παραπάνω έχουν διερευνηθεί ([10], [11]) και υποστηρίζεται το συμπέρασμα ότι η Wikipedia προσφέρεται για την εύρεση σχέσεων μεταξύ εννοιών.

Ταυτόχρονα, με αυτή την επιλογή ικανοποιείται και η προϋπόθεση που θέσαμε προηγουμένως, για άμεση συσχέτιση του γράφου γνώσης με τις λέξεις-κλειδιά που θα απομονώσουμε στα tweet. Η εύρεση λημμάτων της Wikipedia μέσα σε κείμενο έχει διερευνηθεί σημαντικά, και έχουν προταθεί μέθοδοι που χρησιμοποιούν πληροφορίες που δίνει η ίδια η Wikipedia για τη συσχέτιση εννοιών που βρίσκονται στο κείμενο με σελίδες της ([12], [13]), όπως αυτή που επιλέξαμε να χρησιμοποιήσουμε. Ωστόσο, υπάρχουν πολλές ενδιαφέρουσες αντιμετωπίσεις σε αυτό το θέμα, όπως η χρήση μηχανικής μάθησης πάνω στη Wikipedia για την καλύτερη αναγνώριση όρων στο κείμενο [14].

2.3. Περίληψη περιεχομένου στο Twitter

Το θέμα της περίληψης περιεχομένου στο Twitter έχει απασχολήσει πολλούς ερευνητές και έχει αντιμετωπιστεί με πληθώρα διαφορετικών μεθόδων, προκειμένου να καταφέρουν να λύσουν όλα τα ζητήματα που αναφέραμε στην παράγραφο 1.2 ότι δυσκολεύουν πολύ αυτή τη διαδικασία.

Το ίδιο το βασικό πρόβλημα της περίληψης περιεχομένου στο Twitter έχει πολλές προεκτάσεις, και εμφανίζεται με διάφορες παραλλαγές στη διεθνή βιβλιογραφία, όπως η περίληψη και προσωποποίηση του timeline του χρήστη [15], η περίληψη της γνώμης (opinion) των χρηστών πάνω σε κάποιο γεγονός [16], η περίληψη της αλληλεπίδρασης των χρηστών [17] και η αναγνώριση νέων γεγονότων [18] μεταξύ άλλων. Εδώ θα εστιάσουμε μόνο στην ερευνητική δουλειά που αφορά συγκεκριμένα την περίληψη συγκεκριμένων γεγονότων/θεμάτων στο Twitter μέσω της επιλογής αντιπροσωπευτικών tweet.

Σε πολλές περιπτώσεις, επιδιώκεται η εύρεση θεμάτων (topics) και λέξεων-κλειδιών (keywords) στα tweet, όπως έγινε και στην παρούσα εργασία. Βέβαια, παρατηρείται μεγάλη διαφοροποίηση στο πώς βρίσκονται οι λέξεις-κλειδιά, αλλά και στο πώς αξιοποιούνται στη συνέχεια για την εξαγωγή της περίληψης. Πολύ δημοφιλής είναι η χρήση της μεθόδου TF-IDF για αυτό το σκοπό, με διάφορες παραλλαγές [19] και στα πλαίσια διάφορων μεθόδων.

Συνήθης είναι η χρήση clustering στα tweet, όπως για παράδειγμα σε μέθοδο που χρησιμοποιεί το μοντέλο LDA για να ταξινομήσει τα tweet σε cluster με βάση το θέμα τους, από τα οποία διαλέγει αντιπροσωπευτικά tweet με βάση λεξικό γράφο και χρήση PageRank αλγορίθμου [20]. Σε άλλη περίπτωση, προτείνεται το χρονικό clustering των tweet, ώστε αυτά να αντιμετωπιστούν ως αρχεία και να εφαρμοστεί σε αυτά TF-IDF για την επιλογή του πιο αντιπροσωπευτικού από κάθε αρχείο/cluster [21].

Η χρήση γράφου έχει δοκιμαστεί σε αρκετές περιπτώσεις και με πολύ διαφορετικούς τρόπους. Έτσι έχει προταθεί μέθοδος που εξάγει την περίληψη κάνοντας clustering σε γράφο που έχει ως κόμβους λέξεις-κλειδιά που εντοπίστηκαν στα tweets με την μέθοδο TF-IDF [22]. Μια άλλη μέθοδος προτείνει την επιλογή tweet με βάση MMR αλγόριθμο, από γράφο που αποτελείται από όλα τα tweet ως κόμβους, και ακμές μεταξύ όσων θεωρούνται ανόμοια (με βάση την

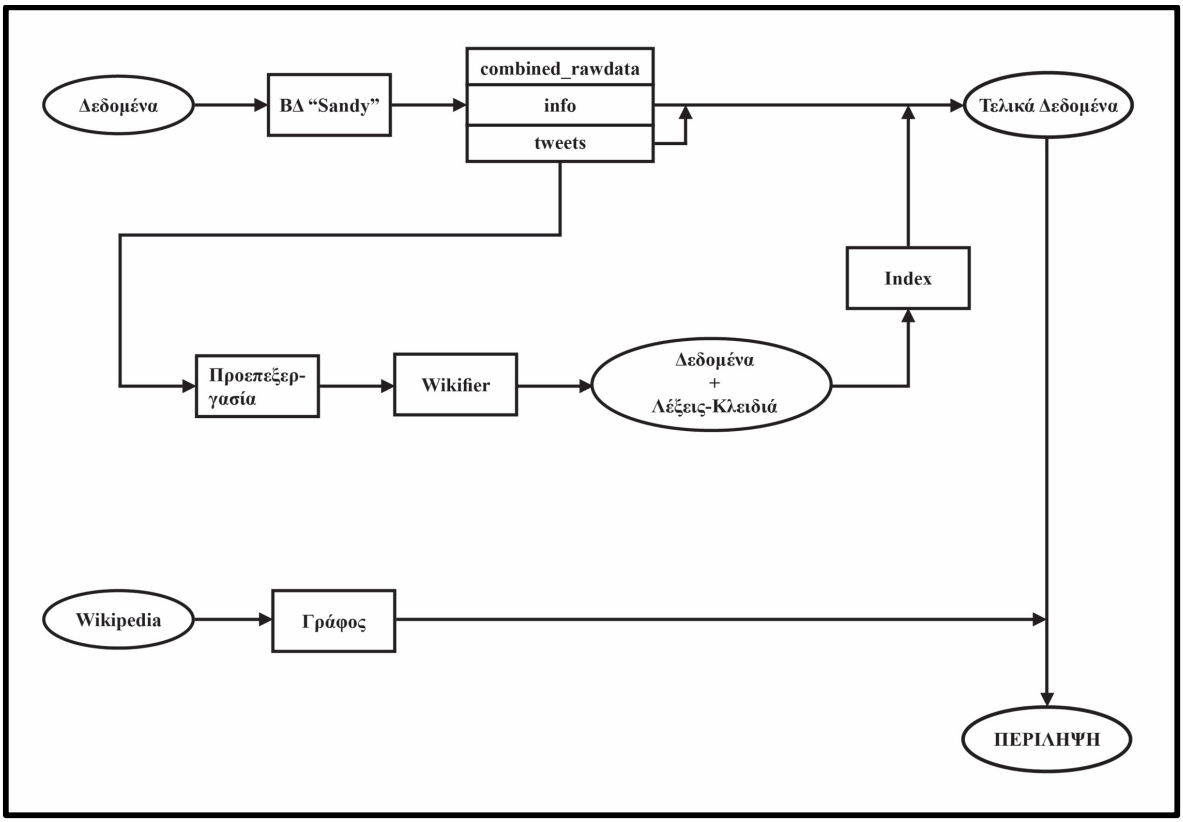
ύπαρξη όρων που εντοπίστηκαν με τη μέθοδο TF-IDF) και βάρη το salience score τους [23]. Τέλος, υπάρχει μέθοδος παρόμοια με τις δύο προηγούμενες, όπου κατασκευάζεται γράφος που αναπαριστά την ομοιότητα μεταξύ διαφορετικών ζευγών tweet, στον οποίο γίνεται clustering με τεχνικές εύρεσης κοινοτήτων, και εντέλει επιλέγονται αντιπροσωπευτικά tweet από κάθε cluster [24].

3. Ανάλυση Συστήματος

3.1. Αρχιτεκτονική του συστήματος

Το σύστημα που αναπτύχθηκε για τους σκοπούς της παρούσας εργασίας φαίνεται στην Εικόνα 1. Σημειωτέον, εδώ παρουσιάζονται μόνο τα βασικά στοιχεία και ενέργειες που έγιναν, και όχι πώς ακριβώς αυτά πραγματοποιήθηκαν, τι προβλήματα προέκυψαν και πώς ξεπεράστηκαν.

Μεγάλη σημασία, πέρα από τη λειτουργία κάθε επιμέρους υποσυστήματος, έχει και η αλληλεπίδρασή του με τα υπόλοιπα στοιχεία, όπως φαίνεται και στο Σχήμα 1.



Εικόνα 1 - Μπλοκ διάγραμμα της αρχιτεκτονικής του συστήματος

Το σύστημα έχει ουσιαστικά δύο εισόδους: τα tweet που δίνονται για την εξαγωγή της περίληψης, καθώς και το σύνολο της Wikipedia στη γλώσσα που θέλει ο χρήστης να γίνει η περίληψη (για τους σκοπούς της εργασίας θεωρούμε ότι είναι de facto η αγγλική, ωστόσο σε μια μελλοντική επέκταση θα μπορούσε να είναι οποιαδήποτε γλώσσα). Επίσης, μπορούμε να θεωρήσουμε ότι δίνεται μια επιπλέον είσοδος από το χρήστη, αυτή που αφορά το μέγεθος της ζητούμενης περίληψης, καθώς όπως εξηγείται και σε επόμενα κεφάλαια, είναι στην επιλογή του χρήστη και όχι κάτι που επιτάσσεται από τα δεδομένα.

Επιπλέον, η είσοδος των tweet δεν αποτελείται μόνο από το κείμενό τους, αλλά και από στοιχεία όπως ο αριθμός retweet που έχουν δεχτεί, η ώρα και η μέρα (timestamp) που δημιουργήθηκαν, και άλλα στοιχεία που μπορεί να παρέχει το Twitter API ούτως ή άλλως για κάθε tweet. Το ζήτημα ποια ακριβώς είναι αυτά τα στοιχεία που ζητάμε δεν εξαντλείται εδώ: θα δούμε σε παρακάτω κεφάλαια ότι στα πλαίσια αυτής της εργασίας και με το σύνολο δεδομένων που επιλέξαμε αυτά ήταν καθορισμένα, ωστόσο σε μια ενδεχόμενη εφαρμογή που θα αξιοποιεί αυτή τη μέθοδο θα μπορούμε να αξιοποιήσουμε τα συμπεράσματα της παρούσας εργασίας ώστε να ζητήσουμε μόνο τα στοιχεία που μας ενδιαφέρουν.

Αντιθέτως, η έξοδος του συστήματος είναι μία, και συγκεκριμένα τα tweet τα οποία αποτελούν τη ζητούμενη περίληψη. Για την έξοδο δεν κρίνεται αναγκαίο να παραθέσουμε άλλα στοιχεία για τα tweet πέρα από το κείμενό τους, ωστόσο αυτά είναι διαθέσιμα και μπορούν να διατεθούν σε περίπτωση που το επιθυμεί ο χρήστης, αν σε μια ενδεχόμενη εφαρμογή θέλαμε να δώσουμε αυτή τη δυνατότητα.

3.2. Ανάλυση των επιμέρους συστημάτων

3.2.1.Βάση Δεδομένων

Εδώ περιλαμβάνονται όλες οι αλλαγές που έγιναν στη βάση δεδομένων που περιείχε το αρχικό σύνολο δεδομένων, ώστε να το επεξεργαστούμε, αλλά και να βοηθηθούμε στην επιλογή των tweet που αποτελούν την τελική περίληψη. Έτσι, από τον αρχικό πίνακα που περιείχε τα δεδομένα, «combined_rawdata»,

δημιουργήσαμε δύο ακόμα: τον πίνακα tweets, που περιέχει όσες καταχωρήσεις του αρχικού πίνακα αναφέρονται σε πραγματικά tweet, και όχι retweet, και τον πίνακα info, όπου υπάρχει ο αριθμός retweet που έχει δεχτεί κάθε tweet (εάν έχει δεχτεί).

Τον πίνακα tweets αξιοποιούμε στο στάδιο της επεξεργασίας των δεδομένων, ενώ για την εξαγωγή της περίληψης χρήσιμοι είναι και οι δύο νέοι πίνακες που εισάγαμε (tweets και info), καθώς περιέχουν στοιχεία χρήσιμα για το «φιλτράρισμα» του συνόλου των δεδομένων, και κατά συνέπεια των αποτελεσμάτων που θα βγάλει η μέθοδος.

3.2.2. Επεξεργασία των tweets

Σε αυτό το στάδιο περιλαμβάνονται όλες οι παρεμβάσεις που έγιναν στο σύνολο των δεδομένων μας ώστε να εντοπιστούν σε αυτό οι λέξεις-κλειδιά. Έτσι, τα tweet περνούν από ένα στάδιο «προεπεξεργασίας» (preprocessing), όπου αφαιρούνται από το κείμενό τους στοιχεία που δεν υπάρχει περίπτωση να αποτελέσουν λέξεις-κλειδιά, και στη συνέχεια τροφοδοτούνται σε ένα λογισμικό που εντοπίζει λήμματα της Wikipedia στο κείμενο (Wikifier).

Επιπλέον, τα αποτελέσματα αυτά αποθηκεύονται με χρήση ενός index, το οποίο επιτρέπει την εύκολη επικοινωνία και συσχέτιση με τα δεδομένα που υπάρχουν στη βάση δεδομένων, κάτι αναγκαίο, καθώς όπως εξηγήσαμε και στην προηγούμενη παράγραφο τα στοιχεία που έχουμε διατηρήσει στη βάση μας βοηθούν να «φιλτράρουμε» το σύνολο των δεδομένων, όπως αυτό προκύπτει μετά την επεξεργασία που περιγράψαμε.

3.2.3. Κατασκευή του γράφου

Για την κατασκευή του γράφου γνώσης επιλέξαμε να βασιστούμε στην Wikipedia. Από αυτή, και συγκεκριμένα την αγγλική της έκδοση, κατασκευάσαμε το γράφο γνώσης αντιστοιχίζοντας το κάθε λήμμα σε έναν κόμβο, και συνδέοντας

τους κόμβους μεταξύ τους (δηλαδή φτιάχνοντας ακμές) με βάση τις συνδέσεις προς άλλα λήμματα που βρίσκονται στο κείμενο της Wikipedia. Έτσι, προσπαθήσαμε να εντοπίσουμε τις σχέσεις που μπορεί να έχουν μεταξύ τους οι έννοιες που αναπαριστούν οι κόμβοι του γράφου.

