INTE2512 – Object-Oriented Programming

INTE2512: News Scraper Application
Software Requirements Specification

Instructor and tutor: Mr. Minh Vu Thanh

Team name: *needName?*

Students:

Nguyen Nam Cuong (s3891758)

Nguyen Ngoc Minh (s3907086)

Phan Nhat Minh (s3904422)

Tran Mai Nhung (s3879954)

Nguyen Vu Thuy Duong (s3865443)

Date: *05/17/2022*

1. Project description

News Scraper Application is a desktop application that scrapes and displays new articles from the five following Vietnamese online newspapers: VnExpress, ZingNews, TuoiTre, ThanhNien, and NhanDan. The program is designed to have a convenient interface that allows the users to view articles in ten different categories with ease, including Newest, Covid, Politics, Business, Technology, Health, Sports, Entertainment, World, and Others. Additionally, each category layouts fifty article headers, which means those fifty headers are equally allocated into a five-page pagination system. Since the newest category is fixed as the application's main page, every scraped piece of news is designed to display from the latest to the oldest published date on this category. For each header, there is a possibility of it being assigned to more than one category. The desktop application is quickly responsive as well as significantly efficient in scraping and displaying new articles from the mentioned outlets.

2. Running Instructions

Ctrl + Click to use the website shortcut

Guide For Window:

Step 1: Install IntelliJ IDEA Community Edition: https://www.jetbrains.com/idea/download/#section=windows

Figure 2.1 Download Intellij IDEA

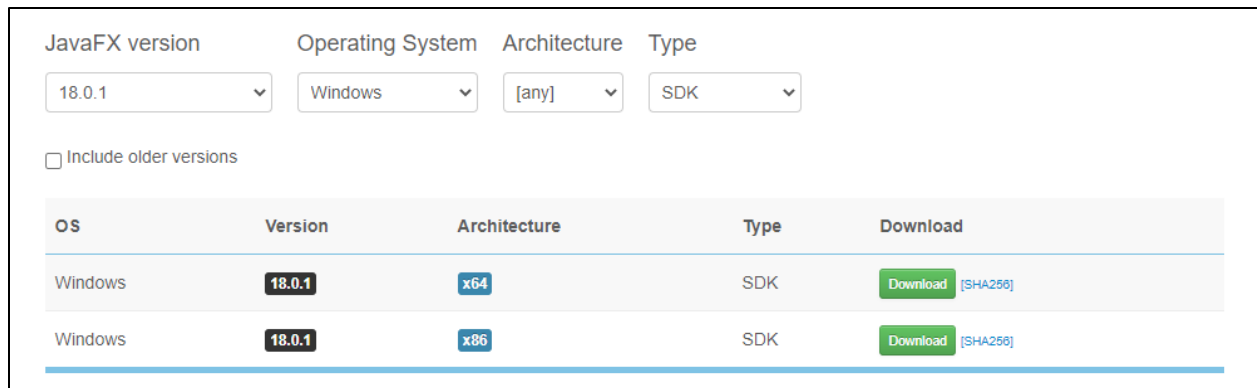**Step 2:** Download JavaFX SDK 18: https://gluonhq.com/products/javafx/

Figure 2.2 Choose JavaFx

Step 3: Download and extract the group project source code folder

Step 4: Open the source code with IntelliJ IDEA. Choose Open -> project folder path -> OK
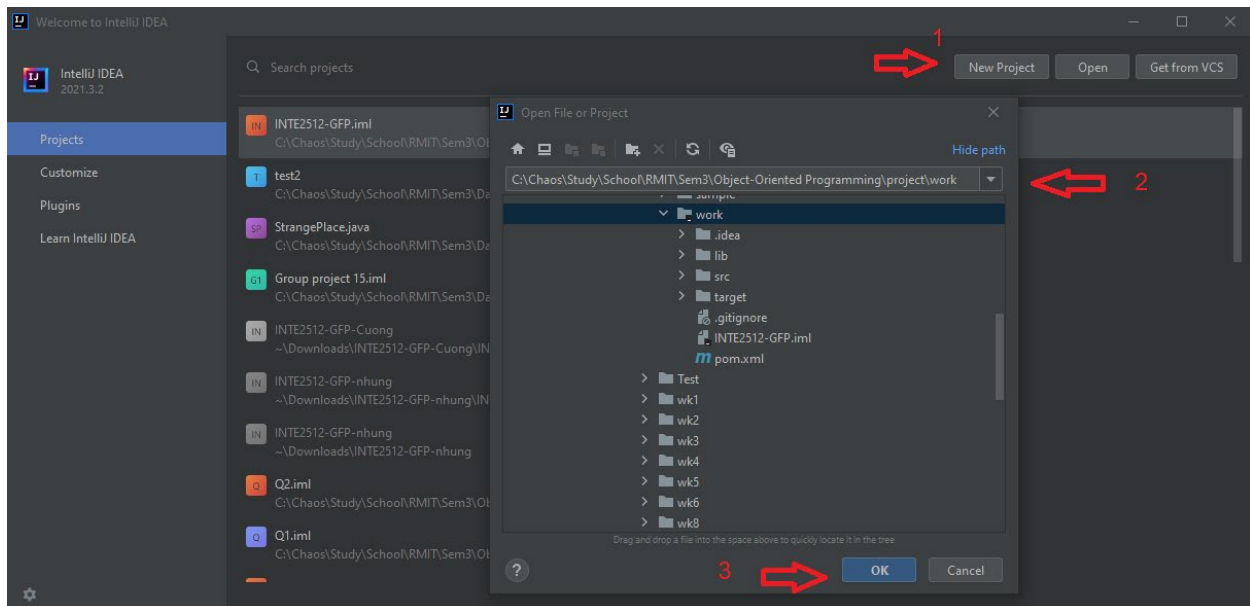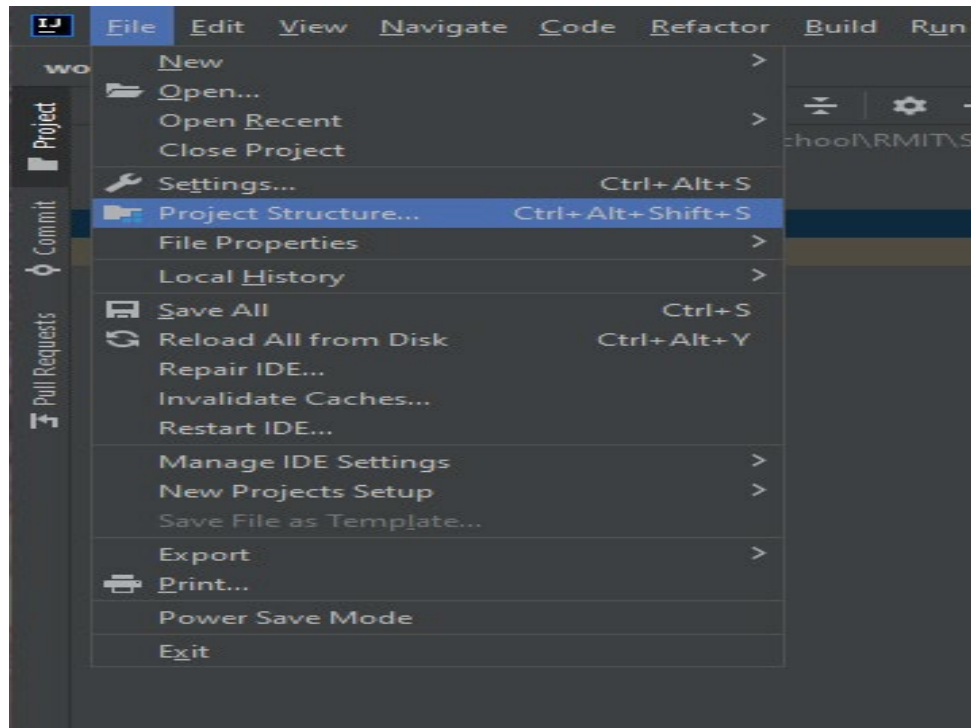

