

CSCI 5408: Assignment 2: Problem 2:

Waleed R. Alhindi (B00919848)

Overview:

The program submitted constitutes fetches article data from an API based on keywords, extracts articles' data from the API response, reformats and cleans the data, then stores it in files and in a MongoDB database collection. As such the program consists of the following components:

- Main.java – The runnable interface class that starts the sequence described above. It initializes a list of keywords, which is fed into the NewsExtract class to fetch articles pertaining to each keyword from the News API.
- NewsExtract.java – This class addresses “Code A” in the assignment instructions. For each keyword supplied, it makes a call to the News API (<https://newsapi.org/>). It then extracts the information of each article in the response into a NewsArticle object. After fetch and extracting each keyword response, it will end up with a collection of NewsArticle objects, each encapsulating an article from a response which it then passes to the FileInterface class to store into files and then to the MongoInterface class to store into a MongoDB database collection.
- NewsArticle.java – A class to store and encapsulate each article extracted from an API response. In other words, each article the NewExtract class extracts from a response is stored in an individual object of this class. This makes accessing, operating, and passing articles between classes and methods more efficient.
- FileInterface.java – This class writes articles to text files that reside under the “Output” folder of this project. It is fed a collection of NewsArticle objects by the NewsExtract class, then either appends article(s) to an existing file or creates and writes to a new file if the latest file already contains 5 articles. This is to ensure that each file contains no more than 5 articles. Furthermore, the class writes each article to a file in the following format:

```
Title: <title>
Source: <source>
Keyword: <keyword>
Author: <author>
Description: <description>
URL: <url>
Image URL: <imageUrl>
Published At: <publishedAt>
Content: <content>
<empty line>
```

- MongoInterface.java – This class interfaces with a Mongo Atlas cluster to store extracted news articles into a MongoDB database collection (database = MyMongoNews, collection = News). Thus, it is passed a collection of extracted NewsArticle objects by the

NewsExtract class. For each of these objects, it first checks whether that news article is an identical duplicate of an existing document in the database collection; and only inserts that article if it is not a duplicate. This means that running this program several times will not result in the database collection being overrun with duplicates.

Program Design:

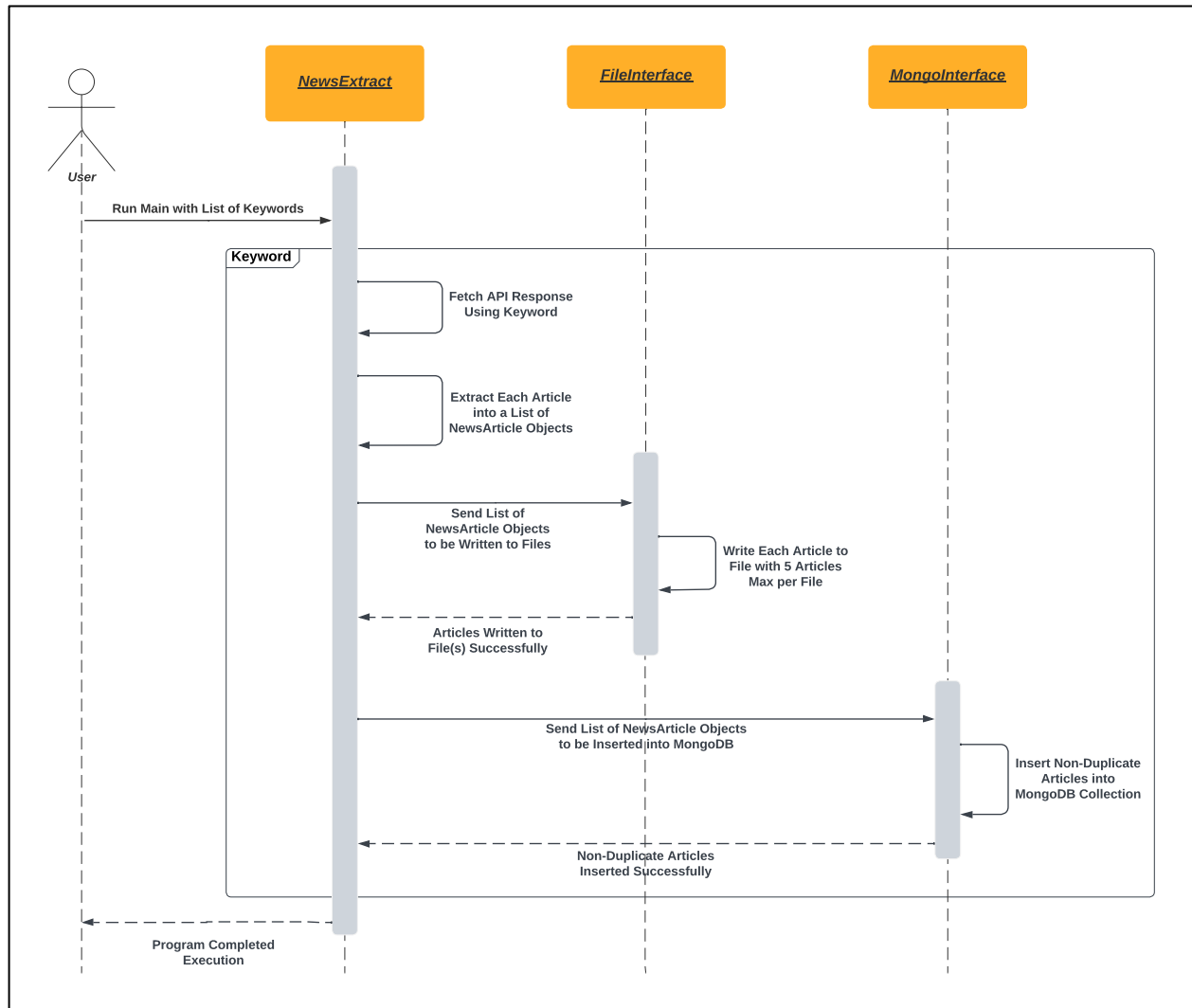


Fig 1. Program Sequence Diagram

Note that the above sequence diagram has also been uploaded in the zipped submission folder and can be found under the “A2” folder as “Sequence Diagram.pdf”

GitLab Repository:

The program has been uploaded to GitLab and can be found at the following link:
https://git.cs.dal.ca/alhindi/csci5408_w23_b00919848_waleed_alhindi/-/tree/main/A2

OR

Under the “A2” folder at the following link:

https://git.cs.dal.ca/alhindi/csci5408_w23_b00919848_waleed_alhindi/-/tree/main/

Furthermore, this repository was shared with the following emails as instructed in the “Submission Guidelines” PowerPoint:

- saurabh.dey@dal.ca
- vs439755@dal.ca
- ar260217@dal.ca
- pr514457@dal.ca
- sh495601@dal.ca

Code A: API Content Extraction:

“Code A” in the assignment instructions refers to the process of fetching news articles based on keywords, then extracting their content. This is handled by the NewsExtract and NewsArticle classes. The NewsExtract class is passed a list of keywords, for each of which it makes an API call to the News API. Then, it parses the response to extract the data of article and stores the data of each article into an object of the NewsArticle class. This process is illustrated in the pseudocode below:

```
Initialize a base-URL string for the news API endpoint
For each keyword in the list of keywords
    Append that keyword to the base-URL
    Fetch the response from that modified URL
    If the response contains any articles
        Stringify the response into a string variable (i.e., use a scanner to read the
        response into a string variable)
        Create a new list to store NewsArticle objects
        Split the response into a list of individual articles using a regex
        For each article in that list
            Extract that article's source using a regex
            Extract that article's author using a regex
            Extract that article's title using a regex
            Extract that article's description using a regex
            Extract that article's URL using a regex
            Extract that article's imageURL using a regex
            Extract that article's publishedAt using a regex
            Extract that article's content using a regex
            Create a new NewsArticle object to store the article's data
            Add this new object to the list of NewsArticles
        Pass the list of extracted NewsArticle objects to the FileInterface class so
        they can be written to file(s)
        Pass the list of extracted NewsArticle object to the MongoInterface class
        to be inserted into the database collection
    Else
        Continue onto the next keyword API call
```

