

Assignment-2 Problem 2:

CODE A: To create a Java program to extract data from NewsAPI.

PseudoCode:

```
String[] keywords = {"Canada", "University", "Dalhousie", "Halifax", "Canada Education",  
"Moncton", "hockey", "Fredericton", "celebration"};
```

```
// Join all the keywords with "OR" and encode the resulting query string
```

```
String query = joinWithOr(keywords);
```

```
String encodedQuery = urlEncode(query);
```

```
// Construct the URL for fetching the news from the API
```

```
String apiUrl = "https://newsapi.org/v2/everything?q=" + encodedQuery +  
"&apiKey=ab3a6355793f4756bcda3e9f15b8b792";
```

```
URL url = new URL(apiUrl);
```

```
// Open a connection to the URL and fetch the news data
```

```
URLConnection connection = (URLConnection) url.openConnection();
```

```
connection.setRequestMethod("GET");
```

```
BufferedReader bufferedReader = new BufferedReader(new  
InputStreamReader(connection.getInputStream()));
```

```
StringBuilder response = new StringBuilder();
```

```
String inputLine;
```

```
while ((inputLine = bufferedReader.readLine()) != null) {
```

```
    response.append(inputLine);
```

```
}
```

```
bufferedReader.close();
```

```
// Pass the news data to the DataProcessingEngine for processing
```

```
DataProcessingEngine dataProcessingEngine = new DataProcessingEngine();
```

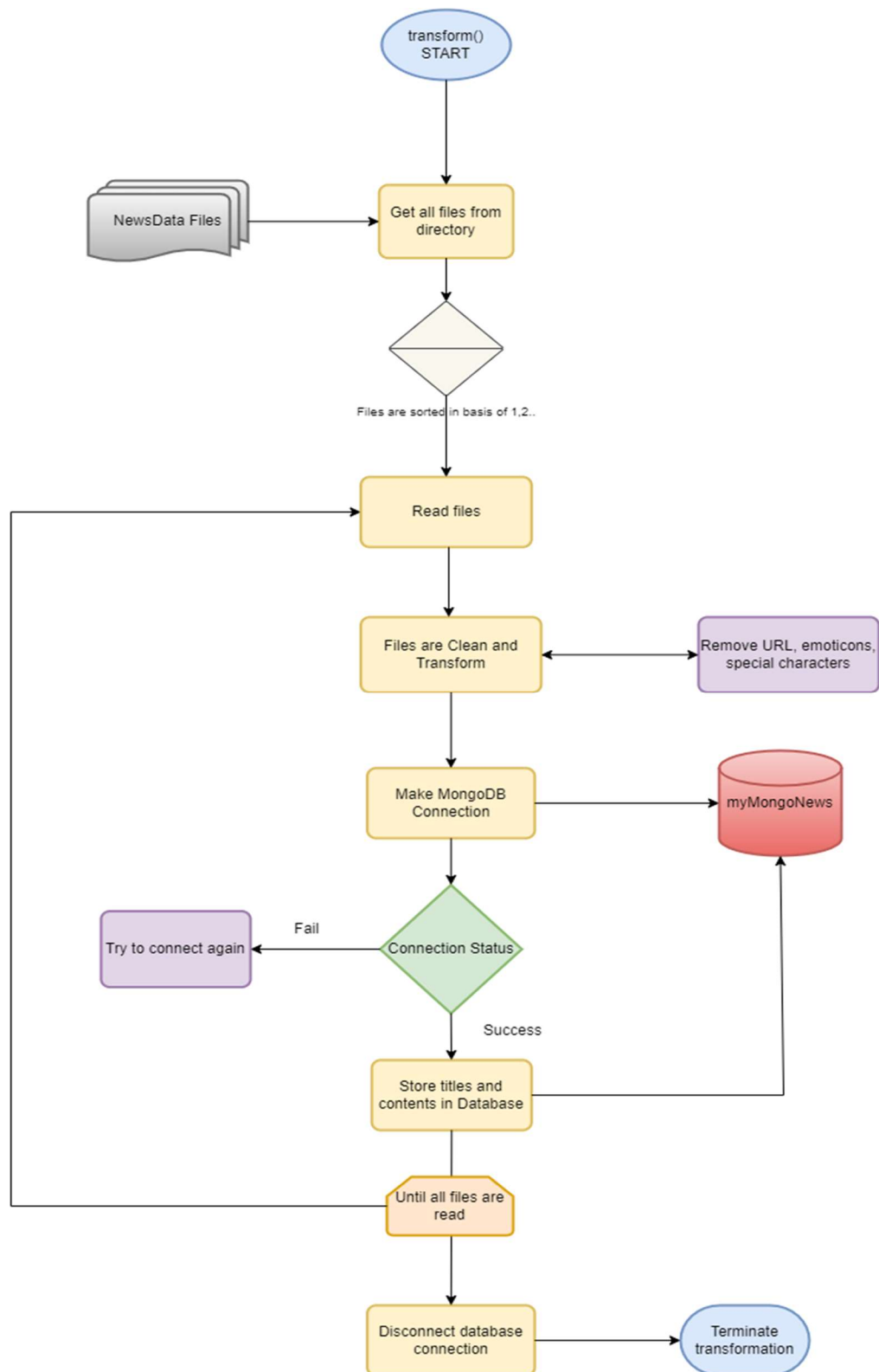
```
dataProcessingEngine.process(response.toString());
```