Μια βασική παρατήρηση είναι ότι ο γράφος γνώσης είναι τελείως ανεξάρτητος από τα δεδομένα που δέχεται ως είσοδο το σύστημα, και άρα για μια δεδομένη γλώσσα αρκεί να κατασκευαστεί μόνο μία φορά. Επιπλέον, καθίσταται σχετικά εύκολη η μετατροπή του συστήματος για να λειτουργεί και σε άλλες γλώσσες, ιδιαίτερα αν ενδεχομένως αξιοποιηθεί σαν εφαρμογή.

3.2.4. Εξαγωγή της Περίληψης

Έχοντας διαθέσιμα τα αποτελέσματα που περιγράφηκαν στις τρεις προηγούμενες παραγράφους, προχωράμε στην εξαγωγή της περίληψης. Τα βήματα που ακολουθούμε είναι τα εξής:

1. Επιλέγουμε έναν κόμβο αντιπροσωπευτικό του θέματος που θέλουμε να συνοψίσουμε, ο οποίος θα αποτελεί τον κόμβο αναφοράς.
2. Βρίσκουμε όλους τους κόμβους που αντιστοιχούν σε λήμματα που εντοπίστηκαν από το λογισμικό Wikifier σε tweets, και με βάση το γράφο τους ταξινομούμε με βάση την απόστασή τους από τον κόμβο αναφοράς. Μας ενδιαφέρουν οι κόμβοι για τους οποίους εντοπίζουμε τις μέγιστες αποστάσεις.
3. Επιλέγουμε tweet που να περιλαμβάνουν τις λέξεις-κλειδιά που αντιπροσωπεύουν οι παραπάνω κόμβοι και είναι στα αγγλικά, σε συνδυασμό με κριτήρια για τον αριθμό retweet που έχουν ή συγκεκριμένους όρους που περιλαμβάνουν (για κάποια από αυτά τα στοιχεία χρειάζεται να ανατρέξουμε στη βάση δεδομένων). Σημειωτέον, σε κάθε κόμβο αντιστοιχούμε το πολύ ένα tweet.
4. Κρατάμε όσα tweet έχουν προσδιοριστεί ότι πρέπει να αποτελούν την τελική περίληψη, ξεκινώντας από το tweet που έχει αντιστοιχηθεί με τον πιο απομακρυσμένο κόμβο και προχωρώντας σε πιο κοντινούς στον κόμβο αναφοράς, μέχρι να εντοπιστεί ο αριθμός tweet που χρειαζόμαστε.

Με βάση τα παραπάνω καταλήγουμε σε ένα σύνολο tweets που αποτελούν τη ζητούμενη περίληψη, για τα οποία ο στόχος είναι να περιγράφουν πλήρως το θέμα που εξετάζεται, και να μην υπάρχουν επαναλήψεις στο περιεχόμενο. Το κατά πόσο επιτεύχθηκε αυτός ο στόχος εξετάζεται στα τελευταία κεφάλαια της παρούσας εργασίας.

4. Βάση Δεδομένων

Στο κεφάλαιο αυτό περιγράφονται όλες οι ενέργειες που αφορούν το αρχικό σύνολο δεδομένων, το οποίο ανακτήσαμε σε βάση δεδομένων, καθώς και οι ενέργειες που έγιναν στη συνέχεια στη βάση.

4.1. Αρχικά Δεδομένα

4.1.1. Εύρεση συνόλου δεδομένων από το Twitter

Για τη δοκιμή της μεθόδου που προτείνουμε, είναι αναγκαίο να βρούμε ένα όσο το δυνατόν μεγαλύτερο σύνολο δεδομένων, πάνω σε κάποιο γνωστό γεγονός, ώστε να μπορούμε να αξιολογήσουμε ποιοτικά τα αποτελέσματα που θα παραχθούν.

Το να εξετάσουμε τη μέθοδο αυτή σε δεδομένα αντλημένα απευθείας από το API του Twitter δεν κρίθηκε δυνατό για δύο λόγους. Πρώτον, όπως αναφέρθηκε και προηγουμένως, ο καλύτερος τρόπος να αξιολογήσουμε τα αποτελέσματα ήταν να αφορούν την εξέλιξη κάποιου γνωστού γεγονότος, ώστε να μπορούμε με ευκολία να κρίνουμε την εγκυρότητά τους. Επιπλέον, το δημόσιο API του Twitter έχει περιορισμούς, καθώς επιτρέπει την πρόσβαση σε δεδομένα μόνο των τελευταίων 7-9 ημερών, δίνει περιορισμένη πρόσβαση στα τρέχοντα tweet και μπορεί να μην επιστρέψει το σύνολο των αποτελεσμάτων κάποιας αναζήτησης.

Έτσι, η καλύτερη λύση είναι να βρεθεί ένα έτοιμο σύνολο από tweet. Σε αυτό το σενάριο υπάρχουν δύο επιλογές: να αγοραστεί κάποιο σύνολο δεδομένων από το Twitter, ή να βρεθεί κάποιο σύνολο που να έχει συλλεχθεί για ακαδημαϊκούς σκοπούς και να διανέμεται δωρεάν. Σε αυτή την περίπτωση, το Twitter και πάλι έχει θέσει περιορισμούς, καθώς πλέον δεν επιτρέπει την διανομή τέτοιων συνόλων εφ' όσον αποτελούνται από tweet, αλλά μόνο τη διανομή των tweet ID τους, τα οποία θα αξιοποιηθούν για να συλλεχθούν από το Twitter τα ζητούμενα tweet.

Ωστόσο, υπάρχουν ορισμένα (λιγοστά) σύνολα tweet τα οποία είχαν συλλεχθεί πριν την αλλαγή της πολιτικής του Twitter, που έθεσε τους παραπάνω όρους και τα οποία είναι πλήρη στο περιεχόμενό τους. Με αυτό το σκεπτικό αναζητήσαμε και βρήκαμε το σύνολο δεδομένων που παρουσιάζεται παρακάτω.

4.1.2. Το σύνολο δεδομένων για τον τυφώνα Sandy

Τα δεδομένα που χρησιμοποιήθηκαν για την επαλήθευση και τον έλεγχο της προτεινόμενης μεθόδου συλλέχθηκαν στα πλαίσια εργασίας που μελέτησε την αποτελεσματικότητα των μέσων κοινωνικής δικτύωσης στην ανίχνευση καταστάσεων έκτακτης ανάγκης, και συγκεκριμένα του Twitter για την παρακολούθηση του τυφώνα Sandy [25].

Ο τυφώνας Sandy ήταν ο μεγαλύτερος τυφώνας που παρατηρήθηκε το 2012, και ένας από τους πιο καταστροφικούς στην ιστορία των ΗΠΑ. Σχηματίστηκε στις 22 Οκτωβρίου 2012, 500 χιλιόμετρα νότια-νοτιοδυτικά του Kingston, που βρίσκεται στην Τζαμάικα. Αρχικά έπληξε την Τζαμάικα στις 24 Οκτωβρίου, στη συνέχεια την Κούβα στις 25 Οκτωβρίου, πέρασε από τις Μπαχάμες και κατευθύνθηκε προς τις ακτές των ΗΠΑ, μεγαλώνοντας σε μέγεθος και φτάνοντας στο μέγιστο στις 29 Οκτωβρίου, περίπου 350 χιλιόμετρα νοτιοανατολικά του Atlantic City. Την επόμενη μέρα, ο τυφώνας αποδυναμώθηκε και έφτασε στο Brigantine, New Jersey. Εκτιμάται ότι πάνω από 8.5 εκατομμύρια άνθρωποι επηρεάστηκαν από blackout, που διήρκησαν ακόμα και εβδομάδες σε κάποιες περιοχές. Στις ΗΠΑ, ο τυφώνας Sandy προκάλεσε 147 θανάτους και καταστροφές πάνω από 50 δις.

Το σύνολο των tweet συλλέχθηκε μέσω της εταιρίας Topsy Labs και αποτελείται από τα διαδικτυακά «ίχνη» του τυφώνα Sandy στο Twitter, κάτι που επιτεύχθηκε με δύο τρόπους:

1) τη συλλογή των tweet που περιέχουν το hashtag «#sandy», και δημοσιεύθηκαν το διάστημα μεταξύ 15 Οκτωβρίου και 12 Νοεμβρίου 2012, δηλαδή ξεκινώντας πριν το σχηματισμό του τυφώνα και καταλήγοντας μετά την άφιξή του στις ΗΠΑ, και,

2) τη συλλογή των tweet που περιέχουν λέξεις-κλειδιά σχετικές με τον τυφώνα και τις συνέπειές του (“sandy”, “hurricane”, “storm”, “superstorm”, “flooding”, “blackout”, “gas”, “power”, “weather”, “climate”, κλπ), το ίδιο χρονικό διάστημα με αυτό που αναφέρθηκε παραπάνω.

Στο σύνολο των δεδομένων για κάθε tweet περιέχεται πληθώρα στοιχείων, όπως το κείμενο του tweet (text), id για το tweet, τον χρήστη, και το αν είναι retweet, θέση, timestamp και άλλα. Έτσι, προκύπτουν 52.55 εκατομμύρια μηνύματα από 13.75 εκατομμύρια ξεχωριστούς χρήστες για επεξεργασία. Σημειωτέον, στα πλαίσια της εν λόγω εργασίας εφαρμόστηκαν πολλαπλά φίλτρα στο αρχικό σύνολο, ώστε να κρατήσουν μόνο όσα tweets προέρχονται από χρήστες που η τοποθεσία τους συμπίπτει με την εξέλιξη του τυφώνα Sandy (πρωτογενείς πηγές), και επίσης να κρατήσουν μόνο τα tweets που ήταν όντως σχετικά με το θέμα. Παρ’ όλα αυτά, στη διάθεσή μας έχουμε μόνο το αρχικό, αφιλτράριστο σύνολο δεδομένων.

Το γεγονός αυτό δεν είναι αρνητικό για τους σκοπούς της παρούσας εργασίας. Αυτό γιατί η περίληψη περιεχομένου στο Twitter, όπως εξηγήσαμε, έχει ακριβώς να λύσει το ζήτημα του μεγάλου σε όγκο και χαμηλού σε πληροφορία περιεχομένου. Αν ελέγχαμε τη μέθοδο που προτείνουμε σε ένα σύνολο δεδομένων το οποίο είχε σε ένα βαθμό φιλτραριστεί, ώστε να είναι πιο πιθανό να περιέχει tweets σχετικά με το θέμα, δεν θα μπορούσαμε να κρίνουμε την απόδοσή της.

Συνολικά, τα χαρακτηριστικά που περιγράφονται κάνουν το σύνολο αυτό ιδιαίτερα χρήσιμο για την παρούσα εργασία, και μάλιστα για δύο λόγους. Πρώτον, γιατί καλύπτει ένα μεγάλο και γνωστό γεγονός από την αρχή ως το τέλος του, κάτι που βοηθάει για την αξιολόγηση των αποτελεσμάτων της μεθόδου που προτείνουμε. Δεύτερον, γιατί τα παραπάνω στοιχεία που δίνει (χρόνος, τοποθεσία κτλ.) μπορούν επίσης να αξιοποιηθούν για την αξιολόγηση των αποτελεσμάτων, αλλά και για να επιλεχθούν με βάση αυτά μέρη του συνόλου των tweet που θα τροφοδοτηθούν στο σύστημα και θα δώσουν αποτελέσματα.

4.2. Για την αποθήκευση του συνόλου των δεδομένων

Το σύνολο των tweet το κατεβάσαμε σε αρχείο .backup, το οποίο ανακτήθηκε με βάση τις οδηγίες των δημιουργών του σε postgres, με χρήση του λογισμικού pgAdmin. Επιλέξαμε να το διατηρήσουμε σε αυτή τη μορφή, και να κάνουμε και αρκετές παρεμβάσεις στη βάση δεδομένων, καθώς η μορφή αυτή έχει αρκετά πλεονεκτήματα τα οποία είναι χρήσιμα για την παρούσα εργασία.

Συγκεκριμένα, η δυνατότητα να εκτελεστούν queries στα δεδομένα με ταχύτητα, κάτι που θα ήταν αδύνατο εκτός της βάσης δεδομένων, είναι ο βασικός λόγος που τα διατηρούμε σε αυτή την μορφή. Αυτό γιατί, ενώ ο βασικός όγκος της επεξεργασίας των δεδομένων γίνεται εκτός της βάσης, ωστόσο για την τελική επιλογή των tweet που αποτελούν την περίληψη, χρειάζεται να εντοπίζουμε αυτά που πληρούν σε κάθε περίπτωση συγκεκριμένες προϋποθέσεις (π.χ. να έχουν ένα ορισμένο αριθμό retweet ή να έχουν παραχθεί σε ένα συγκεκριμένο χρονικό διάστημα) και να τα τροφοδοτούμε στο σύστημα. Σε περίπτωση που δεν είχαμε διατηρήσει όλες αυτές τις πληροφορίες στη βάση, η επιλογή αυτή θα ήταν πολύ πιο χρονοβόρα.

4.3. Επεξεργασία της ΒΔ

Αν και κατά βάση η επεξεργασία των δεδομένων μας έγινε εκτός της ΒΔ, όπως εξηγήθηκε και παραπάνω κρίθηκε χρήσιμη η διατήρηση του αρχικού αντιγράφου σε postgres, για τους λόγους που έχουν ήδη αναλυθεί.

Τα tweet βρίσκονται στον πίνακα combined_rawdata της βάσης δεδομένων sandy, που έχει τις εξής στήλες (columns):

- tweet_id (bigint)
- tweet_text (text)
- created_at (timestamp with time zone)
- user_id (bigint)

- user_followers_count (integer)
- user_friends_count (integer)
- topsy_doc_sentiment_abs (numeric)
- topsy_doc_sentiment (integer)
- retweeted_status_int (bigint)
- time_min (timestamp with time zone)
- lat_final (numeric)
- lng_final (numeric)
- geom (geometry)

Οι στήλες αυτές περιέχουν όλα τα στοιχεία που αναφέρθηκαν ότι παρέχονται για κάθε tweet. Έτσι έχουμε στοιχεία για το ίδιο το tweet, για το χρήστη που έκανε το tweet, για την αξιολόγηση του tweet (θετικό, αρνητικό ή ουδέτερο) και για την τοποθεσία του χρήστη που έκανε το tweet, όπου αυτή είναι διαθέσιμη.

