ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Α.Π.Θ. Τμήμα Πληροφορικής
Εαρινό εξάμηνο 2016-2017

Διδάσκουσα: Καθηγ. Αθηνά Βακάλη
Υπεύθυνος εργασίας: Δημητριάδης Ηλίας

Θέμα εργασίας

Τα κοινωνικά δίκτυα αποτελούν αναπόσπαστο κομμάτι της καθημερινότητας των ανθρώπων. Δεν είναι λίγοι εκείνοι που χρησιμοποιούν τα κοινωνικά δίκτυα για να αναφερθούν σε φλέγοντα ζητήματα της επικαιρότητας, να σχολιάσουν ένα γεγονός που έλαβε χώρα στην περιοχή τους, να αξιολογήσουν μία παράσταση ή μία ταινία. Τα παραπάνω αποτελούν ένα μόνο μικρό δείγμα των δραστηριοτήτων που μπορούν οι χρήστες των κοινωνικών δικτύων να πραγματοποιήσουν. Όπως λοιπόν καθίσταται σαφές ο όγκος των δεδομένων που παράγεται καθημερινά είναι ιδιαίτερα σημαντικός. Για παράδειγμα, σύμφωνα με επίσημα στατιστικά στοιχεία 500M tweets στέλνονται καθημερινά στο κοινωνικό δίκτυο Twitter, ενώ συνολικά υπάρχουν 320M ενεργοί χρήστες.

Ένα κοινωνικό δίκτυο συνιστά μία κοινωνική δομή αποτελούμενη από άτομα (μεμονωμένες οντότητες ή ομάδες) που αναπαρίστανται ως κόμβοι και συνδέονται με έναν ή περισσότερους τύπους αλληλεξάρτησης ανάλογα με το είδος της σχέσης (ρητής ή άρρητης). Λόγω της μεγάλης πληθώρας δεδομένων και πληροφοριών που ανταλλάσσονται ανάμεσα στους χρήστες των κοινωνικών δικτύων, ιδιαίτερο ενδιαφέρον αποτελεί ο εντοπισμός σημαντικών γεγονότων με αυτόματο τρόπο με στόχο την έγκυρη αναγνώριση αυτών και την άμεση πληροφόρηση του ευρύτερου κοινού. Χαρακτηριστικό στοιχείο αποτελεί η στροφή της δημοσιογραφικής κοινότητας προς τα κοινωνικά δίκτυα, και ιδιαίτερα προς τις microblogging υπηρεσίες, τόσο για τη διάδοση μίας πληροφορίας όσο και για τον εντοπισμό καίριων και σημαντικών γεγονότων που λαμβάνουν χώρα ανά τον κόσμο και αναμεταδίδονται από απλούς χρήστες. Ένα γεγονός (**event**) χαρακτηρίζεται από επιμέρους θέματα (**topics**) τα οποία αναδεικνύονται μέσω της αύξησης της σχετικής δραστηριότητας στις microblogging υπηρεσίες, και τα οποία περιγράφουν την εξέλιξη αυτού. Η κατανόηση της εξέλιξης ενός γεγονότος μέσω της ανάλυσης της δραστηριότητας στα κοινωνικά μέσα απαιτεί αρχικά την ανακάλυψη σημαντικών θεμάτων ενδιαφέροντος (**emerging topic detection**), λαμβάνοντας υπόψη τόσο την ένταση της δραστηριότητας, όσο και τη σημαντικότητα και εγκυρότητα του περιεχομένου και κατ' επέκταση θέματος. Για την ανάλυση της δραστηριότητας αναφορικά με ένα γεγονός μπορούν να αξιοποιηθούν πληροφορίες όπως η συχνότητα αναφοράς σε θέματα, η αλληλεπίδραση των χρηστών γύρω από το γεγονός (αναφορά σε χρήστες –mention-, αναμετάδοση πληροφορίας –retweet-, κλπ), η ένταση της δραστηριότητας σε σχέση με τη γεωγραφική θέση των χρηστών, και το είδος και η ποιότητα του περιεχομένου που παράγεται από τους χρήστες. Μία συνήθης μέθοδος ανάλυσης που εφαρμόζεται για τη διεξαγωγή μιας πιο λεπτομερούς ανάλυσης σε σχέση με ένα γεγονός είναι η αναγνώριση της στάσης των χρηστών (**sentiment analysis**) αναφορικά με το εκάστοτε θέμα. Μία ακόμη μέθοδος ανάλυσης αποτελεί η προσπάθεια εξαγωγής της τοποθεσίας των χρηστών, χρησιμοποιώντας το κείμενο και τα μετα-δεδομένα που εμπεριέχονται σε κάθε tweet - post ή αναλύοντας τα χαρακτηριστικά του κοινωνικού δικτύου και τις διασυνδέσεις μεταξύ των χρηστών (**location inference**).

