ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Α.Π.Θ. Τμήμα Πληροφορικής
Εαρινό εξάμηνο 2017-2018

Διδάσκουσα: Καθ. Αθηνά Βακάλη
Υπεύθυνος εργασίας: Δημητριάδης Ηλίας idimitriad@csd.auth.gr

Θέμα εργασίας

Τα κοινωνικά δίκτυα αποτελούν αναπόσπαστο κομμάτι της καθημερινότητας των ανθρώπων. Δεν είναι λίγοι εκείνοι που χρησιμοποιούν τα κοινωνικά δίκτυα για να αξιολογήσουν ένα εστιατόριο, μία παράσταση ή μία ταινία, να ενημερωθούν για τυχόν προσφορές, να ψάξουν για νέες προτάσεις εξόδου στην περιοχή τους ή να εξερευνήσουν νέες περιοχές και τα αξιοθέατα τους. 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 600M tweets στέλνονται καθημερινά στο κοινωνικό δίκτυο Twitter, ενώ συνολικά υπάρχουν 330M 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). Στην εργασία αυτή θα επικεντρωθούμε στην εξαγωγή πληροφορίας από κοινωνικά δίκτυα που έχουν σχηματιστεί στα πλαίσια μιας πληθοποριστικής πλατφόρμας (crowdsourcing) που φιλοξενεί πληροφορίες για επιχειρήσεις διαφόρων τύπων, τοπικές και μη, επιτρέπει στους χρήστες της να αξιολογήσουν αυτές τις επιχειρήσεις (σύμφωνα με την προσωπική τους εμπειρία) και γενικότερα δίνει τη δυνατότητα στον οποιοδήποτε χρήστη να εκμεταλλευτεί την «σοφία» του πλήθους (wisdom of the crowd). Παραδείγματα τέτοιων εφαρμογών είναι το trip advisor (https://www.tripadvisor.com.gr/) , το γνωστό σε όλους amazon (http://www.amazon.com) , το Yelp (https://www.yelp.com/) ή και το Strava (https://www.strava.com/) που αποτελεί μία πλατφόρμα δικτύωσης των απανταχού ερασιτεχνών και επαγγελματιών δρομέων και ποδηλατών, που επιτρέπει στους χρήστες του να αξιολογούν και να μοιράζονται διαδρομές, στατιστικά, κτλ. Η συγκεκριμένη εργασία θα προχωρήσει σε ανάλυση δεδομένων που προέρχονται από την πλατφόρμα Yelp. Το Yelp διατηρεί μια ιδιαίτερα ανοιχή κοινότητα προγραμματιστών κι ερευνητών που αναλύουν ένα υποσύνολο του συνόλου των δεδομένων του, το οποίο είναι διαθέσιμο προς κάθε χρήση.

Τα θέματα που απασχολούν την ερευνητική κοινότητα όσον αφορά την ανάλυση δεδομένων από κοινωνικά δίκτυα και την εξόρυξη πληροφορίας από αυτά είναι ποικίλλα. Μερικά από αυτά είναι τα ακόλουθα:

1. Ο εντοπισμός αναδυόμενων γεγονότων (emerging topic detection)
2. Η αναγνώριση της στάσης και του συναισθήματος σχετικά με κάποιο θέμα (sentiment analysis)
3. Η εξαγωγή της τοποθεσίας των χρηστών και η χωρική κατανομή θεμάτων – προβλημάτων (location inference)
4. Η ανακάλυψη ύποπτων λογαριασμών – απατεώνων (fraud detection)

Αρχικός στόχος της εργασίας, είναι η εκπόνηση μιας θεωρητικής εις βάθους μελέτης της βιβλιογραφίας σχετικά με τα παραπάνω θέματα.