Με βάση το αρχικό σύνολο των tweet, στη βάση δεδομένων έγιναν οι εξής παρεμβάσεις:

4.3.1. Πίνακας tweets

Δημιουργήθηκε νέος πίνακας ο οποίος έχει αποκλειστικά μοναδικά tweet, και όχι retweet (δηλαδή όσα tweet έχουν την ιδιότητα retweeted_status_int ίση με 0). Θεωρούμε ότι δεν χάνεται σημαντική πληροφορία, καθώς τα tweet που απορρίπτουμε έχουν ακριβώς το ίδιο κείμενο με τα αρχικά tweet που κρατάμε.

Ο νέος αυτός πίνακας έχει 33232125 καταχωρήσεις, ή αλλιώς το 63,3% όλων των tweet. Η δομή αυτού του πίνακα είναι ακριβώς η ίδια με αυτή του combined_rawdata, καθώς επιλέξαμε να κρατήσουμε όλες τις στήλες.

Σημειωτέον, τα tweet που επιλέχθηκαν για καταχώρηση στον πίνακα tweets, είναι αυτά που στη συνέχεια τροφοδοτούνται στο σύστημα που κατασκευάσαμε.

4.3.2. Πίνακας info

Δημιουργήθηκε επιπλέον πίνακας ο οποίος εισάγει μια νέα μετρική, τον αριθμό retweet που έχει ένα tweet. Για το σκοπό αυτό, αξιοποιήσαμε την παρατήρηση που αναφέρθηκε παραπάνω για την παράμετρο retweeted_status_int. Έτσι, εστιάζοντας στα tweet για τα οποία η τιμή αυτή παίρνει τιμή διάφορη του 0, παίρνουμε τα retweet που έχουν γίνει. Για αυτά, η παράμετρος αυτή παίρνει τιμή ίση με το tweet_id του αρχικού tweet που έγινε retweet.

Με βάση αυτά, μπορούμε σε sql να βρούμε τον αριθμό retweets που έχει κάθε αρχικό tweet, αναζητώντας πόσες φορές εμφανίζεται το tweet_id του ως retweeted_status_int.

Έτσι, κατασκευάσαμε τον πίνακα info, ο οποίος αποτελείται από 4872517 καταχωρήσεις. Δηλαδή, από τα 33 εκατομμύρια tweet που έγιναν στο διάστημα παρατήρησης με θέμα τον τυφώνα Sandy, μόνο το 14,7% έχει γίνει έστω μια φορά retweet κατά το διάστημα που εξετάζουμε. Από αυτά, ο μέγιστος αριθμός retweets για κάποιο tweet είναι 57662, και ο μέσος όρος retweets είναι 3,95.

4.3.3. Πίνακας tweets_ordered

Στην πορεία της εργασίας προέκυψε η ανάγκη για χρονολογική ταξινόμηση των tweet. Για το λόγο αυτό δημιουργήθηκε ο πίνακας tweets_ordered, στον οποίο οι καταχωρήσεις του πίνακα tweets έχουν ταξινομηθεί με βάση το timestamp τους, και συγκεκριμένα το created_at.

5. Επεξεργασία των tweets

Σε αυτό το κεφάλαιο περιλαμβάνονται όλα όσα αφορούν την επεξεργασία των tweets, ώστε να εντοπιστούν σε αυτά λέξεις-κλειδιά, και συγκεκριμένα λήμματα της Wikipedia. Έτσι, επεξεργαζόμαστε τα διαθέσιμα tweet ώστε να απομακρύνουμε στοιχεία που δεν χρειάζονται, και εξάγουμε από αυτά τις απαραίτητες πληροφορίες για την κατασκευή του γράφου.

Όπως αναλύεται και στις παραγράφους που ακολουθούν, ο μεγάλος όγκος των δεδομένων σε συνδυασμό με την χαμηλή τους ποιότητα δημιούργησαν ποικίλα ζητήματα που έπρεπε να λύσουμε, ώστε να καταλήξουμε στο ζητούμενο αποτέλεσμα.

5.1. Επικοινωνία με την βάση δεδομένων

Όπως έχει ήδη αναφερθεί, η επεξεργασία των δεδομένων θα γίνει σε python. Προκειμένου να αντληθούν τα δεδομένα προς επεξεργασία από την βάση, χρησιμοποιήθηκε η βιβλιοθήκη psycopg2. Η συγκεκριμένη βιβλιοθήκη αποτελεί έναν μετατροπέα της PostgreSQL βάσης δεδομένων για την python [26].

Έτσι, από την βάση, και συγκεκριμένα από τον πίνακα tweets, κατεβάζουμε όλα τα tweet, μαζί με τα αντίστοιχα tweet_id. Τα tweet_id είναι απαραίτητα να τα κρατήσουμε μαζί με τα αντίστοιχα tweet, γιατί αυτά είναι το μοναδικό χαρακτηριστικό που με βεβαιότητα μπορούμε να πούμε ότι τα χαρακτηρίζει (ενώ για παράδειγμα το κείμενο ενός tweet δεν είναι απαραίτητα μοναδικό). Η διατήρηση των tweet_id μαζί με το κείμενο των tweet είναι απαραίτητη για την περαιτέρω επικοινωνία με τη βάση.

Η βιβλιοθήκη psycopg2 φέρνει τα αποτελέσματα του query, και τα καταχωρεί σε μια λίστα από λίστες στην python, κάτι που μας βοηθάει για τα μετέπειτα βήματα.

5.2. Επεξεργασία των tweet

Από τα αρχικά tweets, επιλέγουμε να αφαιρεθούν ορισμένα στοιχεία, τα οποία είναι τα εξής:

1. links
2. mentions

Η αφαίρεση αυτή θα επιτευχθεί με την χρήση κανονικών εκφράσεων για τον εντοπισμό των συνδέσμων (ξεκινά με http ή https και καταλήγει σε κενό) και των αναφορών σε χρήστες (ξεκινά με @ και καταλήγει σε κενό).

Ο λόγος που δεν κρατάμε τα στοιχεία αυτά έχει να κάνει με την επεξεργασία στην οποία θα υποβληθούν. Καθώς ο στόχος είναι να εντοπιστούν λήμματα της Wikipedia στο κείμενο των tweet, μπορούμε να αφαιρέσουμε στοιχεία που γνωρίζουμε ότι δεν μπορούν να αποφέρουν αποτελέσματα.

Δεν θα αφαιρέσουμε τα hashtag, καθώς είναι πιθανό να αποτελούνται από λέξεις που μπορούν να αντιστοιχηθούν σε όρους της Wikipedia, και άρα να έχουν σημασία για την έρευνά μας (π.χ. το hashtag #hurricane). Ωστόσο, αφαιρούμε από αυτά τον χαρακτήρα του hashtag (“#”), ώστε να είναι δυνατό να αναγνωριστούν από το λογισμικό που χρησιμοποιούμε.

Έχοντας υλοποιήσει αυτή τη διαδικασία, έχουμε πλέον στη διάθεσή μας μια λίστα tweet που είναι έτοιμα για επεξεργασία. Τα tweet αυτά τα αποθηκεύουμε σε ένα txt αρχείο, το οποίο θα αποτελέσει το αρχείο εισόδου για το επόμενο στάδιο. Παράλληλα, διατηρούμε σε ένα csv αρχείο λίστα που περιλαμβάνει τα αρχικά tweet, όπως εμφανίζονται στη βάση, και τα αντίστοιχα tweet_id, αντιστοιχισμένα με τα tweet όπως προέκυψαν μετά την παραπάνω επεξεργασία.

5.3. Εύρεση λημμάτων της Wikipedia στο κείμενο των tweet

5.3.1. Για το λογισμικό Wikifier

Η εύρεση λημμάτων της Wikipedia σε ένα κείμενο δεν είναι απλή υπόθεση, και χρειάζεται να λύσει δύο θέματα. Πρώτον, να εντοπίσει έννοιες που μπορούν να συσχετιστούν με τη Wikipedia (πιθανώς όχι μόνο ονόματα, αλλά και ενδιαφέρουσες εκφράσεις). Πέρα από αυτό, χρειάζεται να αντιστοιχίσει τις έννοιες αυτές σωστά σε λήμματα της Wikipedia. Για παράδειγμα, δεν αρκεί να εντοπιστεί ο όρος «Chicago» μέσα στο κείμενο, αλλά χρειάζεται και να διασαφηνιστεί αν αναφέρεται στην πόλη ή στην ταινία.

Το λογισμικό Wikifier χρησιμοποιεί για το σκοπό αυτό μια μέθοδο που προσπαθεί να αντιστοιχίσει ταυτόχρονα όλους τους όρους που έχουν εντοπιστεί στο κείμενο χρησιμοποιώντας τους εσωτερικούς συνδέσμους της Wikipedia για να εκτιμήσει το κατά πόσο η αντιστοίχιση που γίνεται έχει συνοχή, σε αντίθεση με άλλες μεθόδους που αντιστοιχίζουν κάθε όρο ξεχωριστά, χρησιμοποιώντας στοιχεία όπως τις ομοιότητες στο κείμενο μεταξύ του αρχείου που εξετάζεται και την σελίδα της Wikipedia κάθε υποψήφιου λήμματος [12].

Κατά την εφαρμογή του λογισμικού Wikifier, προέκυψαν αρκετά σύνθετα προβλήματα. Το πώς εργαστήκαμε για την επίλυσή τους περιγράφεται παρακάτω.

5.3.2. Μέγεθος αρχείων εισόδου και χρόνος επεξεργασίας

Το λογισμικό Wikifier αποδείχθηκε ότι δεν μπορεί να επεξεργαστεί μεγάλο όγκο δεδομένων ταυτόχρονα, καθιστώντας αναγκαίο τον τεμαχισμό των δεδομένων μας σε πακέτα τα οποία τροφοδοτήσαμε σε αυτό. Μετά από δοκιμές, επιλέχθηκε να μοιραστούν σε «πακέτα» των 1000 tweet, καθώς σε αυτό το μικρό μέγεθος όχι μόνο λειτουργούσε σωστά το λογισμικό, αλλά επίσης ο χρόνος επεξεργασίας ήταν σχετικά μικρός.

Παρ' όλα αυτά, συνολικά η εφαρμογή του λογισμικού στο σύνολο των tweet αποδείχθηκε εξαιρετικά χρονοβόρα (χρειάστηκαν αρκετές μέρες για να πάρουμε αποτελέσματα για το σύνολο των tweet). Βέβαια, λαμβάνουμε υπ' όψιν ότι η επεξεργασία αυτή έγινε σε προσωπικό laptop, και ότι δεν έγινε καμία παρέμβαση στον κώδικα του λογισμικού για βελτιστοποίηση του χρόνου εκτέλεσης. Προφανώς, σε μια ενδεχόμενη εφαρμογή με βάση αυτή τη μέθοδο θα έπρεπε είτε να βελτιστοποιηθεί το λογισμικό, είτε να βρεθεί κάποιο άλλο που να επιτελεί την ίδια λειτουργία (και να επικαιροποιηθεί η έκδοση της Wikipedia που χρησιμοποιεί).

5.3.3. Σφάλματα

Ακόμα και με το μέτρο που περιγράφηκε στην προηγούμενη παράγραφο, σε ορισμένα αρχεία προέκυψαν σφάλματα, με αποτέλεσμα να μη βγουν για αυτά αποτελέσματα. Αυτά τα σφάλματα προέκυπταν από την ύπαρξη ειδικών χαρακτήρων, τους οποίους δεν μπορούσε να επεξεργαστεί σωστά το λογισμικό.

Σε αυτές τις περιπτώσεις, χωρίσαμε τα tweet που περιλαμβάνονται στα εν λόγω αρχεία σε ένα txt αρχείο εισόδου για το καθένα, και τα τροφοδοτήσαμε έτσι στο λογισμικό. Ήταν απαραίτητη αυτή η παρέμβαση, καθώς αφορούσε 3162000 tweet, δηλαδή σημαντικό ποσοστό του συνόλου των tweet.

5.3.4. Παλιά έκδοση της Wikipedia

Το λογισμικό Wikifier που χρησιμοποιήσαμε, προέκυψε από εργασία που εκπονήθηκε το 2011. Σε αυτό, είχαν αξιοποιηθεί τα dumps της Wikipedia, τα οποία προφανώς έχουν ελλείψεις σε σχέση με τα σημερινά. Το πιο χαρακτηριστικό πρόβλημα είναι ότι η λέξη «Sandy» όπου εντοπίστηκε σε tweet δεν μπορούσε να αντιστοιχηθεί σωστά στο λήμμα Hurricane Sandy της Wikipedia, καθώς το 2011 το συγκεκριμένο λήμμα προφανώς δεν υπήρχε.

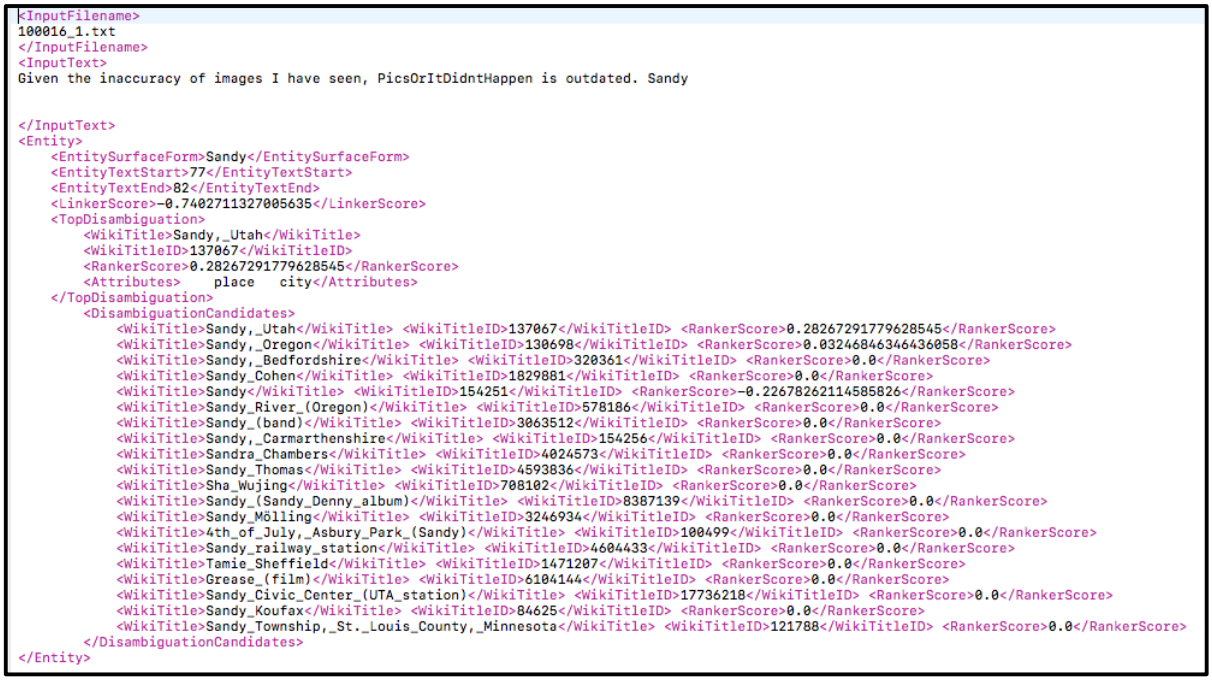
Το θέμα αυτό λύθηκε με παρέμβαση εκ των υστέρων στα αποτελέσματα που προέκυψαν από την εφαρμογή του λογισμικού στα tweet, και συγκεκριμένα με την αλλαγή της αντιστοίχισης της λέξης Sandy στο σωστό λήμμα, όπου αυτή είχε βρεθεί.

