Data Mining Techniques
Spring Semester 2016-2017
1st Exercise, Delivery Date: 02/05/2017
Group Work (2 People)

The objective of the task

The purpose of the work is to familiarize you with the basic stages of the process that you are followed for the application of data mining techniques, namely: collection, pre-treatment / purification, conversion, application of data mining techniques and evaluation. Implementation will be done in the Python programming language using them tools / libraries: jupyter notebook, pandas, gensim and SciKit Learn.

Description

The task is related to the categorization of text data from news articles.
The datasets are .TSV (tab seperated files), ie files in which the fields of records are separated by the '\t' (tab) character. Two files are included:

1. train_set.csv (12267 data points): This file will be used to train your algorithms and include the following fields:

    A. Id: A unique number for the article
    B. Title of article
    C. Content: The content of the article

D. Category: The category to which the article belongs

2. test_set.csv (3068 data points): This file will be used to do forecasts for new data. Contains all fields of the training file off from the 'Category' field. This field will be called to appreciate using it classification algorithms.

The article categories are as follows:
Politics
Film
Football
Business
Technology

Download Dataset

To download the datasets you will need to log in to the address http://195.134.67.98/documents/BigData/Datasets-2016.zip and enter the information you are given the lesson.

Communication Forum

For discussions / queries about the exercise, the piazza will be used:
● Signup link: piazza.com/uoa.gr/spring2017/11

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

Creating WordCloud
At this point you are invited to create a WordCloud for the five categories of articles.
To create a WordCloud you will use the text of all the articles each category. An example of a WordCloud is shown in the following illustration. For creating WordCloud you can use any Python library you wish.
Hint: From the text of the articles you already have the stopwords in its creation wordCloud.

Implementation of Clustering
In this query you will need to implement clustering in the various text files using the K-Means clustering algorithm. The distance function that must to use is Cosine Similarity. The number of clusters for each query will be 5. The K -Means will be applied to the training set. Clustering should to be implemented without using the Category variable.

Note: If you use an implementation ready for K-Means you should report the library / source you used.

- In your query, your code should be output an csv file named: clustering_KMeans.csv
- This file will contain the percentage of data in each category within cluster. The format of the files is shown below:

	Business	Football	Technology	Film	Politics
Cluster1	0.7	0.1	0.1	0.05	0.05
Cluster2					
Cluster3					
Cluster4					
Cluster5					

Classification Implementation
In this question you should try the following Classification methods:

- Support Vector Machines (SVM)
- Random Forests
- Naive Bayes
- K-Nearest Neighbor (your own implementation)

You should also evaluate and record the performance of each method using 10-fold Cross Validation using the following metrics:

- Precision / Recall / F-Measure
- Accuracy

● AUC
● ROC plot

Helpful tips :

1) You should use the technique when pre-processing the data "Latent Semantic Indexing (LSI)". Try a different number of components holding a stable classifier option. Show on a chart the accuracy in relative to the number of components.

2) Try to effectively use the information given by the Title.

3) K-Nearest Neighbor: You will not use an implementation of the algorithm h provided by a library. The implementation of the algorithm should be done by you. In the K-Nearest Neighbor implementation, make the Majority Voting option of the final label.

4) In SVM, experiment with the kernel (rbf, linear), c, and gamma parameters. Parameter selection can also be done with GridSearchCV.

Beat the Benchmark (bonus)

Finally, you should experiment with any method of Classification you want by doing any pre-processing to the data you want in order to overcome as much more you can get your performance on the previous question.You should analyze the steps you have taken. Report your not to
Exceeds 30 pages.