Figure 2.3 Open project

Step 5: In IntelliJ IDEA, choose File->Project Structure or Ctrl + Alts + Shift + S to open Project Structure menu

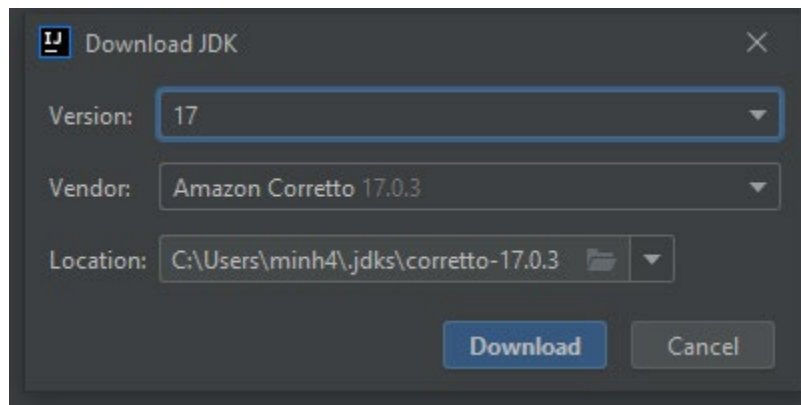*Figure 2.4 Setting Project Structure*

Step 6: In Project Structure menu, drop down to Project Setting -> Project

- In the Project SDK, select Java version 17 from the drop-down menu.
