

Τεχνικές Εξόρυξης Δεδομένων

Εαρινό Εξάμηνο 2015-2016

1η Άσκηση, Ημερομηνία παράδοσης: TBD
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, προεπεξεργασία / καθαρισμός, μετατροπή, εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού Python με την χρήση του εργαλείου SciKit-Learn.

Περιγραφή

Η εργασία σχετίζεται με την κατηγοριοποίηση δεδομένων κειμένου από ειδησεογραφικά άρθρα. Τα Dataset είναι αρχεία CSV του οποίου τα πεδία είναι διαχωρισμένα με τον χαρακτήρα 't'(TAB). Περιέχονται δυο αρχεία:

1. train_set.csv (12267 στοιχεία): 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 (3068 στοιχεία): 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 κατηγορίες των άρθρων είναι 5 και είναι οι εξής:

- Politics
- Film
- Football
- Business
- Technology

Λήψη Dataset

Για να κατεβάσετε τα dataset θα χρειαστεί να συνδεθείτε στην διεύθυνση <http://195.134.67.98/documents/BigData/TED/> και να εισάγετε τα στοιχεία που θα σας δοθούν στο μάθημα.

Δημιουργία WordCloud

Στο σημείο αυτό καλείστε να δημιουργήσετε ένα Wordcloud για τις πέντε κατηγορίες άρθρων. Για την δημιουργία ενός WordCloud θα χρησιμοποιείτε όλα τα άρθρα κάθε κατηγορίας. Παράδειγμα ενός WordCloud παρουσιάζεται στην ακόλουθη εικόνα. Για την δημιουργία του WordCloud μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκης της Python επιθυμείτε.



Υλοποίηση Συσταδοποίησης (Clustering)

Σε αυτό το ερώτημα θα πρέπει να υλοποιήσετε clustering στα διάφορα αρχεία κειμένου χρησιμοποιώντας τον αλγόριθμο clustering K-Means. Η συνάρτηση απόστασης η οποία πρέπει να χρησιμοποιηθεί είναι η **Cosine Similarity**. Ο αριθμός των clusters για κάθε ερώτημα θα είναι 5. Ο K-Means θα εφαρμοστεί στα δεδομένα εκπαίδευσης (training set). Το clustering θα πρέπει να υλοποιηθεί χωρίς να χρησιμοποιήσει η μεταβλητή category.

Σημείωση: Δεν θα πρέπει να χρησιμοποιήσετε κάποια υλοποίηση του αλγορίθμου η οποία παρέχεται από το Scikit-Learn. Η υλοποίηση του αλγορίθμου θα πρέπει να γίνει από εσάς.

- Στο συγκεκριμένο ερώτημα ο κώδικας σας θα πρέπει να βγάζει σαν έξοδο ένα αρχείο csv με τίτλο: clustering_KMeans.csv
- Το αρχείο αυτό θα περιέχει το ποσοστό των δεδομένων κάθε κατηγορίας μέσα στο cluster. Το format των αρχείων φαίνεται παρακάτω:

	Politics	Business	Football	Film	Technology
Cluster1	0.7	0.1	0.1	0	0
Cluster2					
...					
Cluster5					

Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω 4 μεθόδους Classification:

- NaiveBayes (Multinomial και Binomial)
- K-Nearest Neighbor
- Support Vector Machines (SVM)
- Random Forests

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

- Accuracy
- ROC plot

Hint: Κατά την προ- επεξεργασία των δεδομένων θα πρέπει να χρησιμοποιήσετε την τεχνική “[Latent Semantic Indexing \(LSI\)](#)”.

Hint: Προσπαθήστε να χρησιμοποιήσετε αποτελεσματικά την πληροφορία που δίνει ο τίτλος.

Beat the Benchmark

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. Επίσης μπορείτε να χρησιμοποιήσετε ένα συνδυασμό από classifier.

Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε

Χρήσιμα Tutorials:

- [Text classification with Scikit-Learn](#)
- [Scikit-Learn pipeline example](#)

Χρήσιμα εργαλεία (Πέρα από της ανάγκες της άσκησης)

- [SpaCy NLP tool \(easier to use\)](#)
- [NLTK NLP tool](#)
- [GenSim Text-Mining tool](#)

Αρχεία Εξόδου

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

- EvaluationMetric_10fold.csv
- testSet_categories.csv
- roc_10fold.png

To format των αρχείων EvaluationMetric_10fold.csv φαίνεται παρακάτω:

Statistic Measure	Naive Bayes	KNN	SVM	Random Forest	My Method
Accuracy					
ROC					

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

ID	Predicted Category

Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα

Ass1_όνοματεπώνυμο1_AM1_ονοματεπώνυμο2_AM2. 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 συσταδοποίησης δεδομένων.