5.4. Εξαγωγή των αποτελεσμάτων

Το λογισμικό που χρησιμοποιήσαμε επιστρέφει αποτελέσματα σε αρχεία αρκετών μορφών, από τα οποία επιλέγουμε να αξιοποιήσουμε τα xml. Για την ανάγνωση και επεξεργασία των xml αρχείων σε python χρησιμοποιούμε τη βιβλιοθήκη Beautiful Soup.

Η Beautiful Soup είναι μια βιβλιοθήκη της python για την εξαγωγή δεδομένων από αρχεία XML και HTML. Η εξαγωγή αυτή επιτυγχάνεται με την μέθοδο «tree-based parsing», όπου το σύνολο του XML αρχείου παριστάνεται σε μορφή δέντρου (καθώς η XML είναι μια ιεραρχική γλώσσα) και είναι διαθέσιμο εξ' ολοκλήρου στο χρήστη. Ακολουθώντας τη δομή του αρχείου, μπορούμε να έχουμε πρόσβαση σε όλα τα στοιχεία που μας ενδιαφέρουν.

Τα αρχεία που επεξεργαζόμαστε έχουν την εξής, ενδεικτική, μορφή:



Εικόνα 2 - Αρχείο XML που έχει παραχθεί από το λογισμικό Wikifier

Από αυτά, εμείς θέλουμε να βρούμε ποιοι όροι της Wikipedia εντοπίστηκαν σε κάθε tweet, και όχι συνολικά στο κείμενο που δώσαμε ως είσοδο. Για να επιτευχθεί αυτό, μας ενδιαφέρουν τα εξής tag:

1. EntitySurfaceForm, δηλαδή η λέξη ή η φράση από το κείμενο που έχει αντιστοιχηθεί με κάποιο λήμμα της Wikipedia.
2. EntityTextStart, δηλαδή η θέση του κειμένου από την οποία ξεκινάει η παραπάνω λέξη/φράση.
3. EntityTextEnd, δηλαδή η θέση του κειμένου στην οποία τελειώνει η παραπάνω λέξη/φράση.
4. TopDisambiguation, δηλαδή τα στοιχεία του λήμματος της Wikipedia που είναι πιο πιθανό να αντιστοιχεί με την λέξη/φράση που εντοπίστηκε στο κείμενο.

Ένα βασικό ζήτημα που προέκυψε κατά την εξαγωγή των αποτελεσμάτων είχε να κάνει με τη θέση στην οποία βρίσκονται οι όροι που αναγνωρίστηκαν από το λογισμικό, όπως ορίζεται από τις παραμέτρους EntityTextStart και EntityTextEnd.

Συγκεκριμένα, στα αρχικά tweet ορισμένοι χαρακτήρες, δεσμευμένοι στην html (όπως το «&» και τα «<», «>»), εμφανίζονται κωδικοποιημένοι (δηλαδή αντί για «&» έχουμε «&» κ.ο.κ.). Το λογισμικό Wikifier, κατά την επεξεργασία του κειμένου, αναγνώριζε αυτές τις οντότητες και τις μετέτρεπε στους χαρακτήρες που αναπαριστούν, όμως στα αρχεία που επέστρεφε αυτή η μετατροπή δεν φαινόταν, με αποτέλεσμα να μην αντιστοιχούν τα EntityTextStart και EntityTextEnd στις θέσεις που έπρεπε.

Η διαφορά αυτή έπρεπε να ληφθεί υπ' όψιν για να εντοπιστούν σωστά τα tweet στα οποία εντοπίστηκαν τα διάφορα λήμματα της Wikipedia από το λογισμικό, καθώς αναφέρονται σε ενιαία αρχεία κειμένου που περιέχουν 1000 tweet το καθένα.

5.5. Αποθήκευση των αποτελεσμάτων - Index

Έχοντας εξάγει τα λήμματα της Wikipedia που εντοπίστηκαν στα tweet, και έχοντας αντιστοιχήσει τα αποτελέσματα αυτά στο κάθε tweet, έπρεπε όλα αυτά να συνδυαστούν σε ένα τελικό αρχείο που θα περιείχε όλες τις απαραίτητες πληροφορίες για την εξαγωγή της περίληψης.

Έτσι, κρατήσαμε για κάθε tweet το κείμενό του, το tweet ID του (όπως είπαμε είναι το πιο βασικό αναγνωριστικό για κάθε tweet, καθώς πολλές φορές το κείμενο μόνο δεν είναι αρκετό), τα λήμματα της Wikipedia που εντοπίστηκαν στο κείμενό του, καθώς και τον αριθμό των retweet που έχει γίνει το συγκεκριμένο tweet (από τον πίνακα info που φτιάξαμε στην ΒΔ – αν το ζητούμενο tweet ID δεν εντοπιστεί εκεί, προκύπτει ότι το συγκεκριμένο tweet έχει δεχτεί 0 retweet).

Επομένως, μαζί με το αρχείο των λέξεων-κλειδιά που έχουν εντοπιστεί σε κάθε tweet και τα tweet αυτά, έχουμε και ένα index που αντιστοιχίζει κάθε tweet με το ID του, και μπορεί να αξιοποιηθεί σαν κλειδί για να αντληθούν παραπάνω στοιχεία για το tweet αυτό από τη βάση, αλλά και αντίστροφα ώστε να εντοπιστούν tweet μαζί με τα keywords τους που πληρούν συγκεκριμένες προϋποθέσεις. Επιλέξαμε να μην

μεταφέρουμε τον πίνακα αυτό στη βάση, καθώς δεν θα μας πρόσφερε κάποιο πλεονέκτημα.

Σημειωτέον, δεν βρέθηκαν σε όλα τα tweet λέξεις-κλειδιά, κάτι που είναι λογικό αν σκεφτούμε το πόσο φτωχό σε περιεχόμενο μπορεί να είναι ένα tweet, αλλά και το γεγονός ότι αναζητούμε λέξεις στα αγγλικά χωρίς να έχουμε φιλτράρει το σύνολο δεδομένων για να κρατήσουμε μόνο το περιεχόμενο σε αγγλική γλώσσα. Συνολικά, από τα 33232125 tweets που δώσαμε για επεξεργασία, μόνο στα 23254729, ή αλλιώς στο 69,98% εντοπίστηκε κάποιο λήμμα της Wikipedia, και άρα μας απασχολούν για την εξαγωγή της περίληψης.

6. Κατασκευή του γράφου γνώσης

6.1. Επιλογή βιβλιοθήκης για την κατασκευή του γράφου

Η πιο διαδεδομένη βιβλιοθήκη της python για την επεξεργασία γράφων, η networkx, δεν μπορεί να αξιοποιηθεί στα πλαίσια της παρούσας εργασίας. Το μεγάλο μέγεθος του γράφου που θέλουμε να κατασκευάσουμε και στη συνέχεια να επεξεργαστούμε δεν μπορεί να υποστηριχθεί από αυτή τη βιβλιοθήκη.

Στη θέση της, χρησιμοποιήσαμε το εργαλείο graph-tool. Το graph-tool, σε αντίθεση με την networkx που είναι εξ' ολοκλήρου υλοποιημένη σε python, έχει υλοποιηθεί σε C++, με αποτέλεσμα να είναι πολύ πιο γρήγορο. Ενδεικτικά, υπολογίζεται ότι η networkx είναι 20 έως και 170 φορές βραδύτερη από το graph-tool. [27]

Algorithm	graph-tool (4 cores)	graph-tool (1 core)	igraph	NetworkX
Single-source shortest path	0.004 s	0.004 s	0.012 s	0.152 s
PageRank	0.029 s	0.045 s	0.093 s	3.949 s
K-core	0.014 s	0.014 s	0.022 s	0.714 s
Minimum spanning tree	0.040 s	0.031 s	0.044 s	2.045 s
Betweenness	244.3 s (~4.1 mins)	601.2 s (~10 mins)	946.8 s (edge) + 353.9 s (vertex) (~ 21.6 mins)	32676.4 s (edge) 22650.4 s (vertex) (~15.4 hours)

Εικόνα 3 - Σύγκριση της ταχύτητας με την οποία εκτελούνται οι εμφανιζόμενοι αλγόριθμοι για γράφο με 39796 κόμβους και 301498 ακμές

Τα δύο αυτά εργαλεία δεν διαφέρουν μόνο στις ταχύτητες επεξεργασίας που μπορούν να πετύχουν. Επιλέγοντας το graph-tool, χάνουμε πολλές δυνατότητες που προσφέρει η networkx, η βασικότερη από τις οποίες έχει να κάνει με τους κόμβους του γράφου. Ενώ στη networkx ένας κόμβος μπορεί να είναι οτιδήποτε θέλουμε

(στην περίπτωσή μας κάποιος τίτλος σελίδας της Wikipedia), το graph-tool επιτρέπει μόνο την αρίθμηση των κόμβων. Έτσι, Ν κόμβοι θα αριθμηθούν σειριακά από το 0 έως το Ν-1.

Συνολικά, η αναγκαία αυτή επιλογή μας στερεί πολλές ευκολίες που προσέφερε η networkx για την κατασκευή και επεξεργασία του γράφου. Παρ' όλα αυτά, δεν υπήρξε ζήτημα να μην μπορούμε να υλοποιήσουμε λειτουργίες της networkx με το graph-tool, επομένως η επιλογή του ήταν μονόδρομος.

6.2. Για την κατασκευή του γράφου

Για την κατασκευή του γράφου γνώσης, χρειάστηκε να κατεβάσουμε το σύνολο της Wikipedia, ώστε να εξάγουμε την πλήρη λίστα των τίτλων που περιλαμβάνει (page titles), καθώς και τις συνδέσεις μεταξύ των διαφορετικών άρθρων (page links).

Για το σκοπό αυτό, αξιοποιήσαμε τα dumps της Wikipedia, τα οποία ανανεώνονται ανά τακτά χρονικά διαστήματα, και περιλαμβάνουν το σύνολο των πληροφοριών του ιστοτόπου. Από αυτά, επιλέξαμε το αρχείο «enwiki-latest-pages-articles.xml», το οποίο αποτελείται από το σύνολο των σελίδων της Wikipedia, σε xml μορφή.

Το συγκεκριμένο αρχείο, αφού αποσυμπιεστεί, έχει μέγεθος 65,58 GB. Αυτό δημιουργεί προκλήσεις για την επεξεργασία του, καθώς δεν μπορεί να φορτωθεί ολόκληρο, όπως γίνεται από την πλειοψηφία των βιβλιοθηκών που προορίζονται για XML parsing, όπως η BeautifulSoup, που χρησιμοποιήθηκε σε άλλο μέρος της εργασίας. Για το λόγο αυτό, επιλέξαμε να χρησιμοποιήσουμε την βιβλιοθήκη cElementTree, η οποία μας επιτρέπει να φορτώνουμε το αρχείο τμηματικά, με βάση τα tags που μας ενδιαφέρουν, ενώ το γεγονός ότι έχει υλοποιηθεί σε C μας δίνει πολύ ανώτερες ταχύτητες επεξεργασίας και χαμηλότερη κατανάλωση μνήμης.

Με τη χρήση αυτής της βιβλιοθήκης, θα επεξεργαστούμε το αρχείο με την μέθοδο «Event-Based Parsing», όπου αντί για κατασκευή δέντρου, γίνεται αναζήτηση για στοιχεία (elements) της XML. Όταν κάποιο στοιχείο διαβαστεί ολόκληρο (βρεθεί δηλαδή το τέλος του), μπορούμε να το επεξεργαστούμε, και αφού το διαγράψουμε να περάσουμε στο επόμενο. Έτσι, οι απαιτήσεις από τη μνήμη είναι μηδαμινές.

6.2.1.Κόμβοι του γράφου

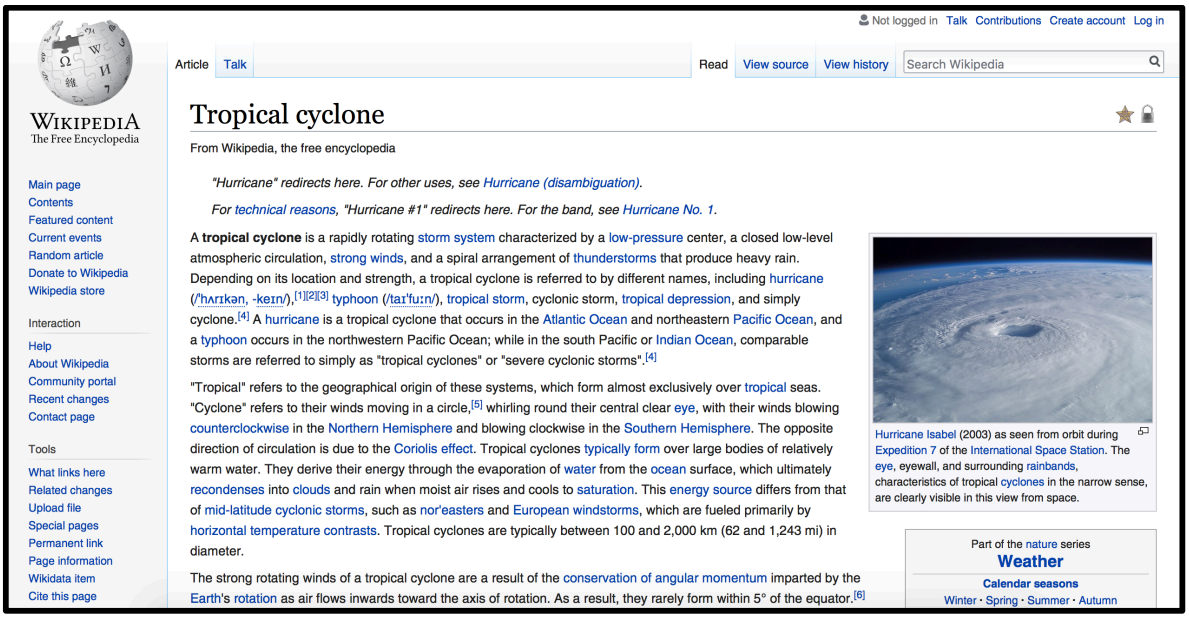
Οι κόμβοι του γράφου είναι οι σελίδες της Wikipedia, επομένως και για λόγους ευκολίας, θα τους αντιστοιχίσουμε με τους πλήρεις τίτλους των σελίδων της Wikipedia. Αυτούς τους βρίσκουμε σε κάθε event με το tag «title», το οποίο ακολουθεί το namespace «{http://www.mediawiki.org/xml/export-0.10/}».