- If the JDK is installed on your computer, select Add JDK and add specify the path to the JDK directory
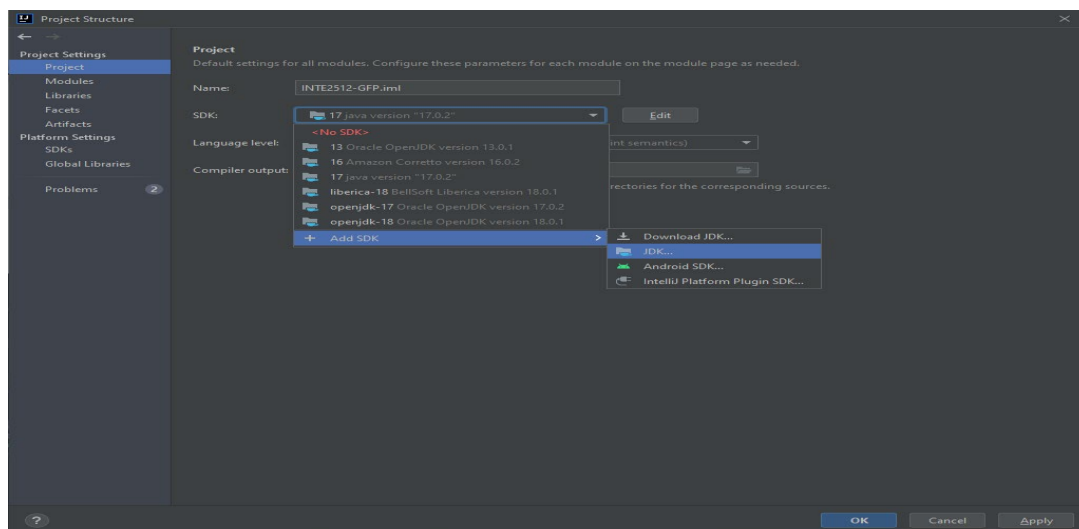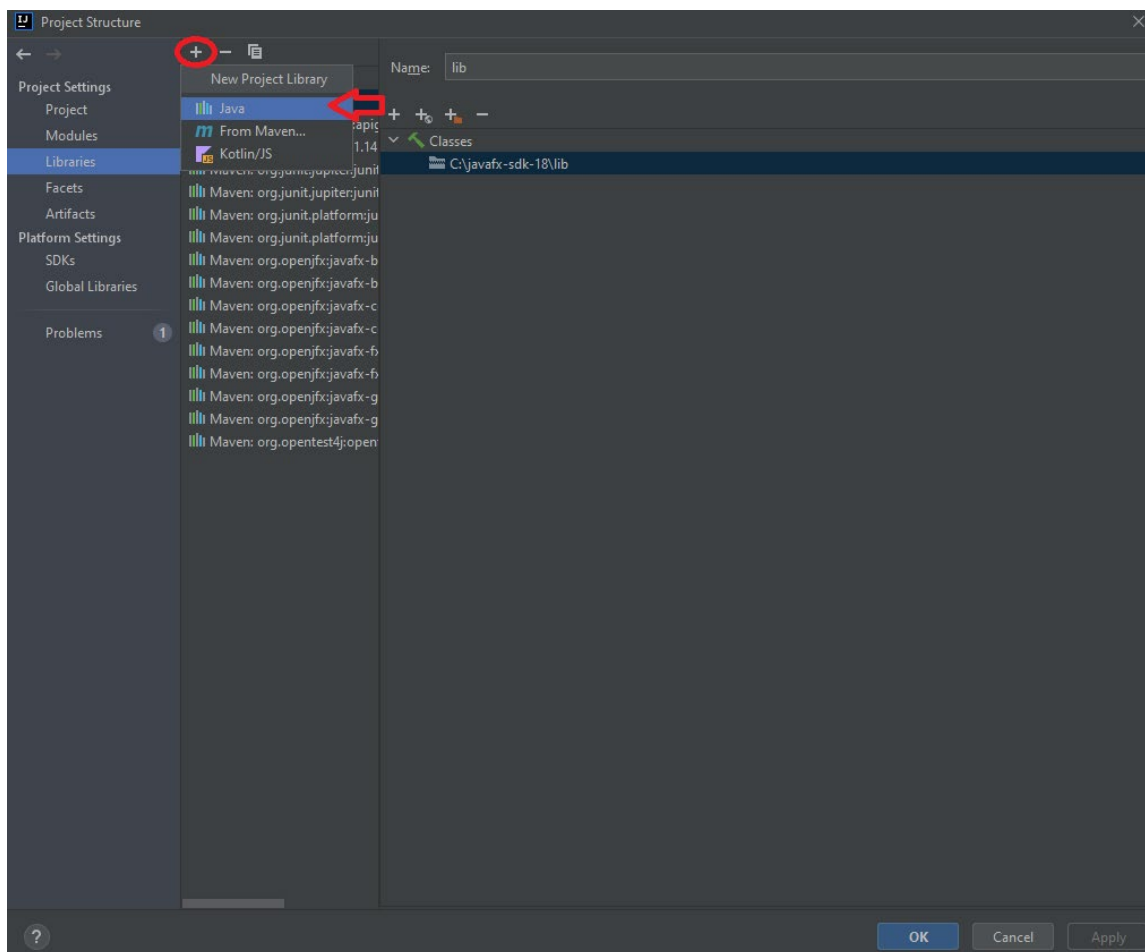

Figure 2.5 Choose SDK on project structure

- If you don't have the necessary JDK, select Download JDK. In the next dialog, select version 17 and change the installation path if needed. Then click Download.
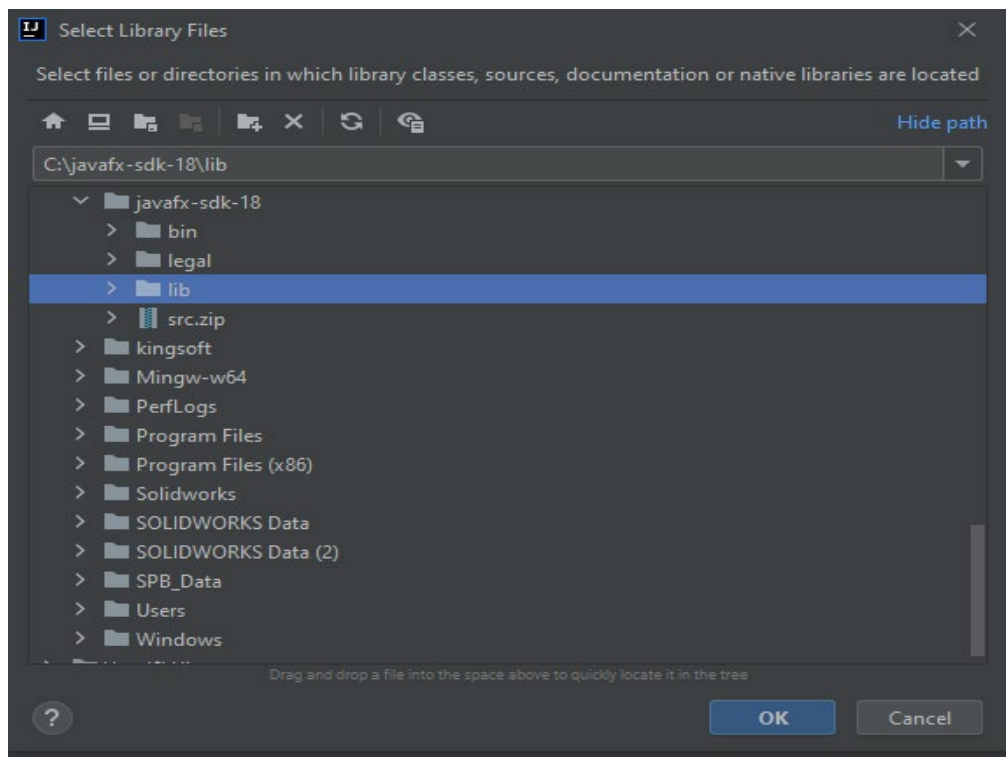

