

Assignment 2 Report

Design Principles used in this Assignment.

Single Responsibility Principle: All the three classes have single Responsibility. For example, the ExtractionEngine is responsible for fetching data from the NewsAPI, the DataProcessingEngine is responsible for parsing the raw JSON data and writing it to files, and the TransformationEngine is responsible for cleaning and transforming the data and uploading it to MongoDB.

Modularity: The code is organized into separate classes and methods, each performing a specific task. This makes the code easier to understand, maintain, and extend.

Abstraction: Methods like parseAndWriteArticles, cleanAndTransformData, and storeTransformedData leave the implementation details and provide a simple interface for performing their respective tasks.

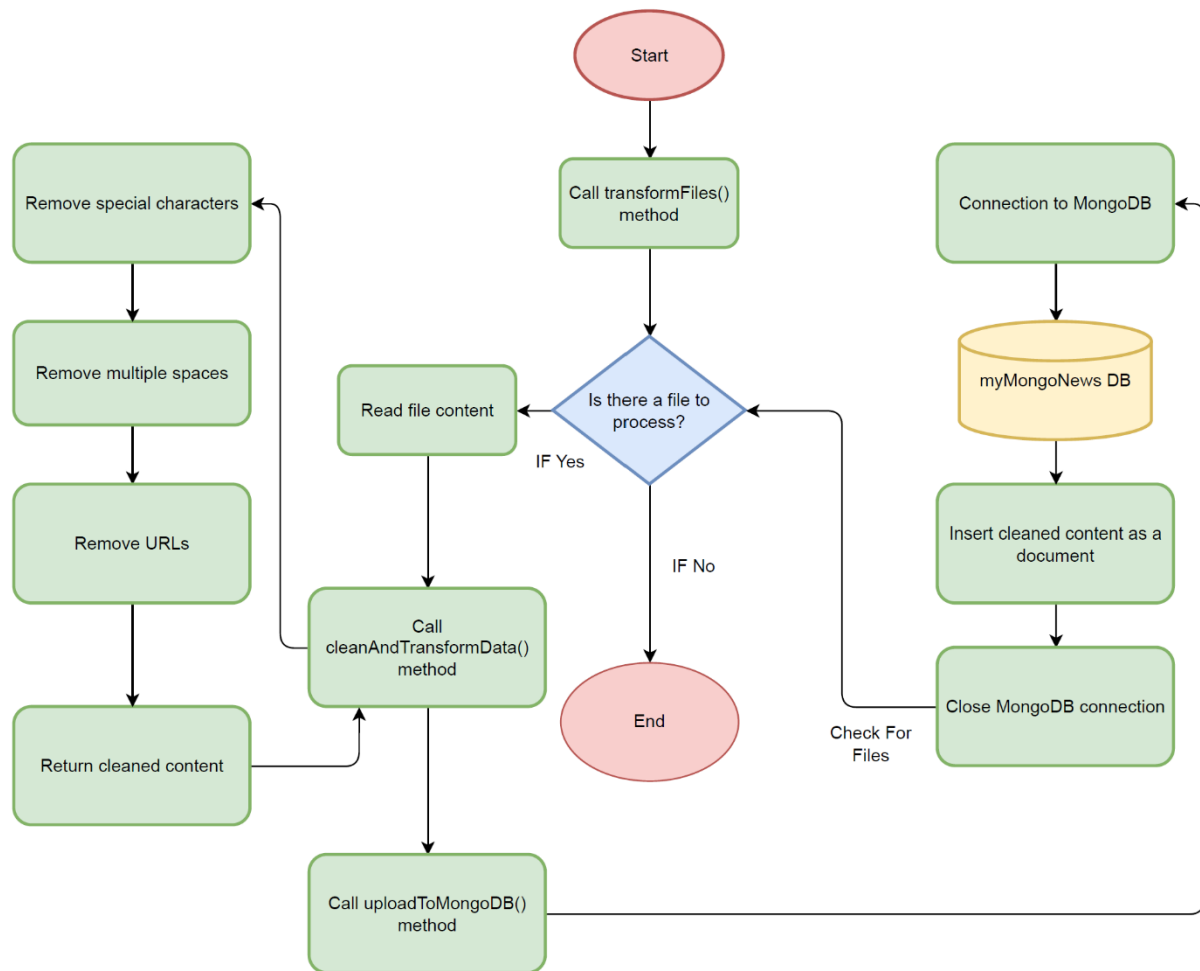
Information Hiding: The implementation details of each class and method are hidden from other classes, allowing for a clean separation of concerns.

Loose Coupling: The classes are loosely coupled, meaning that they can be easily modified or extended without affecting the rest of the system.

