Data Management, Warehousing and Analytics

Assignment 3

Name: Yogish Honnadevipura Gopalakrishna
Banner ID : B00928029
GitLab :
https://git.cs.dal.ca/yogish/csci5408_s23_b00928029_yogish_honnadevipura-gopalakrishna.git

Problem 1A

Algorithm

1. Start the program and define the file paths for the two news files, database name, and collection name.
2. Create a MongoDB client connection using the specified URI and database.
3. Get the collection from the database where the news articles will be stored.
4. Define a method parseAndInsertNews(filePath, collection) to parse the news file and insert the news articles into the MongoDB collection.
5. Open the news file for reading using a BufferedReader.
6. Initialize two StringBuilder objects, titleBuilder and textBuilder, to store the title and text of each news article, respectively.
7. Initialize two boolean variables, isReutersTagOpen and isBodyTagOpen, to track the opening and closing tags for Reuters and Body sections, respectively.
8. Loop through each line in the news file until the end of the file is reached.
9. Check if the line contains the opening tag <REUTERS and update the isReutersTagOpen flag accordingly.
10. If the line contains the <TITLE> tag, extract the title text and append it to the titleBuilder.
11. If the line contains both the opening and closing <BODY> tags, extract the text content between the tags and append it to the textBuilder.
12. If the line contains the opening <BODY> tag but not the closing </BODY> tag, set the isBodyTagOpen flag to true and append the text content after the <BODY> tag to the textBuilder.
13. If the line contains the closing </BODY> tag while the isBodyTagOpen flag is true, append the text content before the </BODY> tag to the textBuilder and set the isBodyTagOpen flag to false.
14. If the line contains the closing </REUTERS> tag and the isReutersTagOpen flag is true, it indicates the end of a news article.
15. Check if the titleBuilder and textBuilder are empty, and if they are, set the title and text variables to empty strings. Otherwise, remove unwanted characters from the titleBuilder and textBuilder using the removePattern() method with appropriate regex patterns.
16. Create a new Document with the title and text extracted from the news article.
17. Insert the Document into the MongoDB collection.
18. Clear the titleBuilder and textBuilder for the next news article.
19. Continue to the next line in the news file.
20. Close the BufferedReader.
21. End the program.

