CSCI 5408

Data Management, Warehousing, And Analytics

Assignment 3 - Problem 2

Build a light-weight analytics engine, that perform custom ETL operation, and specific analysis.

Prepared By

Bhavisha Oza (B00935827)

Problem-2: Sentiment Analysis using BOW model on title of Reuters News Articles

1. Summary:

This Core Java program performs sentiment analysis on news titles using a bag-of-words approach without relying on any additional libraries. In the first step, it creates a bag-of-words representation for each news title by counting the occurrences of words using a simple loop-based counter. The bag-of-words is stored as a map, with words as keys and their respective frequencies as values.

In the second step, the program compares the bag-of-words of each title with lists of positive and negative words downloaded from online sources. It calculates an overall score for each title based on the frequency of positive and negative words. Depending on the score, the program tags each title as "positive," "negative," or "neutral," and includes an additional column to present these findings. By avoiding additional libraries and using only Core Java, the program offers an efficient and self-contained solution for sentiment analysis, making it a practical tool for analyzing the sentiment of news titles.

Requirements & Tasks fulfilled:

1. Create a java program which creates a bag-of-words for each News title using regex [4]
2. Downloaded the text file for the list of positive and negative words from google [3,8]

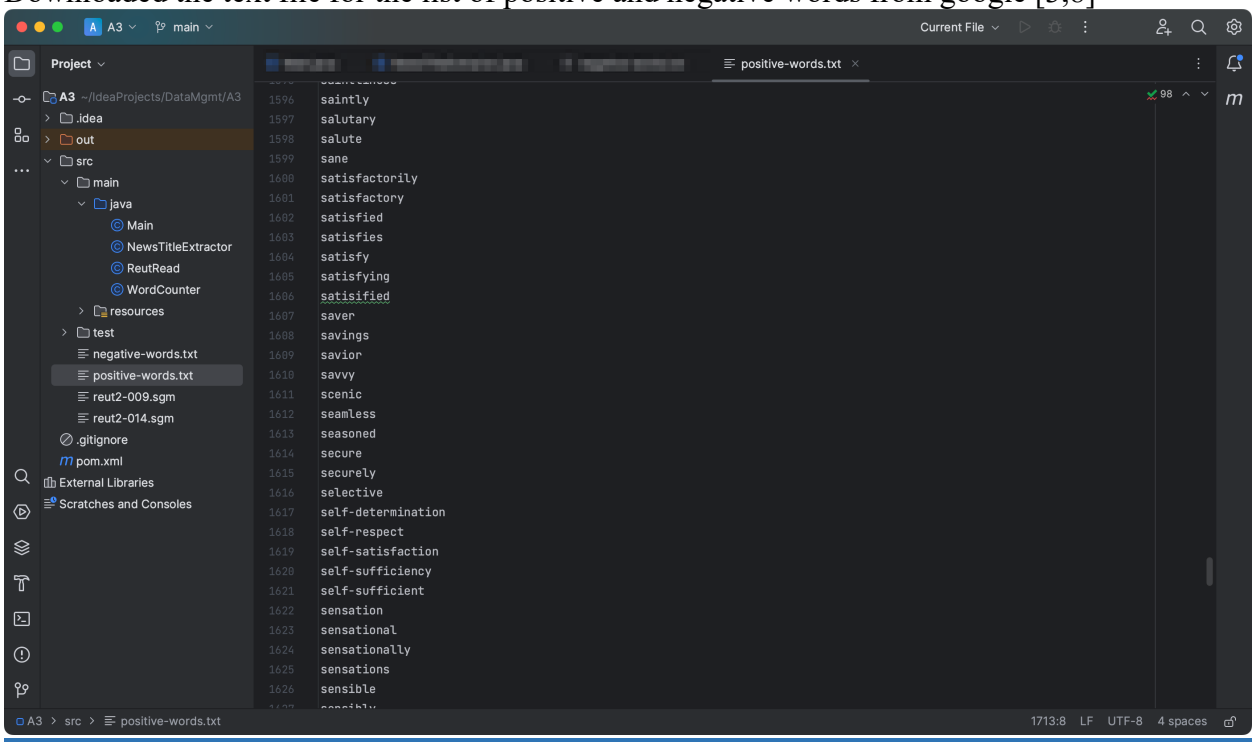

Figure 1: List of positive words

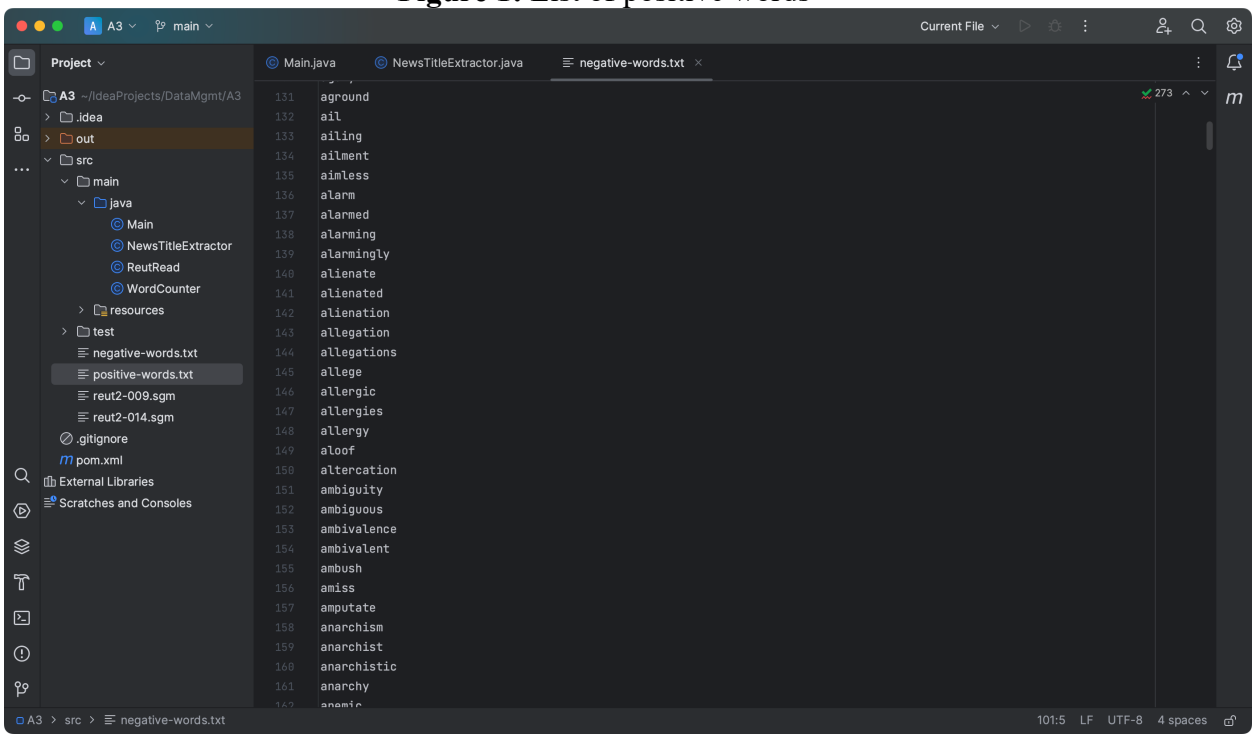

Figure 2: List of negative words

3. Compare each bag-of-words with a list of positive and negative words.
4. Tag each news title as “positive”, “negative”, or “neutral” based on overall score using JTable.

