Α.Π.Θ. Τμήμα Πληροφορικής
Χειμερινό Εξάμηνο 2017-2018

Διδάσκουσα: Καθηγ. Αθηνά Βακάλη
Υπεύθυνοι εργασίας: Γραβάνης Γιώργος
Τσιλιγγίρης Αλέξανδρος
Άρτεμις Ψαλτόγλου

Θέμα εργασίας

Opinion mining με χρήση συλλογών που θα δημιουργηθούν από το Twitter

Γενικά

Τα τελευταία χρόνια, τα social media δίνουν την δυνατότητα στους ανθρώπους να εκφέρουν άφοβα την άποψή τους για πολλά σημαντικά ζητήματα που απασχολούν τους ίδιους ή τις κοινωνίες στις οποίες ανήκουν. Στην εποχή της πληροφορίας στην οποία ανήκουμε, η δημιουργία ενός εργαλείου με το οποίο θα μπορούσε κάποιος να αντιληφθεί τις συσχετίσεις και τις απόψεις του κόσμου γύρω από τα τρέχοντα ζητήματα, θα μπορούσε να βοηθήσει στην ανάπτυξη εφαρμογών ή την κατανόηση των ζητημάτων που απασχολούν την κοινή γνώμη.

Για την υλοποίηση της παραπάνω ιδέας προτείνεται η χρήση του Twitter καθώς είναι ένα μέσο κοινωνικής δικτύωσης, στο οποίο οι χρήστες συμμετέχουν άμεσα στα τρέχοντα ζητήματα και εκφράζουν την άποψή τους με πολύ περιεκτικά μηνύματα (λόγω του περιορισμού των χαρακτήρων σε 140). Επιπλέον θα πρέπει να αναφερθεί ότι το 2 τρίμηνο του 2017 οι ενεργοί χρήστες του twitter αριθμούν 328 εκατομμύρια οι οποίοι παράγουν περίπου 500 εκατομμύρια tweets ανά ημέρα.

Στόχος της εργασίας είναι η δημιουργία συλλογών από tweets για τα 5 πιο viral θέματα που συζητούνται στο twitter την χρονική στιγμή που θα εκπονήσετε την εργασία, με σκοπό την επεξεργασία τους, την ανάλυση του κάθε tweet σε επίπεδο συναισθήματος (sentiment analysis) καθώς επίσης και στην εξαγωγή μετρικών για το σύνολο της κάθε συλλογής. Για την αποθήκευση των συλλογών θα χρησιμοποιηθεί η MongoDB.

Αναλυτική Περιγραφή

1^{ο} μέρος: Συλλογή tweets γύρω από 5 viral θέματα και δημιουργία συλλογών.

Το Twitter δίνει την δυνατότητα ανάκτησης της θεματολογίας με την μεγαλύτερη ροή κατά την δεδομένη χρονική στιγμή που θα του ζητηθεί. Προκειμένου να εξασφαλίσουμε την γρήγορη συγκομιδή tweets θα χρησιμοποιήσουμε αυτή τη δυνατότητα για την γεωγραφική περιοχή των ΗΠΑ.

Από την λίστα με τα θέματα που θα σας επιστραφεί θα χρησιμοποιήσετε τα 5 πρώτα.

Στη συνέχεια θα πρέπει να συνδεθείτε με την ροή του Twitter και να συλλέξετε tweets σχετικά με τα θέματα που έχετε συλλέξει. Για την υλοποίηση του παραπάνω θα πρέπει να χρησιμοποιήσετε το API του Twitter.

Οι συλλογές από tweets που θα δημιουργηθούν θα πρέπει να είναι ξεχωριστές (μία για κάθε θέμα).

Στην άδεια που θα σας δοθεί από το twitter δεν είναι δυνατή η παράλληλη αναζήτηση, συνεπώς η συλλογή των tweets για κάθε θέμα θα πρέπει να γίνει σειριακά ή με την χρήση διαφορετικών κλειδιών.

Για κάθε θέμα θα πρέπει να συλλεχθούν από 1500 tweets. Επιπλέον θα πρέπει αν αποφύγετε την συλλογή retweets.

Επιπλέον, θα πρέπει να αποθηκεύσετε μόνο τα tweets που είναι διατυπωμένα με κείμενο στην αγγλική γλώσσα.

2ο μέρος: Αποθήκευση των δημοφιλών θεμάτων και των tweets σε βάση δεδομένων

Για την καλύτερη διαχείριση των δεδομένων που θα συλλεχθούν απαραίτητη είναι η χρήση βάσης δεδομένων η οποία θα επιτρέπει τη γρήγορη και εύκολη ανάκτησή τους.

