Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2017-2018

1η Άσκηση, Ημερομηνία παράδοσης: 27/04/2018
Ομαδική Εργασία (2 Ατόμων)

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:

- Anaconda [*Python 2.7 version*]
- *Scikit-learn*
- *Pandas*
- NumPy

Περιγραφή Δεδομένων

Η εργασία σχετίζεται με την κατηγοριοποίηση δεδομένων κειμένου από ειδησεογραφικά άρθρα. Τα datasets είναι αρχεία .tsv (tab seperated files), δηλαδή αρχεία στα οποία τα πεδία των εγγραφών είναι διαχωρισμένα με τον χαρακτήρα '\t' (tab). Περιέχονται δυο αρχεία:

1. train_set.csv (*12267* data points): Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας και περιέχει τα εξής πεδία*:
    a. **Id**: Ένας μοναδικό αναγνωριστικό για το άρθρο
    b. **Title**: Ο τίτλος του άρθρου
    c. **Content**: Το περιεχόμενο του άρθρου
    d. **Category**: Η κατηγόρια στην όποια ανήκει το άρθρο

    **Το πεδίο RowNum αγνοήστε το*
2. test_set.csv (*3068* 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.

Οι κατηγόριες των άρθρων είναι οι παρακάτω:

Politics	Film	Football	Business	Technology

Μπορείτε να κατεβάσετε τα δεδομένα από το ακόλουθο link.

1. WordCloud

Στο σημείο αυτό καλείστε να δημιουργήσετε ένα WordCloud για τις πέντε κατηγορίες άρθρων (ένα για την κάθε μια δηλαδή). Για την δημιουργία ενός WordCloud θα χρησιμοποιείστε το κείμενο από όλα τα άρθρα κάθε κατηγορίας. Παράδειγμα ενός WordCloud παρουσιάζεται στην ακόλουθη εικόνα. Για την δημιουργία του WordCloud μπορείτε να χρησιμοποιήσετε την βιβλιοθήκη wordcloud από το εξής link.

Hint: Από το κείμενο των άρθρων να έχετε ήδη βγάλει τα stopwords στην δημιουργία του WordCloud.

2. Υλοποίηση Κατηγοριοποίησης (Classification)

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

- Support Vector Machines – (scikit implementation)
- Random Forests – (scikit implementation)
- Multinomial Naive Bayes – (scikit implementation)
- K-Nearest Neighbor – (**δική σας υλοποίηση**)

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου χρησιμοποιώντας 10-fold Cross Validation υπολογίζοντας τις παρακάτω μετρικές:

- Precision – (scikit implementation)
- Recall – (scikit implementation)
- F-Measure – (scikit implementation)
- Accuracy – (scikit implementation)

Χρήσιμες συμβουλές:

1) Κατά την προ-επεξεργασία των δεδομένων θα πρέπει να χρησιμοποιήσετε την τεχνική “Latent Semantic Indexing (LSI)”. Δοκιμάστε διαφορετικό αριθμό από components

κρατώντας σταθερή επιλογή classifier. Παρουσιάστε σε ένα γράφημα το accuracy σε σχέση με τον αριθμό components.

2) Προσπαθήστε να χρησιμοποιήσετε αποτελεσματικά την πληροφορία που δίνει ο Τίτλος.
3) 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.
4) Στο SVM να πειραματιστείτε με τις παραμέτρους **kernel** (rbf, linear), **C** και **gamma**. Η επιλογή των παραμέτρων μπορεί να γίνει και με GridSearchCV.

3. Beat the Benchmark

Τέλος, θα πρέπει να πειραματιστείτε με όποια μέθοδο κατηγοριοποίησης θέλετε, κάνοντας οποιαδήποτε προ-επεξεργασία στα δεδομένα επιθυμείτε με στόχο να ξεπεράσετε όσο περισσότερο μπορείτε την απόδοση σας στο προηγούμενο ερώτημα. Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε. Το report σας να μην ξεπερνάει τις 30 σελίδες.

4. Αρχεία Εξόδου

Ο κώδικας θα πρέπει για τα ερωτήματα που αφορούν το Classification να δημιουργεί τα παρακάτω αρχεία

- EvaluationMetric_10fold.csv
- testSet_categories.csv

Το format του αρχείου EvaluationMetric_10fold.csv φαίνεται παρακάτω:

Statistic Measure	**Naive Bayes**	**Random Forest**	**SVM**	**KNN**	**My Method**
Accuracy					
Precision					
Recall					
F-Measure					

Το format του αρχείου testSet_categories.csv, το οποίο θα περιέχει τις κατηγορίες των άρθρων που δίνονται στο Test set φαίνεται παρακάτω:

ID	Predicted_Category

Για το αρχείο “testSet_categories.csv” θα πρέπει να χρησιμοποιηθεί αυστηρά η παραπάνω μορφοποίηση διαχωρίζοντας τα δυο πεδία με τον χαρακτήρα TAB (‘\t’) και επίσης θα πρέπει στην πρώτη γραμμή να υπάρχουν οι δυο επικεφαλίδες (ID και Predicted_Category) και ακολούθως οι προβλέψεις του μοντέλου σας στις επόμενες γραμμές διευκρινίζοντας το ID του document από το test set και το αντίστοιχο category (π.χ. Politics).

5. Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα Ass1_όνοματεπώνυμο1_ΑΜ1_ονοματεπώνυμο2_ΑΜ2. Ο φάκελος θα περιέχει:

1. ένα κείμενο με τον σχολιασμό στα πειράματα που κάνατε και στις μεθόδους που δοκιμάσατε σε μορφή PDF. Η αναφορά σας θα πρέπει να περιέχει και τους πίνακες με τα αποτελέσματα των αρχείων εξόδου.
2. τα ζητούμενα αρχεία εξόδου.
3. τα αρχεία κώδικα που γράψατε.

Το εκτενές κείμενο που θα παραδώσετε, θα περιέχει την περιγραφή των δοκιμών σας και οτιδήποτε σκεφτείτε για να δείξετε τι δοκιμές κάνατε, για ποιο λόγο έχουν τα συγκεκριμένα αποτελέσματα οι μέθοδοι που επιλέξατε, πως λειτουργούν αυτές οι μέθοδοι και σχολιασμό των αποτελεσμάτων σας. Όλες οι εργασίες θα αξιολογηθούν στη βάση της **σωστής τεκμηρίωσης** και στο βαθμό που υλοποιούν τα ζητούμενα της εργασίας, όχι με βάση την κατάταξη που επιτυγχάνουν στα αποτελέσματα της κατηγορίας των δεδομένων.

Forum Επικοινωνίας

Για συζητήσεις/απορίες σχετικά με την άσκηση, θα χρησιμοποιηθεί το piazza:

- Signup link: piazza.com/uoa.gr/spring2018/11
- Class link: piazza.com/uoa.gr/spring2018/11/home

Χρήσιμα εργαλεία

- SpaCy NLP tool (easier to use)
- NLTK NLP tool
- Jupyter Notebook
- Gensim