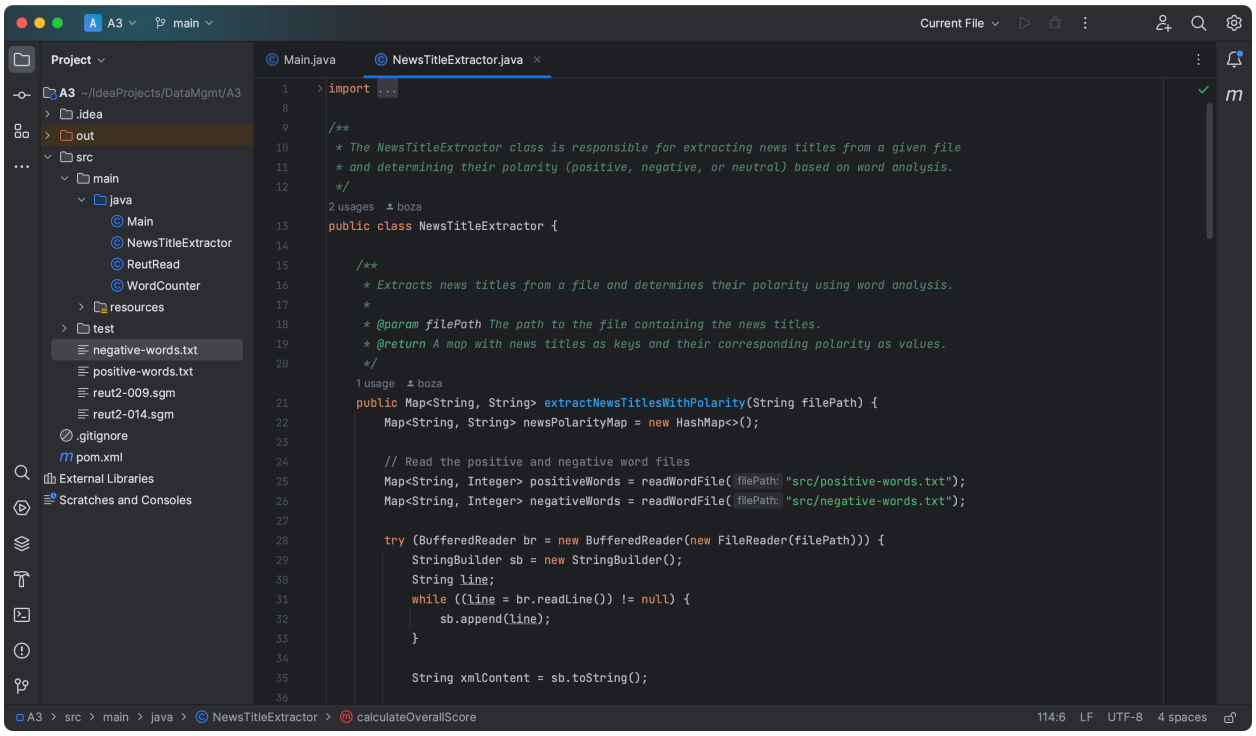

Figure 3: Java program for extracting news titles from a given file and determining polarity.

