# Εργασία για το μάθημα Τεχνικές Εξόρυξης Δεδομένων

## Εαρινό εξάμηνο , Ακ. Έτος 2021-22

*Ανάλυση συναισθημάτων (μπορεί να γίνει και ομαδικά, το πολύ 2 άτομα)*

**Ανάλυση συναισθημάτων (Sentiment Analysis):** η διαδικασία υπολογιστικής ταυτοποίησης και κατηγοριοποίησης των απόψεων που εκφράζονται σε ένα κομμάτι κειμένου, προκειμένου να καθοριστεί εάν η στάση του συγγραφέα έναντι ενός συγκεκριμένου θέματος, προϊόντος κλπ. είναι θετική, αρνητική ή ουδέτερη.


Τα δεδομένα στα οποία θα εργαστείτε προέρχονται από την γνωστή πλατφόρμα κοινωνικής δικτύωσης twitter και αφορούν προσφατα ***tweets σχετικά με τα εμβόλια κατά τις COVID-19*** που χρησιμοποιούνται σε ολόκληρο τον κόσμο.

**Δεδομένα για την εργασία:**

Στο eclass θα βρείτε ένα αρχείο για την εργασία σας. Θα χρειαστεί αρχικά να χωρίσετε το αρχείο σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Πιο συγκεκριμένα χρησιμοποιώντας το αρχείο eclass_all_with_sentiment.pkl **θα δημιουργήσετε**:

Ένα αρχείο **train.tsv** (θα είναι το 80% των συνολικών data points) που θα περιέχει τα δεδομένα που θα χρησιμοποιήσετε για εκπαίδευση των μοντέλων σας. Τα δεδομένα εκπαίδευσης περιέχουν tweets με την ένδειξη positive, negative ή neutral.
Ενα αρχείο **test.tsv** (το 20% των data points) το οποίο θα περιέχει τα δεδομένα που θα χρησιμοποιήσετε για να δοκιμάσετε το μοντέλο σας και να κάνετε μία πρόβλεψη. Πρέπει το μοντέλο σας να αποφασίσει για κάθε ένα από τα tweets που υπάρχουν στο σύνολο των **test** αν εφράζει θετικό, αρνητικό ή ουδέτερο συναίσθημα.

**Ζητούμενα:**

Η εργασία θα γίνει με την γλώσσα προγραμματισμού Python. Στη σχετική ενότητα στο eclass θα βρείτε όλο το υλικό που αφορά την python.

1. **Προεπεξεργασία και καθάρισμα των δεδομένων** (10%)
   Αφαιρούμε τα σημεία στίξης, μετατρέπουμε όλους τους χαρακτήρες σε μικρούς, αφαιρούμε σύμβολα, όπως hashtags, emoticons,emojis, τα links και τα stopwords από το σύνολο των δεδομενων).

2. **Ανάλυση των δεδομένων**. (20%)
   Καλείστε να γράψετε μερικές εντολές σε python που θα σας βοηθήσουν να “μελετήσετε” τα δεδομένα που σας δίνονται και να εξάγετε μερικά συμπεράσματα. Μερικά από τα ερωτήματα που μπορείτε να απαντήσετε είναι τα παρακάτω:

   i. Ποιά είναι η κατανομή των συναισθημάτων positive, negative, neutral σε ολόκληρο το σύνολο των δεδομένων;
   ii. Ποιες είναι οι πιο συνηθισμένες λέξεις σε ολόκληρο το σύνολο των δεδομένων;
   iii. Ποιες είναι οι πιο συνηθισμένες λέξεις στα αρνητικά, στα θετικά και στα ουδέτερα tweets, αντίστοιχα;
   iv. Φτιάξτε ένα υποσύνολο απο τα tweets εκείνα που περιέχουν την λέξη ‘astrazeneca’. Στην συνέχεια εντοπίστε εκείνα που περιέχουν τις λέξεις 'moderna' και 'pfizer' ή 'biontech'. Συγκρίνετε ώς προς το συναίσθημα τα δύο υποσύνολα.
   v. Χρησιμοποιώντας την στήλη date σχεδιάστε το πλήθος των tweets ανά μήνα. Για τους μήνες για τους οποίους υπάρχει μεγάλος αριθμός από tweets ανατρέξτε στην ειδησεογραφία και εντοπίστε αν υπάρχουν σημαντικά γεγονότα που αφορούν την πανδημία για αυτούς τους μήνες (πχ έγκριση ενός νέου εμβολίου, νέα μετάλλαξη, κ.α.) .
   vi. Μπορείτε να σκεφτείτε και κάποιες άλλες παρατηρήσεις που προκύπτουν από τα δεδομένα (τουλαχιστον 2); Παρουσιάστε σχετικά γραφήματα.