PseudoCode with sentence explanation:

1. Create an array of keywords to search for in the news API
2. Join the array of keywords with "OR" and encode the resulting query string using a function called "joinWithOr" and "urlencode" respectively
3. Construct the URL for fetching the news from the API using the encoded query string and API key
4. Open a connection to the URL using HttpURLConnection and set the request method to "GET"
5. Create a BufferedReader to read the response from the API
6. Read the response line by line and append each line to a StringBuilder called "response" until there is no more data to read
7. Close the BufferedReader
8. Create a DataProcessingEngine object
9. Pass the news data, stored in "response" as a string, to the DataProcessingEngine object for processing using the "process" method.

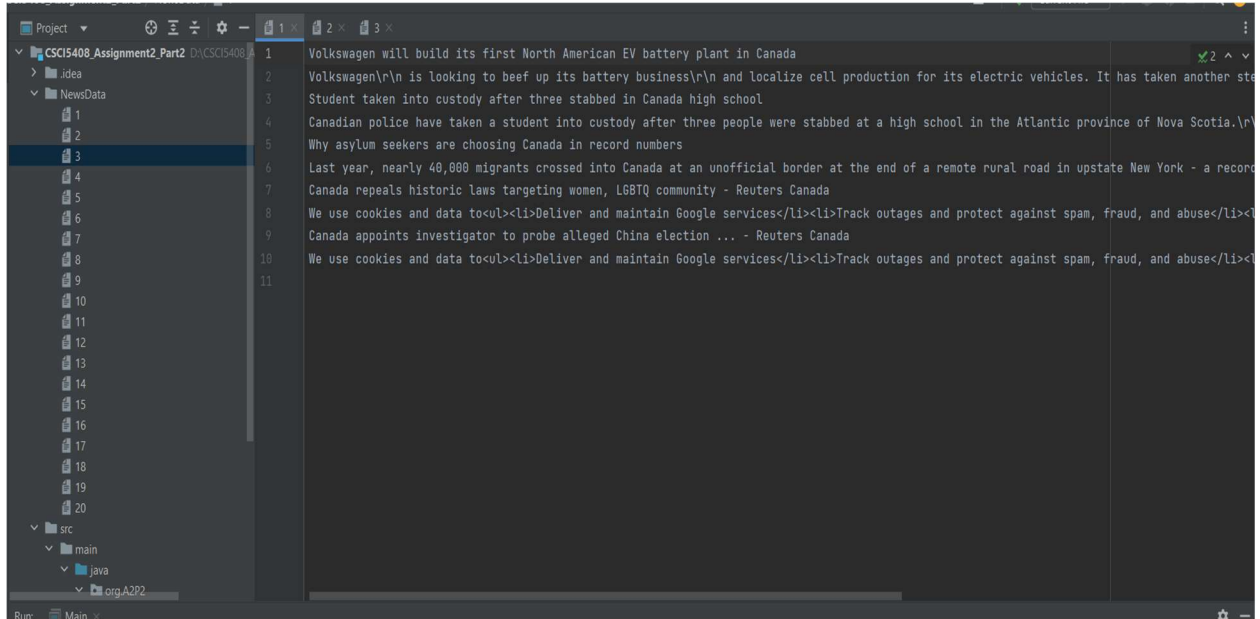
CODE B: Write titles and content in raw files, each files having 5 or less articles.

