Text Mining and Analytics, Spring 2024, Assignment 1

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Due: January 31, 2024

Students in the Computer Science department have decided to organize a movie night. Ray, Leiby, and OJ have selected "The Godfather Parts I and II" for the event. Ray, being a fan of these movies, is eager to learn more about them by reviewing their transcripts. He specifically wants to focus on the character "Michael Corleone" and analyze his lines.

Task 1: To help Ray, you will scrape the scripts for both movies. Two links are provided for you (sample code), however, you can use any other websites to get the scripts. Note that you should get the transcripts from online resources.

After scraping the transcripts, you should extract the lines for Michael Corleone and save them into two files, one per movie. The files should be named “GodFather1.txt” and “GodFather2.txt”. Each line from the movie is written as one line in the file.

Task 2: The next step is to identify the common words used by Michael Corleone in each movie. Utilize word clouds to visualize these common words, creating a word cloud per movie. Place these two plots next to each other (horizontally) and analyze whether the common words have changed between the two movies or if they are largely the same. Ensure that your plots are of high quality, especially when zoomed in (learn how to save the plot as a PDF instead of taking screenshots). Provide a 1–2 paragraph analysis related to what can be observed from the plots.

Task 3: In a paragraph, describe your understanding of WordPiece tokenization. Then, for the following line, compare the tokenization with WordPiece and word tokenization. For WordPiece use a tokenizer from HuggingFace (transformer in Python) with “bert-base-cased” model.

Line: Because they know that no Sicilian will refuse a request on his daughter's wedding day.

Notes for submission:

1. A Python file (only .py is acceptable, not .ipynb) named assignment1.py with codes for all three tasks. Codes should be well-structured with comments to run. (Also, submit your .txt files.)

2. A .pdf file having your plots, analysis, and answer to tasks 2 and 3

3. Any assumptions made by students should be explicitly mentioned in the submitted