**Information Retrieval and Web Analytics**

**Semester 2 - 2020**

**Assignment 3**

Programming Exercise – Python

Weight – 20%

Due Date – 2nd of Oct, 2020.

This is a group assignment. Student should form a group having maximum of 3 students. This assignment is all about application of Text Clustering/Categorization/Classification. Some of the interesting works that groups may undertake but not limited are

* Sentiment Analysis using Twitter data
* Classification of news articles
* Classification of Movie Data using the IMDB reviews
* Spam Identification
* Insight Analysis
* Restaurant/Hotel classification based on customer reviews
* Identification of fake news, hate speech, etc.

Each group should select any machine learning data sets that are freely available in the data science websites or any other useful datasets. Some of the popular machine learning datasets are given below. There are some links given below which shows the previous work done in the are

The group should perform the following activities using the selected datasets.

1. Study the selected data sets.
2. Select a topic
3. Discuss with the lecturer about your work related to selected data sets and your approach to build models.
4. Clean the datasets such as removing any special characters, html tags, etc
5. Tokenize the corpus
6. Apply the Normalization, stemming, Lemmatization for the tokens in order to reduce the term vocabulary (feature selection)
7. Vectorize the tokens
8. Convert the tokens into tf-idf weights then normalize the term vectors.
9. Separate the datasets as Train and Test dataset
10. Build Naïve Bayes model using the Train data
11. Test the model using Test Data
12. Calculate the confusion matrix and identify the model performance.
13. Build K-NN model using the Train data and a random K value
14. Test the model using Test Data
15. Calculate the confusion matrix and identify the model performance
16. Select the best value for K.
17. Build SVM model using Train Data
18. Test the model using the test data
19. Calculate the confusion matrix and identify the model performance.
20. Compare the model performance and select a best model and briefly discuss why this is the best model
21. How can you improve the performance of your models?

**Important**

1. All coding should be done in Python 3.
2. Proper comments should be written for all code snippets.
3. Working python code should be submitted as Jupyter Notebook file.
4. Code snippets should not be copied from Web and from other groups. Code snippets in the web should be used as reference but should be modified and used as per your requirements.

**Links for machine learning datasets**

1. The Best 25 Datasets for Natural Language Processing. <https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/>

# 50 free Machine Learning datasets: Sentiment Analysis. <https://blog.cambridgespark.com/50-free-machine-learning-datasets-sentiment-analysis-b9388f79c124>

1. 10 text mining datasets available on data.world. <https://data.world/datasets/text-mining>
2. 19 Free Public Data Sets for Your Data Science Project. <https://www.springboard.com/blog/free-public-data-sets-data-science-project/>

**Links containing sample python code and explanations for some topics**

<http://disi.unitn.it/moschitti/corpora.htm>

<https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/>

<https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/>

<https://www.geeksforgeeks.org/python-nlp-analysis-of-restaurant-reviews/>

<https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/>

<https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/>

<https://medium.com/analytics-and-data/leveraging-facebook-python-api-for-marketing-analytics-f4372f042112>

<https://www.mattoncode.pl/facebook-data-analysis-using-python-part-1/>

<https://www.oreilly.com/library/view/mining-the-social/9781449368180/>

<http://blog.chapagain.com.np/python-nltk-sentiment-analysis-on-movie-reviews-natural-language-processing-nlp/>

<https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews>

<https://www.kaggle.com/anthonyc1/fake-news-classifier-final-project>

<https://towardsdatascience.com/how-to-build-a-movie-recommender-system-in-python-using-lightfm-8fa49d7cbe3b>