# Προγραμματιστική άσκηση - ΦΑΣΗ 2

Από την πρώτη φάση έχετε έτοιμο ένα μέρος τη εφαρμογής το οποίο συγκεντρώνει tweets σε φακέλους (ή σε ΒΔ) ανά κατηγορία σύμφωνα με το username (tsipras, mitsotakis) και το hashtag (SYRIZA, ND ή ισοδύναμα, π.χ. SYRIZANEL, NEADIMOKRATIA).

Απότι διαπιστώσαμε το πλήθος των tweets με emoticons είναι ελάχιστο επομένως θα αγνοήσουμε αυτή τη κατάταξη στην δεύτερη φάση.

0. Η διαδικασία συγκέντρωσης των tweets θα γίνεται σε ημερήσια και εβδομαδιαία βάση, δηλαδή όλα τα tweets μιας ημέρας θα επεξεργάζονται και στο τέλος της εβδομάδος θα επεξεργάζονται ξανά ώστε να έχουμε μια «περίληψη» αυτών. Κάθε νέα εβδομάδα θα ξεκινά η συγκέντρωση νέων tweets με τις ίδιες όπως παραπάνω κατηγορίες. Έτσι θα έχουμε μια ημερήσια επεξεργασία των tweets και μια εβδομαδιαία ανάλυση για την εξαγωγή της εβδομαδιαίας περίληψης.

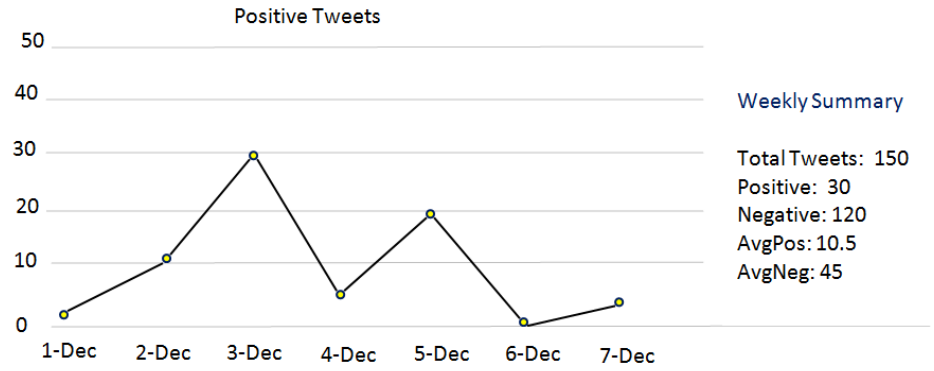
1. Τα κείμενα θα υπόκεινται σε μια επεξεργασία σε ημερήσια βάση. Δηλαδή τα tweets που συγκεντρώθηκαν σε μια ημέρα θα καθαρίζονται από meta-data, σημεία στίξεως, ειδικούς χαρακτήρες, HTTP links και αριθμούς. Θα κρατήσετε μόνο το «καθαρό» περιεχόμενο (κειμενάκι ~140 χαρακτήρων) που περιέχει κάθε tweet. Στη συνέχεια θα μετατρέψετε όλους τους χαρακτήρες σε κεφαλαία χωρίς τόνο. Εφαρμόστε αποκοπή καταλήξεων με την διαδικασία που επισυνάπτεται. Θα μπορούσατε να χρησιμοποιήσετε και τον Greek Analyzer του Lucene μετά το καθάρισμα των tweets.
2. Θα διαπιστώσετε ότι παρόλο που μπορεί να χρησιμοποιήσατε (π.χ. παράμετρο –RT) υπάρχουν ακόμη αρκετά tweets με το ίδιο ακριβώς περιεχόμενο. Αφαιρέστε τα διπλότυπα tweets.
3. Για καθεμιά από τις 4 κατηγορίες (tsipras, mitsotakis, SYRIZA, ND) υπολογίστε τις συχνότητες των θετικών και αρνητικών λέξεων όλων των tweets της ημέρας. Για το σκοπό αυτό χρησιμοποιείστε το λεξικό θετικών λέξεων (PosLex) και το λεξικό αρνητικών λέξεων (NegLex) που σας δίνονται.
4. Υπολογίστε τα positive και negative tweets ημηρεσίως για καθεμιά από τις 4 κατηγορίες. Ένα tweet θεωρείται ως θετικό αν το πλήθος των θετικών λέξεων που περιέχει είναι μεγαλύτερο του πλήθους των αρνητικών λέξεων. Όμοια θεωρείται ως αρνητικό αν το πλήθος των θετικών λέξεων που περιέχει είναι μικρότερο του πλήθους των αρνητικών λέξεων. Σε περίπτωση ισότητας θεωρείστε το tweet ως θετικό ή ως ουδέτερο (όπως θέλετε).
5. Υπολογίστε το συνολικό πλήθος των tweets, το μέσο των θετικών και τυπική απόκλιση των θετικών και αρνητικών tweets εβδομαδιαίως για κάθε κατηγορία.
6. Για τα tweets μιας εβδομάδος ή και περισσότερων εβδομάδων σε περίπτωση που είναι λίγα κατασκευάστε τον πίνακα, X, "*term x document*". Κρατείστε τους όρους που εμφανίζονται τουλάχιστον σε δύο κείμενα. Εφαρμόστε svd ανάλυση στον πίνακα X χρησιμοποιώντας τη συνάρτηση svds του matlab, [U,S,V]=svds(X,k), όπου το k δηλώνει την τάξη προσέγγισης. Τώρα οι όροι παρίστανται με τον πίνακα $U_k$ (mxk), όπου m είναι το πλήθος των όρων. Νορμαλοποιείστε τις γραμμές του πίνακα $U_k$ με την Ευκλείδεια νόρμα. Για κάθε όρο, t, υπολογίστε τους p-πλησιέστερους προς αυτόν όρους $\{t_1, t_2, \ldots, t_p\}$. Αν ο όρος t ανήκει στο θετικό λεξικό χαρακτηρίστε και τους p πλησιέστερους προς αυτόν όρους ως θετικούς και αποθηκεύστε τους σε ένα αρχείο ExtPos(t). Αν ο όρος t ανήκει στο αρνητικό λεξικό χαρακτηρίστε και τους p πλησιέστερους προς αυτόν όρους ως αρνητικούς και αποθηκεύστε τους σε ένα αρχείο ExtNeg(t), διαφορετικά αν ο όρος δεν ανήκει ούτε στο θετικό ούτε στο αρνητικό λεξικό τότε αγνοείστε τον.

