Twitter Community Detection

Πληροφοριακα Συστηματα Παγκοσμιου ιστου

Αναγνωστου αντωνιοσ 2268

Λασκαριδησ Στεφανοσ 2315

{anagnoad,laskstef}@csd.auth.gr

2017

Table of Contents

[Εισαγωγή 3](#_Toc472534595)

[Περιγραφή προβλήματος 4](#_Toc472534596)

[Συλλογή και Αποθήκευση Πρωτογενών Δεδομένων 5](#_Toc472534597)

[Μοντελοποίηση Δεδομένων 6](#_Toc472534598)

[Υπολογισμός Ομοιότητας Χρηστών 7](#_Toc472534599)

[Δημιουργία και Ανάλυση Γράφου 8](#_Toc472534600)

[Δημιουργία Γράφου 8](#_Toc472534601)

[Υπολογισμός Στατιστικών Μέτρων 8](#_Toc472534602)

[Εύρεση Κοινοτήτων 8](#_Toc472534603)

[Σύκριση Δομών Κοινοτήτων 8](#_Toc472534604)

[Αναφορές 9](#_Toc472534605)

# Εισαγωγή

Η παρούσα εργασία εκπονήθηκε στα πλαίσια του μαθήματος «Πληροφοριακά Συστήματα Παγκοσμίου Ιστού» του 7ου εξαμήνου του τμήματος Πληροφορικής ΑΠΘ.

Συγκεκριμένα, σκοπός της είναι η συλλογή και επεξεργασία tweets για μία θεματολογία σε χρονικό διάστημα 3 ημερών καθώς και η ανάλυση της ομοιότητας μεταξύ χρηστών βάσει συγκεκριμένων πεδίων.

Το τεχνικό μέρος της εργασίας αναπτύχθηκε σε Java 1.8, με τη συλλογή από το Twitter Streaming API [1] να χρησιμοποιεί Spark Streaming (v. 2.0.1) [2], ενώ τα υπόλοιπα τμήματα συμβατική Java με χρήση βιβλιοθηκών. Για την αποθήκευση των πρωτογενών δεδομένων, χρησιμοποιήθηκε MongoDB [3] (v3.4.0). Στη συνέχεια εξήχθησαν οι χρήσιμες πληροφορίες σε σχεσιακή βάση δεδομένων PostgreSQL [4] (v. 9.2.18). Για την ανάλυση των γράφων χρησιμοποιήθηκε το πρόγραμμα Gephi [5] (v. 0.9.1). Τέλος, το παραπάνω software εγκαταστάθηκε σε εικονική μηχανή του ~okeanos IaaS [6] με λειτουργικό σύστημα CentOS (v. 7.2.1511).

# Περιγραφή προβλήματος

# Συλλογή και Αποθήκευση Πρωτογενών Δεδομένων

# Μοντελοποίηση Δεδομένων

# Υπολογισμός Ομοιότητας Χρηστών

# Δημιουργία και Ανάλυση Γράφου

## Δημιουργία Γράφου

## Υπολογισμός Στατιστικών Μέτρων

## Εύρεση Κοινοτήτων

## Σύκριση Δομών Κοινοτήτων

# Αναφορές