Στην πλαίσια της εργασίας θα χρησιμοποιηθεί η τεχνολογία βάσεων δεδομένων MongoDB (open source, NoSQL, document-oriented βάση δεδομένων). Συγκεκριμένα, θα πρέπει να δημιουργηθεί μία βάση δεδομένων με τις αντίστοιχες συλλογές για την αποθήκευση των σχετικών tweets.

Για τα tweets, θα πρέπει να διατηρείται όλη η JSON αναπαράσταση όπως επιστρέφεται από το Twitter.

Ο χρόνος που προτείνεται για την εκτέλεση του μέρους 1 & 2 είναι μέχρι τις 27/11

3ο μέρος: Προεπεξεργασία tweets

Προκειμένου να δημιουργηθούν συλλογές κειμένου από τα Tweets, θα πρέπει να υποστούν κάποια προεπεξεργασία:

1. Αρχικά σε περίπτωση που υπάρχουν retweets θα πρέπει να απομακρυνθούν.
2. Για να είναι δυνατή η επεξεργασία των tweets, θα πρέπει να γίνει tokenization, δηλαδή το κείμενο από κάθε tweet να χωριστεί στους όρους που το απαρτίζουν.
3. Απομάκρυνση αριθμών και άλλων συμβόλων. Για κάθε tweet, θα πρέπει να αποθηκευτούν μόνο οι λέξεις.
4. Για να είναι δυνατή η σύγκριση των λέξεων, θα πρέπει να γίνει κανονικοποίηση , δηλαδή οι λέξεις να περιέχουν μόνο πεζά γράμματα.
5. Στη συνέχεια θα πρέπει να αφαιρεθούν οι συχνά εμφανιζόμενες λέξεις όπως άρθρα, προθέσεις, κλπ. (stop words). Η λίστα με τις συχνά εμφανιζόμενες λέξεις, μπορεί να βρεθεί σε διάφορες τοποθεσίες στο διαδίκτυο ή μπορεί να γίνει εισαγωγή στο πρόγραμμά σας από κάποια έτοιμη βιβλιοθήκη (π.χ. nltk.stopwords για την Python). Προσοχή: εκτός από τις προκαθορισμένες λέξεις της stopword λίστας, για κάθε συλλογή θα πρέπει να αφαιρεθούν και οι αντίστοιχες λέξεις – κλειδιά.

4ο μέρος: Μετρικές σε επίπεδο συλλογής

Για την κατανόηση της κάθε συλλογής και την απεικόνιση της κοινής γνώμης την τρέχουσα στιγμή θα χρησιμοποιηθεί ένα σύνολο από μετρικές.

Ένας βασικός δείκτης είναι η εξαγωγή του γενικού συναισθήματος γύρω από το κάθε θέμα. Για τον υπολογισμό του συναισθήματος θα πρέπει να χρησιμοποιηθεί κάποιο web API το οποίο θα επιστρέφει ένα label για το συναίσθημα και τα probability scores (πχ. {"label": "negative", "prob": {"positive": 0.45 , "negative" : 0.65, "neutral":0.4}}. Το http://text-processing.com/docs/sentiment.html είναι μία πρόταση. Αφού υπολογίσετε το συναίσθημα για κάθε tweet (θετικό, αρνητικό, ουδέτερο) θα πρέπει να εμφανίσετε μία πίτα με τα ποσοστά της κάθε κατηγορίας για κάθε συλλογή από tweets.

Επιπλέον θα πρέπει για κάθε tweet να ενημερώσετε την βάση με τα επιπλέον στοιχεία που γνωρίζουμε για αυτό (θα χρειαστεί στο επόμενο βήμα). Με βάση το παράδειγμα που σας δόθηκε, θα πρέπει σε κάθε tweet να προστεθούν τα πεδία label, positive probability, negative probability, neutral probability.

Για κάθε συλλογή, θα πρέπει πριν την αφαίρεση των stop words:

1. να υπολογίσετε τις 50 πιο συχνά εμφανιζόμενες λέξεις και να τις εμφανίσετε σε ιστόγραμμα.
2. να υπολογίσετε το Zipf διάγραμμα για το σύνολο των λέξεων.

Αμέσως μετά την αφαίρεση των stopwords θα πρέπει να δημιουργήσετε ένα νέο ιστόγραμμα με τις 50 πιο συχνά εμφανιζόμενες λέξεις.

Ο χρόνος που προτείνεται για την εκτέλεση του μέρους 3 & 4 είναι μέχρι τις 11/12

5ο μέρος: Μετρικές για χρήστες.

