ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ
ΑΠΘ - ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
2016-2017

Παρουσίαση Εργασίας "**Ανακάλυψη Κοινοτήτων χρηστών στο Twitter**"

Υπεύθυνη Καθηγήτρια: Αθηνά Βακάλη
Υποστήριξη Εργασίας:
- Χατζάκου Δέσποινα
- Φούντα Αντιγόνη-Μαρία

ΣΤΟΧΟΣ

Ανεύρεση κοινοτήτων χρηστών με κοινά χαρακτηριστικά δραστηριότητας και θεματικά ενδιαφέροντα, λαμβάνοντας υπόψιν τις έμμεσες σχέσεις μεταξύ αυτών.

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Κοινότητα: ένα σύνολο κόμβων οι οποίοι είναι πιο πυκνά συνδεδεμένοι μεταξύ τους σε σχέση με το υπόλοιπο δίκτυο,

Ρητές σχέσεις: Σχέσεις ανάμεσα στους χρήστες οι οποίες δηλώνονται άμεσα, είναι εμφανείς

πχ ο χρήστης Α ακολουθεί τον Β, ο Β ανέφερε (mention) τον Α κλπ

Άρρητες σχέσεις: Έμμεσες σχέσεις ανάμεσα στους χρήστες, οι οποίες εξάγονται βάσει ομοιοτήτων στη συμπεριφορά ή στα ενδιαφεροντα χρηστών

πχ στους χρήστες Α και Β αρέσει η νέα δημοσίευση του χρήστη Γ

ΓΕΝΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

Συλλογή & Αποθήκευση Δεδομένων

Twitter Streaming API

MongoDB

Υπολογισμός Ομοιότητας Χρηστών & Δημιουργία Πινάκων Ομοιότητας

Cosine OR Jaccard OR Frequency-based Similarity

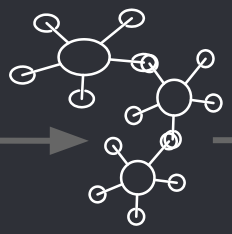
MongoDB OR MySQL OR Neo4j

HTTPS://

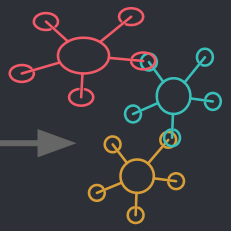
Ανάλυση & Αποθήκευση Οντοτήτων

Hashtags Table	User 1	User 2	User 3
User 1	1	0.5	0.02
User 2	0.5	1	0
User 3	0.02	0	1

x5



Δημιουργία Γράφων Χρηστών



Δημιουργία & Σύγκριση Κοινοτήτων

ΑΝΑΛΥΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΒΗΜΑΤΩΝ

ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ (*11/11/2016*)

Συλλογή tweets μέσω του Twitter Streaming API βάσει συγκεκριμένων λέξεων-κλειδιών.

Προτεινόμενη συνολική διάρκεια: 3 ημέρες

ΛΕΞΕΙΣ - ΚΛΕΙΔΙΑ

Η αναζήτηση στο Streaming API γίνεται με βάση κάποιες λέξεις κλειδιά. Μία προσέγγιση είναι η επιλογή ενός θέματος προτίμησης βάσει του οποίου να γίνει η αναζήτηση. Μία δεύτερη προσέγγιση για τη συγκέντρωση μεγάλου όγκου δεδομένων αποτελεί η επιλογή θέματος από τη λίστα των δημοφιλών θεμάτων του Twitter.

ΔΗΜΟΦΙΛΗ ΘΕΜΑΤΑ

Το Twitter υπολογίζει και ανανεώνει ανά 5 λεπτά λίστες με τα top-10 δημοφιλή θέματα (λέξεις-φράσεις κλειδιά) βάσει της συχνότητας αναφοράς σε αυτά, σε διαφορετικές γεωγραφικές περιοχές ή παγκοσμίως.

Ένας τρόπος συλλογής των πιο δημοφιλών θεμάτων είναι με τακτές κλήσεις στο Twitter REST API ώστε να συλλέγονται τα εκάστοτε θέματα.