Figure 2.6 Download SDK if project does not have required version

Step 7: In Project Structure menu, go to Libraries -> + -> add Java Library
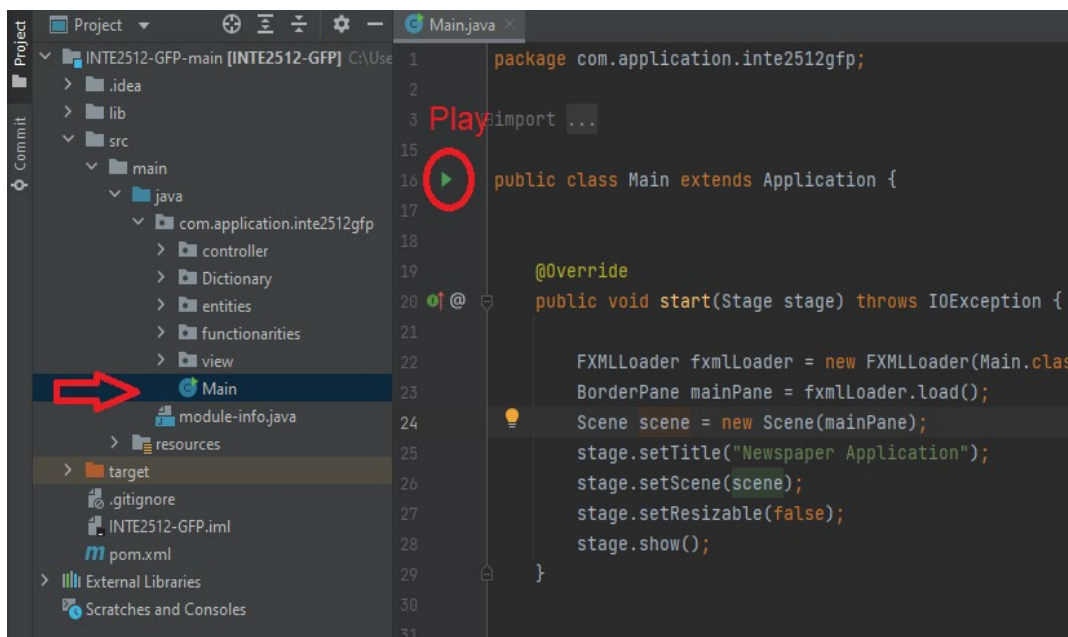

Figure 2.7 Locate JavaFx SDK

Step 8: Locate your JavaFX SDK 18 lib folder and click OK

Figure 2.8 Select Library File

Step 9: Select Apply and OK. Then back to the editor interface, locate Main.java and click play button.

Figure 2.9 Run Application

Guide For MacOS:

Step 1: Install IntelliJ IDEA Community Edition:
https://www.jetbrains.com/idea/download/#section=mac

Figure 2.10 Download Intellij IDEA

Step 2: Download latest version of Liberica JDK: https://bell-sw.com/pages/downloads/#/java-18-current

Figure 2.10 Download Liberica JDK

Step 3: Download and extract the group project source code folder.

Step 4: Open the source code with IntelliJ IDEA. Choose Open -> project folder path -> OK.
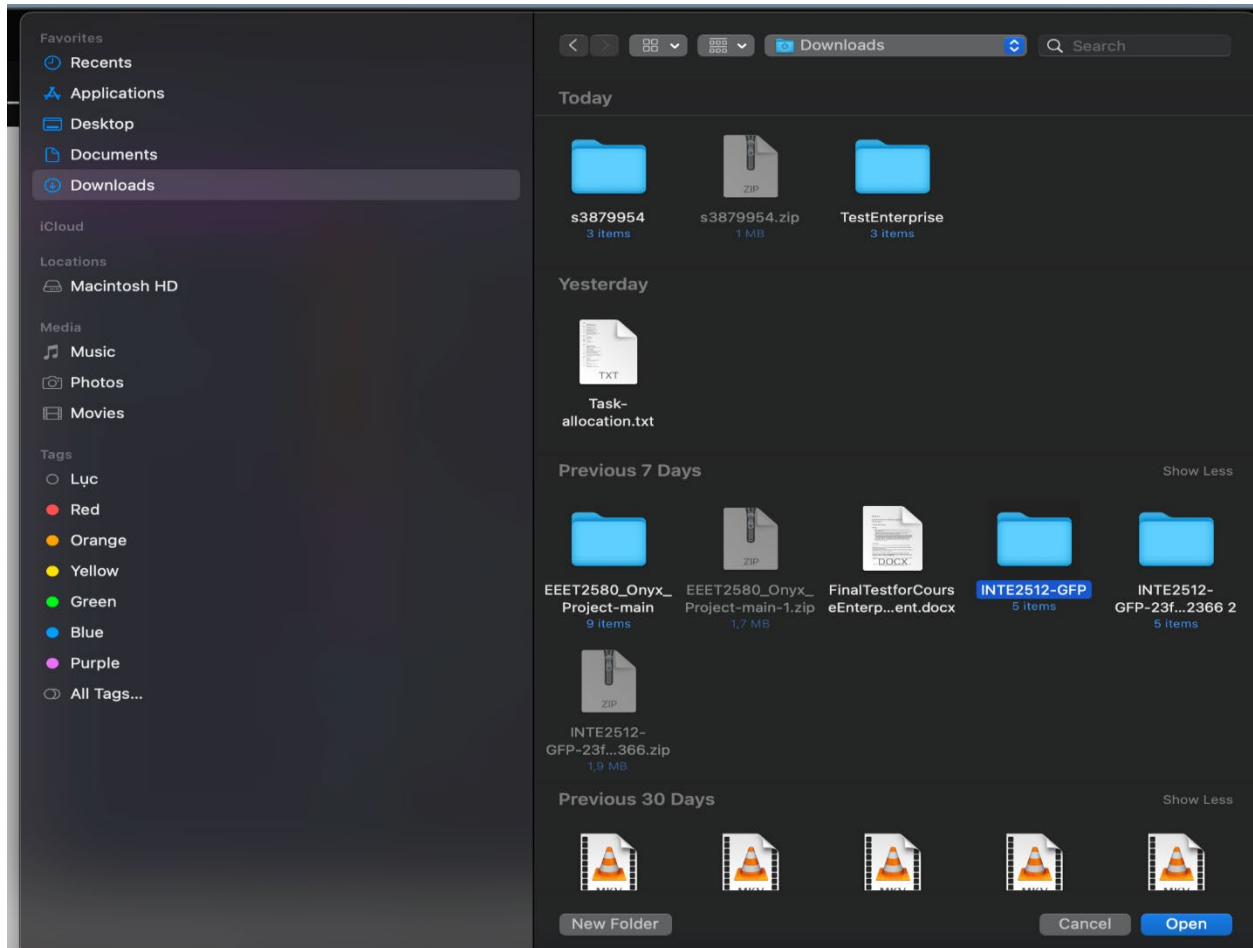
Figure 2.11 Choose project to open