Flowchart

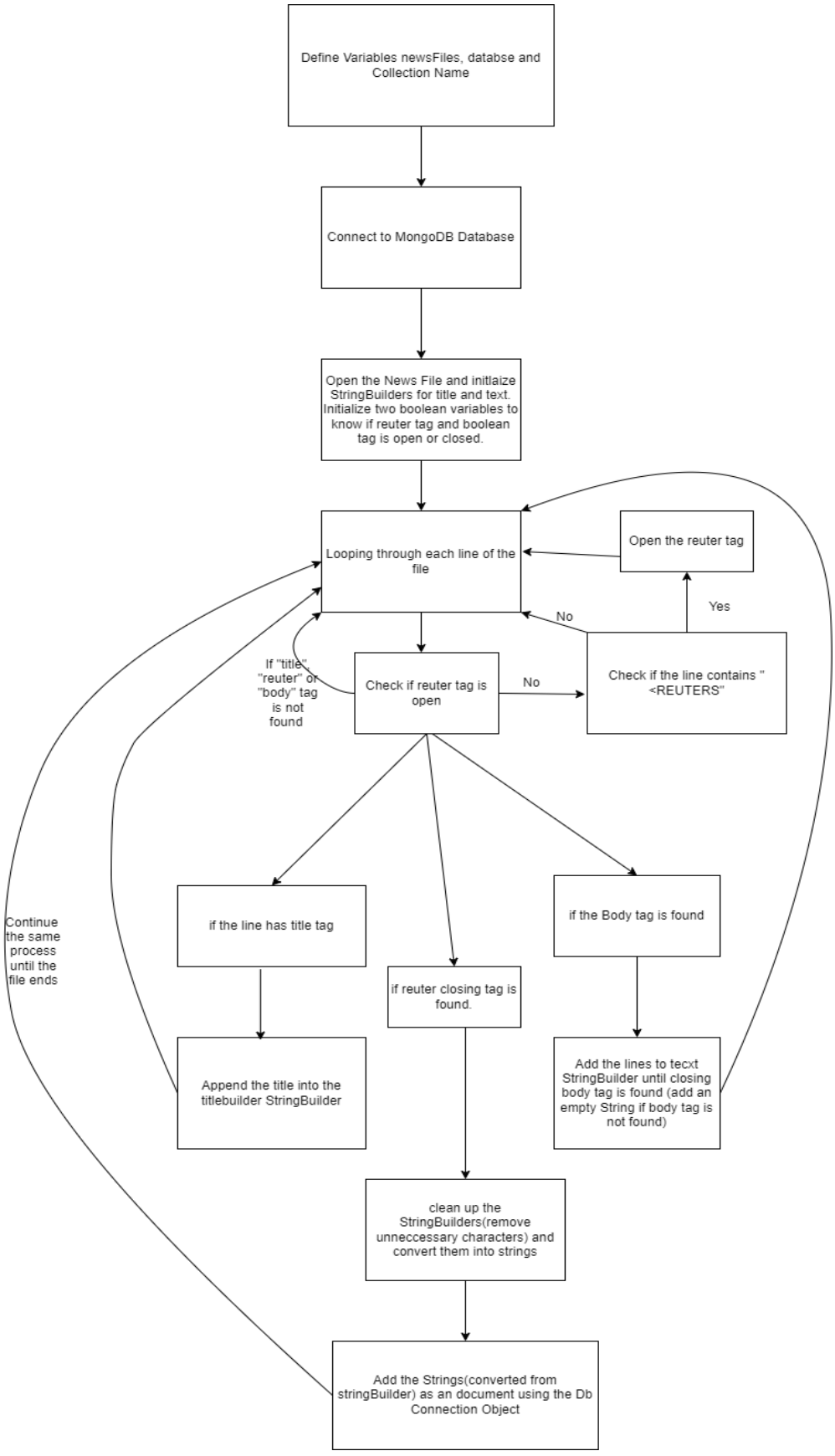

Figure 1.1.1 : Flowchart for Problem 1A

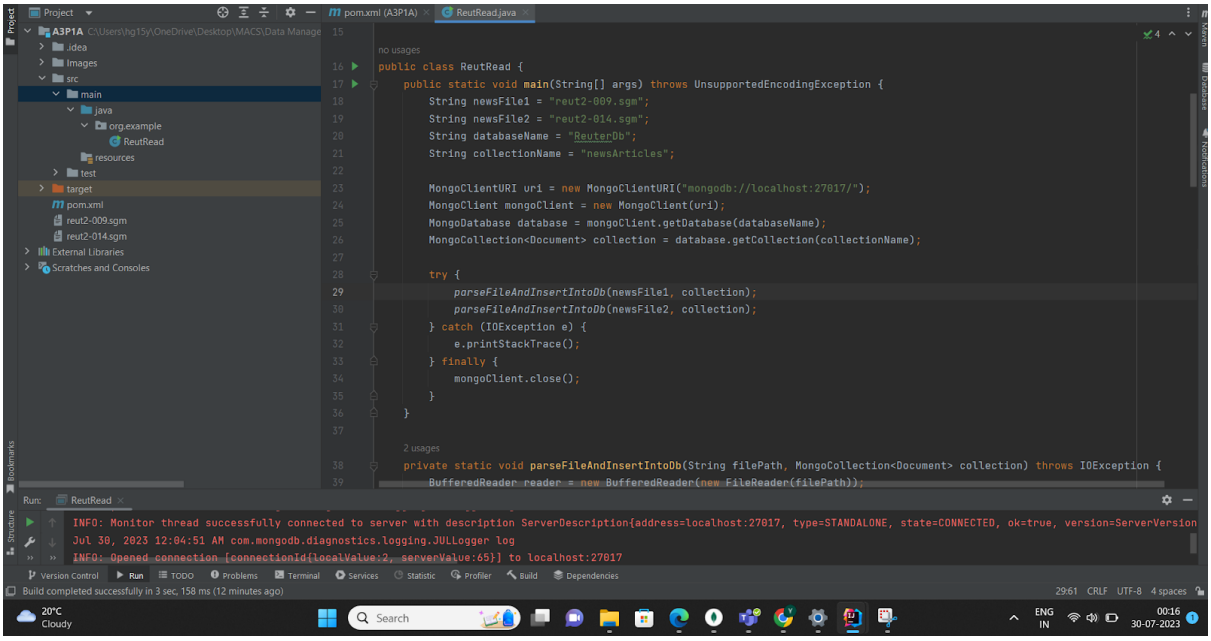

Fig 1.1.2: Program showing the connection to MongoDB localhost

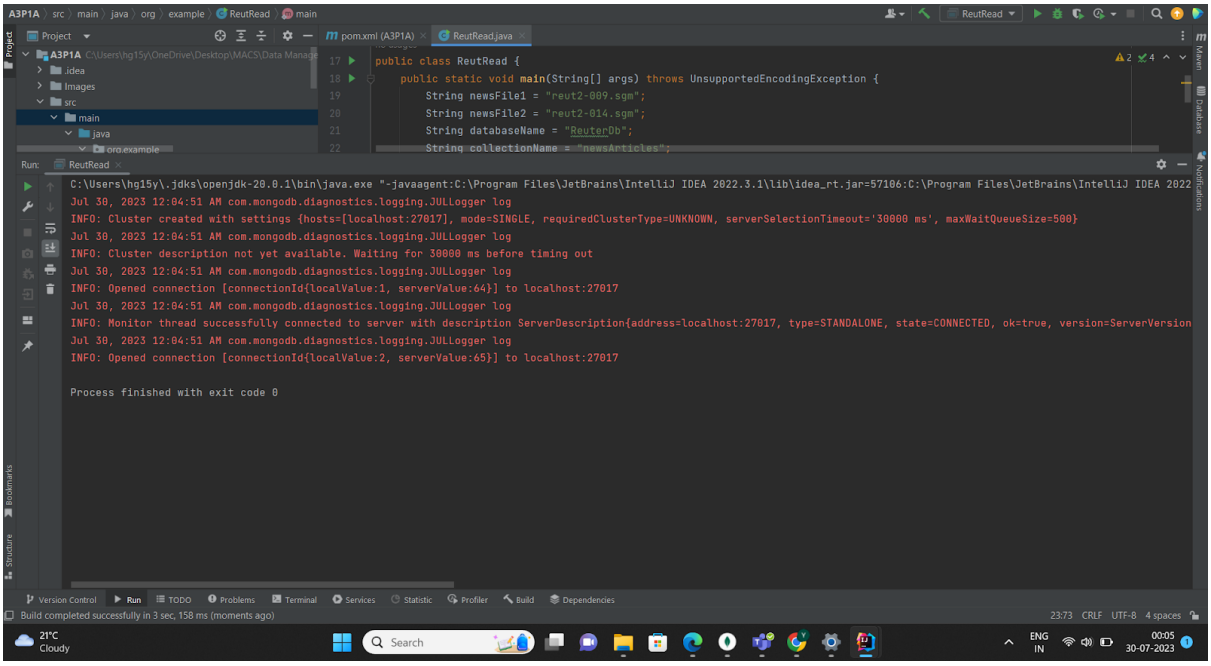

Fig 1.1.3: Logs showing the connection made with MongoDB

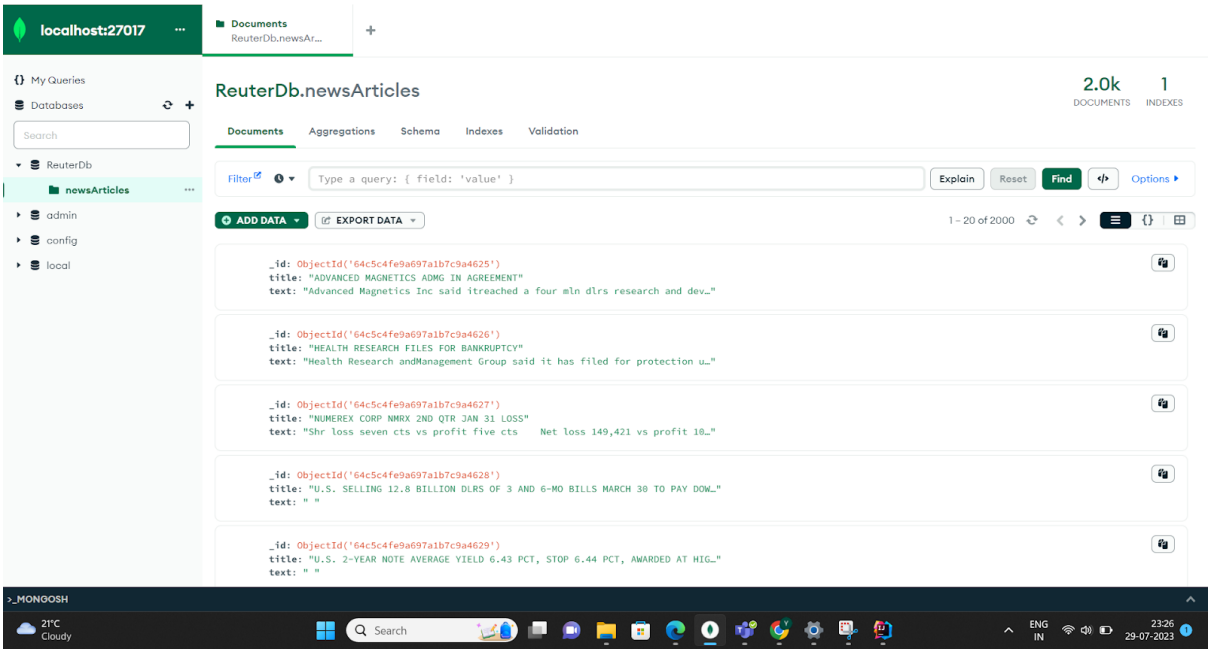

Fig 1.1.4: Output documents generated(2000 documents generated)

Cleaning data process using Regex

removePattern(titleBuilder, "*{6}") : This line removes patterns of six consecutive asterisks ("******") from the titleBuilder. The regular expression "*{6}" matches six occurrences of the asterisk character *.