Με αυτόν τον τρόπο εντοπίστηκαν 16589222 λήμματα, επομένως ισάριθμοι κόμβοι του γράφου. Σημειωτέον, επιλέξαμε να δουλέψουμε με τους τίτλους και όχι τα ID των σελίδων, κι αυτό γιατί και στην περίπτωση των tweets κρατήσαμε τους τίτλους της Wikipedia που βρέθηκαν στο κείμενο. Η επιλογή αυτή έγινε γιατί, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, το λογισμικό Wikifier βασίζεται στην Wikipedia του 2011, επομένως δεν μπορούσαμε να είμαστε σίγουροι ότι θα υπήρχε πλήρης αντιστοιχία μεταξύ των ευρημάτων του, και της τρέχουσας έκδοσης της Wikipedia που χρησιμοποιούμε για την κατασκευή του γράφου.

Μετά τον εντοπισμό των κόμβων του γράφου βέβαια, χρειάζεται ακόμη ένα βήμα, σύμφωνα με τους περιορισμούς που αναφέρθηκαν για τη χρήση του graph-tool. Έτσι, ενώ μας αρκεί να γνωρίζουμε τον αριθμό των κόμβων για να τους κατασκευάσουμε, είναι αναγκαίο να έχουμε αντιστοιχήσει κάθε λήμμα με τον αριθμό του κόμβου που θα το αναπαραστήσει.

6.2.2. Ακμές του γράφου

Για να βρούμε τις ακμές του γράφου, θα εξετάσουμε με ποιες σελίδες «συνδέεται» κάθε σελίδα. Οι συνδέσεις αυτές, στα πλαίσια της Wikipedia, είναι οι σύνδεσμοι που βρίσκονται στο κείμενο και οδηγούν σε άλλες σελίδες :



Εικόνα 4 - Μια σελίδα της Wikipedia. Με μπλε φαίνονται οι σύνδεσμοι στο κείμενο, που οδηγούν σε άλλα λήμματα.

Όπως βλέπουμε και παραπάνω, οι συνδέσεις αυτές είναι πολυάριθμες. Προκειμένου να μειώσουμε το μέγεθος του γράφου, επιλέξαμε να μην κρατήσουμε το σύνολο αυτών των συνδέσεων για κάθε σελίδα, αλλά μόνο όσες εμφανίζονται στο εισαγωγικό κείμενο, το οποίο εκτείνεται μέχρι τον πίνακα περιεχομένων. Έτσι, διατηρούνται οι πιο βασικές συνδέσεις μεταξύ των σελίδων, χωρίς να χρειάζεται να διαχειριστούμε έναν παράλογα μεγάλο όγκο ακμών. Εξάλλου, θεωρούμε ότι στην εισαγωγή θα βρούμε τις πιο σημαντικές πληροφορίες για κάθε λήμμα, επομένως οι βασικές συνδέσεις με άλλες σελίδες θα υπάρχουν με αυτή την επιλογή.

Η ανάκτηση των ακμών από το xml αρχείο της Wikipedia έγινε με τα ίδια εργαλεία που χρησιμοποιήθηκαν και για τους κόμβους. Για τον εντοπισμό των ζητούμενων σελίδων βασιστήκαμε στις εξής παρατηρήσεις:

1. Το κείμενο του άρθρου βρίσκεται με το tag «text».
2. Το τέλος του εισαγωγικού κειμένου ακολουθείται από τα περιεχόμενα της σελίδας, τα οποία εμφανίζονται σε ξεχωριστό πλαίσιο. Αυτό το πλαίσιο ξεκινά με τους χαρακτήρες «==». Επομένως κρατάμε το κείμενο μέχρι την πρώτη εμφάνιση αυτής της συμβολοσειράς.
3. Οι σύνδεσμοι σε άλλες σελίδες της Wikipedia εμφανίζονται μέσα σε διπλά brackets «[[...]]». Ωστόσο, δεν αρκεί μόνο αυτό για τον εντοπισμό τους, καθώς με τον ίδιο τρόπο εμφανίζονται και εικόνες, αρχεία και urls. Επομένως, απορρίπτουμε τις περιπτώσεις που εντός των brackets εμφανίζονται οι όροι «Image», «File», «http». Τέλος, κρατάμε μόνο το κομμάτι μέχρι τον χαρακτήρα «|», καθώς μέχρι εκεί εμφανίζεται ο τίτλος του λήμματος με τον οποίο συνδέεται η σελίδα μας, ενώ μετά από αυτόν έχουμε το πώς εμφανίζεται στο κείμενο της σελίδας (π.χ. ο τίτλος του λήμματος «Tropical Cyclone» στη σελίδα μπορεί να εμφανίζεται ως «hurricane»).

Με βάση αυτά, και κάνοντας χρήση κανονικών εκφράσεων, συλλέξαμε για κάθε τίτλο/σελίδα της Wikipedia τις σελίδες με τις οποίες συνδέεται, με βάση την εισαγωγή του λήμματος.

Αρχικά, τα στοιχεία αυτά συλλέχθηκαν σε μορφή λίστας, όπου κάθε λήμμα της Wikipedia παρατίθεται ακολουθούμενο από αυτά με τα οποία συνδέεται, βάσει της έρευνάς μας. Ωστόσο, αυτή η μορφή δεν ήταν χρήσιμη για την κατασκευή του γράφου, καθώς το ζητούμενο ήταν να βρεθούν τα ζεύγη των λημμάτων που συνδέονται, οι αλλιώς των κόμβων που σχηματίζουν ακμές. Έτσι, τα παραπάνω αποτελέσματα μετατράπηκαν σε ζεύγη, οπότε και διαπιστώθηκε ότι συνολικά, ο γράφος έχει 85763367 ακμές.

Όπως και στην περίπτωση των κόμβων, χρειάστηκε άλλο ένα στάδιο επεξεργασίας, λόγω των περιορισμών του graph-tool. Έτσι, έχοντας την αντιστοίχιση λημμάτων – αριθμών που περιγράφηκε στην προηγούμενη παράγραφο, τα ζεύγη των λημμάτων μετατράπηκαν σε ζεύγη αριθμών/κόμβων. Η

λίστα αυτή είναι που τροφοδοτήθηκε στο graph-tool, κάνοντας δυνατή την κατασκευή του γράφου γνώσης.

7. Εξαγωγή της Περίληψης

Για να καταλήξουμε στα αποτελέσματα της μεθόδου, βασιζόμαστε στον υπολογισμό των ελάχιστων αποστάσεων από τον κόμβο «hurricane», που περιγράφηκε στο προηγούμενο κεφάλαιο. Με βάση αυτές, εντοπίζουμε τις λέξεις-κλειδιά που εντοπίστηκαν σε tweet και απέχουν το περισσότερο από την βασική έννοια που θέλουμε να περιγράψουμε.

Έχοντας αυτό ως δεδομένο, μπορούμε να προχωρήσουμε στην επιλογή των tweet που αποτελούν την τελική περίληψη των γεγονότων του τυφώνα Sandy, όπως αυτά καταγράφηκαν στο Twitter.

7.1. Επεξεργασία του γράφου

Έχοντας κατασκευάσει το γράφο, με βάση τα όσα περιγράφηκαν στις προηγούμενες παραγράφους, θέλουμε να προχωρήσουμε στην επεξεργασία του, ώστε να αξιοποιηθεί για την αξιολόγηση των tweet. Το σκεπτικό, όπως έχουμε εξηγήσει, είναι ότι απομακρυσμένοι κόμβοι του γράφου αναπαριστούν θέματα τα οποία έχουν μεγάλη διαφοροποίηση μεταξύ τους. Έτσι, για τους σκοπούς της εργασίας, θέλουμε να εντοπίσουμε τέτοιους κόμβους, με αποδοτικό τρόπο.

Για το λόγο αυτό, και πριν αρχίσουμε να υπολογίζουμε τις αποστάσεις μεταξύ των κόμβων που μας ενδιαφέρουν (το θέμα αυτό αναλύεται στην επόμενη παράγραφο), επιλέξαμε να βρούμε το minimum spanning tree του γράφου, καθώς ο υπολογισμός των αποστάσεων με βάση αυτό αντί για τον αρχικό γράφο είναι πολύ πιο γρήγορος λόγω της μείωσης των ακμών (θα δούμε και παρακάτω ότι ο αλγόριθμος που χρησιμοποιήσαμε για τον υπολογισμό των αποστάσεων εξαρτάται από τον αριθμό των κόμβων και των ακμών του γράφου).

Ωστόσο, τίθεται το ερώτημα αν αυτή η επιλογή μπορεί να δώσει ικανοποιητικά αποτελέσματα, δεδομένου ότι με αυτή την επιλογή αφαιρούμε από το γράφο πολλές ακμές, οι οποίες αντιπροσωπεύουν σχέσεις μεταξύ των εννοιών, και θεωρητικά είναι ακριβώς αυτό που μας ενδιαφέρει. Ο λόγος που θεωρούμε ότι αυτή η επιλογή είναι δόκιμη, είναι ότι εμείς αξιοποιούμε το γράφο γνώσης όχι για να εντοπίσουμε συγκεκριμένες σχέσεις μεταξύ εννοιών, αλλά να βρούμε συνολικά έννοιες που διαφοροποιούνται σημαντικά μεταξύ τους (το πώς δεν μας ενδιαφέρει), με κριτήριο την μεγάλη απόσταση που έχουν οι κόμβοι που τις αντιπροσωπεύουν στο γράφο.

Το κριτήριο αυτό δεν χάνεται αν πάρουμε το minimum spanning tree. Μπορεί οι αποστάσεις μεταξύ των κόμβων σε αυτό να μην είναι ίδιες με αυτές που θα βρίσκαμε στον αρχικό γράφο, όμως και από αυτό μπορούμε να εντοπίσουμε κόμβους που απέχουν σημαντικά μεταξύ τους.

7.2. Ελάχιστες αποστάσεις

Από το τελικό αρχείο που περιγράψαμε στην παράγραφο 4.6, μπορούμε να καταρτίσουμε μια λίστα με όλα τα λήμματα της Wikipedia που εμφανίζονται έστω και μία φορά σε όλο το σύνολο των tweet, δηλαδή μια λίστα από λέξεις-κλειδιά που μας ενδιαφέρουν για την εξαγωγή της περίληψης. Συνολικά, έχουμε 266517 λέξεις-κλειδιά που εμφανίζονται στα tweet. Για κάθε ένα από αυτά, υπολογίζουμε την ελάχιστη απόσταση (shortest path) από τον κόμβο «hurricane».

Ο κόμβος αυτός αποτελεί ουσιαστικά τον κόμβο αναφοράς, με βάση τον οποίο θέλουμε να μετρήσουμε όλες τις αποστάσεις. Ο λόγος που διαλέξαμε αυτόν τον κόμβο, και δεν το αφήσαμε στην τύχη, έχει να κάνει με το γεγονός ότι θέλουμε να βρούμε έννοιες που διαφοροποιούνται από τη βασική έννοια που συνοψίζουμε, εν προκειμένω τον τυφώνα Sandy. Γενικά, ο στόχος πρέπει να είναι να επιλέγεται κόμβος αναφοράς από το hashtag, τη λέξη ή τη φράση που θέλουμε να συνοψίσουμε. Αν αυτό για κάποιο λόγο δεν είναι δυνατό (δεν βρίσκεται από αυτά κάποιος όρος της Wikipedia), τότε μπορεί να επιλεγεί από τις λέξεις-κλειδιά που

βρέθηκαν στα tweet αυτή που εμφανιζόταν με τη μεγαλύτερη συχνότητα. Και με αυτόν τον τρόπο, στη συγκεκριμένη περίπτωση, θα καταλήγαμε στην ίδια επιλογή για τον κόμβο αναφοράς.

Σημειωτέον, η διαδικασία αυτή είναι αρκετά χρονοβόρα (ο αλγόριθμος ολοκληρώνεται σε O(V+E) χρόνο, όπου V είναι οι κόμβοι και E οι ακμές), ωστόσο χρειάζεται να ολοκληρωθεί μόνο μια φορά για μια συγκεκριμένη περίληψη, καθώς ο γράφος γνώσης δεν εξαρτάται από το δείγμα των tweet που επιλέγουμε κάθε φορά.

Έχοντας υπολογίσει για όλες τις λέξεις-κλειδιά το μέγεθος του shortest-path από τον κόμβο που αντιστοιχούν στον κόμβο αναφοράς, μπορούμε να τις ταξινομήσουμε με βάση την απόσταση. Μας ενδιαφέρουν οι κόμβοι, άρα και οι λέξεις κλειδιά, που έχουν τις μεγαλύτερες αποστάσεις από τον κόμβο αναφοράς. Έτσι, θέλουμε να εντοπίσουμε τις λέξεις-κλειδιά που βρέθηκαν σε tweet και νοηματικά έχουν μεγάλη διαφοροποίηση από τη βασική έννοια που θέλουμε να περιγράψουμε.

Παρακάτω παρατίθενται οι 20 πιο απομακρυσμένοι όροι (σημειωτέον δεν μπορεί να βγει κάποιο ουσιαστικό συμπέρασμα μόνο από αυτούς, ωστόσο έχει ενδιαφέρον σαν ενδιάμεσο αποτέλεσμα):

1. fourtet
2. château citran
3. nicu ceauşescu
4. château cheval blanc
5. arthur richards, 1st baron milverton
6. kensico reservoir
7. lake chabot
8. gironde
9. malani
10. olive
11. pleistocene
12. spirituality

13. teleprinter
14. vinland
15. walkman
16. toltec
17. flounder
18. regions of new zealand
19. yeti
20. cusco

Η εξαγωγή της περίληψης επιτυγχάνεται με την επιλογή tweet που να περιλαμβάνουν τις λέξεις-κλειδιά που εντοπίσαμε ότι «απέχουν» περισσότερο από το βασικό θέμα που θέλουμε να περιγράψουμε, και μάλιστα επιλέγοντας μόνο ένα tweet για κάθε τέτοια λέξη, ώστε να αποφύγουμε το ενδεχόμενο να επιλέξουμε tweet που να επαναλαμβάνονται. Έτσι, για κάθε λέξη-κλειδί που εντοπίστηκε, έχουμε ένα tweet το οποίο θα μπορούσε να αποτελεί μέρος της περίληψης. Σε γενικές γραμμές, για κάθε λέξη-κλειδί θα κρατήσουμε το tweet που την περιέχει και είναι το πιο δημοφιλές, δηλαδή έχει τα περισσότερα retweet. Ωστόσο, χρειάζεται να λάβουμε υπ' όψιν και άλλες παραμέτρους, όπως αναλύουμε στις επόμενες παραγράφους.