Step 5: In IntelliJ IDEA, choose File->Project Structure to open Project Structure menu

Step 6: In Project Structure menu, drop down to Project Setting -> Project

- Select Add JDK and add specify the path to the Liberica JDK directory.
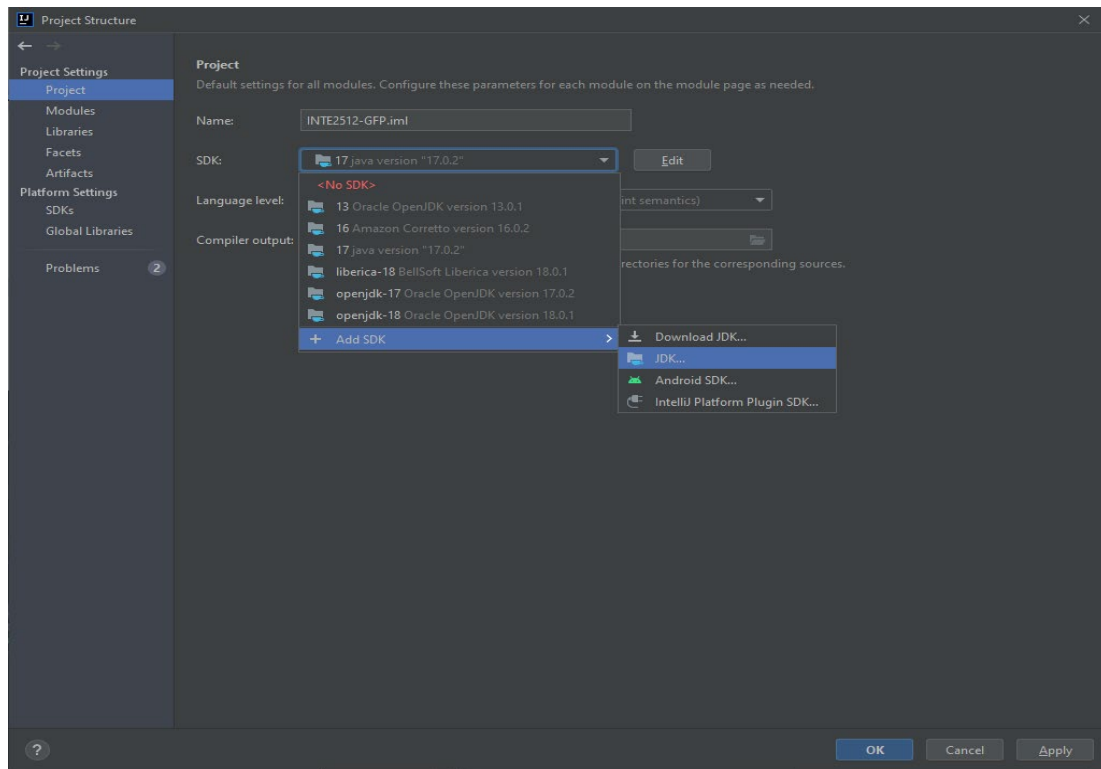
Figure 2.12 Set up project structure
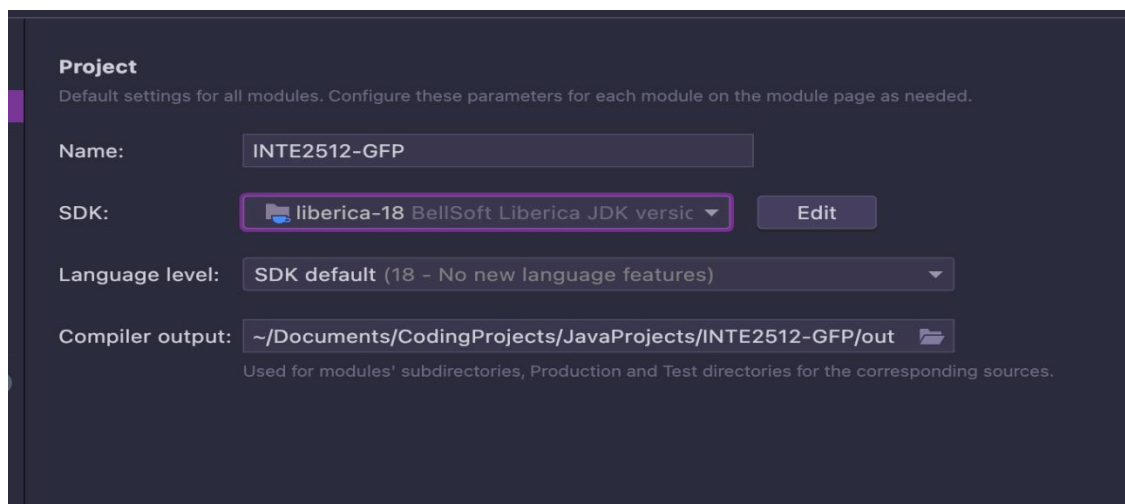
Figure 2.13 Set up SDK