Pseudocode for Data Extraction Code

1. Define the API key, base URL, and search keywords.
2. Create a search query by joining the search keywords with "OR".
3. Encode the search query.
4. Construct the request URL using the base URL, encoded search query, and API key.
5. Create a URL object with the request URL.
6. Initialize an empty string builder for storing the JSON data.
7. Try the following steps:
 - a. Open an HTTP connection to the URL.
 - b. Set the request method to "GET".
 - c. Get the response code from the connection.
 - d. If the response code is HTTP_OK (200):
 - i. Create a buffered reader for reading the input stream from the connection.
 - ii. Read lines from the buffered reader and append them to the JSON data string builder.
 - iii. Close the buffered reader.
 - iv. Print the response JSON data.
 - v. Call the DataProcessingEngine.parseAndWriteArticles() method with the JSON data as an argument.
 - vi. Otherwise, print an error message indicating that the data could not be fetched.
8. Catch any exceptions that may occur during the process, and print the error stack trace.

Flowchart of the transformation engine




Transformation engine Flowchart

How data is stored in MongoDB myMongoNews Database:


myMongoNews.cleanedArticles

STORAGE SIZE: 28KB LOGICAL DATA SIZE: 32.1KB TOTAL DOCUMENTS: 100 INDEXES TOTAL SIZE: 20KB

Find Indexes Schema Anti-Patterns 0 Aggregation Search Indexes

Filter  Type a query: { field: 'value' }


QUERY RESULTS: 1-20 OF MANY



_id: ObjectId('6429e5dc12ed94112ab3fdce')

title: "Volkswagen will build its first North American EV battery plant in Can..."


content: "Volkswagen r n is looking to beef up its battery business r n and loca..."



_id: ObjectId('6429e5dd12ed94112ab3fdcf')

title: "Student taken into custody after three stabbed in Canada high school"


content: "Canadian police have taken a student into custody after three people w..."




_id: ObjectId('6429e5dd12ed94112ab3fdd0')

title: "Why asylum seekers are choosing Canada in record numbers"

content: "Last year, nearly 40,000 migrants crossed into Canada at an unofficial..."



_id: ObjectId('6429e5dd12ed94112ab3fdd1')

 PREVIOUS 1-20 of many results

How Text files are generated:



How the data has been stored In text files:

```

Title: Volkswagen will build its first North American EV battery plant in Canada
Content: Volkswagen\r\n is looking to beef up its battery business\r\n and localize cell produ
-----

Title: Student taken into custody after three stabbed in Canada high school
Content: Canadian police have taken a student into custody after three people were stabbed at
-----

Title: Why asylum seekers are choosing Canada in record numbers
Content: Last year, nearly 40,000 migrants crossed into Canada at an unofficial border at the
-----

Title: Canada repeals historic laws targeting women, LGBTQ community - Reuters Canada
Content: We use cookies and data to<ul><li>Deliver and maintain Google services</li><li>Track
-----

Title: Canada appoints investigator to probe alleged China election ... - Reuters Canada
Content: We use cookies and data to<ul><li>Deliver and maintain Google services</li><li>Track
-----

```

Manual testing of functionality or validation testing

1 . Test Case: Validate the ExtractionEngine's functionality

Steps:

- Run the ExtractionEngine class.
- Observe the printed request URL and response code.
- Check if the JSON response is printed.

Expected Result:

- The request URL should be properly formed with the search query and API key.
- The response code should be 200 (HTTP_OK).
- The JSON response should contain the news articles data.

2. Test Case: Validate the DataProcessingEngine's functionality

Steps:

- Run the ExtractionEngine class, which triggers the DataProcessingEngine.
- Check if the news_articles_.txt files are generated in the project directory.
- Open the generated files and verify their content.

Expected Result:

- Each file should contain 5 news articles or less.
- The content of the files should match the data from the JSON response.

3. Test Case: Validate the TransformationEngine's functionality

Steps:

- Run the ExtractionEngine and DataProcessingEngine classes to generate the news articles files.
- Run the TransformationEngine class.
- Check the content stored in the MongoDB collection.

Expected Result:

- The cleaned and transformed content should be stored in the MongoDB collection.
- The content in the collection should not contain any URLs, special characters, or emoticons.
- Multiple spaces should be replaced with a single space.

Assumption

If you run the program first time, It will create several articles files, having maximum 5 articles in each text files, If you run the program again, it will append the news articles at the last or existing article in the particular article text file, So you can remove all the articles files, after you run the program. With this each time news articles files are created with maximum number of 5 articles.

References and Tools used:

Diagrams.net - <https://app.diagrams.net/>

Atlas MongoDB - <https://www.mongodb.com/atlas/database>