Θεωρητικό μέρος

Στα πλαίσια του θεωρητικού μέρους καλείστε να μελετήσετε τη σχετική βιβλιογραφία έτσι ώστε κάθε ομάδα να επιλέξει ένα (1) από τα παραπάνω θέματα. Θα πρέπει να προχωρήσετε σε μία πλήρη ανάλυση των μοντέλων που προτείνονται για τη διαχείριση των παραπάνω προβλημάτων και να παρουσιάσετε τις βασικές αρχές που τα διέπουν. Για την καλύτερη κατανόηση του προβλήματος θα πρέπει να μελετήσετε αρθρογραφία η οποία θα καλύπτει σφαιρικά το γενικότερο πλαίσιο του κάθε θέματος , ενώ στη συνέχεια ζητείται να εντρυφήσετε στις υπάρχουσες μεθόδους που χρησιμοποιούνται για την αντιμετώπιση του κάθε προβλήματος.

Για τη θεωρητική μελέτη κάθε ομάδα θα πρέπει να μελετήσετε το λιγότερο 6-8 άρθρα για το κάθε θέμα. 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.

Για παράδειγμα, αν επιλέξετε την ενότητα 2, θα πρέπει 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, neutral 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) και η μέθοδος που βασίζεται στη χρήση λεξικών (lexicon-based techniques). Τα τελευταία χρόνια μεγάλη έμφαση έχει δοθεί σε τεχνικές μηχανικής μάθησης που συμπεριλαμβάνουν Deep Learning ή εργαλεία όπως το Word2Vec. Στο θεωρητικό κομμάτι θα πρέπει να παρουσιάσετε εργασίες σχετικές με την πρώτη προσέγγιση δίνοντας μεγαλύτερη έμφαση σε τεχνικές μηχανικής μάθησης μιας και είναι σαφώς πιο αποδοτικές και οι SOTA (State Of The Art) προσεγγίσεις υλοποιούν τέτοιες τεχνικές.

Η δομ της εργασίας θα πρέπει να περιλαμβάνει ενδεικτικά:

- Εισαγωγή
  - Τι είναι το θέμα με το οποίο θα ασχοληθείτε
  - Ποια είναι τα βασικά χαρακτηριστικά στην σύγχρονη εποχή
  - Ποια είναι τα ανοιχτά ζητήματα που εντοπίζονται για αυτό το θέμα
- Θεμελιώδεις έννοιες και βασικά στοιχεία
  - Διάφορους βασικούς και χρήσιμους ορισμούς που αναδεικνύονται στις εργασίες που μελετήθηκαν
- Μοντέλα
  - Ανάδειξη των μοντέλων ή/και τεχνικών που χρησιμοποιήθηκαν στις εργασίες που μελετήσατε, να παρουσιαστούν σε πίνακες ή σχήματα
- Σύνολα Δεδομένων και πειραματική διαδικασία
  - Περιγραφή του dataset που χρησιμοποιήθηκε
  - Καταγραφή των πειραματικών αποτελεσμάτων με χρήση διαγραμμάτων.
- Συμπεράσματα
  - Συμπεράσματα και προτάσεις για μελλοντική εργασία
- Αναφορές
  - Σύμφωνα με το σύστημα αναφορών του template που σας έχει δοθεί.

Η εργασία θα πρέπει να έχει έκταση 10-12 σελίδες και να συμπεριλαμβάνει τουλάχιστον 10 αναφορές. Οι αναφορές αυτές θα πρέπει να χρησιμοποιούνται σε όλο το άρθρο. Επιπλέον η εργασία θα πρέπει να συνοδεύεται από μία παρουσίαση 20 διαφανειών όπου θα τονίζονται τα πιο σημαντικά της στοιχεία.