7.3. Γλώσσα tweets

Ως τώρα δεν μας είχε απασχολήσει το γεγονός ότι στο σύνολο δεδομένων μας περιέχονται tweets γραμμένα σε διάφορες γλώσσες, και όχι μόνο στην αγγλική (η επιλογή τους, όπως περιγράφηκε στην παράγραφο 4.1 εστίαζε στην ύπαρξη ορισμένων λέξεων ή του hashtag, επομένως δεν επαρκεί για να εξασφαλίσει ότι όλα είναι στα αγγλικά).

Ωστόσο, η περίληψη που εξάγουμε δεν μπορεί να είναι σε πολλές γλώσσες. Έτσι, κατά την επιλογή των tweet που θα αποτελέσουν την περίληψη, εξετάζουμε αν είναι γραμμένα στην αγγλική γλώσσα, με χρήση της βιβλιοθήκης langdetect της Python, η οποία εξετάζει κάθε tweet και επιστρέφει «en» αν είναι στα αγγλικά.

Έτσι, κατά την επιλογή των tweet που αντιστοιχούν σε κάθε λέξη-κλειδί, προσθέτουμε ένα παραπάνω κριτήριο από τον αριθμό retweet, τη γλώσσα, και άρα αγνοούμε όσα δεν είναι στα αγγλικά.

Σε μια μελλοντική εφαρμογή που βασίζεται σε αυτή τη μέθοδο, ο χρήστης θα μπορούσε να έχει τη δυνατότητα να επιλέξει τη γλώσσα αποτελεσμάτων που θα ήθελε να βλέπει.

7.4. Μέγεθος περίληψης

Δεδομένης της ιδιομορφίας του περιεχομένου στο Twitter, έχουμε ήδη εξηγήσει ότι το αποτέλεσμα της περίληψης θα είναι ένα σύνολο αντιπροσωπευτικών tweet. Το πόσα θα είναι αυτά τα tweet δεν είναι συγκεκριμένο, ούτε υπάρχει κάποιος κανόνας που να ορίζει τι ποσοστό των αρχικών tweet θα είναι μια καλή περίληψη.

Κατά βάση, η επιλογή του μεγέθους της περίληψης επαφίεται στις προτιμήσεις του χρήστη. Έτσι, σε μια εφαρμογή που βασίζεται σε αυτή τη μέθοδο, αυτό το νούμερο θα μπορούσε να το επιλέγει ο χρήστης. Εδώ, έχουμε επιλέξει αυθαίρετα να παραθέτουμε σε κάθε περίπτωση 20 tweet.

Συνολικά, ικανά να συμπεριληφθούν στην περίληψη είναι όλα τα tweet που έχουν αντιστοιχηθεί με κάποιο κόμβο με τον τρόπο που περιγράψαμε παραπάνω, άσχετα από το πόσο απέχουν αυτοί οι κόμβοι από τον κόμβο αναφοράς. Επομένως, η μέγιστη περίληψη που θα μπορούσαμε να πάρουμε θα είχε μέγεθος ίσο με τον αριθμό των λέξεων-κλειδιών που βρέθηκαν στα tweet (στην παρούσα εργασία, 266517 tweet).

7.5. Χρονολογική Σειρά

Προκειμένου τα αποτελέσματα να είναι πιο «ευανάγνωστα» και «εύληπτα» για το χρήστη, θεωρούμε ότι είναι προτιμότερο να εμφανίζονται σε χρονολογική σειρά. Έτσι, έχοντας δημιουργήσει τον πίνακα tweets_ordered στη βάση δεδομένων, αρκεί να αντιπαραβάλουμε τη σειρά με την οποία εμφανίζονται τα tweet ID στον πίνακα αυτόν, με τα ID των tweet που έχουμε επιλέξει για την περίληψη, και να ταξινομήσουμε αυτά με βάση τη σειρά των πρώτων.

Έτσι, ο χρήστης μπορεί να αποκτήσει μια καλύτερη αντίληψη όχι απλά για το γεγονός, αλλά και για την εξέλιξή του, κάτι που είναι πολύ σημαντικό ιδιαίτερα σε περιπτώσεις όπως το σύνολο δεδομένων που εξετάζουμε, που έχει πολύ μεγάλη χρονική διάρκεια.

Σημειωτέον, έχουμε επιλέξει η παρουσίαση των αποτελεσμάτων να γίνεται ακριβώς αντίθετα από τη σειρά που χρησιμοποιεί το Twitter. Στο Twitter ο χρήστης βλέπει με αντίστροφη χρονική σειρά τα tweet που τον ενδιαφέρουν, ωστόσο θεωρούμε ότι η περίληψη ενός γεγονότος δεν μπορεί να παρουσιάζεται έτσι.

8. Αξιολόγηση και βελτίωση συστήματος

8.1. Μέθοδος

Έχοντας διαθέσιμα τα δεδομένα για τις αποστάσεις των κόμβων που μας ενδιαφέρουν, μπορούμε να προχωρήσουμε στην επιλογή των tweet που αποτελούν την περίληψη. Το βασικό ζήτημα είναι πώς θα βελτιστοποιήσουμε αυτή την επιλογή, ώστε να εξασφαλίσουμε όσο το δυνατόν πιο πλήρη κάλυψη του θέματος, με δεδομένο βέβαια τον περιορισμό του αριθμού των tweet που θα αποτελούν την περίληψη. Αυτό το λέμε, γιατί μπορεί εδώ να παραθέτουμε μόνο 20 tweet ως περίληψη, ωστόσο είναι πολύ πιθανό ένας χρήστης που θα ζητούσε την περίληψη ενός τόσο μεγάλου συνόλου δεδομένων να ήθελε ως περίληψη και μεγαλύτερο όγκο tweet. Σε κάθε περίπτωση, το ζητούμενο είναι από τη μία, να έχουμε tweet που πιάνουν διαφορετικές πτυχές, από την άλλη να αφορούν όντως το βασικό θέμα, και πάνω σε αυτούς τους άξονες θα προσπαθήσουμε να αξιολογήσουμε ποιοτικά τα αποτελέσματα.

Σε όλες τις μεθόδους που προτείνουμε, το βασικό κριτήριο επιλογής είναι η ύπαρξη στο tweet λήμματος της Wikipedia που έχει τη μεγαλύτερη απόσταση (biggest shortest path) από τον κόμβο που αντιστοιχεί στο λήμμα «hurricane». Ωστόσο, το κριτήριο αυτό δεν αρκεί, καθώς κάθε λήμμα έχει εντοπιστεί σε πολλά tweets. Δεύτερο κριτήριο, σε κάθε περίπτωση, είναι ο αριθμός των retweet, όπως αυτά έχουν υπολογιστεί στο κεφάλαιο 4. Το κριτήριο της δημοφιλίας είναι σημαντικό, καθώς είναι ένας καλός δείκτης για το κατά πόσο έχουμε εντοπίσει μια άποψη που είναι διαδεδομένη, μια άποψη που περισσότεροι χρήστες επιλέγουν να διαδώσουν, χωρίς βέβαια από μόνο του να είναι αρκετό. Στις παραγράφους 8.1.2 – 8.1.5 προσθέτουμε παραπάνω κριτήρια για την επιλογή tweet, προκειμένου να έχουμε πιο ουσιαστικά αποτελέσματα.

Σε κάθε περίπτωση, παίρνουμε τα αποτελέσματα της επεξεργασίας των δεδομένων μας, όπως περιγράψαμε στην παράγραφο 5.5, και τα αντιπαραβάλουμε με τους όρους που έχουν προκύψει από την επεξεργασία του γράφου, με βάση τις

ελάχιστες αποστάσεις. Για κάθε όρο, εντοπίζουμε τα tweet που τον περιέχουν, και εξετάζουμε αν κάποιο από αυτά πληροί όλους τους όρους που έχουμε θέσει (στην περίπτωση του χρονικού βήματος χρειάζεται να ανατρέξουμε και στη βάση δεδομένων). Αφού επιλέξουμε τα tweet που αποτελούν την περίληψη, τα ταξινομούμε χρονολογικά, όπως περιγράφηκε στην παράγραφο 7.5, και έτσι καταλήγουμε στην τελική έξοδο του συστήματος.

8.2. Αποτελέσματα

8.2.1. Επιλογή tweet από το αρχικό σύνολο δεδομένων

Αρχικά, παραθέτουμε τα αποτελέσματα αν κρατήσουμε για τους πιο απομακρυσμένους όρους το tweet με το μέγιστο αριθμό retweet που τους περιλαμβάνει:

1. Danny Dunn and the #Weather Machine No 10 Review http://t.co/NfsuLnvo #mountain #crystal #crystalmountainweather

2. @drbarq General Dyer was the one who actually ordered the firing on the people gathered in Jallianwala Bagh. Michael O'Dwyer was Governor.

3. Richebourg Features With Cheval Blanc at Trotter’s Sale http://t.co/UKFkxWCQ #AuctionNews @BloombergNews #NewYork @ChristiesInc #BWC

4. @Samsung_India #BeCreative with #QuickCommand I would book tkts frm #IRCTC in Malani Express ;nw thts real power

5. @rosepowell the author William Hope Hodgson not only survived a hurricane at sea, but took photos during it,

6. I'm a little disappointed with my Christian social media friends that haven't posted any fresh snow forgiveness analogies with the weather.

7. Healing myself with the power of Neo Citran and Pac-Man

8. Court-martial of Lt-Gen Gu Junshan wld become the biggest mil. corruption scandal since Communists swept to power. http://t.co/Ubds1zJD

9. If i had Natural Magic, Sandy would die, Bitch! #HurricaneSandy

10. Jamie xx has remixed Fourtet's Lion. It's a pretty good soundtrack for dark, rainy, hurricaney weather: http://t.co/P4DF96RJ

11. Tree and a collum fell down at my appartment... the storm isn't even here yet,..

12. wow randolph still isn't 100% but ccm has power #fuckery

13. STORM OF THE CENTURY, WITH THE POWER OF OVER 9000 SUNS, has been downgraded to a tropical storm..

14. Kensico is out of power . Please don't come over here

15. No hurricane hangover in Media where we are having our 1st rehearsal of Dr Doolittle. Opening November 20 th at Media Theatre.

16. Despite this storm I still have to be The Black Unicorn for Halloween! @1987Marty ❤💙 #NYG

17. From a 'small, sad girl' to chaos in the corridors of power http://t.co/E8LJBNSN via @smh MM shld have kept her role as the JWH Dragonslayer

18. @Bred_Red Expendable goods in march 2 retain power.They-big govt-didn't forget.2 bad NE banked on big govt.$ goes 2 big govt,$ stays there

19. Such a good weather! (@ Lake Chabot Trail Challenge) on #Yelp http://t.co/G5Ujy7Qd

20. Who was the first governor of a unified Nigeria?

A-Sir Arthur Richards

B-Sir Hugh Clifford... http://t.co/sDQwHglz

Παρατηρούμε ότι πολλά tweet δεν είναι σχετικά με το θέμα που θέλουμε να συνοψίσουμε, και άρα η περίληψη δεν είναι αυτή που θα θέλαμε. Παρακάτω θα επιχειρηθεί να προστεθούν παραπάνω φίλτρα στην επιλογή, ώστε να γίνει πιο έγκυρη η περίληψη.

Παρ' όλα αυτά, από τώρα μπορούμε να παρατηρήσουμε ότι ο βασικός στόχος που είχαμε βάλει με την χρήση του γράφου γνώσης, δηλαδή να εντοπίσουμε tweet που δεν επαναλαμβάνεται το περιεχόμενό τους, έχει επιτευχθεί. Οι παρεμβάσεις, επομένως, που προτείνονται στις επόμενες παραγράφους έχουν στόχο να βελτιώσουν την εγκυρότητα και την πληρότητα των αποτελεσμάτων.

8.2.2. Επιλογή tweet μόνο εφ' όσον έχουν πάνω από τον μέσο όρο των retweet

Σε αυτή τη δοκιμή, αγνοήσαμε τους όρους που ενώ είχαν μεγάλη απόσταση από τον αρχικό κόμβο, και περιλαμβάνονταν σε κάποιο tweet, αυτά είχαν λιγότερα από 4 retweet (όπως περιγράφηκε στην παράγραφο 4.3.1 ο μέσος όρος retweet είναι 3,95). Θέλαμε έτσι να αποφύγουμε αποτελέσματα που μπορεί να ήταν άσχετα με το βασικό θέμα (άρα είναι πιθανό να ήταν λιγότερο δημοφιλή συγκριτικά με το σύνολο των tweet τις μέρες που εξελισσόταν ένα τόσο σημαντικό γεγονός).

1. @spudgun01 Blair was in power when Waterhouse reported, and during Dunblane cover up too. All in this together!! #coverups

2. Court-martial of Lt-Gen Gu Junshan wld become the biggest mil. corruption scandal since Communists swept to power. http://t.co/Ubds1zJD

3. Dolphins dealing with 25' waves right now in Hurricane Sandy. In my book Dolphin Diaries I describe 30% loss in pod. Lets hope 4 best

4. Amber Alert now issued for black Nissan Altima MD tags 8L7618M Children inside (5, 4, & 1 yr old) taken from #SeatPleasant gas station

5. Congrats to Storm ninth grader Hayley Haakenstad on winning the #MSHSL State Tennis Consolation Championship! Way to go Hayley! #tennisstar

6. The Village Voice named Everyman Espresso's new Soho location as the Best Espresso Bar in New York.... http://t.co/dq36kkqu

7. Ken Gray WVU- "We are monitoring the Weather situation but class is still in session so if you die we dont care"

8. Romaine said one of first goals will be to get power back in Brookhaven, "even if we have to blow up LIPA to do it." #LongIsland #Brookhaven

9. STORM OF THE CENTURY, WITH THE POWER OF OVER 9000 SUNS, has been downgraded to a tropical storm..

10. Power restored to Lothian ES, Magothy River MS, Severn River MS, Southern MS. All four schools will reopen on time TOMORROW, Nov. 2.

11. As Power and Subways Return to New York, Normalcy Is Still A Long Way Off - by @ahess247 http://t.co/1ByHtJpC

12. Love George RR Martin's books? Enjoy the tales of David Gemmell, you may enjoy Whispers of a Storm http://t.co/8OgpftXr #fantasy #kindle

13. According to the National Oceanic and Atmospheric Administration, Oakland has the best climate in the US. Shh, don't tell the neighbors...