removePattern(titleBuilder, "<") : This line removes the HTML entity representation of the less-than symbol ("<") from the titleBuilder. The pattern < represents the character "<" in HTML entities.

removePattern(titleBuilder, "[<>]") : This line removes all angle brackets ("<" and ">") from the titleBuilder. The pattern [<>] is a character class that matches any occurrence of the characters "<" or ">".

removePattern(textBuilder, "Reuter&#3;") : The line removePattern(textBuilder, "Reuter&#3;"); will remove occurrences of the pattern "Reuter" from the textBuilder string.

Problem 1B

Creation of GCP instance:

- Login to the console.cloud.google.com
- Search for dataproc in the search bar. It is an apache hadoop cluster
- Selected cluster on computer engine
- Configure the cluster
    - Changed the location to us-east-1 as it is the nearest location
    - Change the configuration to single node
- Create the cluster
- Write the java program of mapreduce
- Generate the Jar file
- Connect to the instance through the SSH terminal
- Upload the JAR file
- Run the hadoop command so that it generates the output

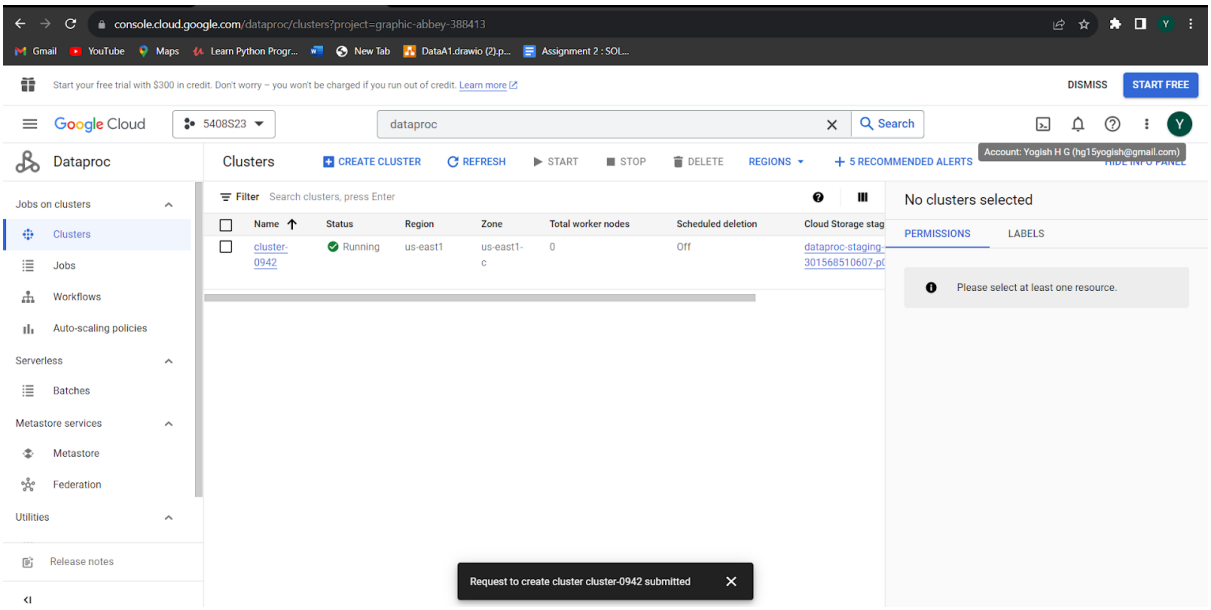

Fig 1.2.1: Creation of GCP Instance

Algorithm of Map reduce function

Mapper Phase:

Each input line is read from the input file as a Text object.
The TokenizerMapper class splits the input line into individual words using the regular expression \\s+, which matches one or more whitespace characters (spaces, tabs, etc.).
For each word in the line, the word is converted into a Text object, and the value 1 is assigned to the one variable, representing the occurrence count of that word.

The word and one pair is emitted as an intermediate key-value pair from the mapper.

Shuffle and Sort Phase:
The MapReduce framework takes care of shuffling and sorting the intermediate key-value pairs outputted by the mappers. This phase groups all occurrences of the same word together and sorts them by key (word) in preparation for the reduce phase.

Reducer Phase:
The IntSumReducer class receives the grouped and sorted key-value pairs from the shuffle and sort phase.
For each unique word (key), the reducer iterates through the list of occurrence counts (values) associated with that word.
It calculates the total count for each word by summing up all the occurrence counts.
The word and its final count are written as the output key-value pair using the context.write() method.

Combiner Optimization:
The job is configured to use the IntSumReducer class as a combiner as well. The combiner performs a local reduce task on the mapper side before sending the data to the reducers. It helps to minimize the data that needs to be transferred over the network and optimize the performance of the job.
In this case, the combiner is the same as the reducer, and it helps to aggregate word counts locally on the mapper side.

Input and Output Paths:
The input path for the MapReduce job is set to /input, indicating the directory where input files are located.
The output path is set to /output, indicating the directory where the results will be stored.

Job Execution:

The main method sets up the MapReduce job by configuring various aspects such as input and output formats, mapper and reducer classes, input and output key-value types, etc.
It then submits the job to the Hadoop cluster for execution and waits for its completion.
The job execution status is returned as an exit code, where 0 indicates success, and 1 indicates failure.

Frequency of Highest words : As with Frequency of 1
Frequency of lowest words : The with frequency of 6183

```
hg15yogish@problem1b-m:~$ hdfs dfs -cat word-frequency-output/part-r-00000
"AS     1
"America's      2
"An     1
"At     2
"B"     12
"Big    1
"Brazil's       1
"Citibank       2
"Day-to-day     1
"Disaster       1
"Every  1
"Financial      1
"First  1
"For    3
"His    1
"I      47
"If     11
"It's   9
"Marcos 1
"No-one 1
"None   1
"Our    4
"Over   1
"People 1
"Pizza  1
"Quite  1
"Stockholders   1
"That's 1
"The    64
"These  2
"Today,"        1
"We're  6
"While  2
"With   3
```

Fig 1.2.2 : Output showing words and their frequencies

Problem 2

This Java program is designed to process news files in the Reuters dataset (represented as ".sgm" files) and analyze the occurrence of words in the news titles. It will then classify each news title as "Positive," "Negative," or "Neutral" based on the presence of positive and negative words from two external files: "PositiveWords" and "NegativeWords." The program will output the analysis results into a file named "output.txt."

Description of each method in the program:

main method:

- This is the entry point of the program.
- It initializes two news file paths: "reut2-009.sgm" and "reut2-014.sgm."
- The parseFileAndInsertIntoTitleFile method is called twice to extract the titles from these two news files and insert them into a file named "titleFile."
- Then, it reads the "titleFile" and processes each title using the outputFileCreator method.

isWordPresentInFile method:

- This method checks if a given word (targetWord) is present in a given file (filePath).
- It reads the file line by line and checks for an exact match with the target word.
- If the word is found, it returns true; otherwise, it returns false.

outputFileCreator method:

- This method analyzes the frequency of each word in a news title and determines its polarity (positive, negative, or neutral) based on external lists of positive and negative words.
- It takes the count (news number) and line (news title) as input.
- The method opens the "PositiveWords" and "NegativeWords" files for word lists.
- It splits the news title (line) into individual words and counts the occurrences of each word, storing them in a HashMap.
- It then calculates the total score by summing the frequencies of positive words and subtracting the frequencies of negative words.
- Based on the total score, the title's polarity is determined.
- The result is written into the "output.txt" file, including the news number, title, word, frequency, and polarity.