1. Θα πρέπει να υπολογισθεί το συνολικό συναίσθημα των tweets κάθε χρήστη γύρω από ένα θέμα. Με τον τρόπο που θα γίνει η συλλογή των tweets δεν μπορούμε να εξασφαλίσουμε ότι θα συλλέξουμε μόνο ένα tweet από κάθε χρήστη. Θα πρέπει λοιπόν με βάση το ID του κάθε χρήστη να υπολογίσουμε το συνολικό του score για το συγκεκριμένο topic.
2. Επιπλέον για κάθε χρήστη θα πρέπει να υπολογίσετε και τον λόγο μεταξύ followers και friends.
3. Στη συνέχεια θα πρέπει τυπώσετε το Cumulative Distribution Frequency διάγραμμα για τον λόγο που υπολογίσατε έτσι ώστε να ελέγξετε αν υπάρχει κάποια ανωμαλία στην συλλογή σας.

Ο χρόνος που προτείνεται για την εκτέλεση του μέρους 5 είναι μέχρι τις 18/12

Παράδοση εργασίας

Κάθε ομάδα θα πρέπει να παραδώσει ένα συμπιεσμένο φάκελο με όνομα τα ΑΕΜ των φοιτητών της ομάδας (π.χ. 1234_2341_3214.zip). Ο φάκελος θα περιλαμβάνει:

1. Τον πηγαίο κώδικα σε Java ή Python που αναπτύχθηκε για τη συλλογή δεδομένων, την προεπεξεργασία των δεδομένων και την ανάλυση των αποτελεσμάτων.
2. Τα εκτελέσιμα .jar αρχεία για τη συλλογή δεδομένων.
3. Την βάση δεδομένων που έχετε δημιουργήσει. Να γίνει export η βάση (η οποία θα ονομαστεί με βάση τους ΑΜ σας - 123_456_789) μαζί με τα collections και τα tweets.
4. Μία αναφορά (.docx αρχείο) στην οποία θα περιέχεται η τεχνική περιγραφή της υλοποίησης καθώς επίσης και τα διαγράμματα που θα έχετε δημιουργήσει.
5. Ένα README.txt που θα δίνει γενικές οδηγίες για την εκτέλεση του προγράμματος.

Τα παραδοτέα θα πρέπει να αποσταλούν στη διεύθυνση ηλεκτρονικού ταχυδρομείου: ggravanis@csd.auth.gr με cc στο alextsil@csd.auth.gr και στο artemisap@csd.auth.gr . Κάθε ομάδα και όλα τα μέλη της θα εξεταστούν προφορικά σε ημερομηνία που θα ορισθεί από τους διδάσκοντες. Η προφορική εξέταση λαμβάνει το 10% του συνολικού βαθμού. Αν κάποιο μέλος απουσιάσει αδικαιολόγητα από την προφορική εξέταση χάνει αυτό το 10%. Σε περίπτωση παράδοσης της εργασίας μετά το πέρας της καταληκτικής ημερομηνίας, η ομάδα χάνει το 20% του συνολικού βαθμού της. Οι ομάδες θα αποτελούνται από 3-4 άτομα.

Προθεσμία υποβολής εργασίας 08/01/2018

Χρήσιμοι σύνδεσμοι

Γενικά

1. https://www.mongodb.com/
2. https://developer.twitter.com/en/docs/api-reference-index
3. http://text-processing.com/
4. https://www.programmableweb.com/
5. https://en.wikipedia.org/wiki/Zipf%27s_law

For Java

6. http://twitter4j.org/en/
7. http://opennlp.apache.org/
8. http://knowm.org/open-source/xchart/
9. https://hc.apache.org/

For Python

10. http://www.tweepy.org/
11. http://www.nltk.org/
12. https://matplotlib.org/
13. http://docs.python-requests.org/en/master/

Αναφορές

[1] S. A. Anwar Hridoy, M. T. Ekram, M. S. Islam, F. Ahmed, and R. M. Rahman, “Localized twitter opinion mining using sentiment analysis,” *Decision Analytics*, vol. 2, no. 1, Dec. 2015.

[2] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining.,” in *LREc*, 2010, vol. 10.

[3] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and As. Perera, “Opinion mining and sentiment analysis on a twitter data stream,” in *Advances in ICT for emerging regions (ICTer), 2012 International Conference on*, 2012, pp. 182–188.

[4] J. A. Balazs and J. D. Velásquez, “Opinion Mining and Information Fusion: A survey,” *Information Fusion*, vol. 27, no. Supplement C, pp. 95–110, Jan. 2016.