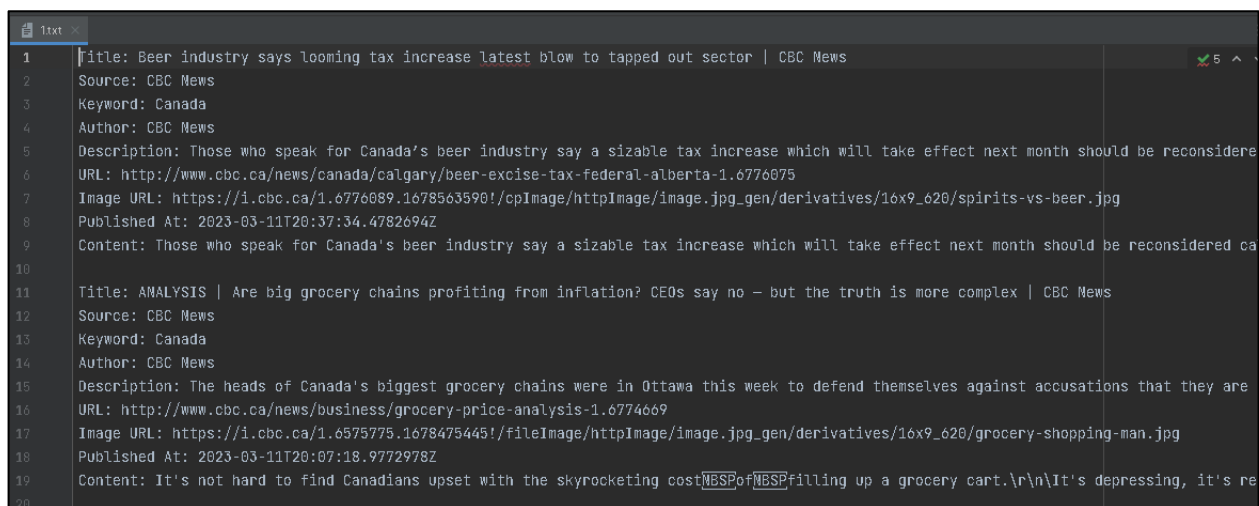
Additionally, the screenshot below showcases the code used to extract articles' data from API response:

```
private List<NewsArticle> parseResponse(String responseStr, String keyword){
    List<NewsArticle> newsArticles = new ArrayList<>();
    String[] articles = responseStr.split( regex: "\\\"source\\\"");
    for(String article: articles){
        if(!article.contains("id")){
            continue;
        }
        String source = article.split( regex: "\\\"name\\\"")[1].split( regex: "\\\"\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String author = article.split( regex: "\\\"author\\\"")[1].split( regex: "\\\"\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String title = article.split( regex: "\\\"title\\\"")[1].split( regex: "\\\"description\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String description = article.split( regex: "\\\"description\\\"")[1].split( regex: "\\\"url\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String url = article.split( regex: "\\\"url\\\"")[1].split( regex: "\\\"urlToImage\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String imageUrl = article.split( regex: "\\\"urlToImage\\\"")[1].split( regex: "\\\"publishedAt\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String publishedAt = article.split( regex: "\\\"publishedAt\\\"")[1].split( regex: "\\\"content\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        String content = article.split( regex: "\\\"content\\\"")[1].split( regex: "\\\"\\\"")[0].replaceAll( regex: "\\\"", replacement: "").trim();
        NewsArticle newsArticle = new NewsArticle(source, author, title, description, url, imageUrl, publishedAt, content, keyword);
        newsArticles.add(newsArticle);
    }
    return newsArticles;
}
```

Fig 2. Article Data Extraction Code Snippet

Code B: Processing and Writing API Content to File(s):

“Code B” in the assignment instructions refers to the reformatting and storage of extracted articles into files, where each file can store no more than 5 articles. This is handled by the FileInterface class, which receives a list of extracted NewsArticle objects from the NewsExtract class, then stores each NewsArticle’s data into a text file under the “Output” directory of this project as seen in the screenshots below:



```
1 Title: Beer industry says looming tax increase latest blow to tapped out sector | CBC News
2 Source: CBC News
3 Keyword: Canada
4 Author: CBC News
5 Description: Those who speak for Canada's beer industry say a sizable tax increase which will take effect next month should be reconsidere
6 URL: http://www.cbc.ca/news/canada/calgary/beer-excise-tax-federal-alberta-1.6776075
7 Image URL: https://1.cbc.ca/1.6776089.1678563590!/cpImage/httpImage/image.jpg_gen/derivatives/16x9_620/spirits-vs-beer.jpg
8 Published At: 2023-03-11T20:37:34.4782694Z
9 Content: Those who speak for Canada's beer industry say a sizable tax increase which will take effect next month should be reconsidered ca
10
11 Title: ANALYSIS | Are big grocery chains profiting from inflation? CEOs say no – but the truth is more complex | CBC News
12 Source: CBC News
13 Keyword: Canada
14 Author: CBC News
15 Description: The heads of Canada's biggest grocery chains were in Ottawa this week to defend themselves against accusations that they are
16 URL: http://www.cbc.ca/news/business/grocery-price-analysis-1.6774669
17 Image URL: https://1.cbc.ca/1.6575775.1678475445!/fileImage/httpImage/image.jpg_gen/derivatives/16x9_620/grocery-shopping-man.jpg
18 Published At: 2023-03-11T20:07:18.9772978Z
19 Content: It's not hard to find Canadians upset with the skyrocketing cost of filling up a grocery cart. It's depressing, it's re
```

Fig 3. Example Text File Output Containing Extracted Article Data

To achieve this, for each NewsArticle object in the list passed by NewsExtract, the FileInterface class first checks whether the latest file already contains 5 articles. If not, it appends the current article to that file. However, if the file does already store 5 articles, then it creates a new file and writes the article’s data into that newly created file. This ensures that a file will only store a maximum of 5 articles.

Code C: Transformation and Storage of API Content into MongoDB:

“Code C” in the assignment instructions refers to the cleaning and subsequent mongoDB storage of articles extracted from the API response. This is handled by the MongoInterface class, which is passed a list of extracted NewsArticle objects. Since the article’s data has already been parsed into instance variables in each NewsArticle object, this class simply needs to clean the data (i.e., remove special characters and the like), then insert it as a new document in the MongoDB MyMongoNews database’s News collection.

However, in order to prevent the insertion of identical duplicates, this class first checks whether an extracted article is identical to an existing document in the collection; only inserting the article if it is not a duplicate.