Στόχος της συγκεκριμένης εργασίας είναι η ανεύρεση σημαντικών σημείων αναφοράς σε σχέση με την εξέλιξη ενός γεγονότος, αξιοποιώντας τη δραστηριότητα των χρηστών στο κοινωνικό δίκτυο Twitter, και η παρουσίασή τους μέσω web εφαρμογής, με στόχο την καλύτερη επισκόπηση του γεγονότος.

Στα πλαίσια του **πρώτου μέρους της εργασίας** ζητείται να πραγματοποιηθεί βιβλιογραφική μελέτη σχετικά με τις υπάρχουσες προσεγγίσεις στο ζήτημα της επισκόπησης σημαντικών γεγονότων (even summarization) από τα κοινωνικά μέσα. Για την καλύτερη κατανόηση του προβλήματος θα πρέπει να μελετήσετε αρθρογραφία η οποία θα καλύπτει σφαιρικά το γενικότερο πλαίσιο της επισκόπησης σημαντικών γεγονότων από κοινωνικά μέσα (Ενότητα 1), ενώ στη συνέχεια ζητείται να εντρυφήσετε στις υπάρχουσες μεθόδους για την αναγνώριση της γνώμης/συναισθημάτων των χρηστών σε σχέση με ένα γεγονός.

Η **Ενότητα 1** είναι κοινή για όλες τις ομάδες και περιλαμβάνει ζητήματα σε σχέση με: μεθόδους που χρησιμοποιούνται για την ανεύρεση αναδυόμενων και δημοφιλών θεμάτων στα κοινωνικά δίκτυα. Μια συνήθης βασική κατηγοριοποίηση είναι σε μεθόδους που βασίζονται στην ομαδοποίηση κειμένων [1, 2] και σε μεθόδους που αξιοποιούν συχνότητα εμφάνισης όρων [3, 4, 5, 6]. Κατά την επισκόπηση γεγονότων χρησιμοποιούνται διάφορα μοντέλα για την αναπαράσταση της σημασιολογικής εγγύτητας όρων/κειμένων (π.χ. βάσει γράφου [3, 6, 7]), ενώ επιπρόσθετα δεδομένα που μπορούν να αξιοποιηθούν είναι το είδος του περιεχομένου του κειμένου [8] ή η συνδεσμικότητα [9] και επιρροή χρηστών [6]. Η μελέτη θα πρέπει να καλύπτει και τρόπους αξιολόγησης των προτεινόμενων μεθόδων και αξιοποίησης αυτών σε εφαρμογές.