5. Output of the written program:

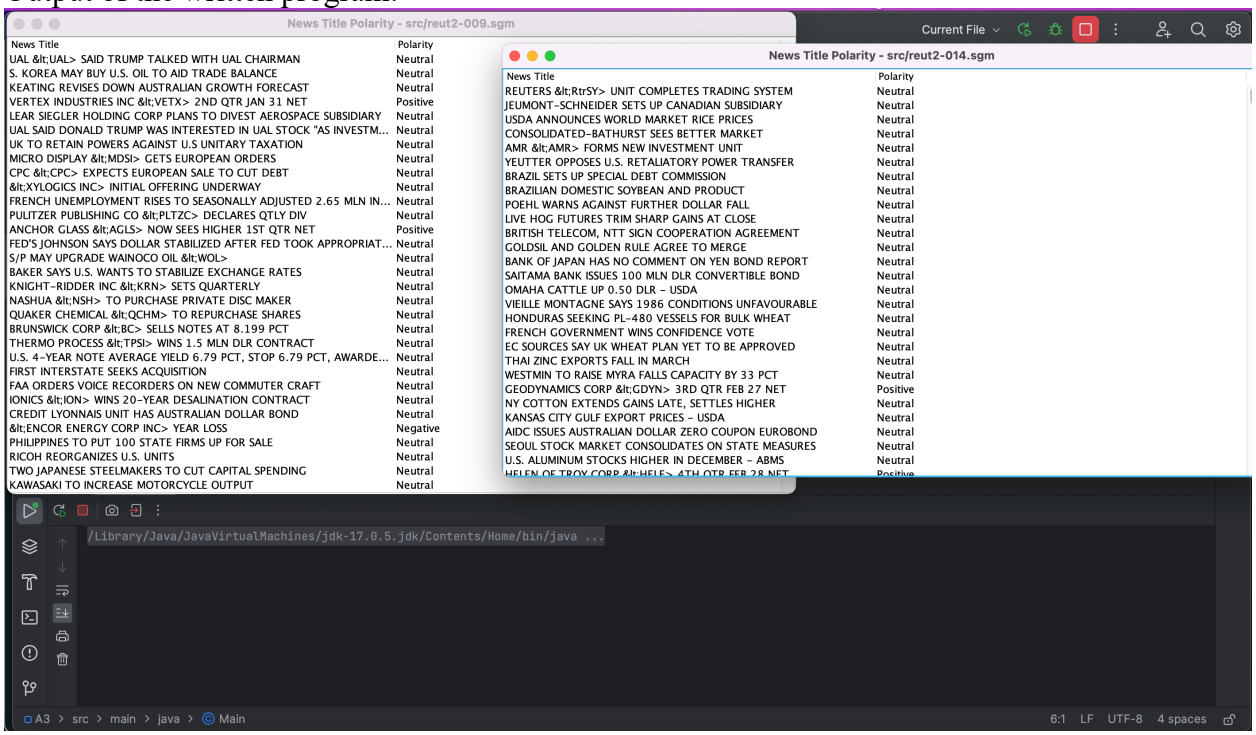

Figure 4: Output of Java program for extracting news titles from a given file and determining polarity.

6. An algorithm of Sentiment Analysis using BOW model on title of Reuters News Articles:

1. Initialize an empty `newsPolarityMap` to store the news titles and their polarities.
2. Read the positive and negative word files and store the words in separate `positiveWords` and `negativeWords` maps, respectively.
3. Open the input file specified by `filePath` and read its content line by line.
4. For each line of content, append it to a `StringBuilder` to create the `xmlContent` string.
5. Create a regular expression pattern to match the news title within `<TITLE>` and `</TITLE>` tags.
6. Use a matcher to find all occurrences of the news title pattern in the `xmlContent` string.
7. While iterating over the matched titles:
   - Extract the current news title from the group(1) of the matcher.
   - Create an empty `wordCountMap` to store the word frequencies for the current title.
   - Split the title into individual words using whitespace as a delimiter and update the `wordCountMap` with the word frequencies.
8. Calculate the overall score for the current news title by iterating over the entries in the `wordCountMap`:
   - For each word in the title:
     - Retrieve the positive and negative word frequencies from the `positiveWords` and `negativeWords` maps.
     - Calculate the word's contribution to the overall score as (positiveCount - negativeCount) * wordFrequency.
     - Sum up the contributions to get the overall score for the title.
9. Determine the polarity of the current news title based on its overall score:
   - If the overall score is greater than 0, set the polarity as "Positive."
   - If the overall score is less than 0, set the polarity as "Negative."
   - Otherwise, set the polarity as "Neutral."
10. Add the current news title and its polarity to the `newsPolarityMap`.
11. Continue to the next matched title and repeat steps 7 to 10 until all titles have been processed.
12. Return the `newsPolarityMap` containing news titles as keys and their corresponding polarities as values.

Full code can be found in the given repository: https://git.cs.dal.ca/boza/csci-5408/-/tree/main/A3.

References:

[1] "Download intellij idea – the leading Java and Kotlin IDE," *JetBrains* [Online]. Available: https://www.jetbrains.com/idea/download/?section=mac [Accessed: 01 July 2023].

[2] "The most popular database for modern apps", MongoDB, 2020. [Online]. Available: https://www.mongodb.com/ [Accessed: 20 July 2023].

[3] Positive-words.txt, *Gist*. Available: https://gist.github.com/mkulakowski2/4289437 [Accessed: 31 July 2023].

[4] "Build, test, and debug regex," *regex101* [Online]. Available: https://regex101.com/ [Accessed: 20 July 2023].

[5] "Lab-5," *Brightspace Dalhousie University* [Online]. Available: https://dal.brightspace.com/d2l/le/content/271677/viewContent/3661458/View [Accessed: July 28, 2023].

[6] "Lab-7," *Brightspace Dalhousie University* [Online]. Available: https://dal.brightspace.com/d2l/le/content/271677/viewContent/3681262/View [Accessed: July 28, 2023].

[7] "Flowchart Maker & Online Diagram Software," *Draw.io* [Online]. Available: https://app.diagrams.net [Accessed: 29 July 2023].

[8] Negative-words.txt, *Gist*. Available: https://gist.github.com/mkulakowski2/4289441 [Accessed: 31 July 2023].