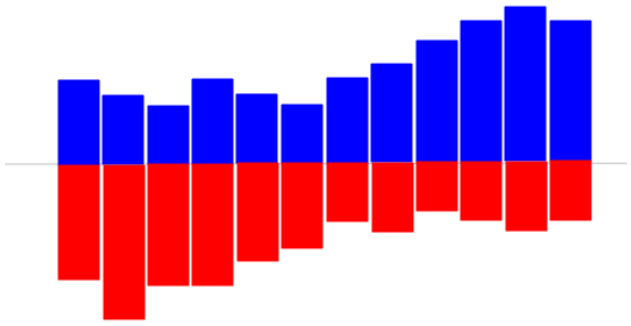
7. Για κάθε όρο $t_i$ υπολογίστε το πλήθος των όρων ($NP_i$) του ExtPos($t_i$) που ανήκουν στο PosLex. Όμοια το πλήθος των όρων ($NN_i$) του ExteNeg($t_i$) που ανήκουν στο NegLex. Τυπώστε τους μέσες τιμές από όλους τους όρους:

$$\overline{NP} = \frac{\sum NP_i}{m} \qquad \overline{NN} = \frac{\sum NN_i}{m}$$ (m=το πλήθος των όρων)

8. Τυπώστε τα σύνολα (a) newPos={t: t∈ExtPos and t∉PosLex} και (b) newNeg={t: t∈ExtNeg and t∉NegLex}
9. Επαναλάβετε τα 7 (μετά τη λύση svd), 8 και 9 με τιμές του p=1, 2, 4, 5, 10.
10. Γράψτε τα συμπεράσματά σας από το 10, π.χ. για ποια τιμή του p θεωρείτε ότι τα newPos, newNeg είναι τα καλύτερα?.

Η δημιουργία UI για τα 4, 5, και θα εκτιμηθεί ιδιαίτερα. Ένα παράδειγμα οπτικοποίησης των αποτελεσμάτων φαίνεται στον πίνακα. Η εβδομαδιαία περίληψη εμφανίζει τις τιμές εφόσον έχει συμπληρωθεί η εβδομάδα. Τα εβδομαδιαία αποτελέσματα αποθηκεύονται κατάλληλα και είναι έτοιμα για την εμφάνισή τους στην οθόνη.



Το σύστημα θα μπορούσε να εμφανίζει διαχρονικά τους μέσους των θετικών και αρνητικών tweets

*Βοήθεια στο 7. Η λύση με το matlab είναι αρκετά απλή. Μπορούμε να βρούμε τους p-πλησιέστερους γειτόνους κάθε όρου ως εξής.*

*Υπολογίζουμε τον πίνακα C=Uk*Uk'; όπου Uk' δηλώνει τον ανάστροφο πίνακα του Uk. Ο πίνακας C είναι mxm και το στοιχείο $C_{ij}$ δηλώνει την ομοιότητα του όρου $t_i$ με τον όρο $t_j$. (τα διαγώνια στοιχεία του C θα είναι 1). Επομένως σε κάθε όρο, $t_i$, εκχωρούμε τους όρους που αντιστοιχούν στα p μεγαλύτερα στοιχεία της γραμμής (διαφορετικά της διαγωνίου).*