Πρακτικό μέρος

Όπως αναφέραμε νωρίτερα, η εργασία αυτή θα προχωρήσει σε ανάλυση δεδομένων που προέρχονται από την πλατφόρμα Yelp. Επιλέξαμε το yelp γιατί έχει δημοσιοποιήσει μέρος του συνόλου των δεδομένων του στα πλαίσια του Yelp Dataset Challenge (https://www.yelp.com/dataset/challenge). Περισσότερες πληροφορίες σχετικά με αυτό το διαγωνισμό, θα βρείτε στο link που έχουμε παραθέσει. Στο παρακάτω μπορείτε να βρείτε πληροφορίες για τους παλιούς νικητές (https://www.yelp.com/dataset/challenge/winners) και σε αυτό

(https://scholar.google.com/scholar?q=citation%3A+Yelp+Dataset&btnG=&hl=en&as_sdt=0%2C5) μια συλλογή των paper που έχουν δημοσιευτεί στα πλαίσια αυτού του διαγωνισμού, στα οποία μπορείτε να ανατρέξετε για περισσότερες λεπτομέρειες. Σε αυτή την εργασία σας ζητείται να αναπτύξετε ένα μοντέλο στο οποίο θα εισάγετε μία αξιολόγηση σε μορφή κειμένου και θα σας επιστρέφει μία εκτίμηση της αριθμητικής αξιολόγησης. Είναι ουσιαστικά μία διαδικασία sentiment analysis, όπου αντί για positive, neutral, negative θα έχουμε 5, 4, 3, 2, 1 αστέρια. Επίσης σας ζητείται να προχωρήσετε σε μία τύπου στατιστική ανάλυση, προσπαθώντας να εντοπίσετε κοινά χαρακτηριστικά ή αλληλεπιδράσεις μεταξύ δεδομένων που δεν είναι ορατές με το γυμνό μάτι. Τέλος θα πρέπει να παρουσιάσετε τα αποτελέσματα σας σε μία ιστοδελίδα, χρησιμοποιώντας όποια εργαλεία κρίνετε εσείς ως κατάλληλα.

1ο μέρος: Σχετικά με το dataset: Το dataset που μας παρέχει το Yelp είναι πολυδιάστατο κι αρκετά μεγάλο. Συνολική εικόνα σχετικά με αυτό μπορείτε να έχετε εδώ: https://www.yelp.com/dataset/documentation/json. Λόγω του μεγέθους του (6+ GB) , έχουμε επιλέξει κι εξάγει ένα υποσύνολο του συνολικού dataset το οποίο περιορίζεται μόνο σε επιχειρήσεις που ανήκουν στην κατηγορία των εστιατορίων (Restaurants) και βρίσκονται στην ευρύτερη περιοχή του Las Vegas. Το σύνολο των δεδομένων στο οποίο θα πειραματιστείτε, χωρίζεται στις εξής τρεις κατηγορίες:

- yelp_restaurant_Las_Vegas: στοιχεία για 5899 εστιατόρια στην περιοχή του Las Vegas. Περιέχει πληροφορίες σχετικά με μέση αξιολόγηση, ακριβείς συντεταγμένες, διεύθυνση, ώρες λειτουργίας, πόλη, ταχυδρομικό κώδικα, αριθμό συνολικών αξιολογήσεων, κατηγορία εστιατορίου (π.χ. Ιταλικό) κ.α.
- reviews_Las_Vegas_restaurants: 906K αξιολογήσεις που αφορούν εστιατόρια στην περιοχή του Λας Βέγκας. Περιέχει πληροφορίες όπως, ημερομηνία αξιολόγησης, χρήστης που την πρόσθεσε, stars (αριθμητική αξιολόγηση), αξιολόγηση σε μορφή κειμένου, κ.α.
- users_restaurant_reviews_Las_Vegas: 337K χρήστες που έχουν κάνει αξιολογήσεις σε εστιατόρια του Λας βέγκας. Περιέχει πληροφορίες όπως το id του κάθε χρήστη, το σύνολο των αξιολογήσεων που έχει κάνει, τους φίλους του (άλλοι χρήστες της πλατφόρμας) και πολλά άλλα χαρακτηριστικά που μπορείτε να δείτε στο documentation link παραπάνω.

Είναι βασικό να μελετήσετε όσο το δυνατόν καλύτερα το dataset που σας παρέχεται ώστε να κατανοήσετε τις συνδέσεις που υπάρχουν μεταξύ των δεδομένων και να σκεφτείτε πως αυτές μπορούν να αξιοποιηθούν καλύτερα.