14. Japanese firms today announce purchases of a UK nuclear power plant and the Branston pickle brand. And I thought my shopping lists were odd.

15. Max Abelson tells how Wall Street's elite survived #Sandy with fine wine and Monopoly: http://t.co/HoOJH6yr via @BloombergNews @maxabelson

16. Is the weather depressing you? Cheer yourself up with a #SciFi #Fantasy #Adventure #Series The Guardians

17. RT Diesel: The Lifeblood of the Recovery Effort: Generators power data centers, much of lower Manhattan. http://t.co/yzJPp72F #sandy

18. Oy amish living continues w no power at the farm. Will they believe me at Rolex that I cracked my watch on a manure spreader in the dark??!

19. Deborah Lippmann reminisces on the transformative power of porcelain nails: http://t.co/ATPFSMQi

20. Total voting activity in the HUB today: 6091 total votes: 3622 for President Obama (59%), 2257 for Governor Romney (37%).

Και εδώ, παρατηρούμε ότι υπάρχουν περιθώρια βελτίωσης, καθώς ακόμα η μέθοδος βρίσκει tweet τα οποία δεν θα έπρεπε να περιλαμβάνονται στην περίληψη (π.χ. τα αποτελέσματα 12 και 16).

8.2.3. Επιλογή tweet που περιέχουν τη λέξη «hurricane»

Ως τώρα δουλέψαμε με το σύνολο των tweet. Ωστόσο, όπως είχαμε περιγράψει στην παράγραφο 3.1, το σύνολο των δεδομένων μας περιέχει tweet που αντλήθηκαν με την αναζήτηση πολλαπλών λέξεων-κλειδιών. Ωστόσο, το σύνηθες στην περίληψη περιεχομένου στο Twitter είναι αυτή να γίνεται με βάση μια λέξη, φράση ή ένα hashtag, και όχι περισσότερα από ένα, και αντίστοιχα το σύνολο των tweet θα αντλείται με κριτήριο να περιέχει το συγκεκριμένο.

Μάλιστα, αν εξετάσουμε προσεκτικά τα αποτελέσματα των προηγούμενων δύο παραγράφων, θα δούμε ότι τα tweet που δεν θα έπρεπε να απαρτίζουν την περίληψη, περιλαμβάνουν όρους που είχαν αξιοποιηθεί για τη συλλογή του συνόλου των δεδομένων γιατί είναι εν δυνάμει σχετικοί με το βασικό θέμα, τον τυφώνα Sandy, αλλά στις συγκεκριμένες περιπτώσεις αναφέρονται σε κάτι άλλο. Είναι λογικό, βέβαια, το σύστημα να εντοπίζει και να ξεχωρίζει τέτοια tweet, καθώς όντως το περιεχόμενό τους διαφοροποιείται σημαντικά από το βασικό θέμα (και με αυτό τον τρόπο δηλαδή φαίνεται η συνεισφορά του γράφου γνώσης). Ωστόσο, χρειάζεται να βελτιωθεί το σύστημα, ώστε αυτά τα tweet στο τέλος να απορρίπτονται, και ο πιο σίγουρος τρόπος για να εξασφαλιστεί αυτό είναι να απαιτήσουμε τα τελικά tweet να περιλαμβάνουν το βασικό θέμα προς περίληψη.

Έχοντας συμπεράνει από τις προηγούμενες δύο παραγράφους ότι παίρνουμε καλύτερα αποτελέσματα όταν κρατάμε tweet που έχουν παραπάνω από το μέσο όρο retweet αποκλειστικά, θα συνεχίσουμε έτσι και εδώ. Έτσι, επιλέξαμε τα αποτελέσματα με τον ίδιο τρόπο που περιγράφηκε στην παράγραφο 8.1.2, ωστόσο αυτή τη φορά κρατήσαμε μόνο όσα περιλάμβαναν την λέξη «hurricane».

1. Image of Hurricane Sandy off the US east coast, with maximum sustained wind speeds of 140 km/h, captured by #MetopA. http://t.co/13EHhrH9

2. Latest track from National Hurricane Center has Sandy with a direct hit on New Jersey. Kathy Orr has the latest on #CBS3 at 11.

3. Dolphins dealing with 25' waves right now in Hurricane Sandy. In my book Dolphin Diaries I describe 30% loss in pod. Lets hope 4 best

4. Hurricane Sandy 25 miles from Great Exuma Island, in Bahamas. http://t.co/LMc3hqwV

5. Lower Manhattan during Hurricane Donna in 1960. #sandy #frankenstorm http://t.co/8cePUV4W

6. I hope this hurricane blows away all the stinkbugs

7. Is Bryan Stork excited for Miami week? "Yeah, it's hurricane season, man."

8. Video: Meteorologist David Bernard with latest on Hurricane Sandy http://t.co/yCAmk9mo

9. #NationalTextYourExDay is rather go lay in a beach chair right in the path of #HurricaneSandy

10. Homa 3andohom hurricane esmo sandy we7na 3andena nawa esmaha amsheer w nerga3 ne2ool leh homa 3andohom el mozaz

11. Hurricane Hunters confirm #Sandy is now the most intense hurricane ever north of NC, beating The Great Hurricane of 1938.

12. My condolences go out to Brooklyn this morning, as Hurricane LeBron le mando pa la pinga last night.

13. Hope you all reason with hurricane season ok! (@ Frankenstorm Apocalypse - Hurricane Sandy w/ 2886 others) http://t.co/UrHlFvkc

14. Hurricane Essentials: Beef Jerky, Guiness, Bacon Hotsauce, Slap Yo Mama rub, Brisket, Kimber 1911. @InfidelJustice @trentj1 @CagedSoutherner

15. "hurricane" "#sandy" is just an elaborate ploy by the left to give #obummer more time in office. not even real waves. ever heard of cgi?

16. @SillyBdilly ill bet Kathy Bertrand is having a field day with this hurricane

17. 2.0 magnitude quake rocks northern New Jersey. It was centered in Ringwood, which is still trying to recover from Hurricane #Sandy.

18. Hurricane Sandy covers an area from South America to the Sargasso Sea. An eventual threat to Northeast USA (Mon) http://t.co/QIWph9TE

19. The Cove Restaurant in Cape May has been destroyed by Hurricane Sandy. http://t.co/LDYI8wV0

20. Obama is not the reason for Hurricane Sandy, he is not the reason for Firebaugh's lockdown. Your ignorance is the reason for your stupidity.

8.2.4.Επιλογή tweet που περιέχουν τις λέξεις «hurricane» και «sandy»

Στη συνέχεια των αποτελεσμάτων της προηγούμενης παραγράφου, που ήδη είναι σημαντικά πιο προσανατολισμένα στο ζητούμενο θέμα, προσθέτουμε σαν παράμετρο την ύπαρξη και της λέξης «sandy» για την επιλογή των tweet.

1. Image of Hurricane Sandy off the US east coast, with maximum sustained wind speeds of 140 km/h, captured by #MetopA. http://t.co/13EHhrH9

2. Latest track from National Hurricane Center has Sandy with a direct hit on New Jersey. Kathy Orr has the latest on #CBS3 at 11.

3. Dolphins dealing with 25' waves right now in Hurricane Sandy. In my book Dolphin Diaries I describe 30% loss in pod. Lets hope 4 best

4. Hurricane Sandy 25 miles from Great Exuma Island, in Bahamas. http://t.co/LMc3hqwV

5. Lower Manhattan during Hurricane Donna in 1960. #sandy #frankenstorm http://t.co/8cePUV4W

6. Instead of hating on Kelsey or the girl who made the video, focus on a bigger picture. A hurricane just destroyed the East Coast. #sandyhelp

7. Region I and Region II family, please be safe in this weather! #HurricaneSandy

8. Video: Meteorologist David Bernard with latest on Hurricane Sandy http://t.co/yCAmk9mo

9. Louisville's UPS Worldport hub about to become much busier, thanks to Hurricane Sandy http://t.co/k5EQw5bz

10. #NationalTextYourExDay is rather go lay in a beach chair right in the path of #HurricaneSandy

11. Homa 3andohom hurricane esmo sandy we7na 3andena nawa esmaha amsheer w nerga3 ne2ool leh homa 3andohom el mozaz

12. Hurricane Hunters confirm #Sandy is now the most intense hurricane ever north of NC, beating The Great Hurricane of 1938.

13. Hope you all reason with hurricane season ok! (@ Frankenstorm Apocalypse - Hurricane Sandy w/ 2886 others) http://t.co/UrHlFvkc

14. "hurricane" "#sandy" is just an elaborate ploy by the left to give #obummer more time in office. not even real waves. ever heard of cgi?

15. 2.0 magnitude quake rocks northern New Jersey. It was centered in Ringwood, which is still trying to recover from Hurricane #Sandy.

16. Sandy McNeill would never let me down like hurricane Sandy did...@ginamcneill_

17. Hurricane Sandy covers an area from South America to the Sargasso Sea. An eventual threat to Northeast USA (Mon) http://t.co/QIWph9TE

18. Gerald Flurry analyzes hurricane Sandy from God's perspective in his new KoD Web Excl. to be released this evening at http://t.co/wKmHurfU.

19. The Cove Restaurant in Cape May has been destroyed by Hurricane Sandy. http://t.co/LDYI8wV0

20. Obama is not the reason for Hurricane Sandy, he is not the reason for Firebaugh's lockdown. Your ignorance is the reason for your stupidity.

Παρατηρούμε ότι πλέον τα αποτελέσματα είναι σαφέστατα προσανατολισμένα στην κάλυψη του βασικού θέματος, επομένως μπορούμε ασφαλώς να συμπεράνουμε ότι έχουμε βελτιώσει με αυτά τα μέτρα την εγκυρότητα των αποτελεσμάτων του συστήματος.

8.2.5. Επιλογή tweet με χρονικό βήμα

Ως τώρα ο χρόνος μας έχει απασχολήσει μόνο για την τελική παρουσίαση των tweet. Ωστόσο, δεν έχουμε εξετάσει πώς τα tweet που επιλέγουμε κατανέμονται στο χρονικό διάστημα που εξετάζουμε.

Τίθεται το ερώτημα: αν τα tweet που επιλέγονται, που όντως φωτίζουν διαφορετικές πτυχές του βασικού θέματος, καταφέρνουν να αποδώσουν με μια σχετική πληρότητα (σχετική γιατί εξετάζουμε μια περίληψη μεγέθους 20 tweet πάνω σε αρχικό σύνολο πάνω από 30 εκατομμύρια tweet) το θέμα στην εξέλιξή του, για το διάστημα που εξετάζουμε. Είναι χαρακτηριστικό, ότι τα tweet που παρουσιάστηκαν στην προηγούμενη παράγραφο, κατανέμονται στο χρόνο που εξετάζουμε ως εξής: την πρώτη εβδομάδα δεν υπάρχει κανένα, τη δεύτερη δύο, την τρίτη 17 και την τέταρτη ένα.

Έτσι, σαν τελευταία βελτίωση προτείνεται η επιλογή tweet που να πληρούν όλα τα κριτήρια που χρησιμοποιήθηκαν στην παράγραφο 8.2.4 (καθώς τα αποτελέσματα αυτής ήταν, ποιοτικά, αυτά που καλύτερα ικανοποιούσαν τους στόχους που είχαμε θέσει εξαρχής), αλλά επιπλέον απαιτούμε να είναι ίδιος ο αριθμός των tweet ανά εβδομάδα εξέτασης. Με αυτόν τον τρόπο επιδιώκουμε να έχουμε πληρέστερη κάλυψη του θέματος, καθώς έτσι δεν θα χαθεί χρήσιμη πληροφορία που μπορεί στα προηγούμενα βήματα να αγνοήθηκε.

Το σύνολο δεδομένων που εξετάζουμε χωρίζεται σε 4 εβδομάδες. Προκειμένου να διαλέξουμε tweet με χρονικό βήμα, θα μετατρέψουμε κάθε αρχή και τέλος της εβδομάδας σε timestamp, μεταξύ των οποίων θα ζητήσουμε tweet από τη βάση δεδομένων, και στη συνέχεια θα εφαρμόσουμε τα κριτήρια που περιγράφηκαν στις προηγούμενες παραγράφους, ώστε για κάθε εβδομάδα να διαλέξουμε 5 tweet.

1. 11pm Advisory on Hurricane #Sandy: 75mph winds, pressure has dropped slightly to 969mb. Chris Bradley has the… http://t.co/jxc3gtB3

2. Statement Regarding Changes to Open Market Operations Due to Hurricane Sandy: The Federal Reserve Bank of New Yo… http://t.co/PQwpfzQ9

3. Passing this along: Cut-a-Thon to Benifit Hurricane Sandy Victims, with Brian O'Halloran from Clerks - Toms River, http://t.co/5BOaZw6V

4. Due to the extenuating circumstances with Hurricane Sandy, the Alchemy Songwriting Competition deadline is extended to November 7.

5. Support your local 7 Mile Island businesses. Many of our merchants have undergone Hurricane Sandy damage but have... http://t.co/wFJw72XC

6. get your "Crippler" Shirt, I will donate $5 per shirt we sell to victims of #HurricaneSandy

@ufc @danawhite http://t.co/EOJsZSTx

7. Grand Bahamas airport as well as Stella Maris, The Bight and Arthur's Town are open after Hurricane Sandy.

8. Jazz for Hurricane Sandy Benefit Tonight Le Poisson Rouge 6-2am

$20 @JAZZFOUNDATION @AwildaRivera @jazzsea12 @WBGO http://t.co/AGH12vtN

9. Image of Hurricane Sandy off the US east coast, with maximum sustained wind speeds of 140 km/h, captured by #MetopA. http://t.co/13EHhrH9

10. Latest track from National Hurricane Center has Sandy with a direct hit on New Jersey. Kathy Orr has the latest on #CBS3 at 11.

11. Dolphins dealing with 25' waves right now in Hurricane Sandy. In my book Dolphin Diaries I describe 30% loss in pod. Lets hope 4 best

12. Hurricane Sandy 25 miles from Great Exuma Island, in Bahamas. http://t.co/LMc3hqwV

13. Video: Meteorologist David Bernard with latest on Hurricane Sandy http://t.co/yCAmk9mo

14. #NationalTextYourExDay is rather go lay in a beach chair right in the path of #HurricaneSandy

15. Homa 3andohom hurricane esmo sandy we7na 3andena nawa esmaha amsheer w nerga3 ne2ool leh homa 3andohom el mozaz

16. Hurricane Hunters confirm #Sandy is now the most intense hurricane ever north of NC, beating The Great Hurricane of 1938.

17. 2.0 magnitude quake rocks northern New Jersey. It was centered in Ringwood, which is still trying to recover from Hurricane #Sandy.