   Μπορείτε να παρουσιάσετε τα παραπάνω αποτελέσματα είτε σε διαγράμματα είτε με ένα Word cloud (όπου είναι εφικτό).

3. **Vectorization - εξαγωγή χαρακτηριστικών** (30%)

   Ακολουθήστε τις οδηγίες που παρουσιάσαμε στο φροντιστήριο και ετοιμάστε τα χαρακτηριστικά για κάθε tweet (στήλη text) χρησιμοποιώντας:

   1. Bag-of-words
   2. Tf-idf
   3. word embeddings

   Χρησιμοποιήστε τη βιβλιοθήκη pickle της Python για να αποθηκεύσετε τα χαρακτηριστικά σε αρχεία *.pkl . Με αυτό τον τρόπο δεν χρειάζεται να υπολογίζονται από την αρχή τα χαρακτηριστικά κάθε φορά που τρέχετε το

πρόγραμμά σας, αλλά μπορείτε μόνο να τα φορτωνεται στην μνήμη χρησιμοποιώντας την αντίστοιχη μέθοδο *load*.

**4**. **Δοκιμάζουμε ταξινομητές (SVM, KNN, Random Forests)** (30%)

a. SVM
b. Random Forests
c. KNN

Δοκιμάστε τους ταξινομητές σας με τα χαρακτηριστικά BOW, TFID, και word embeddings. Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου χρησιμοποιώντας 10-fold Cross Validation χρησιμοποιώντας τις παρακάτω μετρικές:

- Precision / Recall / F-Measure
- Accuracy

**4**. **Μοντελοποίηση θεμάτων LDA (Latent Dirichlet Allocation).** (10%)

Σε αυτό το ερώτημα θα χρησιμοποιήσετε τον αλγόριθμο LDA για να εντοπίσετε στα δεδομένα κειμένου που σας έχουν δοθεί, τα θέματα από τα οποία απαρτίζονται.

Βήματα που θα ακολουθήσετε είναι τα παρακάτω

- Tokenizing
- Stop word removal
- Lemmatization
- Stemming
- Εφαρμογή bag-of-words

Στη συνέχεια χρησιμοποιήστε την μέθοδο LDA για να εξάγετε τα topics από του τίτλους. Στις παραμέτρους του LDA κρατήστε αρχικά σταθερό τον αριθμό των topics σε 10.

Ο αλγόριθμος LDA ανήκει στην κατηγορία αλγορίθμων που εφαρμόζουν μάθηση χωρίς επίβλεψη. Συνεπώς δεν είναι εφικτό απο πριν να γνωρίζουμε πόσα topics μπορούν να εξαχθούν από τα δεδομένα μας. Το εργαλείο pyLDAvis μπορεί να βοηθήσει σε αυτή την κατεύθυνση. Για την αξιολόγηση των αποτελεσμάτων όσον αφορά τα θέματα που εξάγει ο αλγόριθμος LDA συχνα χρησιμοποιείται η μετρική της συνεκτικότητας ή topic coherence. Δύο έγγραφα είναι συνεκτικά αν υποστηρίζουν το ένα το άλλο, εάν μπορούν, δηλαδή, να ερμηνευθούν σε ένα κοινό πλαίσιο. Η μετρική της συνεκτικότητας ή ***topic coherence***, αποδίδει ένα

σκορ σε κάθε θέμα με βάση την σημασιολογική ομοιότητα των κυριότερων λέξεων που έχουν αποδοθεί σε αυτό. Έτσι, γίνεται ένας διαχωρισμός ανάμεσα σε θέματα που παρουσιάζουν σημασιολογική συνοχή και είναι ερμηνεύσιμα, και σε θέματα που έχουν, απλώς, προκύψει από την εφαρμογή των στατιστικών μεθόδων.