2ο μέρος: Αποθήκευση και μοντελοποίηση 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). Αφού κατεβάσετε τα δεδομένα που έχουμε ανεβάσει, από εδώ: https://www.dropbox.com/sh/hqam9656lx5va82/AACTqPqvpu4nk29aeZpraaQga?dl=0 θα πρέπει να τα εισάγετε σε μία τοπική βάση MongoDB και εφ'όσον σας χρειαστεί να τα διαμορφώσετε ανάλογα με τις ανάγκες σας. Η διαδικασία αυτή είναι αρκετά απλή, αφού τα δεδομένα σας προσφέρονται σε μορφή json που είναι και η default για τη MongoDB.

3ο μέρος: Προεπεξεργασία και μοντελοποίηση κειμένου αξιολογήσεων (reviews). Το βήμα αυτό περιλαμβάνει την προεπεξεργασία των αξιολογήσεων με στόχο την προετοιμασία τους για την περαιτέρω ανάλυση. Η διαδικασία αυτή συνήθως περιλαμβάνει βήματα όπως: i) το φιλτράρισμα των χαμηλής ποιότητας αξιολογήσεων, ii) το διαχωρισμό του κειμένου σε λεξιλογικές μονάδες (tokenization) και την αναγωγή στη ρίζα τους (stemming), iii) 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 Yelp και iv) τη μετατροπή των όρων σε μικρά γράμματα. Φυσικά εδώ μπορείτε να πειραματιστείτε και να προχωρήσετε χρησιμοποιώντας κι άλλες μεθόδους επεξεργασίας ή και λιγότερες.

Μελετήστε καλά τα δεδομένα για να τα κατανοήσετε πλήρως ώστε να χρησιμοποιησετε και να αξιοποιήσετε κάποια μετα-δεδομένα που ίσως προκύψουν από την πιθανή διασύνδεση τους.

4ο μέρος: Εξαγωγή συναισθηματικής πληροφορίας: Το 4ο μέρος της εργασίας περιλαμβάνει τον εντοπισμό του εκφραζόμενου γενικού συναισθήματος μέσω μίας τεχνικής βασισμένης σε μηχανική μάθηση. Η ταξινόμηση των αξιολογήσεων θα γίνει σε 2 κατηγορίες: positive , negative. Τα βήματα που θα πρέπει να ακολουθηθούν είναι τα εξής:

Α.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:

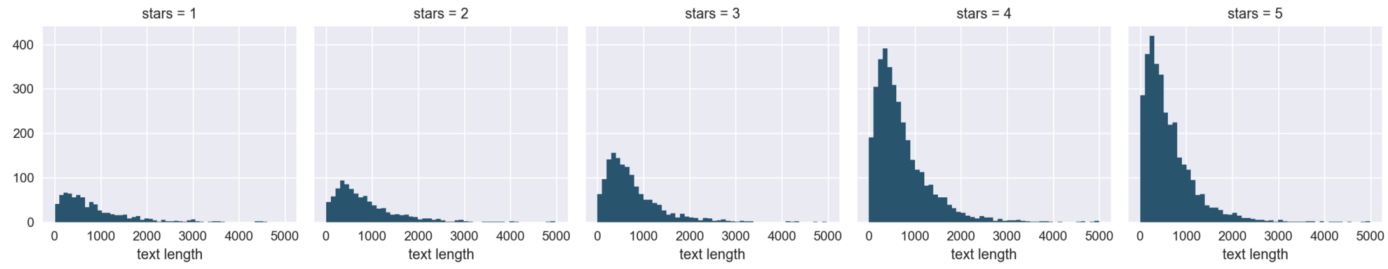
a. Η απλή εμφάνιση ή όχι ενός όρου
b. Οι λέξεις βαθμονομημένες βάσει της συχνότητας εμφάνισης τους
c. Σημεία στίξης
d. emoticons

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 στις αξιολογήσεις που έχουμε στη διάθεση μας. Η λίστα από τα features που αναφέραμε είναι φυσικά ενδεικτική, καθώς μπορείτε να εισάγετε όσα ακόμη θέλετε ή και να αφαιρέσετε. Καλό θα ήταν πριν την τελική επιλογή των features να έχετε ολοκληρώσει το 5ο μέρος της άσκησης που θα σας βοηθήσει να καταλάβετε αν υπάρχει κάποια συσχέτιση (correlation) μεταξύ των δεδομένων που έχουμε διαθέσιμα.