Στην **Ενότητα 2** καλείστε **αρχικά** να επικεντρωθείτε σε μεθόδους που επιτρέπουν την κατανόηση των απόψεων και των συναισθημάτων των χρηστών μέσω της ανάλυσης κειμένων. Υπάρχουν δύο επίπεδα ανάλυσης: (i) ανάλυση της γνώμης (opinion mining), όπου τα κείμενα χαρακτηρίζονται ως προς τη θετικότητα ή την αρνητικότητα που εκφράζουν (positive/negative), (ii) ανάλυση των επιμέρους συναισθημάτων (affective analysis), η οποία χαρακτηρίζει τα κείμενα ως προς συγκεκριμένα συναισθήματα. Στην εργασία αυτή θα επικεντρωθούμε μόνο στο 1ο επίπεδο ανάλυσης (positive or negative). Για την πραγματοποίηση μίας τέτοια ανάλυσης σε οποιοδήποτε από τα παραπάνω επίπεδο δύο συχνά χρησιμοποιούμενες προσεγγίσεις είναι αυτή της μηχανικής μάθησης (machine learning) [10, 11, 12, 13] και η μέθοδος που βασίζεται στη χρήση λεξικών (lexicon-based techniques) [14, 15]. Τα τελευταία χρόνια μεγάλη έμφαση έχει δοθεί σε τεχνικές μηχανικής μάθησης που συμπεριλαμβάνουν Deep Learning ή εργαλεία όπως το Word2Vec [19, 20, 21, 22] . Στο θεωρητικό κομμάτι θα πρέπει να παρουσιάσετε εργασίες σχετικές με την πρώτη προσέγγιση δίνοντας μεγαλύτερη έμφαση σε τεχνικές μηχανικής μάθησης. Στην **συνέχεια** καλείστε να μελετήσετε μεθόδους που χρησιμοποιούνται για την εκτίμηση της τοποθεσίας του εκάστοτε χρήστη. Υπάρχουν δύο κυρίως κατηγορίες μεθόδων: (i) εξαγωγή τοποθεσίας χρησιμοποιώντας το περιεχόμενο του κάθε tweet και των μετα-δεδομένων που εμπεριέχονται σε αυτό [23, 24] και (ii) αναλύοντας το κοινωνικό δίκτυο και τις σχέσεις μεταξύ των χρηστών χρησιμοποιώντας τεχνικές ανάλυσης γράφων [25, 26, 27]. Η δεύτερη κατηγορία έχει επιδείξει καλύτερα αποτελέσματα, είναι όμως πιο περίπλοκη και απαιτεί σίγουρα μεγαλύτερους υπολογιστικούς πόρους.

Για τη θεωρητική μελέτη κάθε ομάδα θα πρέπει να μελετήσει το λιγότερο 5 άρθρα για την Ενότητα 1, καθώς και το λιγότερο 6 άρθρα για την ενότητα 2. Προτείνονται ως σημεία εκκίνησης τα άρθρα που παρατίθενται στην ενότητα **Αναφορές**, ενώ απαιτείται και η συλλογή επιπρόσθετης αρθρογραφίας. Η βιβλιογραφική μελέτη θα πρέπει να καταγραφεί υπό μορφή δημοσίευσης συνεδρίου και συγκεκριμένα σύμφωνα με το πρότυπο των ACM SIG Proceedings [L1], στην αγγλική ή ελληνική γλώσσα.

Στα πλαίσια του **δεύτερου μέρους της εργασίας** απαιτούνται τα ακόλουθα :

1ο μέρος: Συλλογή δεδομένων από το Twitter. Για τη συλλογή των δεδομένων από το Twitter θα γίνει χρήση του Twitter Streaming API [L2], το οποίο υποστηρίζει τη συνεχή παροχή νέων tweets βάσει ορισμένων αρχικών κριτηρίων. Η κάθε ομάδα θα πρέπει αρχικά να επιλέξει ένα διαφορετικό θέμα/γεγονός, και στη συνέχεια τους κατάλληλους όρους-κλειδιά ή/και χρήστες σε συνεργασία με την υπεύθυνη για την υποστήριξη της εργασίας. Η συλλογή των δεδομένων θα πρέπει να γίνει για επαρκές χρονικό διάστημα σε σχέση με το γεγονός και την αναμενόμενη δραστηριότητα των χρηστών. Για τη συλλογή των δεδομένων από το Twitter προτείνεται η χρήση της Java βιβλιοθήκης Twitter4j [L3] ή η Python βιλιοθήκη tweepy [L20] οι οποίες διευκολύνουν σημαντικά τη διαδικασία. (**Προσοχή:** για τα tweets που θα συλλεχθούν θα πρέπει να διατηρείτε την πλήρη JSON μορφή που επιστρέφει το Twitter Streaming API).