18. Hurricane Sandy covers an area from South America to the Sargasso Sea. An eventual threat to Northeast USA (Mon) http://t.co/QIWph9TE

19. The Cove Restaurant in Cape May has been destroyed by Hurricane Sandy. http://t.co/LDYI8wV0

20. Obama is not the reason for Hurricane Sandy, he is not the reason for Firebaugh's lockdown. Your ignorance is the reason for your stupidity.

Σε μια ενδεχόμενη επέκταση της παρούσας εργασίας, να σημειωθεί ότι θα μπορούσε ο χρήστης να επιλέξει συγκεκριμένο αριθμό tweet για κάθε χρονική περίοδο (π.χ. να ζητήσει περισσότερα tweet για την εβδομάδα που έφτασε ο τυφώνας Sandy σε μια συγκεκριμένη χώρα), ή να περιορίσει όλα τα αποτελέσματα σε ένα υποσύνολο των tweet με βάση το χρόνο.

8.3. Πλεονεκτήματα της μεθόδου

Για να μπορέσουμε να εντοπίσουμε τα πλεονεκτήματα της μεθόδου μας, μπορούμε να αξιοποιήσουμε ένα παράδειγμα μιας άλλης μεθόδου περίληψης. Παρακάτω φαίνονται τα αποτελέσματα της μεθόδου που χρησιμοποιεί σαν κριτήριο τη δημοφιλία των tweet για να εξάγει περίληψη. Συγκρίνουμε με τα αποτελέσματα του κριτηρίου της δημοφιλίας, πρώτον γιατί είναι μια αποδεκτή μέθοδος περίληψης περιεχομένου στο Twitter [21], και δεύτερον γιατί μας επιτρέπει να αξιολογήσουμε τη συνεισφορά (ή μη) του γράφου γνώσης στην επιλογή των tweet.

Συγκεκριμένα, χρησιμοποιήσαμε το ίδιο σύνολο δεδομένων και η μέθοδος επέστρεψε ως περίληψη τα πιο δημοφιλή tweets, δηλαδή αυτά που αναπαράχθηκαν περισσότερο (περισσότερα retweet) και πήραμε τα εξής αποτελέσματα:

1. This London weather is class don't you think :/ x

2. Everyone in the path of the hurricane should head to their second or third home to safety #Sandy #RomneyStormTips

3. What if Gangnam Style was actually a giant rain dance and we've brought this on ourselves? #sandy

4. everyone dealing with the hurricane up north be safe

5. RETWEET IF YO LUCKY ASS STILL GOT POWER

6. RETWEET IF YO LUCKY ASS STILL GOT POWER

7. Karens boobs are probably going crazy with all this rain. #HurricaneSandy

8. WHAT IF GANGAM STYLE WAS ACTUALLY JSST A GIANT RAIN DANCE AND WE BROUGHT THIS HURRICANE ON OURSELVES?

9. FOR EVERY 100 RETWEETS, WE WILL BE DONATING $1,000 TO HELP REBUILT COMMUNITIES DAMAGED BY HURRICANE SANDY. PLEASE RETWEET!!!

10. Everyone stay safe tonight!

11. Help #Sandy survivors. Donate to the @RedCross via @iTunesMusic at http://t.co/o1B7TY1W. Out of US, visit http://t.co/nhdkPvqA

12. Gonna try to work with Red Cross to have u guys all help donate for those affected by hurricane sandy

13. Obama won Massachusetts the state where Romney was Governor. If they didn't vote for him, that should tell you something. #Swerve #Obama2012

14. Obama won Massachusetts the state where Romney was Governor. If they didn't vote for him, that should tell you something. #Swerve #Obama2012

15. Good rehearsal this morning. Storm is coming.

16. We decided that for the Jersey, Brooklyn, and 2 MSG shows every ticket sold a dollar will be donated to Hurricane Sandy Relief. #GIVEBACK

17. Today the #BELIEVEtour hits New Jersey and I am proud that we will donate a portion of ALL tix sold to Hurricane Sandy Relief. #GIVEBACK

18. thank u everyone for coming out last night and supporting in New Jersey...we raised money for Hurricane Sandy relief. #GIVEBACK #GREATSHOW

19. I hear some of you guys are camping outside the today show! Make sure u stay safe and warm! And see ya tuesday! Xx

20. Ok we're ready! Rehearsals done for the Today Show!! New york you ready?

Παρατηρώντας τα tweets που απαρτίζουν την περίληψη, φαίνεται ότι, ενώ η μέθοδος επιτυγχάνει να επιστρέφει περιεχόμενο σχετικό με το θέμα (και με αντίστοιχες βελτιώσεις που προτάθηκαν για τη δική μας μέθοδο θα μπορούσε να βελτιωθεί πιθανώς κι άλλο η εγκυρότητα του περιεχομένου), αυτό είναι επαναλαμβανόμενο, με αποτέλεσμα η περίληψη να είναι πολύ φτωχή από άποψη πληροφορίας (π.χ. τα τελευταία 5 tweet αναφέρονται σε συναυλία του Justin Bieber εκείνη την περίοδο, ενώ βλέπουμε ακόμα και ολόιδια tweet να έχουν επιλεχθεί).

Σε αντίθεση με αυτήν την προσέγγιση, η μέθοδος που προτείνουμε καταφέρνει να παράγει περιλήψεις που:

- είναι σχετικές με το θέμα ενδιαφέροντος, αφού χρησιμοποιούν λέξεις κλειδιά σχετικά με το θέμα και έχουν αναπαραχθεί στο δίκτυο από πολλούς χρήστες
- δεν περιγράφουν το θέμα μονόπλευρα, καθώς αξιοποιώντας τον γράφο γνώσης, περιλαμβάνουν σημασιολογικά διαφοροποιημένο περιεχόμενο. Αυτό έχει ως αποτέλεσμα οι παραγόμενες περιλήψεις να μην υποφέρουν από επαναλήψεις και να αντανακλούν πληρέστερα την εξέλιξη ενός γεγονότος, αφού περιγράφουν διαφορετικές πτυχές του.

9. Επίλογος

9.1. Συμπεράσματα

Ο μεγάλος όγκος των δεδομένων που παράγονται καθημερινά στα μέσα κοινωνικής δικτύωσης καθιστά αδύνατη την παρακολούθησή τους από τους χρήστες χωρίς εξειδικευμένα εργαλεία. Τα ίδια τα μέσα, προσπαθούν να λύσουν αυτό το ζήτημα με την εφαρμογή αλγορίθμων προτάσεων, ώστε οι χρήστες να βλέπουν προσωποποιημένα timeline με περιεχόμενο που, θεωρητικά, τους ενδιαφέρει περισσότερο.

Ωστόσο, αυτό έχει σαν αποτέλεσμα οι χρήστες που στρέφονται στα μέσα κοινωνικής δικτύωσης για ενημέρωση να δυσκολεύονται να αποκτήσουν άμεσα, εύκολα και γρήγορα αντίληψη των γεγονότων που συζητιούνται σε αυτά. Επομένως, καθίσταται επιτακτική η ανάπτυξη μεθόδων που θα παράγουν σύντομες και πλήρεις περιλήψεις πάνω στο περιεχόμενο που ενδιαφέρει τους χρήστες.

Σε αυτή την εργασία, προτείναμε μια μέθοδο περίληψης περιεχομένου στο Twitter. Η διαφορά της συγκεκριμένης μεθόδου, σε σχέση με άλλες υπάρχουσες πάνω στο ίδιο θέμα, έγκειται στην αξιοποίηση ενός γράφου γνώσης για την αξιολόγηση της σημασιολογικής διαφοροποίησης των tweet. Ταυτόχρονα, αξιοποιούνται κριτήρια όπως η δημοφιλία των tweet (όπως αυτή μετριέται σε retweet, δηλαδή πόσες φορές άλλοι χρήστες επέλεξαν να το αναπαράγουν) και η ύπαρξη συγκεκριμένων λέξεων-κλειδιών, αντιπροσωπευτικών του βασικού θέματος προς περίληψη, για να εξασφαλιστεί ότι τα tweet θα είναι σχετικά με το θέμα.

Τα αποτελέσματα δείχνουν ότι η μέθοδος αυτή προσφέρεται για περαιτέρω ανάπτυξη, καθώς επιτυγχάνει σε σημαντικό βαθμό τους σκοπούς που θέσαμε, δηλαδή να επιλέγει tweet τα οποία να αποδίδουν με πληρότητα και ποικιλομορφία το εξεταζόμενο γεγονός, χωρίς επαναλήψεις. Φάνηκε από τα αποτελέσματα ότι η συνεισφορά του γράφου γνώσης για την αξιολόγηση του περιεχομένου των tweet είναι καθοριστική.

9.2. Μελλοντικές Επεκτάσεις

Η μέθοδος αυτή θα μπορούσε σε μελλοντική της επέκταση να βελτιωθεί σημαντικά, τόσο ως προς την απόδοσή της, όσο και ως προς την ποικιλία των επιλογών που παρέχει στο χρήστη.

Καταρχάς, μπορούν εύκολα να δοθούν στο χρήστη του συστήματος πολλές επιλογές για την παραμετροποίηση της περίληψης που θα λάβει, όπως:

- Επιλογή του μεγέθους της περίληψης, συνολικά και ανά συγκεκριμένο χρονικό παράθυρο
- Επιλογή των στοιχείων που θα έχει για κάθε tweet η περίληψη (όπως για το χρήστη που έκανε το tweet, το χρόνο που έγινε κ.ά.)
- Επιλογή των λέξεων-κλειδιών που θεωρεί βασικά ως προς την περίληψη και πρέπει να περιλαμβάνονται σε κάθε tweet που την απαρτίζει (μπορεί να είναι και κάποιο hashtag)
- Επιλογή της γλώσσας των tweet που θα απαρτίζουν την περίληψη (αρχικά μπορεί να επιλέξει αν θα φιλτράρει τα αποτελέσματα για να κρατήσει μόνο τα αγγλικά ή όχι, στη συνέχεια και με την εύρεση εφαρμογών wikifier για άλλες γλώσσες θα μπορούσε να επεκταθεί σε αυτές)

Τέλος, ενδιαφέρουσα θα ήταν η επέκταση της μεθόδου για streaming δεδομένα. Στην περίπτωση αυτή θα μπορούσαν να αξιοποιηθούν εργαλεία big data όπως το Storm, το ElasticSearch κ.ά.

10. Βιβλιογραφία

[1] Twitter, "Twitter Help Center - Getting Started," [Online]. Available: https://help.twitter.com/en/twitter-guide. [Accessed September 2018].

[2] Twitter, "Twitter trends FAQs," [Online]. Available: https://help.twitter.com/en/using-twitter/twitter-trending-faqs. [Accessed September 2018].

[3] "Internet Live Stats," [Online]. Available: http://www.internetlivestats.com/twitter-statistics/.

[4] T.-Y. Kim, J. Kim, J. Lee and J.-H. Lee, "A Tweet Summarization Method Based on a Keyword Graph," in *ICUIMC '14 Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, New York, NY, USA, 2014.

[5] Twitter, "About your Twitter timeline," [Online]. Available: https://help.twitter.com/en/using-twitter/twitter-timeline. [Accessed September 2018].

[6] Twitter, "Twitter search FAQs," [Online]. Available: https://help.twitter.com/en/using-twitter/top-search-results-faqs. [Accessed September 2018].

[7] L. V. Nair, "How Is Twitter Ranking Your Tweets for Hashtag Searches, Exactly?," 10 May 2018. [Online]. Available: https://www.zoomowl.com/twitter-hashtag-ranking-algorithm/. [Accessed September 2018].

[8] K. Garimella, G. De Francisci Morales, A. Gionis and M. Mathioudakis, "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *WWW 2018: The 2018 Web Conference*, Lyon, France, 2018.

[9] D. Pla Karidi, Y. Stavrakas and Y. Vassiliou, "A Personalized Tweet Recommendation Approach Based on Concept Graphs," in *Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*, 2016.

[10] Z. S. Syed, T. Finin and A. Joshi, "Wikipedia as an Ontology for Descibing Documents," in *International Conference on Weblogs and Social Media*, Seattle, Washington, 2008.

[11] M. Strube and S. P. Ponzetto, "WikiRelate! Computing Semantic Relatedness Using Wikipedia," in *AAAI-06*, Boston, Massachusetts, 2006.

[12] L. Ratinov, D. Roth, D. Downey and M. Anderson, "Local and Global Algorithms for Disambiguation to Wikipedia," in *ACL*, 2011.

[13] M. Rada and A. Csomai, "Wikify! Linking Documents to Encyclopedic Knowledge," in *CIKM'07*, Lisboa, Portugal, 2007.

[14] D. Milne and I. Witten, "Learning to Link with Wikipedia," in *CIKM'08*, Napa Valley, California, 2008.

[15] Z. Ren, S. Liang, E. Meij and M. De Rijke, "Personalized Time-Aware Tweets Summarization," in *SIGIR'13*, Dublin, Ireland, 2013.

[16] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li and H. Wang, "Entity-Centric Topic-Oriented Opinion Summarization in Twitter," in *KDD'12*, Beijing, China, 2012.

[17] Y. Chang, X. Wang, Q. Mei and Y. Liu, "Towards Twitter Context Summarization with User Influence Models," in *WSDM'13*, Rome, Italy, 2013.

[18] S. Petrovic, M. Osborne and V. Lavrenko, "Streaming First Story Detection with application to Twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010.

[19] B. Sharifi, M. A. Hutton and J. Kalita, "Experiments in Microblog Summarization," in *2010 IEEE Second International Conference on Social Computing*, 2010.

[20] M. Asif Hossain Khan, D. Bollegala, G. Liu and K. Sezaki, "Multi-Tweet Summarization of Real-Time Events," in *2013 International Conference on Social Computing*, 2013.

[21] N. Alsaedi, P. Burnap and O. Rana, "Automatic Summarization of Real World Events Using Twitter," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 2016.

[22] T.-Y. Kim, J. Kim and J. Lee, "A Tweet Summarization Method Based on a Keyword Graph," in *IMCOM (ICUIMC)'14*, Siem Reap, Cambodia, 2014.

[23] X. Liu, Y. Li, F. Wei and M. Zhou, "Graph-based Multi-tweet Summarization Using Social Signals," in *Proceedings of COLING 2012: Technical Papers*, Mumbai, 2012.

[24] S. Dutta, S. Ghatak, M. Roy, S. Ghosh and A. K. Das, "A Graph Based Clustering Technique for Tweet Summarization," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015.

[25] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck and M. Cebrian, "Performance of Social Network Sensors during Hurricane Sandy," *PLoS ONE 10(2): e0117288,* 2015.

[26] "Psycopg – PostgreSQL database adapter for Python," [Online]. Available: http://initd.org/psycopg/docs/index.html. [Accessed September 2018].

[27] [Online]. Available: https://graph-tool.skewed.de/performance.