PYTHON PROGRAMMING FOR DATA SCIENCE & TEXT MINING (TEC 640) FINAL PROJECT

Deadline February 28th, 2023

FIRST PART

1. Assignment Description

This assessment evaluates your ability to classify text data from a given corpus. You need to implement a Python program using the modules you learned in the course in order to perform a series of actions. You should deliver the source code along with a report that explains clearly the steps taken to answer each question in the form of a Jupyter notebook. This assignment corresponds to 40% of your final grade.

2. Assessment Criteria & Marking Scheme

The assessment consists of 14 actions that will be evaluated with 30 points. In addition, the notebook's structure, aesthetics, and content will also be evaluated with 10 points.

The perfect score for this assignment is 40 points.

3. Classification

Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often helpful in generating a vast amount of sentiment data upon analysis. This data is useful in understanding people's opinions about various topics.

The given dataset (Tweets.csv) can be used for a sentiment analysis task about the problems of each major U.S. airline. Twitter data was scraped from February 2015, and contributors were asked first to classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). The aim of this assignment is to execute a series of tasks for classification using Scikit-learn and Tweets.csv. You don't need to obtain a high-performance model but to demonstrate your capacity to implement the necessary steps in Python.

You need to perform the following actions in your Jupyter notebook:

1. Load the Tweets.csv file and print the first 10 samples. (1 point)
2. Show which column will be used as input for the classification. (1 point)
    - **Hint:** Show a few examples from the column.
3. Show which column will be used as output for the classification. (1 point)
    - **Hint:** Show a few examples from the column.
4. Transform the values of the output columns from categorical to numerical ones. (2 points)
    - **Hint**: You need to transform neutral=1, positive=2, and negative=3.
5. Perform an exploratory data analysis on the data and create at least five plots of your choice. (5 points)
6. Split the data into a training and a test set. (1 point)
7. Vectorize the training and test sets. (2 points)
8. Create a Decision Tree classifier and train it with the training data. (2 points)
9. Calculate the accuracy of the classifier using the test data. (1 point)
10. Calculate the F-score of the classifier using the test data. (1 point)
11. Repeat steps 8 and 9 by changing the hyperparameters of the classifier. (5 points)
    - Create 3 different combinations for the hyperparameters.
    - **Hint:** Read what hyperparameters can be changed here
        - *https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html*
12. Repeat steps 8 and 10 using the Support Vector Machines classifier. (3 points)
    - **Hint:** Read more information here
        - *https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html*
13. Create the ROC curves of the two classifiers together. (2 points)
14. Elaborate on which of the models you select at the end. (3 points)

4. Hints

Consider the following tips:

1. The *spam-detection.ipynb* and *topic-classification.ipynb* notebooks can be used as an inspiration.
2. Don't forget to use diverse markdown tags in your notebook.

SECOND PART

1. Assignment Description

This assessment evaluates your ability to answer a series of questions on Text Mining. The assignment consists of twenty general questions covering major aspects of the course. You should deliver your answers in the provided Jupyter notebook. This assignment corresponds to 20% of your final grade.

2. Assessment Criteria & Marking Scheme

Each question of the assessment carries equal points (1 point). You will need to provide your answers in the designated area of the provided Jupyter notebook (quiz.ipynb). Your answers should be precise and to the point. Avoid too lengthy responses.

The perfect score for this assignment is 20 points.

3. Deliverables & Deadline

You must submit a link to the GitHub repository to Google Classroom before the deadline on the 28th of February (23:59 pm, Geneva time). Of course, you are welcome to contact me in advance for any questions or feedback.

Good luck!

Dr. Nikos Tsourakis