2ο μέρος: Προεπεξεργασία και μοντελοποίηση κειμένου tweets. Το βήμα αυτό περιλαμβάνει την προεπεξεργασία των tweets με στόχο την προετοιμασία τους για την περαιτέρω ανάλυση. Η διαδικασία αυτή συνήθως περιλαμβάνει βήματα όπως: i) το φιλτράρισμα των tweets χαμηλής ποιότητας, ii) το διαχωρισμό του κειμένου σε λεξιλογικές μονάδες (tokenization) και την αναγνώριση τυχόν hashtags, URLs, αναφορές σε άλλους χρήστες (mentions) ή σε άλλα tweets (retweets) που περιλαμβάνονται σε αυτό, iv) την αφαίρεση κοινών λέξεων (stop words), όπως άρθρα, αντωνυμίες, κλπ (π.χ. the, at, this, is, was, were), καθώς και όρων που εμφανίζονται συχνά στο Twitter (π.χ. RT, via), και v) τη μετατροπή των όρων σε μικρά γράμματα.

Χρήσιμη διαδικασία για την ανεύρεση θεμάτων είναι επίσης η ανακάλυψη ονομαστικών οντοτήτων (name entities) στο κείμενο μέσω τεχνικών ανάλυσης φυσικής γλώσσας (NLP). Η διαδικασία αυτή είναι αλγοριθμικά πιο σύνθετη, όμως μπορεί να υλοποιηθεί εύκολα με τη χρήση εύχρηστων βιβλιοθηκών (π.χ. Apache OpenNLP [L4] ,Stanford NLP [L5] ,NLTK (Python) [L21]).

Μελετήστε προσεκτικά τη μορφή αναπαράστασης των tweets μέσω του Twitter API για να κατανοήσετε καλύτερα ποια (μετα-)δεδομένα μπορείτε να αξιοποιήσετε και πως μπορείτε να τα εξάγετε. Ιδιαίτερη προσοχή θέλει η αναγνώριση των υπερσυνδέσμων που εμπεριέχονται στα tweets: πολύ συχνά χρησιμοποιούνται συντμημένα URLs είτε μέσω της αντίστοιχης υπηρεσίας του Twitter, ή/και μέσω τρίτων υπηρεσιών, όπως είναι η υπηρεσία bit.ly. Στις περιπτώσεις χρήσης συντμημένων URLs θα πρέπει πρώτα να τα ανάγετε στην πλήρη τους μορφή (όπου αυτό καθίσταται δυνατό).

3ο μέρος: Μοντελοποίηση και αποθήκευση δεδομένων. Για τη διαχείριση των αποτελεσμάτων που θα εξαχθούν απαιτείται η χρήση βάσης δεδομένων με κατάλληλο σχήμα που να επιτρέπει την αποδοτική αποθήκευση και ανάκτησή τους. Στην πλαίσια της εργασίας θα χρησιμοποιηθεί η βάση δεδομένων MongoDB [L6] (open source, NoSQL, document-oriented βάση δεδομένων).