ΔΙΑΤΗΡΗΣΗ ΔΕΔΟΜΕΝΩΝ (*11/11/2016*)

Δημιουργία βάσης δεδομένων για τα tweets, διατηρώντας ολόκληρη την JSON αναπαράσταση, με τη χρήση της MongoDB.

MongoDB

- Open source, NoSQL, document-oriented ΒΔ
- Ευέλικτη και χωρίς περιορισμούς αποθήκευση των δεδομένων σε αρχεία τύπου BSON (Binary JSON, JSONlike έγγραφα).
- Διατήρηση των δεδομένων σε ζεύγη κλειδί-τιμή (key-value)
- Κάθε βάση δεδομένων (database) μπορεί να αποτελείται από μία ή περισσότερες συλλογές (collections) - ομαδοποίηση των δεδομένων σε σύνολα.
- Κάθε σύνολο μπορεί να περιέχει μηδέν ή περισσότερα έγγραφα (documents) και κάθε έγγραφο ένα ή περισσότερα πεδία.
- Τα έγγραφα αποθηκεύονται στη βάση με schema-free τρόπο: στην ίδια συλλογή μπορούν να αποθηκευτούν έγγραφα με διαφορετική δομή ή πεδία.

MONGODB EXAMPLE

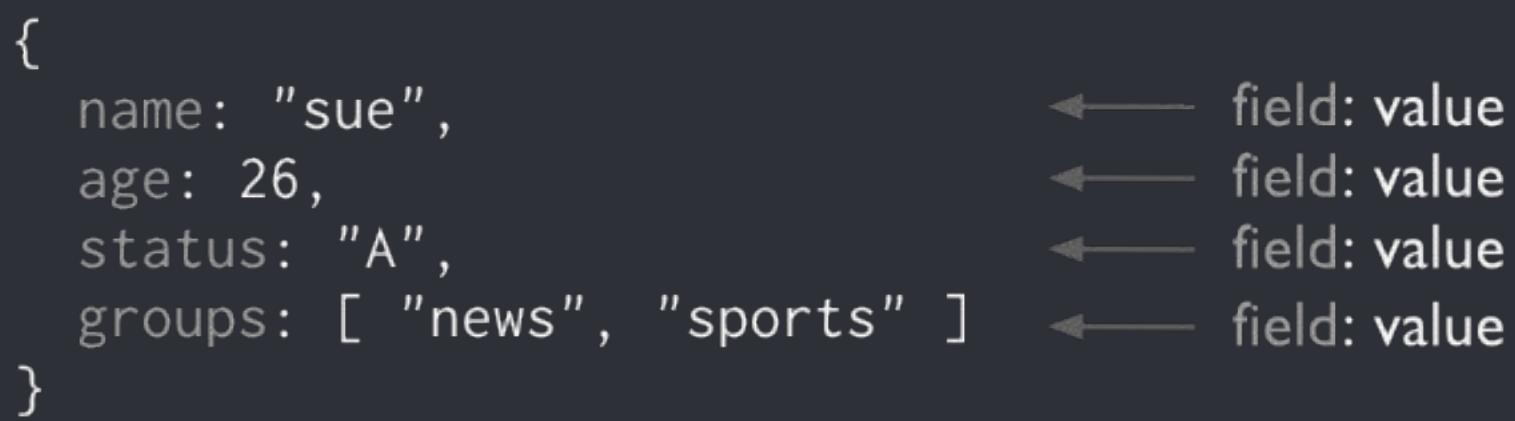


Image Source: https://iamcharmcaster.wordpress.com/2013/12/03/installing-mongodb/

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (*05/12/2016*)

Διαχωρισμός κειμένου σε λεξικολογικές μονάδες (***tokenization***) και αναγνώριση **οντοτήτων.**

Αναγνώριση:

- URLs
- Hashtags: ξεκινούν με τον χαρακτήρα #
- Mentions: ξεκινούν με τον χαρακτήρα @ και στη συνέχεια ακολουθεί το username του χρήστη
- Retweets: δηλώνονται με συγκεκριμένη δομή μέσα στην JSON αναπαράσταση του tweet

Προσοχή στα shortened URLs (t.co / bit.ly κοκ)!

