COMP 6940: Big Data and Visual Analytics - Assignment 2

University of the West Indies, St. Augustine

Due Date: March 15th @ 11:59 PM

1 Text Analysis

The dataset for this section can be downloaded here.
It is based on reviews of 3 Disneyland branches posted by visitors on Trip Advisor.

1.1 Preprocessing, Data Organization and Visualisation (20 marks)

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. [5 marks]
2. Clean the reviews content data and store the cleaned text in a new column 'review_content_clean'. For each step of your text cleaning give a brief explanation of why you chose to perform that method on the text. [10 marks]
3. Visualise aspects of the data to briefly summarise overall trends. [5 marks]

1.2 Text Classification (30 marks)

1. Select a metric to access the performance of your classifier and provide a brief explanation of why you chose that metric. [5 marks]
2. Perform the following classification experiments keeping track of the performance of each classification task for future use: [20 marks]
   (a) Logistic regression model on word count
   (b) Logistic regression model on TFIDF
   (c) Logistic regression model on TFIDF + ngram
   (d) Support Vector Machine model on word count
   (e) Support Vector Machine model on TFIDF
   (f) 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. [5 marks]

1.3 Topic Modeling (20 marks)

1. Using TFIDF and Count Vectorizer models imported for sklearn, perform topic modelling using the following topic modeling algorithms: [10 marks]
   (a) NMF
   (b) LDA
   (c) SVD
2. When choosing the number of topics give a brief explanation of why that number was chosen. [5 marks]
3. Discuss based on the top 10 words each of the algorithms choose for each topic cluster what category the topics fall under. [5 marks]