# Τεχνικές Εξόρυξης Δεδομένων
# Εαρινό Εξάμηνο 2016-2017

1η Άσκηση, Ημερομηνία παράδοσης: 02/05/2017
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 εργαλίων/βιβλιοθηκών: *jupyter notebook*, *pandas*, *gensim* και *SciKit Learn*.

## Περιγραφή

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 unique αριθμός για το άρθρο
    b. Title: Ο τίτλος του άρθρου
    c. Content: Το περιεχόμενο του άρθρου
    d. Category: Η κατηγόρια στην όποια ανήκει το άρθρο
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 |

## Λήψη Dataset

Για να κατεβάσετε τα datasets θα χρειαστεί να συνδεθείτε στην διεύθυνση http://195.134.67.98/documents/BigData/Datasets-2016.zip και να εισάγετε τα στοιχεία που θα σας δοθούν στο μάθημα.

## Forum Επικοινωνίας

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

- Signup link: piazza.com/uoa.gr/spring2017/11
- Class link: piazza.com/uoa.gr/spring2017/11/home

## 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. Παράδειγμα ενός WordCloud παρουσιάζεται στην ακόλουθη εικόνα. Για την δημιουργία του WordCloud μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκης της Python επιθυμείτε.


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

## 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: Αν χρησιμοποιήσετε κάποια υλοποίηση έτοιμη για τον K-Means θα πρέπει να αναφέρετε την βιβλιοθήκη/πηγή που χρησιμοποιήσατε.

- 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 | 0.05 | 0.05 |
| Cluster2 | | | | | |
| Cluster3 | | | | | |
| Cluster4 | | | | | |
| Cluster5 | | | | | |

**Υλοποίηση Κατηγοριοποίησης (Classification)**

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

- Support Vector Machines (SVM)
- Random Forests
- Naive Bayes
- K-Nearest Neighbor (**δική σας υλοποίηση**)

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
- AUC
- ROC plot

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**.

**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.
Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε. Το report σας να μην ξεπερνάει τις 30 σελίδες.

**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

Το format των αρχείων EvaluationMetric_10fold.csv φαίνεται παρακάτω:

| Statistic Measure | **Naive Bayes** | **Random Forest** | **SVM** | **KNN** |
|---|---|---|---|---|
| Accuracy | | | | |
| Precision | | | | |
| Recall | | | | |
| F-Measure | | | | |
| AUC | | | | |

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 (Test_Document_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.

**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 συσταδοποίησης δεδομένων.