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ (*05/12/2016*)

Αποθήκευση των οντοτήτων που βγήκαν απ' το προηγούμενο βήμα και των άμεσων συνδέσεων κάθε χρήστη, σε κατάλληλη δομή / βάση αποθήκευσης δεδομένων.

Πιο συγκεκριμένα πρέπει να αποθηκευτούν:

- Hashtags, URLs, χρήστες και retweets όπως προέκυψαν από την ανάλυση των δεδομένων. Για τα URLs πρέπει να διατηρείται η αντιστοίχιση της πλήρους και συντμημένης μορφής.
- Οι άμεσες συνδέσεις κάθε χρήστη με τις άλλες οντότητες (χρήστες, tweets, hashtags και URLs) & ο χρόνος χρήσης/αναφοράς (timestamp) για κάθε οντότητα, όπως φαίνεται παρακάτω.

user, timestamp, hashtag

user, timestamp, URL

user, timestamp, mentioned_user

user, timestamp, retweeted_tweet

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ (*05/12/2016*)

Η επιλογή της τεχνολογίας / δομής που θα χρησιμοποιηθεί γίνεται κατά βούληση, διαλέγοντας την πλέον κατάλληλη από:

- Document-oriented ΒΔ, πχ MongoDB ή CouchDB
- ΒΔ προσανατολισμένη σε γράφους, πχ Neo4j
- Σχεσιακή ΒΔ, πχ MySQL

Η παραπάνω επιλογή θα πρέπει να δικαιολογηθεί στην Τεχνική Αναφορά!

ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ ΜΕΤΑΞΥ ΧΡΗΣΤΩΝ (*19/12/2016*)

Δημιουργία πινάκων ομοιότητας, όπου θα αποθηκεύεται για κάθε ένα από τα ζεύγη χρηστών το αποτέλεσμα του μέτρου ομοιότητας που έχει επιλεχθεί για κάθε ένα από τα τέσσερα χαρακτηριστικά (hashtags, urls, users & tweets).

Συνολικά θα πρέπει να δημιουργηθούν 5 πίνακες: ένας πίνακας για κάθε χαρακτηριστικό, και 1 πίνακας στον οποίο θα συνδυάζονται τα παραπάνω αποτελέσματα σε ένα συνολικό μέτρο ομοιότητας.

Η επιλογή του μέτρου ομοιότητας μπορεί να γίνει ελεύθερα. Κάποια μέτρα που προτείνονται είναι τα παρακάτω:

- Cosine Similarity
- Jaccard Similarity
- Frequency-based Similarity

Η παραπάνω επιλογή θα πρέπει να δικαιολογηθεί, με σχετική αναφορά στην αρθρογραφία!

Παράδειγμα (Κανονικοποιημένου) Πίνακα Ομοιότητας

Hashtags Table	User 1	User 2	User 3
User 1	**1**	**0.5**	**0.02**
User 2	**0.5**	**1**	**0**
User 3	**0.02**	**0**	**1**

ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ ΜΕΤΑΞΥ ΧΡΗΣΤΩΝ (*19/12/2016*)

Προσοχή! Είναι πιθανό κάποιες από τις οντότητες να εμφανίζονται πολύ συχνά (πχ γενικά hashtag τύπου #tbt). Αυτό θα δυσκολέψει τη διάκριση των επιμέρους ενδιαφερόντων των χρηστών, επομένως καλό είναι να αφαιρεθούν τέτοιες γενικές περιπτώσεις.

1. Υπολογισμός εμπειρικής κατανομής της χρήσης των οντοτήτων στο σύνολο των χρηστών
2. Αφαίρεση των οντοτήτων που χρησιμοποιούνται από έναν πολύ μεγάλο αριθμό χρηστών με βάσει:
    - είτε κάποιο ποσοστό
    - είτε κάποια τεχνική για την εύρεση outliers (πχ boxplots)

ΔΗΜΙΟΥΡΓΙΑ ΓΡΑΦΟΥ ΧΡΗΣΤΩΝ (*26/12/2016*)

Κατασκευή γράφων χρηστών, με βάση τους πίνακες ομοιότητας. Για κάθε πίνακα ομοιότητας θα πρέπει να κατασκευαστεί και ένας γράφος.

