Συστήματα Ανάκτησης Πληροφοριών

Εργασία 2η

Χρύσα Μαυράκη p3130128

Σ’ αυτήν την εργασία χρησιμοποιήσα τα tweets από την πρώτη εργασία και μερικά επιπλέον.

Επέκτεινα την πρώτη εργασία κρατώντας και τη λειτουργικότητά της. Πλεον το συναίσθημα καθορίζεται μετρώντας το πλήθος των θετικών και αρνητικών λέξεων αφού πρώτα αφαιρεθούν τα διπλότυπα και καθαριστεί το κάθε tweet απο URL γίνει stem, μετατραπεί σε κεφαλαία, αφαιρεθούν οι τόνοι καθώς και τα αμιγώς ξενόγλωσσα tweet.

Για να πάρω τις πληροφορίες απο τα excel αρχεία που μας δόθηκαν χρησιμοποίησα την εξωτερική βιβλιοθήκη poi-ooxlm της apache.

Δίνεται η δυνατότητα εκτύπωσης ημερίσιων και εβδομαδιαίων στοιχείων (μέσος όρος, τυπική απόκλιση) για κάθε μια από τις κατηγορίες. Τα αποτελέσματα φαίνονται και στο αρχείο dailyAndWeeklyStats το οποίο έχω συμπεριλάβει.

Για την ανάλυση svd χρησιμοποίησα την εξωτερική βιβλιοθήκη commons-math3 της apache commons.

Εν τέλει για το ερώτημα 10, οι όροι οι οποίοι πληρούσαν τις προϋποθέσεις (να εμφανίζονται σε 2 διαφορετικά tweets) ήταν περίπου 4 χιλιάδες και το svd ήταν αρκετά χρονοβόρο. Για τον λόγο αυτό συμπεριλαμβάνω και τα αποτελέσματα της κονσόλας, καθώς επίσης και το σύνολο των αρχείων που δημιουργήθηκαν, τα οποία περιέχουν για κάθε εκτέλεση με διαφορετικό p τα extended λεξικά. Παρατήρησα ότι όσο αυξάναμε το p τόσο αυξάνονταν και τα λεξικά με αποτέλεσμα κάποια στιγμή να υπάρχουν πολλές ίδιες λέξεις και στο θετικό και στο αρνητικό λεξικό. Επίσης αν το p είναι πολύ χαμηλό (πχ 1) τότε το λεξικό δεν αυξάνεται σημαντικά. Οπότε λογικά, καλύτερη θα είναι μια ενδιάμεση λύση με p=4 ή 5 όπου θα αυξηθεί το λεξικό αλλά δεν θα καταλήξουμε σε υπερβολικό αριθμό συγκρούσεων.