The screenshot below showcases the code used to check whether an extracted NewsArticle object is a duplicate and that inserts NewsArticles that are not duplicates:

```
Bson query = Filters.and(
    Filters.eq( fieldName: "title", sanitize(article.getTitle()),
    Filters.eq( fieldName: "source", sanitize(article.getSource()),
    Filters.eq( fieldName: "keyword", article.getKeyword()),
    Filters.eq( fieldName: "author", sanitize(article.getAuthor()),
    Filters.eq( fieldName: "description", sanitize(article.getDescription()),
    Filters.eq( fieldName: "url", article.getUrl()),
    Filters.eq( fieldName: "imageUrl", article.getImageUrl()),
    Filters.eq( fieldName: "publishedAt", article.getPublishedAt()),
    Filters.eq( fieldName: "content", sanitize(article.getContent())
);
Document existingDoc = collection.find(query).first();
if(existingDoc==null){
    Document insertArticle = new Document();
    insertArticle.append("_id", new ObjectId());
    insertArticle.append("title", sanitize(article.getTitle()));
    insertArticle.append("source", sanitize(article.getSource()));
    insertArticle.append("keyword", article.getKeyword());
    insertArticle.append("author", sanitize(article.getAuthor()));
    insertArticle.append("description", sanitize(article.getDescription()));
    insertArticle.append("url", article.getUrl());
    insertArticle.append("imageUrl", article.getImageUrl());
    insertArticle.append("publishedAt", article.getPublishedAt());
    insertArticle.append("content", sanitize(article.getContent()));
    InsertOneResult insertRes = collection.insertOne(insertArticle);
}
```

Fig 4. Duplication Verification and Insertion of Non-Duplicates Code Snippet

Furthermore, the screenshot below showcases the documents in the database collection after this program has completed execution:

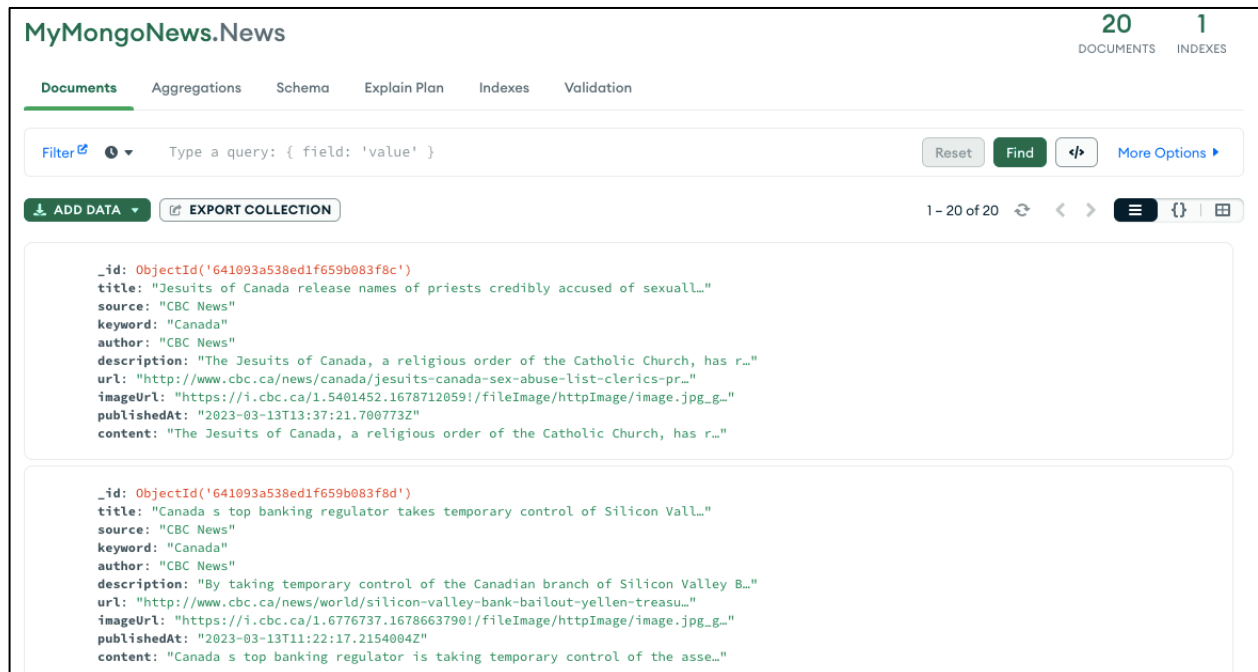


Fig 5. Successful Insertion of Articles into MongoDB Collection

Data Extraction, Transformation, and Storage Flowchart:

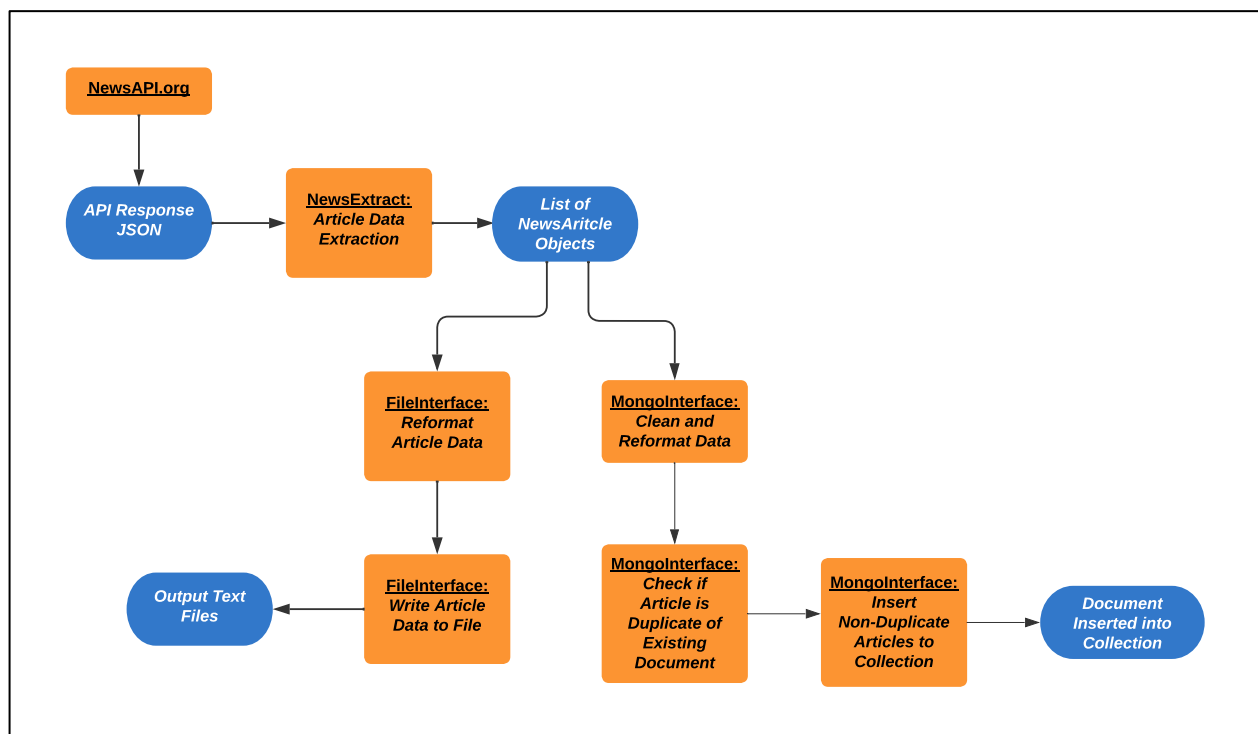


Fig 6. Data Fetching, Extraction, Reformatting, Cleaning, and Storage Flowchart

Note that the above sequence diagram has also been uploaded in the zipped submission folder and can be found under the “A2” folder as “Flowchart.pdf”

Test Cases:

1. Ensuring Articles are Extracted from API Response and Written to File:

This is validated by simply running the program and observing the changes made inside the “Output” directory of this project. Before running the program, we can observe that the “Output” directory is empty as seen below:

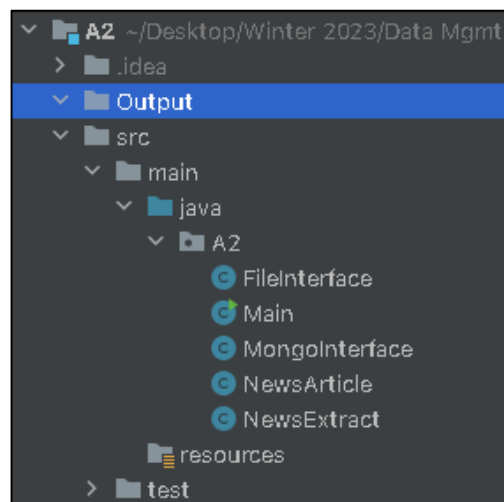


Fig 7. File Output Directory Before Running Program

Once, the program has completed execution, we observe that files have indeed been created in that directory, and that article data has been extracted correctly and written into the created files as seen below:

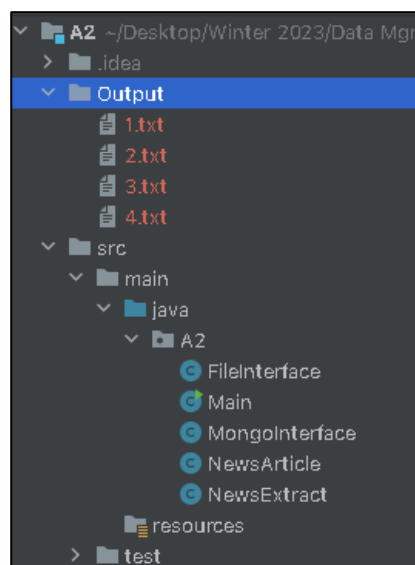
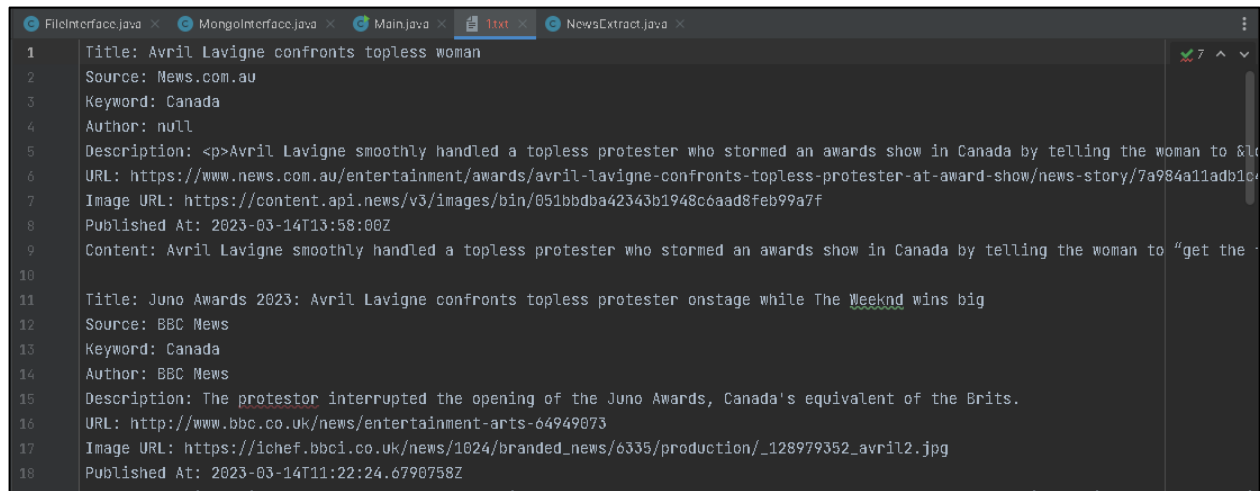


Fig 8. Files Created in File Output Directory After Program Execution



```
1 Title: Avril Lavigne confronts topless woman
2 Source: News.com.au
3 Keyword: Canada
4 Author: null
5 Description: <p>Avril Lavigne smoothly handled a topless protester who stormed an awards show in Canada by telling the woman to &l
6 URL: https://www.news.com.au/entertainment/awards/avril-lavigne-confronts-topless-protester-at-award-show/news-story/7a984a11adb1c
7 Image URL: https://content.api.news/v3/images/bin/051bbdba42343b1948c6aad8feb99a7f
8 Published At: 2023-03-14T13:58:00Z
9 Content: Avril Lavigne smoothly handled a topless protester who stormed an awards show in Canada by telling the woman to "get the
10
11 Title: Juno Awards 2023: Avril Lavigne confronts topless protester onstage while The Weeknd wins big
12 Source: BBC News
13 Keyword: Canada
14 Author: BBC News
15 Description: The protester interrupted the opening of the Juno Awards, Canada's equivalent of the Brits.
16 URL: http://www.bbc.co.uk/news/entertainment-arts-64949073
17 Image URL: https://ichef.bbci.co.uk/news/1024/branded_news/6335/production/_128979352_avril2.jpg
18 Published At: 2023-03-14T11:22:24.6790758Z
```