- Μία ακμή ανάμεσα σε δύο χρήστες έχει ως βάρος την τιμή του αντίστοιχου μέτρου ομοιότητας.
- Απομάκρυνση συνδέσεων μεταξύ χρηστών που έχουν πολύ μικρή ομοιότητα (μπορεί να γίνει με τον ίδιο τρόπο που περιγράφηκε νωρίτερα η αφαίρεση οντοτήτων).

Απεικόνιση των γράφων που δημιουργήθηκαν, με το εργαλείο Gephi.

- Για την εισαγωγή των γράφων θα πρέπει πρώτα να αναπαρασταθούν σε κάποιο από τα formats που υποστηρίζει το Gephi (πχ CSV, GEXF).
- Δοκιμή ανάμεσα στα διάφορα layouts για την καλύτερη οπτικοποίηση του εκάστοτε γράφου.

Προαιρετικά: Αποθήκευση των γράφων στη βάση που αναπτύχθηκε νωρίτερα.

ΥΠΟΛΟΓΙΣΜΟΣ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΤΡΩΝ (*02/01/2017*)

Υπολογισμός των παρακάτω μετρικών, οι οποίες σχετίζονται με τη δομή και τη συνεκτικότητα του γράφου και σχολιασμός των αποτελεσμάτων συγκριτικά για όλους τους γράφους.

Μετρικές:

- Average Clustering Coefficient
- Diameter
- Graph Density

ΑΝΕΥΡΕΣΗ ΚΟΙΝΟΤΗΤΩΝ (*02/01/2017*)

Εφαρμογή του αλγορίθμου Louvain στους γράφους (στο Gephi εμφανίζεται ως modularity) ώστε να προκύψουν κοινότητες.

- Να ληφθούν υπόψιν τα βάρη των ακμών
- Να γίνει χρωματισμός των κόμβων με βάση την κοινότητα στην οποία ανήκει
- Να γίνει επιλογή του κατάλληλου layout ώστε να αναδειχθούν οι κοινότητες
- Προσοχή στην τιμή του modularity (ιδανικά από 0.3 μέχρι 0.7)

ΣΥΓΚΡΙΣΗ ΔΟΜΩΝ ΚΟΙΝΟΤΗΤΩΝ (*16/01/2017*)

Τελικός στόχος: παρατήρηση διαφοράς που παρουσιάζουν οι διάφορες ομαδοποιήσεις ανάλογα με το εκάστοτε μέτρο ομοιότητας.

Για να επιτευχθεί αυτό θα πρέπει:

- Να γίνει εξαγωγή από το Gephi του αρχείου στο οποίο περιγράφεται η ανάθεση κόμβων σε κοινότητες για κάθε γράφο.
- Να γίνει υπολογισμός του δείκτη NMI (Normalized Mutual Information) ανα ζεύγος μέτρων, για τις κοινότητες που ανακαλύφθηκαν.

Τέλος, θα πρέπει να σχολιαστούν τα αποτελέσματα.

ΠΑΡΑΔΟΤΕΑ

Πηγαίος Κώδικας

Ο κώδικας που αναπτύχθηκε για την υλοποίηση της εργασίας

(Υλοποίηση σε Java)

Τεχνική Αναφορά

Σύντομη τεχνική περιγραφή της υλοποίησης και σχολιασμός των αποτελεσμάτων

Gephi files

Τα αρχεία εισόδου δεδομένων στο Gephi & τα αντίστοιχα Gephi projects

Προθεσμία: 16 Ιανουαρίου 2017 !

ΕΠΙΚΟΙΝΩΝΙΑ

Για επικοινωνία/απορίες μπορείτε να στέλνετε στα παρακάτω email:

deppych@csd.auth.gr
founanti@csd.auth.gr

Τα **email** θα πρέπει να ξεκινάνε με **πρόθεμα**: [P_PSPI_2016]

ΧΡΗΣΙΜΟΙ ΣΥΝΔΕΣΜΟΙ

- Twitter Streaming API: https://dev.twitter.com/streaming/overview
- MongoDB: http://www.mongodb.org/
- CouchDB: http://couchdb.apache.org/
- Neo4j: https://neo4j.com/
- MySQL: https://www.mysql.com/
- Gephi: https://gephi.org/