Step 7: In Project Structure menu, go to Libraries -> + -> add Java Library
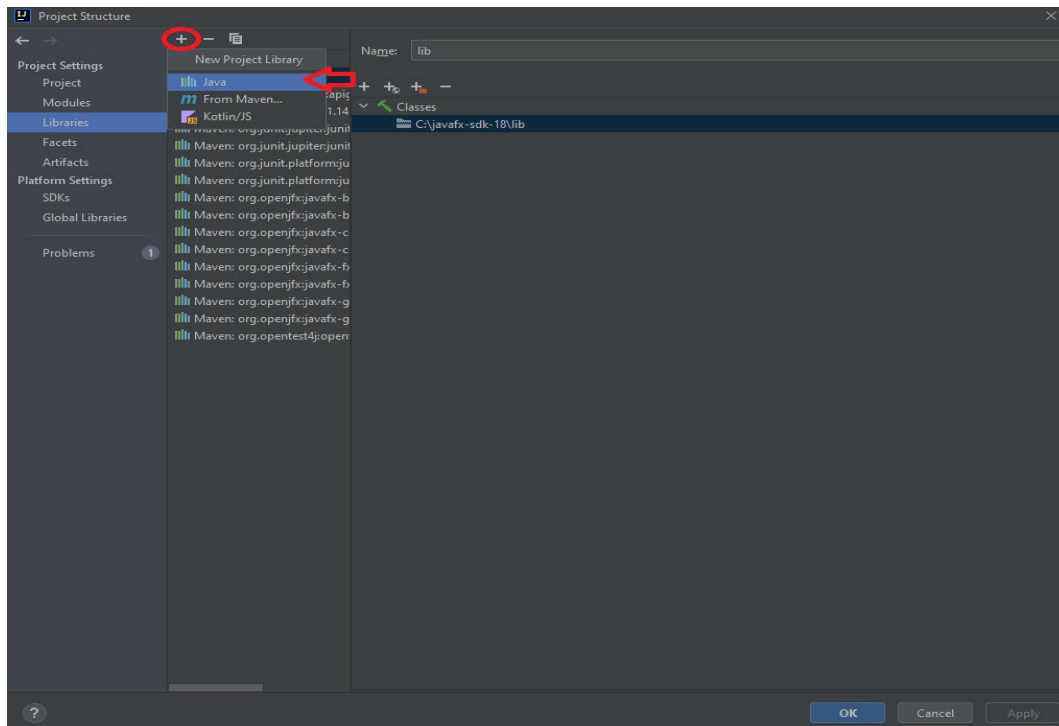
Figure 2.14 Add library

Step 8: Locate your JavaFX SDK 18 lib folder and click OK

Figure 2.15 Select Library File

Step 9: Select Apply and OK. Then back to the editor interface, locate Main.java and click play button.
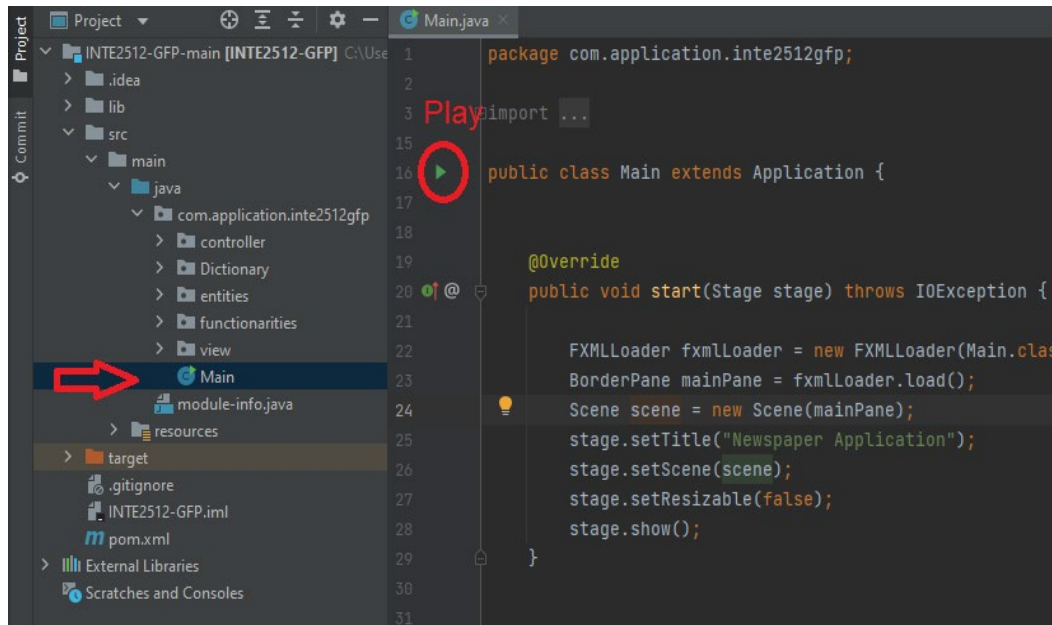
Figure 2.16 Run main application

3. Features Completed

- **The loading screen displays the loading time**
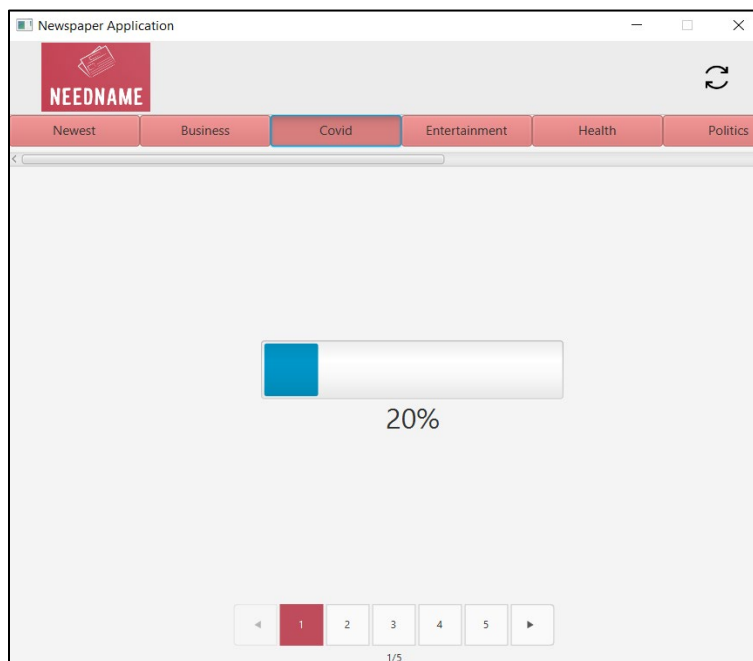
Figure 3.1 Loading Bar Animation

