Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2020-2021
2η Άσκηση

Ομαδική Εργασία (2 Ατόμων-με την ίδια ομάδα που είχατε στην 1η εργασία)

Σκοπός της εργασίας
Σκοπός της εργασίας είναι η εξοικείωσή σας με τα βασικά στάδια της διαδικασίας που ακολουθούνται για την εφαρμογή τεχνικών εξόρυξης δεδομένων, ήτοι: συλλογή, προ-επεξεργασία/καθαρισμός, μετατροπή, εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού Python με την χρήση των εργαλείων/βιβλιοθηκών: jupyter notebook, pandas, gensim και SciKit Learn.

<u>Περιγραφή</u>

Η εργασία σχετίζεται με την κατηγοριοποίηση δεδομένων κειμένου από ειδησεογραφικά άρθρα και συγκεκριμένα τον εντοπισμό των ψευδών ειδήσεων. Το αρχείο δεδομένων με το οποίο θα ασχοληθείτε (και το οποίο βρίσκεται στην ηλεκτρονική τάξη στην ενότητα Έγγραφα) περιέχει 2 αρχεία σε μορφή csv (Fake.csv, True.csv).
Σε ένα jupyter notebook θα χρειαστεί να απαντήσετε στα παρακάτω ερωτήματα.

1. Προεπεξεργασία/καθάρισμα
Αυτό το βήμα είναι ***προαιρετικό*** και ενδέχεται να βοηθήσει την απόδοση. Εδώ θα ακολουθήσετε τις οδηγίες και τις ιδέες που έχουμε αναλύσει στα φροντιστήρια (αφαίρεση των γραμμών που έχουν null τιμές, αφαίρεση σημείων στίξης, μετατροπή σε πεζά γράμματα). Συμβουλεύουμε να πειραματιστείτε σε αυτό το βήμα.

2. Μελέτη των δεδομένων.
Τα δεδομένα που σας δίνονται περιέχουν 4 στήλες (title, text, subject,date). Καλείστε να εξερευνήσετε τα δεδομένα απαντώντας στα παρακάτω ερωτήματα.

α. Ο τίτλος καθορίζει αρκετά το περιεχόμενο ενός άρθρου. Προσπαθήστε να οπτικοποιήστε τους τίτλους των fake και των non-fake news έτσι ώστε να μπορεί κάποιος να εντοπίσει τα κεντρικά θέματα τα οποία διαπραγματεύονται τα άρθρα.
β. Σχεδιάστε γραφήματα που να δείχνουν το μέσο όρο των χαρακτήρων στα fake και στα true news αντίστοιχα. Κάντε το ίδιο και για τη στήλη text.
γ. Σχεδιάστε την κατανομή του αριθμού των λέξεων για τη στήλη title και για τη στήλη text (και για τα δύο αρχεία)
δ. Επαναλλάβετε το προηγούμενο ερώτημα αφαιρώντας τα stopwords.

ε. Ποιά είναι τα πιο συχνά (πχ 10 ή 20) bigrams στους τίτλους και ποιά στο κυρίως άρθρο;

3. Δημιουργία συνόλου εκμάθησης και δοκιμής

Σε ένα δικό σας ξεχωριστό αρχείο θα επιλέξετε ένα υποσύλο απο τα fake και ένα από τα true, το οποίο θα το ονομάσετε **train.csv** και τα υπόλοιπα θα τα βάλετε σε άλλο αρχείο, το **test.csv**. Το νέο αρχείο train πρέπει να έχει τις εξής στήλες: title, text, subject,date, label. Η νέα στήλη label θα έχει την τιμή 1 αν το άρθρο προέρχεται από το true.csv και 0 αν προέρχεται απο το fake.csv. Αντίστοιχα θα φτιάξετε και το test.csv. Το train και το test προσπαθήστε να έχουν ίδια ποσόστωση μεταξύ fake και true άρθρων.

Σημείωση
Επειδή τα αρχεία είναι μεγάλα, για τις ανάγκες της εργασίας θα ήταν καλό αρχικά να πειραματιστείτε με ένα υποσύνολο των δεδομένων σας (10% του συνόλου) προτού τρέξετε τους αλγόριθμους σε όλα τα δεδομένα.

2. Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω μεθόδους Classification:

- Logistic Regression
- Naive Bayes
- Support Vector Machines (SVM, να πειραματιστείτε με τις παραμέτρους kernel (rbf, linear), c και gamma. Η επιλογή των παραμέτρων μπορεί να γίνει και με GridSearchCV)
- Random Forests

Η κατηγοριοποίηση να γίνει στις εξής διαφορετικές αναπαραστάσεις των κειμένων:

1. Στον αντίστοιχο πίνακα document-words που θα προκύψει από την BoW αναπαράσταση των κειμένων τόσο σε απλά counts, όσο και ξεχωριστά στον tf-idf μετασχηματισμό των counts.
2. Στον αντίστοιχο πίνακα document-vectors που θα προκύψει από το word2vec (μπορείτε να δοκιμάσετε και pre-trained embeddings, είτε από το word2vec, είτε glove ή το fast-text)
   Σημείωση
   Θα πρέπει να παράξετε τις αναπαραστάσεις των documents/tweets με βάση τα embeddings των λέξεων που εμφανίζονται σε κάθε ένα (π.χ. Να πάρετε το average των διανυσμάτων).

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου στα test δεδομένα χρησιμοποιώντας τις παρακάτω μετρικές:

- Accuracy
- F1 score

Σχολιάστε τα αποτελέσματα σας, καταγράψτε οτιδήποτε έχει ενδιαφέρον από τα βήματα που ακολουθήσατε και αναφέρετε πιθανούς λόγους αστοχίας.

3. Beat the Benchmark (bonus)

Τέλος θα πρέπει να πειραματιστείτε με όποια μέθοδο Classification θέλετε, κάνοντας οποιαδήποτε άλλη προ-επεξεργασία στα δεδομένα επιθυμείτε με στόχο να ξεπεράσετε όσο περισσότερο μπορείτε την απόδοση σας στο προηγούμενο ερώτημα.