parseFileAndInsertIntoTitleFile method:

- This method reads a news file and extracts the titles from it.
- It takes the filePath of the news file as input.
- It opens the "titleFile" for writing.
- The method parses the file line by line.
- When it encounters the start of a Reuters news item ("<REUTERS"), it starts collecting the title text until it reaches the end of the news item ("</REUTERS").
- The title text is extracted and cleaned (removing unwanted patterns such as asterisks, HTML tags, etc.).
- The cleaned title is then written into the "titleFile" on a new line.

removePattern method:

- This method removes specified patterns (given as a regular expression) from a StringBuilder.
- It takes the stringBuilder and regex as input.
- It compiles the regular expression pattern, creates a matcher, and replaces all occurrences of the pattern with an empty string (effectively removing them).
- The cleaned StringBuilder is returned.

Testing

1. Checking polarity calculation functionality by sending test data

```
        outputFileCreator( count: 1,  line: "best best laptops in canada");
    }
```

Fig 2.1.1 : Test data to test polarity functionality

File Edit View

News#	Title	Word	Frequency	Polarity
1	best best laptops in canada	canada	1	Neutral
1	best best laptops in canada	in	1	Neutral
1	best best laptops in canada	best	2	Positive
1	best best laptops in canada	laptops	1	Positive

Fig 2.1.2 : Output for Test data of calculating polarity functionality

2. Testing whether positive and negative words are recognised correctly.

```
        System.out.println(isWordPresentInFile( filePath: "PositiveWords",  targetWord: "abound"));
        System.out.println(isWordPresentInFile( filePath: "PositiveWords",  targetWord: "hello"));
        System.out.println(isWordPresentInFile( filePath: "NegativeWords",  targetWord: "abort"));
        System.out.println(isWordPresentInFile( filePath: "NegativeWords",  targetWord: "abound"));
    }
```

Fig 2.2.1 : Inputs for checking negative and positive words

```
Run: Main
C:\Users\hg15y\.jdks\openjdk-20.0.1\bin\java.exe "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA 2022.3.1\lib\idea_rt.jar=57615:C:\Program Files\JetBrains\IntelliJ IDEA 2022
true
false
true
false
```

Fig 2.2.2 : Verifying whether correct outputs are generated

3. Testing whether titles are retrieved and whether the polarity is generated for the actual data.

```
ADVANCED MAGNETICS ADMG IN AGREEMENT
HEALTH RESEARCH FILES FOR BANKRUPTCY
NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS
U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS
U.S. 2-YEAR NOTE AVERAGE YIELD 6.43 PCT, STOP 6.44 PCT, AWARDED AT HIGH YIELD 85 PCT
COMMODORE CBU, ATARI IN SETTLEMENT
BALDRIGE SUPPORTS NIC TALKS ON CURRENCIES
TRIANGLE TRI BEGINS EXCHANGE OFFER
SOUTHMARK SM UNIT IN PUBLIC OFFERING OF STOCK
EASTMAN KODAK CO TO SELL HOLDINGS IN ICN PHARMACEUTICALS AND VIRATEK INC
FEUD PERSISTS AT U.S. HOUSE BUDGET COMMITTTEE
TREASURY BALANCES AT FED ROSE ON MARCH 23
FARM CREDIT SYSTEM SEEN NEEDING 800 MLN DLRS AID
USX X USS UNIT RAISES PRICES
UNIONIST URGES RETALIATION AGAINST JAPAN
EXXON (XON) GETS 99.2 MLN DLR CONTRACT
EATON (ETN) GETS 53.0 MLN DLR CONTRACT
ZAIRE AUTHORIZED TO BUY PL 480 RICE - USDA
MCDONNELL DOUGLAS GETS 30.6 MLN DLR CONTRACT
MIDIVEST ACQUIRES ASSETS OF BUSINESS AVIATION
U.S. WHEAT CREDITS FOR JORDAN SWITCHED
DOLLAR EXPECTED TO FALL DESPITE INTERVENTION
U.S. TO SELL 12.8 BILLION DLRS IN BILLS
```