Every time the user switches between categories or clicks to view an article, it is time-consuming to fetch the articles from the outlets. Therefore, a loading bar is implemented to display the percentage of the said progress.
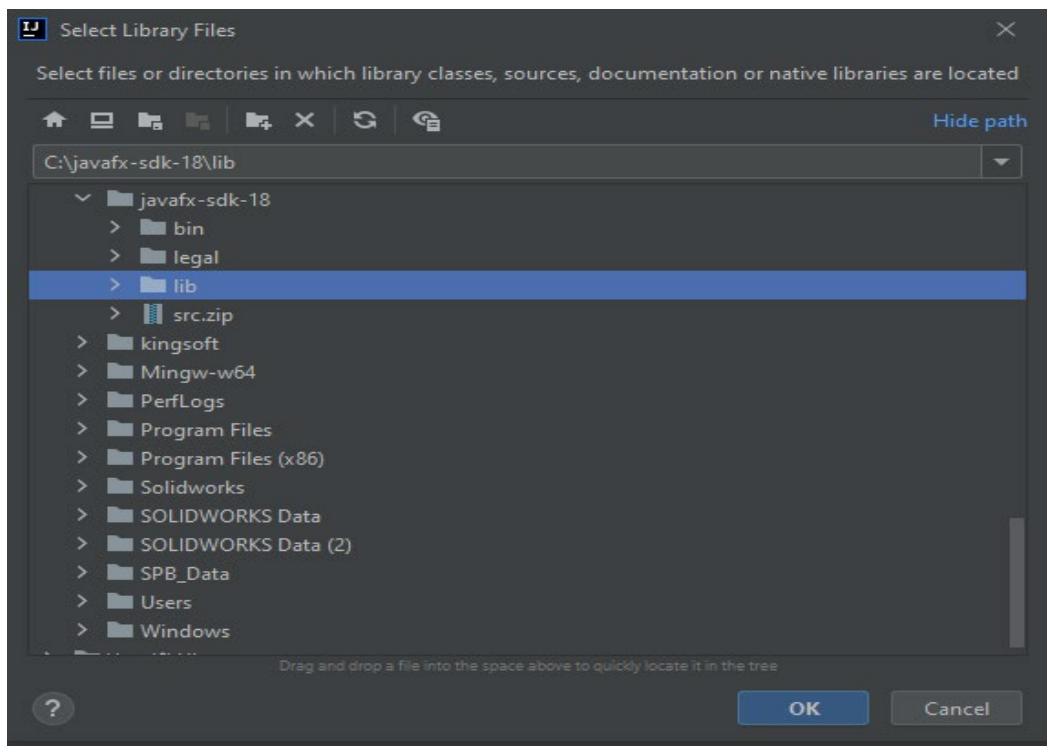
- **Up to date articles from VnExpress, TuoiTre, ThanhNien, ZingNews, and NhanDan up to from 10 categories (each article have at least 2 categories)**
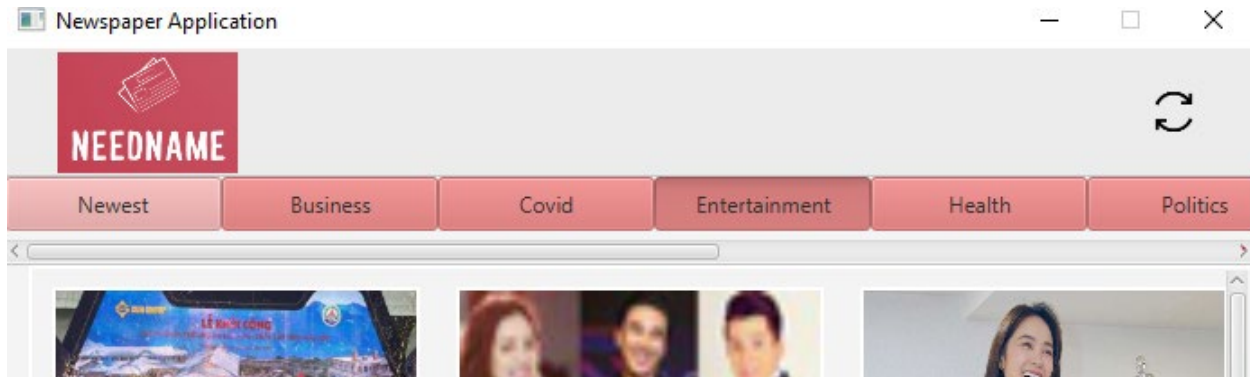
Figure 3.2 Category buttons feature
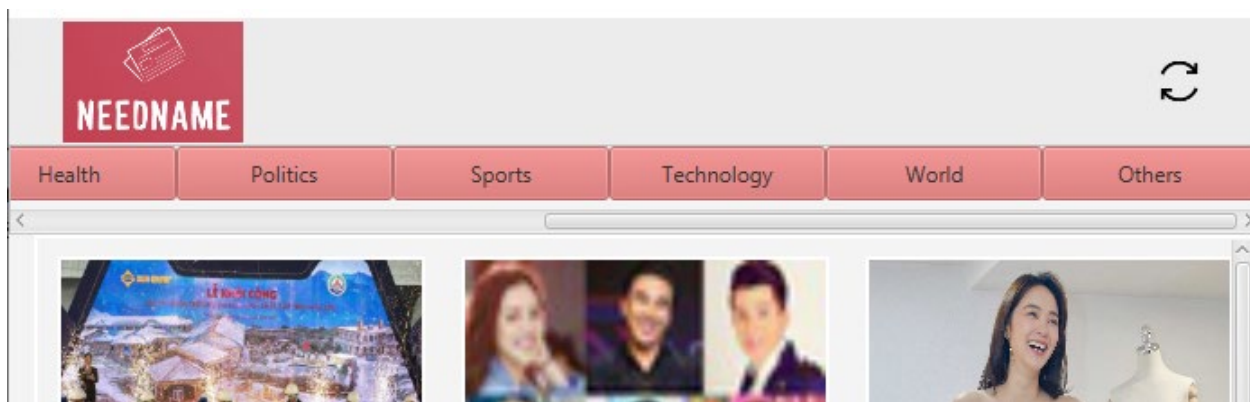
Figure 3.3 Horizontal bar for category buttons

The application provides ten categories on the top horizontal bar. In order to view all categories, the user needs to drag the horizontal scroll bar below the category buttons.