4ο μέρος: Ανάλυση των tweets και ανεύρεση αναδυόμενων θεμάτων. Για την ανεύρεση θεμάτων ενδιαφέροντος καλείστε να αξιοποιήσετε το περιεχόμενο και τα μεταδεδομένα των tweets και να εφαρμόσετε είτε κάποια από τις μεθόδους που μελετήσατε στο μέρος Α, είτε συνδυασμό περισσότερων μεθόδων, ή ακόμα και να αναπτύξετε κάποια δική σας επέκταση/μέθοδο βάσει όσων μελετήσατε. Λάβετε υπόψη σας ότι ένα αναδυόμενο θέμα σε σχέση με κάποιο γεγονός χαρακτηρίζεται κυρίως από απότομη αύξηση της δημοφιλίας του σε σχέση με τις προηγούμενες χρονικές στιγμές. Γενικά, συνήθως εφαρμόζεται κάποια μέθοδος κατακερματισμού του χρόνου και χρησιμοποιούνται μετρικές δημοφιλίας όρων (ή συνδυασμών όρων) (popularity), σε συνδυασμό με μετρικές που αποτυπώνουν την αλλαγή της συχνότητας εμφάνισής τους στο χρόνο (burstiness), αξιοποιώντας το κείμενο των tweet. Επιπλέον χαρακτηριστικά που αξιοποιούνται σχετίζονται με την αξιοπιστία/κύρος (authority) του χρήστη που αναμεταδίδει την πληροφορία και την αξιολόγηση της ποιότητας του περιεχομένου. Ο αλγόριθμό σας θα πρέπει να αναθέτει σε κάθε θέμα που ανακαλύπτει μια σύντομη περιγραφή και να καταγράφει τη χρονική στιγμή (ή τις χρονικές στιγμές) στην (στις) οποία(ες) αυτό αναφέρεται.

5ο μέρος: Εξαγωγή συναισθηματικής πληροφορίας. Το 5ο μέρος της εργασίας περιλαμβάνει τον εντοπισμό του εκφραζόμενου γενικού συναισθήματος μέσω μίας τεχνικής βασισμένης σε μηχανική μάθηση. Η ταξινόμηση των tweets θα γίνει σε 2 κατηγορίες: positive , negative. Τα βήματα που θα πρέπει να ακολουθηθούν είναι τα εξής:

1.Επιλογή χαρακτηριστικών (features): Η επιλογή των χαρακτηριστικών εξαρτάται κυρίως από τα διαθέσιμα datasets (σύνολα δεδομένων). Αν το annotated dataset που έχουμε στη διάθεση μας είναι ανεπεξέργαστο, μπορούμε να διαλέξουμε ένα σύνολο από διαφορετικά features,όπως:

a. Η απλή εμφάνιση ή όχι ενός όρου
b. Οι λέξεις βαθμονομημένες βάσει της συχνότητας εμφάνισης τους
c. Σημεία στίξης
d. emoticons

Κατά την εφαρμογή των χαρακτηριστικών που αναφέρθηκαν παραπάνω αυτό που μελετάται είναι η ύπαρξη ή όχι κάποιας λέξης (ή η συχνότητα εμφάνισης της εκάστοτε λέξης), σημείου στίξης ή emoticon. Αφού έχουν εξαχθεί τα attributes που θα χρησιμοποιηθούν κατά την εκπαίδευση του αλγορίθμου, θα πρέπει να μελετηθεί η εμφάνιση ή όχι αυτών στα tweets που έχουν γίνει annotated.

2. Εφαρμογή διαφορετικών αλγορίθμων ταξινόμησης (π.χ. Naïve Bayes, Decision Trees, Multinomial Logistic Regression) και επιλογή αυτού που επιστρέφει τα καλύτερα αποτελέσματα. Για την εφαρμογή των

παραπάνω αλγορίθμων δεν υπάρχει περιορισμός στο εργαλείο που θα χρησιμοποιηθεί (π.χ. Weka [L7], Matlab [L8], Sklearn [L22], Keras [L23]).

Για την εκπαίδευση του μοντέλου θα δοθεί σε όλες τις ομάδες ένα σύνολο από tweets τα οποία θα πρέπει αρχικά να χαρακτηριστούν βάσει του εκφραζόμενου συναισθήματος (σε μία από τις 2 κατηγορίες που περιγράφηκαν παραπάνω). Στη συνέχεια, θα συγκεντρωθούν τα αποτελέσματα από όλες τις ομάδες με σκοπό τη δημιουργία ενός συνόλου δεδομένων που θα περιλαμβάνει annotated tweets. Διαφορετικά κάθε ομάδα θα χρησιμοποιήσει ένα από τα πολλά διαθέσιμα ανοιχτά σύνολα δεδομένων που είναι ήδη σχολιασμένα (annotated).