Fig 2.3.1: First part of the titles generated

```
MASON BEST FORMS ENERGY HOLDING COMPANY
CXR TELCOM CORP CXRL 3RD QTR MARCH 31 NET
PROXMIRE OUTLINES INSIDER TRADING LEGISLATION
HELEN OF TROY CORP HELE 4TH QTR FEB 28 NET
BANKERS TRUST BT PUTS BRAZIL ON NON-ACCRUAL
FIRST MERCANTILE CURRENCY FUND INC 1ST QTR NET
 UK INTERVENTION BD SAYS EC SOLD 118,350 TONNES WHITE SUGAR AT REBATE 46.496 ECUS.
STOLTENBERG SEES MOVES TO STRENGTHEN PARIS ACCORD
U.K. INTERVENTION BOARD DETAILS EC SUGAR SALES
FORD EXTENDS INCENTIVE PROGRAM ON LIGHT TRUCKS TO APRIL 30 FROM APRIL SIX
NORANDA TO SELL 150 MLN DLRS IN DEBENTURES
BACHE SECURITIES CANADA BUYS TORONTO EXCHANGE SEAT FOR 301,000 DLRS
MAFINA BOND WITH WARRANTS SET AT 250 MLN SFR
MEAD MEA EXPECTS IMPROVED EARNINGS THIS YEAR
ENDOTRONICS SEEKS TO ESTABLISH 2ND QTR RESERVE
AMERTEK INC ATEKF 1ST QTR NET
COMSTOCK GROUP CSTK SELLS PREFERRED STOCK
QVC NETWORK QVCN CLARIFIES AGREEMENT
ALEX BROWN INC ABSB 1ST QTR MARCH 27 NET
EQUITABLE RESOURCES EQT FILES UNIT OFFERING
TOWN AND COUNTRY JEWELRY MANUFACTURING TCJC
```

Fig 2.3.2: Last part of the titles generated(1694 titles)

output

File Edit View

News#	Title	Word	Frequency	Polarity
1	ADVANCED MAGNETICS ADMG IN AGREEMENT	admg	1	Neutral
1	ADVANCED MAGNETICS ADMG IN AGREEMENT	agreement	1	Neutral
1	ADVANCED MAGNETICS ADMG IN AGREEMENT	advanced	1	Positive
1	ADVANCED MAGNETICS ADMG IN AGREEMENT	in	1	Positive
1	ADVANCED MAGNETICS ADMG IN AGREEMENT	magnetics	1	Positive
2	HEALTH RESEARCH FILES FOR BANKRUPTCY	for	1	Neutral
2	HEALTH RESEARCH FILES FOR BANKRUPTCY	health	1	Neutral
2	HEALTH RESEARCH FILES FOR BANKRUPTCY	files	1	Neutral
2	HEALTH RESEARCH FILES FOR BANKRUPTCY	bankruptcy	1	Neutral
2	HEALTH RESEARCH FILES FOR BANKRUPTCY	research	1	Neutral
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	loss	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	numerex	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	corp	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	nmrx	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	jan	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	2nd	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	qtr	1	Negative
3	NUMEREX CORP NMRX 2ND QTR JAN 31 LOSS	31	1	Negative
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	6-mo	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	u.s.	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	dlrs	2	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	pay	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	down	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	march	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	3	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	billion	2	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	1.2	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	and	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	of	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	selling	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	bills	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	to	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	12.8	1	Neutral
4	U.S. SELLING 12.8 BILLION DLRS OF 3 AND 6-MO BILLS MARCH 30 TO PAY DOWN 1.2 BILLION DLRS	30	1	Neutral
5	U.S. 2-YEAR NOTE AVERAGE YIELD 6.43 PCT, STOP 6.44 PCT, AWARDED AT HIGH YIELD 85 PCT	pct	1	Neutral

Fig 2.3.3 : output after finding polarity of the titles.

References

[1] SimpliCode, “Mongodb Connection in Java | how to connect MongoDB with Java | MongoDB tutorial | SimpliCode,” 28-Jan-2023. [Online]. Available: https://www.youtube.com/watch?v=axgM35lUnOk. [Accessed: 01-Aug-2023].

[2] “MapReduce word count Program in Java,” *Educative: Interactive Courses for Software Developers*. [Online]. Available: https://www.educative.io/answers/mapreduce-word-count-program-in-java [Accessed: 01-Aug-2023].

[3] Marcin, *Negative-words.Txt*. . Available : https://gist.github.com/mkulakowski2/4289441

[4] Marcin, *Positive-words.Txt*. Available : https://gist.github.com/mkulakowski2/4289437