- **Google style paging system**

Figure 3.4 Pagination

With the implementation of a modern-style pagination system, the user can easily view ten articles on each page at a time by pressing on the desired page number.

- **Refresh application**
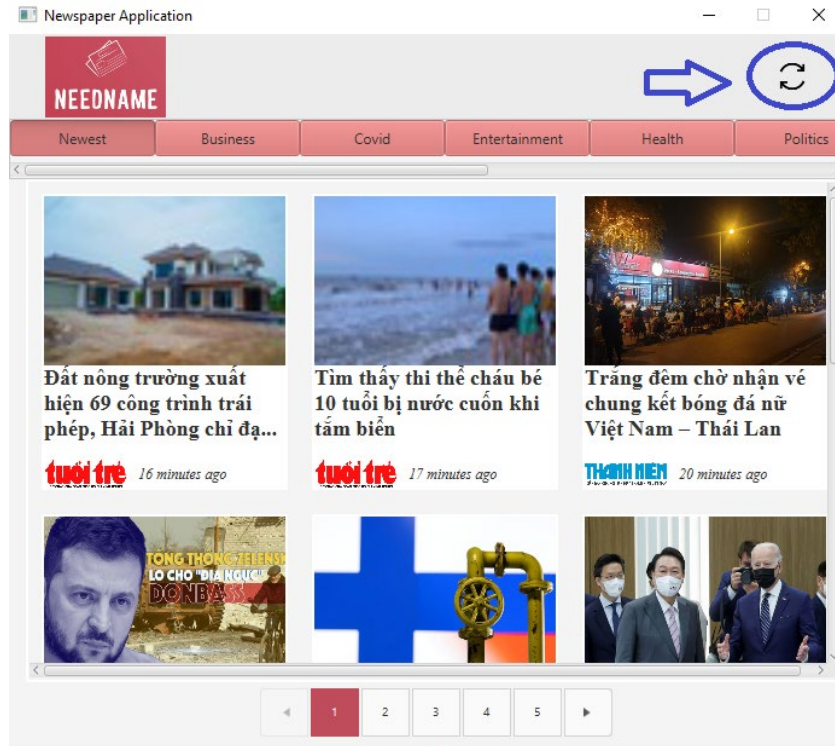
Figure 3.5 Reload button

The primary purpose of the refresh button is to fetch newly published articles at the moment the button is pressed. So that the program can display the latest articles as the users wish.

- **Summary display of article (Title, image, publisher, publish time)**

Figure 3.5 Summary display of articles

The summary display of an article shows basic information such as an image, a bold header, a source, and the duration calculated since the article has been published.

- **Article display**

Figure 3.6 Display article content

Regarding the article content, the application displays the header, the exact published date, the content, along with images. In one exceptional case, specifically in the content of all articles from TuoiTre outlets, a list of other category options and many different links that lead to other articles are displayed by the program. Unfortunately, it is unavailable to view when clicked.

- **Go back button inside article**

Figure 3.7 Back to main button

The "back to main" button is crucial since it redirects the user from an article's content back to the application's main page.

4. Known Bugs and Work-Around

- Unable to interact with videos of some newspapers due to WebView limitations. To handle this, we try to remove any video content as much as possible.
- Null content for some articles such as Live Tonight News, especially for this SEA Games season. Therefore, we changed our approach by eliminating unrelated tags or IDs and displaying the article throughout WebView Engine.
- Slow loading time (about 10 to 15 seconds on average) depending on the Internet transmission capacity. It is recommended to press the reload button to refresh the application. Furthermore, we also specifically let articles on Newest belong to multiple categories to reduce searching categories for articles based on title. This resulted in reducing the loading time before displaying articles.
- Mouse-based application. It is impossible to interact with keyboards due to our early app version.
- Sometimes the new sites change the class, tags, or IDs of HTML elements, or there are some special cases that the program cannot fetch, so the application is not capable of scraping enough required articles. Also, we try to rerun as much as possible to find more of these hidden bugs.
- The main menu is fixed size. The user can only expand the article view.

5. Demo Video (link to presentation video)

The demonstration video is published on YouTube. Please click to the link below to watch the demonstration of our group:

https://youtu.be/e9CLszMQjpE

6. Schedule

In this project, we use the Scrum model as our schedule plan because this project topic is new to us, so the Scum model allows us to make any change when we recognize our current plan is wrong.

- **Sprint 1 (1 Week)**
    - o Analyzing the project topic.
    - o Learn JavaFX, jsoup and Maven.
- **Sprint 2 (2 Weeks)**
    - o Design and draw the class diagram.
    - o Design the application UIs.
    - o Continuing to learn JavaFX, jsoup and Maven.
- **Sprint 3 (2 Weeks)**
    - o Fetch articles summary of required categories in all news sites.

- Develop the main menu UI with required feature such as progress indicators, categories navigation and pagination system.
- Load the article summary to the main menu UI.
- Perform testing on these features.

- **Sprint 4 (1 Week)**
  - Fetch and scrape the article content of all news sites.
  - Load article content using JavaFX WebView library.
  - Perform testing on these features.
- **Sprint 5 (1 Week)**
  - Implement the multithread to speed up the loading time.
  - Optimize code algorithms.
  - Perform test in the entire applications.
  - Fix most of the bugs.
- **Sprint 6 (1 Week)**
  - Comment all the methods and classes in the source code.
  - Write the report for the application.
  - Demonstrate the application.
  - Submit the project.