Β. Εφαρμογή διαφορετικών αλγορίθμων ταξινόμησης (π.χ. Naïve Bayes, Decision Trees, Multinomial Logistic Regression, SVM, Convolutional Neural Networks) 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]). Σας προτείνουμε να δοκιμάσετε να εφαρμόσετε πολλούς διαφορετικούς αλγορίθμους μηχανικής μάθησης, ώστε να εξοικειωθείτε και να βγάλετε πιο δομημένα αποτελέσματα.

Για την εκπαίδευση του μοντέλου θα χρησιμοποιήσετε ως βασικό input τις αξιολογήσεις των χρηστών με αριθμό αστεριών 1 (negative) και 5(positive). Σε περίπτωση που κάποια ομάδα θέλει να πειραματιστεί με παραπάνω από δυό κλάσεις (1-2-3-4-5) μπορεί φυσικά να το κάνει. Η επιπλέον αυτή προσπάθεια θα εκτιμηθεί αναλόγως στην τελική βαθμολόγηση της εργασίας.

5ο μέρος: Στατιστική ανάλυση και εκτενέστερη εξερεύνηση του συνόλου των δεδομένων: Η οπτικοποίηση των δεδομένων και των διαφόρων τους features μπορεί να βοηθήσει ιδιαίτερα στην καλύτερη χρησιμοποίησή τους. Σε αυτό το μέρος καλείστε να αποδώσετε σε μορφή διαγραμμάτων τις διάφορες συσχετίσεις μεταξύ των δεδομένων μιας ίδιας συλλογής ή μιας συλλογής που προκύπτει από την ένωση μιας ή παραπάνω συλλογών. Θα πρέπει να παρουσιάσετε τουλάχιστον 5 γραφήματα και να εξηγήσετε την ερμηνεία τους. Θα πρέπει να είναι τέτοια ώστε η ερμηνεία τους να προσθέτει γνώση που δεν είναι εκ των προτέρων εμφανής. Ένα χαρακτηριστικό παράδειγμα είναι η συσχέτιση του μήκους (text length) με τον αριθμό των αστεριών των αξιολογήσεων. Ένα τέτοιο γράφημα θα μπορούσε να μοιάζει με αυτο:



Στο παραπάνω παράδειγμα παρατηρούμε ότι η κατανομή του αριθμού των λέξεων είναι παραπλήσια για κάθε κλάση (1* -5*) οπότε η χρήση του text_length στον classifier μας κατά πάσα πιθανότητα δεν θα οδηγούσε σε κάποια βελτίωση στην απόδοση του μοντέλου μας. Παρόλα αυτά βλέπουμε ότι ο αριθμός των αξιολογήσεων με 4 και 5 αστέρια είναι σφώς μεγαλύτερος από το πλήθος των αξιολογήσεων για 1,2 ή 3 αστέρια. Κάτι τέτοιο θα οδηγούσε πιθανώς σε αποτελέσματα χαμηλότερης ακρίβειας, λόγω του υψηλού bias. Η βιβλιοθήκη Pandas (Python) έχει αρκετά έτοιμα εργαλεία που θα σας βοηθήσουν να εξάγετε πιθανές συσχετίσεις μεταξύ δεδομένων. Το παραπάνω γράφημα όπως είπαμε είναι ενδεικτικό, τα γραφήματα που θα εξάγεται μπορεί και πρέπει να είναι και διαφορετικού τύπου, έτσι ώστε η ερμηνεία τους να γίνεται πιο εύκολο κατανοητή.

