MSBA7012 Individual Assignment

Deadline: Sunday, February 28, 2021 11:59pm

Datasets:

- FBPosts.csv contains all posts submitted by the official Facebook page accounts for 182 movies released in the United States in 2012. The content of each post is stored in the "message_and_description" column.
- Bing Liu's Opinion Lexicon: negative-words.txt and positive-words.txt

Questions:

1. Zipf's law states that the frequency of a word appearing in a large text corpus is inversely proportional to its rank. Make a plot in Python to illustrate the Zipf's law using the words in all Facebook posts in the FBPosts.csv file. The x-axis of the plot is the rank of a word and the y-axis is the frequency of a word. Word frequency is defined as the number of times a word appear in all posts. The number of distinct words may be large; you can consider the top 1,000 words only. Based on the plot you create, discuss whether the Zipf's law is supported by this dataset and explain why it is supported or not. Limit your answer to 100 words. (5 marks)
2. In Python, visualize the top 15 words with the highest tf-idf score for each of the following 4 movies: Avengers (imdb_id= tt0848228), The Dark Knight Rises (tt1345836), The Hunger Games (tt1392170), and The Twilight Saga (tt1673434). Briefly summarize the insights you gain from this analysis. Limit your answer to 100 words. (5 marks)
3. In Python, visualize the top 15 bigrams with the highest tf-idf score for each of the same 4 movies in Question #2. Compare the results you obtain for Questions #2 and #3 and comment on what additional insights you have gained from analyzing the bigrams in addition to the unigrams. Limit your answer to 100 words. (5 marks)
4. Identify the top 20 most common positive and negative words based on Bing Liu's opinion lexicon in all page posts and visualize the word frequencies in a bar chart (one for top 20 positive words and one for top 20 negative words). (5 marks)
5. Does the sentiment of Facebook page posts help predict the opening box office revenue? Interpret the economic significance of your result and explain why the sentiment of Facebook page posts helps or does not help predict the opening box office revenue. You may define sentiment in the following three ways: (1) fraction of positive words, (2) fraction of negative words, and (3) fraction of positive words - fraction of negative words. Feel free to use any analytics techniques (e.g., visualization, regression, machine learning, etc.) to provide an answer to this question. Since it is a prediction problem, you should only utilize the posts created before each movie's release date. Limit your answer to one A4 page, including any text summary, figures, or tables. (10 marks)

Deliverables:

- A Word document (.docx) containing all the answers including plots or figures for the first 4 questions and a one-page writing for your answer to the last question.
- Source code of your programs for all questions in one file (either .py or .ipynb). Add comments to your code to improve readability. Make sure the grader can easily identify the source code for each of the five questions.
- A readme.txt file describing the package/environment requirements to run your programs.
- Compress the above three files into a zip file named with your student ID, e.g., 123456.zip.
- You should not make any modifications to the three input files: FBPosts.csv, negative-words.txt, and positive-words.txt. They are the raw data input to your programs. Also, DO NOT include these three files in your zip file.