COMP 3610 – BIG DATA ANALYTICS
The University of the West Indies, St. Augustine
Assignment 4

Date Distributed: April 13th, 2020 **Date Due:** April 29th, 2020 (11:59pm)

Section 1: Text Analysis

Warm Up (20 marks)

You are given the following list of "Documents" where there exist only 3 words: "Apple", "Orange", and "Banana". Every sentence or document is made up of these words and they become the basis of a 3 dimensional vector space. The "sentence" or "document" is simply a linear combination of these vectors where the number of appearances of the words is the coefficient along that dimension:

```
corpus = ['Apple Orange Orange Apple', # [ 2.  0.  2.]
  'Apple Banana Apple Banana', # [ 2.  2.  0.]
  'Banana Apple Banana Banana Banana Apple', # [ 2.  4.  0.]
  'Banana Orange Banana Banana Orange Banana', #[ 0.  4.  2.]
  'Banana Apple Banana Banana Orange Banana'] # [ 1.  4.  1.]
```

Write a Python function (from scratch) which takes the corpus and creates a vector representation of the corpus. The pandas or numpy structure may be used. The comments at the end of each line represent the corresponding vector representation.
The function should output the following based on the corpus given above:

```
[[ 2.  0.  2.]
 [ 2.  2.  0.]
 [ 2.  4.  0.]
 [ 0.  4.  2.]
 [ 1.  4.  1.]]
```

Preprocessing and Data Organization (20 marks)

The dataset can be downloaded here.

Tasks

1. Create a new column in the dataframe called 'sentiment'. Using appropriate existing columns, populate the new column with 0's and 1's where 0 refers to a negative sentiment and 1 refers to a positive sentiment.
2. Clean the subtitles data and store the cleaned text in a new column 'subtitle_clean'.
    i. For each step of your text cleaning give a brief explanation of why you chose to perform that method on the text.
3. Use TFIDFVectorizer and CountVectorizer to encode the cleaned subtitles.

Text Classification (30 marks)

1. When choosing a metric to access the performance of your classifier provide a brief explanation of why you chose that metric.

2. Perform the following classification experiments keeping track of the performance of each classification task for future use:

    i. Logistic regression model on word count
    ii. Logistic regression model on TFIDF
    iii. Logistic regression model on TFIDF + ngram
    iv. Support Vector Machine model on word count
    v. Support Vector Machine model on TFIDF
    vi. Support Vector Machine model on TFIDF + ngram

    You may use the SVM classifier from sklearn

3. Plot a bar graph showing the performance of each of the experiments.

Topic Modeling (20 marks)

1. Using TFIDF and Count Vectorizer models imported for sklearn, perform topic modelling using the following topic modeling algorithms:
    i. NMF
    ii. LDA
    iii. SVD.

2. When choosing the number of topics give a brief explanation of why that number was chosen.

3. Discuss based on the top 10 words each of the algorithms choose for each topic cluster what category the topics fall under.

Visualization (10 marks)

Choose the clusters obtained from a topic model algorithm from above and plot a word cloud for each of the clusters. For example, if the number of topics chosen was 10 and the topics were obtained from the SVD algorithm, 10 word clouds should be plotted.

Section 2: Time Series Analysis

Literature Review (30 marks)

In no more than **500 words** write a literature review of the following paper: Load Forecasting using Deep Neural Networks

Forecasting (40 marks)

Context

To better follow energy consumption, the government wants energy suppliers to install smart meters in every home in England, Wales and Scotland. There are more than 26 million homes for the energy suppliers to get to, with the goal of every home having a smart meter by 2020.

This roll out of meter is lead by the European Union who asked all member governments to look at smart meters as part of measures to upgrade our energy supply and tackle climate change. After an initial study, the British government decided to adopt smart meters as part of their plan to update our ageing energy system.

For this assignment you will use a refactorised version of this dataset from the London data store, that contains the energy consumption readings for a sample of 5,567 London Households that took part in the UK Power Networks led Low Carbon London project between November 2011 and February 2014. The data from the smart meters seems associated only to the electrical consumption.

There is information on the ACORN classification details that you can find in this report or the website of CACI.

Weather data for London area was added, the data was collected from the darksky api.

Dataset

There are two datasets that will be used for this assignment:

1. Daily Weather data for the period November 2011 until April 2014.
2. Energy Consumption data for an area for the period November 2011 until April 2014.

Tasks

The **"temperatureHigh" and "time"** data from the weather dataset should be used for analysis. The **"energy_sum" and "day"** data from the energy dataset should be used for analysis.

Ensure that there is only one row for a given date.

1. Identify any trends in the datasets and discuss if trends in weather are related to trends in energy consumption.
2. Identify Seasonal patterns in the datasets and discuss if seasonality in weather are related to seasonality in energy consumption.
3. Forecast using simple exponential smoothing the expected **temperature** for 2014, use the RMS metric to indicate the accuracy of your forecast.
    i. Choose the alpha that results in the lowest RMS
    ii. Plot a graph showing the actual value and the forecasted values
4. Forecast using simple exponential smoothing the expected **energy consumption** for 2014, use the RMS metric to indicate the accuracy of your forecast.

   i. Choose the alpha that results in the lowest RMS
   ii. Plot a graph showing the actual value and the forecasted values
5. Combine the initial datasets using an appropriate column as the index. Indicate why you chose that index.
6. Plot separate graphs showing the **"temperatureHigh"** and **"energy_sum"** data from the combined dataset.
7. Discuss what observations can be made from the plot with the combined data vs the plots from the individual datasets.

Prediction (30 marks)

This section relies on the combined dataset from Forecasting question 5.

Tasks

1. Create a new column in the dataset, derived from the **"energy_sum"** values, which indicates if energy consumption was high or low. An appropriate threshold should be used to determine whether or not energy consumption was high or low. Justify the threshold value you chose.
2. Split the dataset into a train and test set. The train set should be all the data before 2014 and the test set should be all data from 2014.
3. Choose the appropriate columns to be used for training. Give a brief explanation why those columns were chosen.
4. Use an appropriate classifier to predict the energy consumption for 2014.
5. Show the performance of the classifier using an appropriate metric.

Submission Details:

1. Ensure that your name and ID and present in the notebooks. Separate notebooks should be used for each section.
2. Export/download your files from Jupyter notebook.
3. Email your files to ssooklal27@gmail.com. Put as the title of the email COMP 3610 A4.
4. There would be a 10% penalty per day for late submissions, up to 5 days.