6ο μέρος: Εξαγωγή γεωγραφικής πληροφορίας. Η εξαγωγή γεωγραφικής πληροφορίας είναι πολύ εύκολη όταν ο χρήστης μοιράζεται τις ακριβείς του συντεταγμένες (τύπου check-in). Δυστυχώς (ή ευτυχώς) αυτό συμβαίνει μόνο για το ~2% του συνόλου των tweets που ποστάρονται καθημερινά. Παρόλα αυτά, το συγκεκριμένο ποσοστό στο σύνολο ενός μεγάλου αριθμού από tweets είναι επαρκές για να βγουν κάποια σχετικά συμπεράσματα. Συγκεκριμένα, αναλύοντας tweets που συμπεριλαμβάνουν γεωγραφική τοποθεσία μπορεί κανείς να εξάγει ένα σύνολο λέξεων που είναι ενδεικτικά μιας τοποθεσίας (παράδειγμα η λέξη howdy είναι πιο πιθανό να χρησιμοποιείται από χρήστες που έχουν κάποια σχέση με το Τέξας). Η εμφάνιση τέτοιων λέξεων μπορεί να αποτελεί ένα feature στην προσπάθεια μας να εξάγουμε την τοποθεσία κάποιου χρήστη. Ένα άλλο feature μπορεί να είναι η ζώνη ώρας του χρήστη, στοιχείο διαθέσιμο για κάθε tweet. Χρησιμοποιώντας λογική όμοια με αυτή του 5ου μέρους μπορούμε να πετύχουμε έναν τρόπο προσδιορισμού της τοποθεσίας του χρήστη, αν και δεν θα οδηγηθούμε σε συμπεράσματα μεγάλης ακρίβειας. Εναλλακτικά, μπορούμε αναλύοντας τις σχέσεις μεταξύ χρηστών. Σύμφωνα με τη βιβλιογραφία (28) , οι σχέσεις μεταξύ χρηστών είναι ενδεικτικές της απόστασης μεταξύ των χρηστών. Για παράδειγμα, έστω ότι ο χρήστης John ακολουθεί την χρήστη Jennifer, η οποία με τη σειρά της ακολουθεί κι αυτή τον John. Αυτή η bi-directional (αμφίδρομη) σχέση είναι μία ένδειξη ότι οι χρήστες John και Jennifer είναι γνωστοί και στην πραγματικότητα, υποδηλώνοντας ότι ίσως βρίσκονται και σε κοντινές τοποθεσίες. Φυσικά, αυτό μπορεί να ενισχυθεί όταν ο χρήστης Χ κάνει mention τον χρήστη Υ, κτλ. Γνωρίζοντας την τοποθεσία του χρήστη Χ ΄μπορούμε να κάνουμε μία εκτίμηση και για την τοποθεσία του χρήστη Υ. Χρησιμοποιώντας έναν αλγόριθμο διάδοσης της πληροφορίας, μπορούμε έχοντας γεωγραφικές πληροφορίες για λίγους χρήστες να κάνουμε μία εκτίμηση για την τοποθεσία πολλών περισσότερων.

7ο μέρος: Δημιουργία web εφαρμογής για την απεικόνιση των αποτελεσμάτων ανάλυσης. Τα αποτελέσματα της ανάλυσής σας θα πρέπει να παρουσιαστούν σε web εφαρμογή με ελκυστικό τρόπο για τον χρήστη. Η εφαρμογή θα πρέπει να παρουσιάζει κάποια βασικά στατιστικά για το σύνολο των δεδομένων που αναλύσατε (π.χ. ρυθμός παραγωγής tweets, αριθμός χρηστών, κλπ), και τα σημαντικά θέματα που ανακαλύψατε σε σχέση με το χρόνο εξέλιξης του γεγονότος μαζί με ενδεικτικά tweets και πιθανόν εικόνες που περιλαμβάνονται σε αυτά. Επιπλέον, η εφαρμογή θα πρέπει να παρουσιάζει τα αποτελέσματα της συναισθηματικής ανάλυσης. Συστήνεται η χρήση βιβλιοθηκών που προσφέρουν διαδραστικότητα, όπως: Google Charts [L9], TimelineJS [L10], d3.js [L11], charts.js [L12]. Επίσης, μπορείτε να πάρετε ιδέες από υπάρχουσες εφαρμογές που παρουσιάζουν αποτελέσματα ανάλυσης από δεδομένα του Twitter [16, 17, 18, L13, L14, L15, L16].

