**CS5590 APL - Python Programming**

**LAB2**

**Deadline: 3/13/2019**

The following assignment focus on to make one familiar with Python and Machine Learning

\*\*\*\*\***Do not use any datasets that you have already worked on and the dataset chosen should be unique for each question\*\*\*\*\***

1. Pick any dataset from the dataset sheet in the class sheet or online which includes both numeric and non-numeric features

a. Perform exploratory data analysis on the data set (like Handling null values, removing the features not correlated to the target class, encoding the categorical features, …)

b. Apply the three classification algorithms Naïve Baye’s, SVM and KNN on the chosen data set and report which classifier gives better result.

2. Choose any dataset of your choice. Apply K-means on the dataset and visualize the clusters using matplotlib or seaborn.

a. Report which K is the best using the elbow method.

b. Evaluate with silhouette score or other scores relevant for unsupervised approaches (before applying clustering clean the data set with the EDA learned in the class)

3. Write a program in which take an Input file, use the simple approach below to summarize a text file:

Link to input file: <https://umkc.box.com/s/7by0f4540cdbdp3pm60h5fxxffefsvrw>

a. Read the data from a file

b. Tokenize the text into words and apply lemmatization technique on each word.

c. Find all the trigrams for the words.

d. Extract the top 10 of the most repeated trigrams based on their count.

e. Go through the text in the file

f. Find all the sentences with the most repeated tri-grams

g. Extract those sentences and concatenate

h. Print the concatenated result

4. Create Multiple Regression by choosing a dataset of your choice (again before evaluating, clean the data set with the EDA learned in the class). Evaluate the model using RMSE and R2 and also report if you saw any improvement before and after the EDA.

**LAB Submission Guidelines (for both In Class and Online students):**

1. LAB submission must be in a group of three students.

2. Submit your source code and documentation to GitHub and represent the work through wiki page properly (submit your screenshots as well. The screenshot should have both the code and the output)

3. Comment your code appropriately

4. Video Submission (2 – 3 min video showing the demo of the LAB, with a brief voice over on the code explanation)

5. Submit **only** report at Turnitin in UMKC blackboard

6. Remember that similarity score should be less than **15%**

7. Use this link to submit your LAB#:

<https://docs.google.com/forms/d/e/1FAIpQLSfHq5Sw1Q75lwzo7WSwieKwmX0lLComYexOFDjFVc_WO8drXQ/viewform>

8. The report should include below details

I. Introduction

II. Objectives

III. Approaches/Methods

IV. Workflow

V. Datasets (if applicable)

VI. Parameters

VII. Evaluation & Discussion

VIII. Conclusion

**LAB Evaluation Criteria:**

1. Report similarly score (should be less than **15%**)

2. Report Quality (check the below example reports for reference)

3. Time (should submit before due time)

4. Wiki page

**Example Reports:**

<https://github.com/stratospark/food-101-keras>

<https://github.com/matterport/Mask_RCNN>

<http://blog.stratospark.com/deep-learning-applied-food-classification-deep-learning-keras.html>

**University Policy on Student Conduct:**

Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: <https://catalog.umkc.edu/special-notices/academic-honesty/>

<https://www.umsystem.edu/ums/rules/collected_rules/programs/ch200/200.010_standard_of_conduct>