Fig 9. File Content After Program Execution

2. Ensuring a File Stores a Maximum of 5 Articles:

To verify that each article stores a maximum of 5 articles, the program is run, then each generated file is observed to verify that it contains 5 or fewer articles. The screenshot below showcases the contents of a file after the program has completed execution where only 5 articles have been stored in that file:


```

Title: Avril Lavigne confronts topless woman
Source: News.com.au
Keyword: Canada
Author: null
Description: <p>Avril Lavigne smoothly handled a topless protester who stormed an awards show in Canada by telling the woman to &ldquo;get the f*** off&rdquo; the stage.</p>
URL: https://www.news.com.au/entertainment/awards/avril-lavigne-confronts-topless-protester-at-award-show/news-story/7a984a11adb1c431a8a598c2c5a54b29
Image URL: https://content.api.news/v3/images/bin/051bbdba42343b1948c6aad8feb99a7f
Published At: 2023-03-14T13:58:00Z
Content: Avril Lavigne smoothly handled a topless protester who stormed an awards show in Canada by telling the woman to "get the f*** off" the stage.\r\nThe singer had been introducing performer AP Dhillon dur_ [+1607 chars]

Title: Juno Awards 2023: Avril Lavigne confronts topless protester onstage while The Weeknd wins big
Source: BBC News
Keyword: Canada
Author: BBC News
Description: The protester interrupted the opening of the Juno Awards, Canada's equivalent of the Brits.
URL: http://www.bbc.co.uk/news/entertainment-arts-64949073
Image URL: https://ichef.bbci.co.uk/news/1024/branded_news/6335/production/_128979352_avril2.jpg
Published At: 2023-03-14T11:22:24.6790758Z
Content: Avril Lavigne confronted a topless environmental protester onstage at the Juno Awards, the Canadian equivalent of the Brits.\r\nAs the singer introduced a performance at the start of Monday's ceremony,_ [+2626 chars]

Title: Canada: Truck ploughs into pedestrians, killing two
Source: BBC News
Keyword: Canada
Author: BBC News
Description: Two people have died and nine have been injured after being hit by a truck in Quebec province.
URL: http://www.bbc.co.uk/news/world-us-canada-64947528
Image URL: https://ichef.bbci.co.uk/news/1024/branded_news/E200/production/_128977085_gettyimages-1247337078.jpg
Published At: 2023-03-14T09:52:21.7423444Z
Content: Two people have been killed and nine others injured after pedestrians were hit by a pick-up truck in the town of Amqui, in Canada's northern Quebec.\r\nThe driver of the vehicle, a 38-year-old local ma_ [+1544 chars]

Title: Foreign nurses out $24,000 – and left with no recourse – after job offers in N.L. disappear | CBC News
Source: CBC News
Keyword: Canada
Author: CBC News
Description: A Niagara Falls, Ont., mother who paid a Toronto-based employment agency $24,000 for immigration services to bring her two daughters, who are foreign nurses, to Canada discovered there was nowhere to turn after a dispute arose with the agency and the job offer.
URL: http://www.cbc.ca/news/canada/immigration-agency-apex-1.6777620
Image URL: https://i.cbc.ca/1.6777626.1678740901!/fileImage/httpImage/image.jpg_gen/derivatives/16x9_620/april-nuval-abrey-nuval-and-joy-thompson.jpg
Published At: 2023-03-14T08:07:22.1970991Z
Content: Joy Thompson has a dream of reuniting her family and having her daughters finally join her in Canada.\r\nThompson came here in 2004 as a domestic worker to help support her children and put them throug_ [+14685 chars]

Title: Junos 2023: Watch Canada's biggest night in music | CBC News
Source: CBC News
Keyword: Canada
Author: CBC News
Description: The 52nd Juno Awards will be broadcast live tonight and hosted by Marvel star Simu Liu for the second year in a row.
URL: http://www.cbc.ca/news/entertainment/junos-2023-livestream-1.6777611
Image URL: https://i.cbc.ca/1.6775057.1678475474!/fileImage/httpImage/image.jpg_gen/derivatives/16x9_620/2023-juno.jpg
Published At: 2023-03-13T23:22:19.2620751Z
Content: The 52nd Juno Awards will be broadcast live tonight and hosted by Marvel star Simu Liu (Shang-Chi and the Legend of the Ten Rings) for the second year in a row.\r\nThe Canadian music awards handed out _ [+965 chars]

```

Fig 10. Output File Contains Only 5 Articles (Separated by Empty Line)

3. Verifying that Extracted Articles are Inserted into the MongoDB Database Collection:

To verify that articles extracted from the News API have been successfully inserted into the MyMongoNews database's News collection, the program is run while observing the state of the collection before and after the execution of the program.

As seen below, before running the program, the collection is empty and contains no documents:

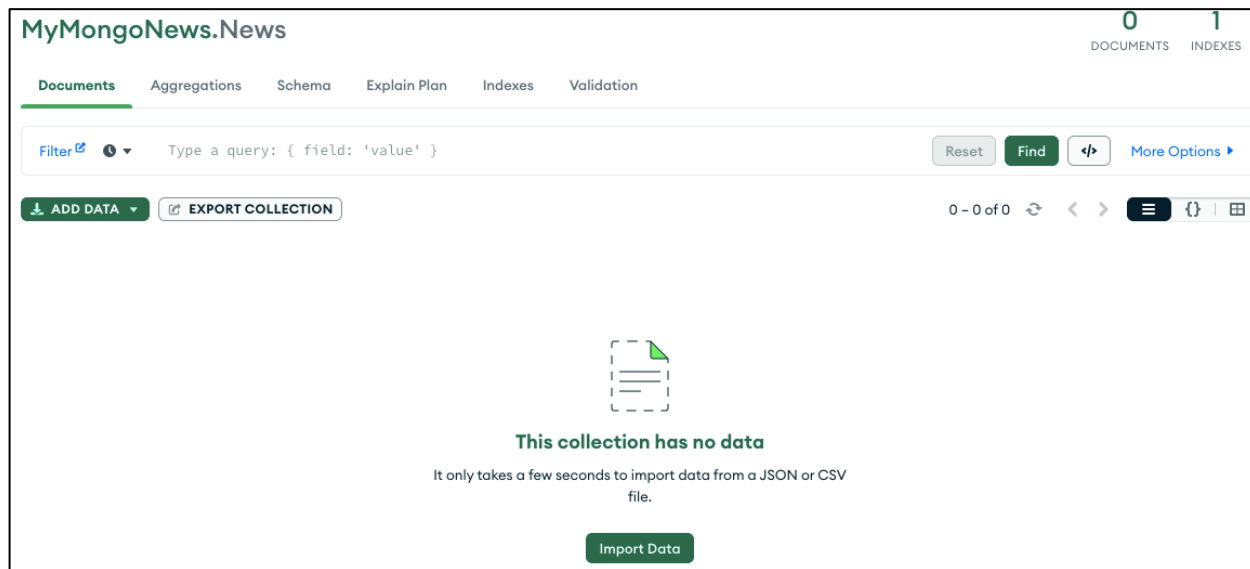


Fig 11. MyMongoNews' News Collection Before Running Program

However, once the program has completed execution, we can now observe that 20 documents have been inserted successfully:

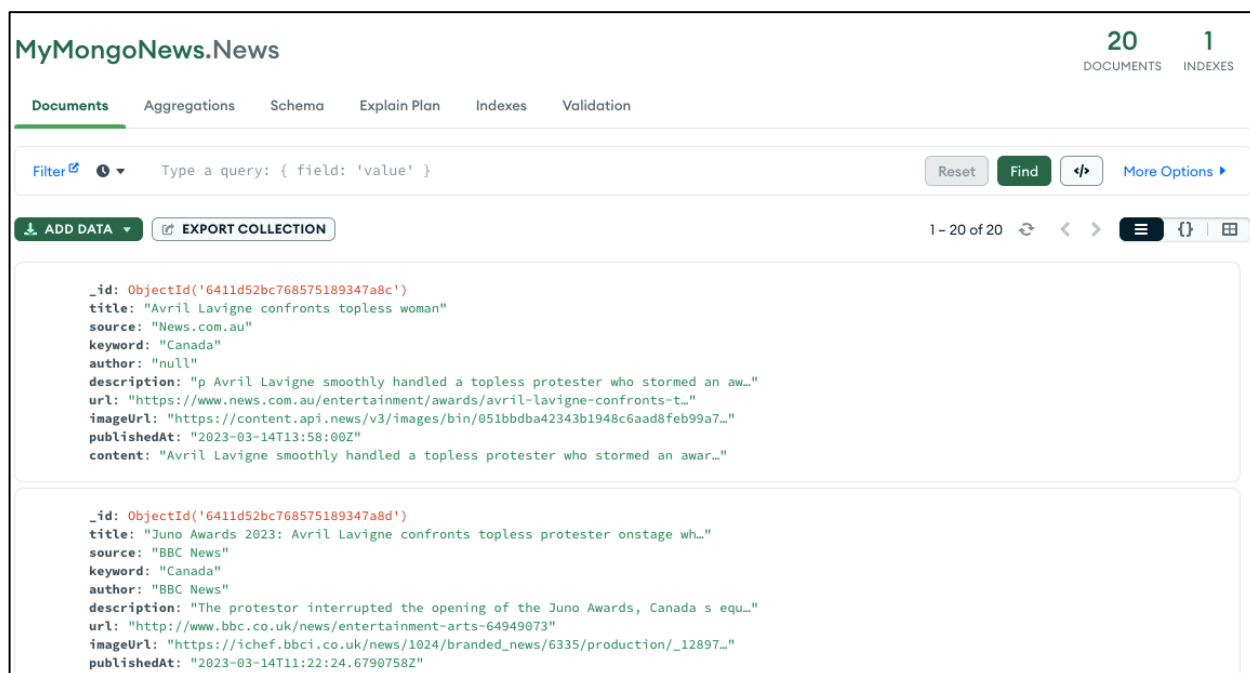


Fig 12. MyMongoNews' News Collection After Running Program

Conclusion:

In conclusion, this assignment simulates how data is collected, processed, and stored. Namely, through the use of APIs, programs to clean and process data, and NoSQL databases. The program submitted aims to replicate this process by fetching news articles from a News API, extracting the data from the response, reformatting and cleaning the extracted data, and storing it into files and a MongoDB database collection.

References:

- [1] “News API – Search News and Blog Articles on the Web,” *News API*. [Online]. Available: <https://newsapi.org/> [Accessed: March 12, 2023].
- [2] “MongoDB atlas: Multi-cloud Developer Data Platform,” *MongoDB*. [Online]. Available: <https://www.mongodb.com/atlas> [Accessed: March 12, 2023].
- [3] “MongoDB Compass,” *MongoDB*. [Online]. Available: <https://www.mongodb.com/products/compass> [Accessed: March 12, 2023].
- [4] “Intelligent Diagramming,” *LucidChart*. [Online], Available: <https://www.lucidchat.com/> [Accessed: March 13, 2022].