Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2019-2020
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.

Περιγραφή

Η εργασία σχετίζεται με την κατηγοριοποίηση δεδομένων κειμένου από ειδησεογραφικά άρθρα. Το dataset με το οποίο θα ασχοληθείτε περιέχει ειδησεογραφικά αρχεία σε μορφή .txt (στο αρχείο fulltext.zip) τα οποία ανήκουν σε 5 κατηγορίες (business, entertainment, politics, sport, tech).

Θα χρειαστεί να δημιουργήσετε τα παρακάτω αρχεία .TSV (tab separated files), δηλαδή αρχεία στα οποία τα πεδία των εγγραφών είναι διαχωρισμένα με τον χαρακτήρα '\t' (tab). Σε κάθε .txt αρχείο η πρώτη γραμμή είναι ο τίτλος τους άρθρου.

1. train_set.csv (θα είναι το 80% των συνολικών data points): Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας και περιέχει τα εξής πεδία:
   a. Id: Ένας unique αριθμός για το άρθρο
   b. Title: Ο τίτλος του άρθρου
   c. Content: Το περιεχόμενο του άρθρου
   d. Category: Η κατηγορία στην οποία ανήκει το άρθρο
2. test_set.csv (το 20% των data points): Το αρχείο αυτό θα χρησιμοποιηθεί για να κάνετε προβλέψεις για νέα δεδομένα. Περιέχει όλα τα πεδία του αρχείου εκπαίδευσης εκτός από το πεδίο 'Category'. Το πεδίο αυτό θα κληθείτε να το εκτιμήσετε χρησιμοποιώντας αλγόριθμους κατηγοριοποίησης.

1. Δημιουργία WordCloud

Στο σημείο αυτό καλείστε να δημιουργήσετε ένα WordCloud για τις πέντε κατηγορίες άρθρων. Για την δημιουργία ενός WordCloud θα χρησιμοποιείστε το κείμενο από όλα τα

άρθρα κάθε κατηγορίας. Για την δημιουργία του WordCloud μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκης της Python επιθυμείτε.

2. Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω μεθόδους Classification:
● 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
● Naive Bayes
● K-Nearest Neighbor (Δεν θα χρησιμοποιήσετε κάποια υλοποίηση του αλγορίθμου η οποία παρέχεται από βιβλιοθήκη. Η υλοποίηση του αλγορίθμου θα πρέπει να γίνει από εσάς. Στην υλοποίηση του K-Nearest Neighbor να γίνει με Majority Voting η επιλογή του τελικού label.)

Η κατηγοριοποίηση να γίνει στις εξής διαφορετικές αναπαραστάσεις των κειμένων:
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.

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
● ROC plot

3. Beat the Benchmark (bonus)

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.

4. Υλοποίηση Συσταδοποίησης (Clustering)

Σε αυτό το ερώτημα θα πρέπει να υλοποιήσετε clustering στα διάφορα αρχεία κειμένου. Ο αριθμός των clusters για κάθε ερώτημα θα είναι 5. Η συσταδοποίηση θα γίνει με χρήση του αλγορίθμου clustering K-Means. Η συνάρτηση απόστασης η οποία πρέπει να χρησιμοποιηθεί είναι η Cosine Similarity. Ο K-Means θα εφαρμοστεί στα δεδομένα

εκπαίδευσης (training set). Το clustering θα πρέπει να υλοποιηθεί χωρίς να χρησιμοποιήσει η μεταβλητή Category. Η συσταδοποίηση να γίνει στις εξής διαφορετικές αναπαραστάσεις των κειμένων:

- 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)
- Στον αντίστοιχο πίνακα document-embeddings που θα προκύψει χρησιμοποιώντας pre-trained embeddings (ένα εκ των word2vec, glove,fast-text).

Επίσης, να οπτικοποιήσετε την κατανομή των κειμένων στον χώρο και να είναι εμφανής η ομάδα (cluster) που ανήκουν καθώς και η πραγματική κλάση (category) τους. Για να μπορέσετε να προβάλετε τα σημεία σε 2d χώρο, χρησιμοποιήστε μια μέθοδο συμπίεσης εκ των Principal Component Analysis (PCA), Singular Value Decomposition (SVD) ή Independent Component Analysis (ICA) (αν χρησιμοποιήσετε και τις 3 έχετε bonus). Τη μέθοδο συμπίεσης θα την εφαρμόσετε και στις δύο αναπαραστάσεις που αναφέρονται προηγουμένως.