Θα πρέπει να γίνει χρήση κάποιου παρόχου δωρεάν διαδικτυακής φιλοξενίας της ιστοσελίδας σας [π.χ. L17, L18, L19]. Για να βρείτε δημοσιεύσεις που είναι διαθέσιμες (κατόπιν πληρωμής), ένας πολύ χρήσιμος ιστότοπος είναι ο L24. Τέλος, μία καλή μηχανή αναζήτησης δημοσιεύσεων βάσει θέματος είναι η L25.

Οι εργασίες θα υλοποιηθούν από ομάδες 3-4 ατόμων

Παραδοτέα	Προθεσμία
1.**Βιβλιογραφική αναφορά:** σύμφωνα με τις προδιαγραφές που δόθηκαν στο Μέρος Α. 2.**Αρχείο παρουσίασης** της βιβλιογραφικής αναφοράς.	**24/4/17**
3.**Λειτουργικός κώδικας** υλοποίησης του Β Μέρους με επαρκή σχολιασμό. Για την ανάπτυξη του κώδικα θα γίνει χρήση της γλώσσας Java ή της γλώσας Python. 4.**Δεδομένα** από το Twitter που χρησιμοποιήθηκαν για πειραματισμό στο Β Μέρος. 5.**Τεχνική αναφορά** (~ 10 σελίδες) η οποία θα περιλαμβάνει: α) την περιγραφή του μοντέλου που χρησιμοποιήθηκε για την αναπαράσταση των δεδομένων, β) την περιγραφή της υλοποίησης των μεθόδων επεξεργασίας και υπολογισμού, γ) την παράθεση και το σχολιασμό ενδεικτικών αποτελεσμάτων, και δ) την περιγραφή της εφαρμογής που αναπτύχθηκε.	**12/6/17**

Αναφορές

1. Saša Petrovid, Miles Osborne, and Victor Lavrenko, “Streaming first story detection with application to Twitter”, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT ‘10), Association for Computational Linguistics, Stroudsburg, PA, USA, 181-189. (document based).

2. Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: news in tweets. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS ‘09). ACM, New York, NY, USA, 42-51. (document based).

3. H. Sayyadi, L. Raschid. “A Graph Analytical Approach for Topic Detection”, ACM Transactions on Internet Technology (TOIT), 2013 (term based, graph model).

4. Jianshu Weng, Bu-Sung Lee: Event Detection in Twitter. ICWSM 2011 (term based – signal analysis wavelets).

5. Dehong Gao; Wenjie Li; Xiaoyan Cai; Renxian Zhang; You Ouyang, “Sequential Summarization: A Full View of Twitter Trending Topics,”. In IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22, no.2, pp.293,302, Feb. 2014. (term based, time segments).

6. Mario Cataldi, Luigi Di Caro, Claudio Schifanella: Personalized emerging topic detection based on a term aging model. ACM TIST 5(1): 7 (2013). (term based, social features).

7. NIFTY: A System for Large Scale Information Flow Tracking and Clustering by C. Suen, S. Huang, C. Eksombatchai, R. Sosic, J. Leskovec. ACM International Conference on World Wide Web (WWW), 2013. (http://snap.stanford.edu/nifty/index.php) (graph, application).

8. Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. TwiNER: named entity recognition in targeted twitter stream. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR ‘12). ACM, New York, NY, USA, 721- 730 (NER).

9. Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Structural trend analysis for online social networks. Proc. VLDB Endow. 4, 10 (July 2011), 646-656. (based on term and social relationships, graph).

10. D. Chatzakou, N. Passalis, A. Vakali. MultiSpot: Spotting Sentiments with Semantic Aware Multilevel Cascaded Analysis. Big Data Analytics and Knowledge Discovery (DaWaK), volume 9263, pages 337-350, Springer, 2015.

11. R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. InProceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 151-161.

12. Georgios Paltoglou and Mike Thelwall. 2012. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. ACM Trans. Intell. Syst. Technol. 3, 4, Article 66 (September 2012), 19 pages. (sentiment analysis, lexicon-based methodology).

13. B. Pang et al. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. Of the 42nd annual meeting on Association for Computational Linguistics, 271, 2004.

14. Yan Dang, Yulei Zhang, and HsinChun Chen. 2010. A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. IEEE Intelligent Systems 25, 4 (July 2010), 46-53. (sentiment analysis, lexicon-based methodology combined with machine learning).

15. Chatzakou, D.; Koutsonikola, V.; Vakali, A.; Kafetsios, K., "Micro-blogging Content Analysis via Emotionally-Driven Clustering," Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on , vol., no., pp.375,380, 2-5 Sept. 2013.

16. Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. 2013. STED: semi-supervised targeted-interest event detectionin in twitter. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13). ACM, New York, NY, USA, 1466-1469. (application).

17. Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. 2013. EvenTweet: online localized event detection from twitter. Proc. VLDB Endow. 6, 12 (August 2013), 1326-1329. (application).

18. K. McKelvey, F. Menczer. Truthy: Enabling the Study of Online Social Networks. CSCW 2013 Demonstration, 25 February 2013. http://www.truthy.indiana.edu/ (application).

19. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014, June). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL (1)* (pp. 1555-1565).

20. Severyn, A., & Moschitti, A. (2015, June). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado* (pp. 464-469).

21. Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING* (pp. 69-78).

22. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

23. Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. "Where Is This Tweet From? Inferring Home Locations of Twitter Users." *ICWSM* 12 (2012): 511-514.

24. Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010.

25. Jurgens, David, et al. "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice." *ICWSM*. 2015.

26. Compton, Ryan, David Jurgens, and David Allen. "Geotagging one hundred million twitter accounts with total variation minimization." *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014.

27. Kong, Longbo, Zhi Liu, and Yan Huang. "Spot: Locating social media users based on social network context." *Proceedings of the VLDB Endowment* 7.13 (2014): 1681-1684.

28. Backstrom, Lars, Eric Sun, and Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

ΙΣΤΟΤΟΠΟΙ ΥΠΟΣΤΗΡΙΚΤΙΚΟΥ ΥΛΙΚΟΥ

L1. http://www.acm.org/sigs/publications/proceedings-templates.

L2. https://dev.twitter.com/docs/streaming-apis

L3. http://twitter4j.org/en/index.html

L4. https://opennlp.apache.org/

L5. http://nlp.stanford.edu/

L6. http://www.mongodb.org/

L7. http://www.cs.waikato.ac.nz/ml/weka/

L8. http://www.mathworks.com/products/matlab/

L9. https://developers.google.com/chart/

L10. https://timeline.knightlab.com/

L11. https://d3js.org/

L12. http://www.chartjs.org/

L13. http://tweettracker.fulton.asu.edu/ (application)

L14. http://twitris.knoesis.org/yolandastorm2013/ (application)

L15. http://topsy.com/ (application)

L16. http://mediafinder.eurecom.fr/ (application)

L17. http://www.hostinger.gr/

L18. http://freehostingnoads.net/

L19. http://freehostingnoads.ga/

L20. http://www.tweepy.org/

L21. http://www.nltk.org/

L22. http://scikit-learn.org/stable/

L23. https://keras.io/

L24. http://sci-hub.cc/

L25. http://www.arxiv-sanity.com/