Χρησιμοποιήστε την μετρική της συνεκτικότητας (*topic coherence*) της βιβλιοθήκης **gemsim** για να βρείτε τον βέλτιστο αριθμό των θεμάτων που μπορεί να εξαχθούν από τα δεδομένα σας. Συγκεκριμένα σχεδιάστε ένα γράφημα που για διάφορες τιμές θεμάτων δείχνει την τιμή του coherence score (c_v). Επιλέγεται, κατ' αυτόν τον τρόπο, ο αριθμός των θεμάτων που μεγιστοποιεί την μετρική της συνεκτικότητας.
Τέλος, χρησιμοποιείστε την βιβλιοθήκη pyLDAvis για να οπτικοποιήσετε τα αποτελέσματα του LDA.

**6. Beat the Benchmark (bonus)** (10%)

Τέλος θα πρέπει να πειραματιστείτε με όποια μέθοδο Classification θέλετε, κάνοντας οποιαδήποτε προ-επεξεργασία στα δεδομένα επιθυμείτε με στόχο να ξεπεράσετε όσο περισσότερο μπορείτε την απόδοση σας στο προηγούμενο ερώτημα. Ενδεικτικά αναφέρουμε μερικά επιπλέον χαρακτηριστικά που μπορείτε να χρησιμοποιήσετε
Πλήθος αναφορών σε άλλους χρήστες μέσα στο tweet
Πλήθος υπερσυνδέσμων (hyperlinks) μέσα στο tweet
Πλήθος λέξεων που εκφράζουν αρνητικό συναίσθημα
Πλήθος λέξεων που εκφράζουν θετικό συναίσθημα
Πλήθος θαυμαστικών (!) μέσα στο tweet
Πλήθος ερωτηματικών (?)
Πλήθος εισαγωγικών, μονά (‘) και διπλά (“).
Πλήθος αναδημοσιεύσεων (retweets)
Πλήθος κεφαλαίων γραμμάτων
Πλήθος πεζών γραμμάτων
Επίσης μπορείτε να πειραματιστείτε και με τις μεθόδους Lemmatization, Stemming για τη βελτίωση της απόδοσης.

**Παρουσίαση αποτελεσμάτων:** Σχεδιάστε έναν πίνακα στο οποίο να φαίνονται τα αποτελέσματά σας για τους διαφορετικούς ταξινομητές που χρησιμοποιήσατε με

βάση τα διαφορετικά  διανύσματα χαρακτηριστικών. Σχολιάστε τα αποτελέσματά σας (πότε παρατηρείται βελτίωση, ποιά πιστεύετε ότι είναι τα καλύτερα χαρακτηριστικά κα) .

**Παραδοτέο**:

Η εργασία μπορεί να εκπονηθεί ατομικά ή σε ομάδες **2 ατόμων**.

Ο φάκελος **scr** της εργασίας είναι ο φάκελος στον οποίο θα γράψετε τον κώδικά σας και είναι και αυτός που θα παραδώσετε (δηλαδή δεν θα παραδώσετε εκ νέου τα δεδομένα εκπαίδευσης/δοκιμής). Θα ανεβάσετε στο eclass ένα φάκελο της μορφής sdixxxx. (όπου sdi το ΑΜ ενός εκ των ατόμων της ομάδας).

Το παραδοτέο σας πρέπει να περιέχει **ΥΠΟΧΡΕΩΤΙΚΑ** ένα **Python notebook** με το οποίο θα μπορεί κάποιος να τρέξει την εργασία σας βήμα-βήμα. Στο notebook μπορείτε σε όποια σημεία κρίνετε απαραίτητο να εισάγετε **visualizations** με τον τρόπο που θα εξηγήσουμε στα φροντιστήρια ώστε να παρουσιάσετε και με ωραίο τρόπο τα αποτελέσματά σας. **Το notebook αποτελεί και την ολοκληρωμένη αναφορά** για την εργασία σας (δεν θα παραδώσετε τίποτα σε doc, pdf) , σχεδιάστε το με προσοχή, να θυμάστε να γράψετε μία περιγραφή σε κάθε βήμα για το τι ποιό ερώτημα απαντάται σε κάθε κελί.