6ο μέρος: Εξαγωγή γεωγραφικής πληροφορίας. Στο συγκεκριμένο dataset η εξαγωγή γεωγραφικής πληροφορίας για την κάθε επιχείρηση είναι πολύ εύκολη καθώς στην συλλογή yelp_restaurant_Las_Vegas συμπεριλαμβάνονται οι ακριβείς συντεταγμένες. Σε αυτό το μέρος σας ζητείται απλά να απεικονίσετε αυτή την πληροφορία σε ένα χάρτη της επιλογής σας. Θα μπορούσατε να χρησιμοποιήσετε την βιβλιοθήκη της Google Maps ή το Open Street Map.

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 και όλα τα γραφήματα τα οποία εξάγατε. Επιπλέον, η εφαρμογή θα πρέπει να παρουσιάζει τα αποτελέσματα της συναισθηματικής ανάλυσης και τον χάρτη τον οποίο δημιουργήσατε στο 6ο μέρος. Συστήνεται η χρήση βιβλιοθηκών που προσφέρουν διαδραστικότητα, όπως: Google Charts, TimelineJS ,d3.js [L11], charts.js [L12].

Θα πρέπει να γίνει χρήση κάποιου παροχέα 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.

1.**Βιβλιογραφική αναφορά:** σύμφωνα με τις προδιαγραφές που δόθηκαν στο Μέρος Α. ***30/4/2018***

2.**Αρχείο παρουσίασης** της βιβλιογραφικής αναφοράς. ***30/4/2018***

3.**Λειτουργικός κώδικας** υλοποίησης του Β Μέρους με επαρκή σχολιασμό. Για την ανάπτυξη του κώδικα θα γίνει χρήση της γλώσσας Java ή της γλώσσας Python. ***4/6/2018***

4.**Δεδομένα** από το Yelp που χρησιμοποιήθηκαν για πειραματισμό στο Β Μέρος. ***4/6/2018***

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. ***4/6/2018***

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. 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).
7. 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).
8. 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).
9. 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).
10. 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).
11. T. Takahashi, R. Tomioka and K. Yamanishi, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 120-130, Jan. 2014.
12. Rill, Sven, et al. "Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis." *Knowledge-Based Systems* 69 (2014): 24-33.
13. Unankard, S., Li, X., & Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web*, *18*(5), 1393-1417.
14. Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, *31*(1), 132-164.
15. Maurya, A., Murray, K., Liu, Y., Dyer, C., Cohen, W. W., & Neill, D. B. (2016). Semantic scan: detecting subtle, spatially localized events in text streams. *arXiv preprint arXiv:1602.04393*.
16. Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*, 42-49.
17. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14-46.
18. 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).
19. 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.
20. 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.

21. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89
22. 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).
23. 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.
24. 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).
25. 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.
26. 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).
27. . Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING* (pp. 69-78).
28. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.
29. Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. "Where Is This Tweet From? Inferring Home Locations of Twitter Users." *ICWSM* 12 (2012): 511-514.
30. 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.
31. Jurgens, David, et al. "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice." *ICWSM*. 2015.
32. 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.
33. 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.
34. Do, T. H., Nguyen, D. M., Tsiligianni, E., Cornelis, B., & Deligiannis, N. (2017). Multiview Deep Learning for Predicting Twitter Users' Location. *arXiv preprint arXiv:1712.08091*.
35. 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
36. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31
37. Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427.
38. Hooi, B., Song, H. A., Beutel, A., Shah, N., Shin, K., & Faloutsos, C. (2016, August). Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 895-904). ACM.
39. Yu, R., He, X., & Liu, Y. (2015). Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *10*(2), 18.
40. Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, *30*(1), 89-116.

41. Rayana, S., & Akoglu, L. (2015, August). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 985-994). ACM.
42. Xu, C., & Zhang, J. (2015, November). Towards collusive fraud detection in online reviews. In *Data Mining (ICDM), 2015 IEEE International Conference on* (pp. 1051-1056). IEEE.
43. Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. *ICWSM*, *13*, 2-11.

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/

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/