**Αποθήκες Δεδομένων και Τεχνικές Ανάλυσης Δεδομένων ΠΜΣ 2024  
Ανάλυση Συναισθημάτων σε δεδομένα κειμένου**

**Γεώργιος Βώβος**

**Αριθμός μητρώου: Μ013124003**

**Νικόλαος Βώβος**

**Αριθμός μητρώου: M013124004**

*Σκοπός της εργασίας είναι η ολοκλήρωση μιας διαδικασίας εξαγωγής και παρουσίασης των συναισθημάτων όπως αυτά αποτυπώνονται σε δεδομένα κειμένου.   
Η υλοποίησή σας θα πρέπει να γίνει στη γλώσσα R.*

*Στο πλαίσιο αυτού του project καλείστε να υλοποιήσετε τα παρακάτω ερωτήματα τα οποία συνιστούν τα επιμέρους στάδια ολοκλήρωσης μιας διαδικασίας ανάλυσης συναισθήματος:*

***1. Δεδομένα****: Θα πρέπει να διαμορφώσετε ένα σύνολο από δεδομένα τα οποία αφορούν απόψεις για ένα συγκεκριμένο θέμα. Μπορείτε είτε να βρείτε κάποιο έτοιμο σύνολο δεδομένων ή κάποιο από τα διαθέσιμα δεδομένα που υπάρχουν στο eclass.   
Στην περίπτωση που υλοποιήσετε data scraper (πχ για το threads) θα έχετε και ένα bonus 0.5 μονάδας. Θα πρέπει να διαμορφώσετε ένα σύνολο από δεδομένα τα οποία αφορούν απόψεις για ένα συγκεκριμένο θέμα. Μπορείτε είτε να βρείτε κάποιο έτοιμο σύνολο δεδομένων ή κάποιο από τα διαθέσιμα δεδομένα που υπάρχουν στο eclass. Στην περίπτωση που υλοποιήσετε data scraper (πχ για το threads) θα έχετε και ένα bonus 0.5 μονάδας.*Για τους σκοπούς αυτής της άσκησης γράψαμε ένα μικρό python script(LoadPostsUsingMastodonAPI.py) που κάτεβάζει δεδομένα(posts) χρησιμοποιώντας το Mastodon API.  
Οδηγίες για το πως να τρέξετε το αρχείο αφού πρώτα πάρετε ένα API Key βρίσκονται στο **ReadMe.md**  
  
<https://en.wikipedia.org/wiki/Mastodon_(social_network)>   
Mastodon is a free and open-source software platform for decentralized social networking with microblogging features similar to Twitter.  
[mastodon.social](https://mastodon.social/explore) is one of the many independent Mastodon servers you can use to participate in the fediverse.   
Ένας άλλος δημοφιλής server είναι ο <https://hachyderm.io/home> τον οποίο χρησιμοποιήσαμε για να κατεβάσουμε τα δεδομένα μας.

Ψάξαμε για δεδομένα με την λέξη κλεδί «Trump» αλλά πολύ εύκολα μπορούμε να διαλέξουμε κάποιο άλλο θέμα αλλάζωντας την γραμμή 73  
LoadPostsUsingMastodonAPI.py

    topic = "Trump"  # Replace with the topic you want to search for

Διαλέξαμε το Mastodon κι όχι μια άλλη πλατφόρμα(e.g Threads) λόγω των γενναιόδωρων Rate Limits.

***2. Καθαρισμός Δεδομένων:*** *Με εφαρμογή κατάλληλων μεθόδων για τον καθαρισμό των δεδομένων θα πρέπει να διαμορφώσετε ένα σύνολο δεδομένων που θα περιέχει όρους κειμένου οι οποίοι έχουν νόημα για την ανάλυσή σας.*Ξεκινήσαμε φιλτράρωντας links, αριθμούς, άρθρα και μερικές επιπλέον λέξεις που επηρέαζαν τα αποτελέσματα ενώ δεν έπρεπε.  
  
Είδαμε ότι η απλή μας λογική είχε πρόβλημα με τα «negations» , οπότε γράψαμε λίγο κώδικα που τα χειρίζεται λίγο καλύτερα (με μια απλή string-replace λογική)  
Μετά από αυτό, τα αποτελέσματα φαίνεται να είναι καλύτερα (ολά όχι τέλεια)

***3. Εξαγωγή συναισθημάτων:*** *Με χρήση ενός από τα διαθέσιμα πακέτα της R, καλείστε να αξιολογήσετε το συναίσθημα που εκφράζεται στα δεδομένα σας σε 2 επίπεδα: α) αρνητικό-θετικό-ουδέτερο και β) σε ένα εύρος συναισθημάτων που θα το ορίσετε εσείς ανάλογα με το πακέτο που θα χρησιμοποιήσετε.*

***4. Κατηγοριοποίηση των χρηστών με βάση το συναίσθημα που εκφράζουν:*** *Με βάση το αποτέλεσμα της αξιολόγησης του συναισθήματος του κάθε χρήστη που πραγματοποιήθηκε στο προηγούμενο βήμα, θα πρέπει να εφαρμόσετε έναν αλγόριθμο ομαδοποίησης προκειμένου να εξάγετε ομάδες χρηστών που παρουσιάζουν ομοιότητες ως προς το συναίσθημα που εκφράζουν (0.5 μονάδες bonus).*

Αυτό το κομμάτι, μιας και προαιρετικό, δεν υλοποιήθηκε.

***4. Οπτικοποίηση των αποτελεσμάτων:*** *Με χρήση κατάλληλων πακέτων να παρουσιάσετε συγκεντρωτικά τα αποτελέσματα της ανάλυσης σας που να δείχνουν την κατανομή των συναισθημάτων στα 2 επίπεδα.*

Social media σαν τους διάφορους Mastodon servers, Threads, BlueSky έγιναν περισσότερο δημοφιλή μετά την εξαγορά του Twitter από τον Musk οπότε θα περιμέναμε θέματα σχετικά με τον Musk/Trump να μάλλον να μην έχουν ιδιαίτερα θετικά συναισθήματα.

A graph with different colored bars

AI-generated content may be incorrect.A pie chart with different colored circles

AI-generated content may be incorrect. A close up of words

AI-generated content may be incorrect.