1. Start of the DataProcessingEngine class
2. Define the process method that takes a StringBuilder object called "response" as input parameter
3. Compile two regex patterns, one for the title and one for the content
4. Initialize variables including fileName, fiveNewsCounter, and path
5. Create a new FileWriter object to write data to a file
6. Loop through all titles and contents and write them to the file
 - a. If there are five articles in a file, create a new file
 - b. Extract the title and content from the response
 - c. Write the title and content to the file
 - d. Increment fiveNewsCounter
7. Close the writer
8. Create a new TransformationEngine object
9. Call the transform method to store data in MongoDB

CODE C: Flowchart for transformation Engine:

Test cases and Screenshots

1. Title and contents store in files.



2. Documents inserted in MongoDB:

```
15:54:22.269 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 45.93 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.270 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.310 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 39.73 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.312 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.358 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 45.36 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.359 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.405 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 45.75 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.407 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.452 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 45.02 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.466 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.563 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 98.65 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.565 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.610 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 44.93 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.613 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.794 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 180.32 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:22.798 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:23.156 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 358.41 ms using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:23.159 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" started on database myMongoNews using a connection with driver-generated ID 7 and server-generated ID 2318
15:54:23.196 [main] DEBUG org.mongodb.driver.protocol.command -- Command "insert" succeeded in 37.77 ms using a connection with driver-generated ID 7 and server-generated ID 2318

Process finished with exit code 0
```

3. Documents (titles and contents) visible in Database with removed URL, emoticons and special characters.

The screenshot shows the MongoDB Atlas web interface. On the left sidebar, there's a '+ Create Database' button and a search bar labeled 'Search Namespaces'. Below that, the database 'myMongoNews' is selected, and the collection 'newsData' is highlighted. The main panel shows the 'myMongoNews.newsData' collection. At the top, it displays statistics: 'STORAGE SIZE: 56KB', 'LOGICAL DATA SIZE: 61.46KB', 'TOTAL DOCUMENTS: 200', and 'INDEXES TOTAL SIZE: 40KB'. Below these are tabs for 'Find', 'Indexes', 'Schema Anti-Patterns', 'Aggregation', and 'Search Indexes'. The 'Find' tab is active. A 'Filter' input field contains the query '{ field: 'value' }'. To the right of the filter are 'Reset', 'Apply', and 'More Options' buttons. Below the filter, it says 'QUERY RESULTS: 1-20 OF MANY'. Two documents are displayed in a list. The first document has a red _id field, a title 'Volkswagen will build its first North American EV battery plant in Can...', and a content field 'Volkswagenrn is looking to beef up its battery businessrn and localize...'. The second document has a red _id field, a title 'Student taken into custody after three stabbed in Canada high school', and a content field 'Canadian police have taken a student into custody after three people w...'. At the bottom, there are navigation buttons: '< PREVIOUS', '1-20 of many results', and 'NEXT >'. An 'INSERT DOCUMENT' button is located in the top right corner of the main panel.