CS 657 Mining Massive Datasets
Fall 2024

Assignment 1: Modified WordCount/Pairs Counting

Dev Divyendh Dhinakaran G01450299
Tejaswi Samineni G01460925

README File

In this folder, we have the following:

- **Code**: Contains the Python code of the modified WordCount program, specifically sotu_analysis_final.py.
- **Sample Output Files**: Contains six sample output files in .csv format:
  1. Calculating_avg_std.txt
  2. cleaned_speeches.txt
  3. frequentpairs_lift_values.txt
  4. flesch_kincaid_scores.txt
  5. high_lift.txt
  6. spiked_words.txt
- **Report**: The detailed report can be found in Assignment1_Report_G01450299_G01460925.pdf.

Introduction

The purpose of this assignment was to modify the WordCount program using PySpark to analyze the State of the Union speeches dataset. This included cleaning the text, removing punctuation, stopwords, and other unnecessary elements, and performing a detailed analysis on word frequencies, readability scores, and word co-occurrences.

Key Output Files

- **Code**: The Python code implementing the updated WordCount, text cleaning, and analysis, named sotu_analysis_final.py, is located in the Code folder.

- **Sample Output Files**: In the Sample Output Files folder, we provide the following output files, each generated by the program:
    - Calculating_avg_std.txt: Contains the average and standard deviation of word counts within 4-year windows.
    - cleaned_speeches.txt: Shows the cleaned speech data after removing stopwords, punctuation, HTML, and URLs.
    - frequentpairs_lift_values.txt: Includes word pairs and their calculated lift values.
    - flesch_kincaid_scores.txt: Contains the Flesch-Kincaid readability scores for each speech.
    - high_lift.txt: Lists the word pairs with a lift greater than 3.0.
    - spiked_words.txt: Displays the words that showed spikes in frequency in subsequent years.

Results and Visuals

For the average, standard deviation, and outstanding words in the year following a presidential period, we present the results in a clear and legible manner. For example, we have provided a list of the 10 most salient words in each case, highlighting the key trends and spikes. We have also included images in the report to visually represent the data and results.