RMIT Vietnam

School of Science and Technology

INTE2512 Object-Oriented Programming.

SYSTEM DOCUMENTATION

OF A

NEWS AGGREGATOR APPLICATION

Lecturer: Mr. Quang Tran Ngoc

Student 1 Name: Nguyen Hung Anh **Student 1 Number:** s3877798

Student 2 Name: Hoang Phuc **Student 2 Number:** s3879362

Student 3 Name: Thai Thuan **Student 3 Number:** s3877024

Student 4 Name: Phong Tan Le **Student 4 Number:** s3877819

Submission Due Date: 19/9/2021

I. Product description:

A news aggregator application that takes news from VnExpress, ZingNews, TuoiTre, ThanhNien, and NhanDan. The application will display the newest news from the five outlets above. There are 10 categories, each of them has 5 pages and every page has 10 articles to read. The newest category will display the newest news from any category, the Others category will display news that cannot be divided into any category, and Covid, Politics, Business, Technology, Health, Sports, Entertainment, World Category will hold news which relates to that category.

II. Running instructions:

- Step 1: Install and set up IntelliJ IDEA.
- Step 2: Install and unzip Liberica JDK 16 [5].
- Step 3: Open Intellij, click on “Get From VCS”, then copy-paste the Github link: https://github.com/Phuc0906/FinalProjectTeamExpected into the URL box.
- Step 4: Select file -> Project Structure. A new window will be displayed, In the “Project SDK”, click on the arrow -> Add SDK -> JDK-> choose the Liberica JDK 16 (the one was unzipped earlier).
- Step 5: Click on “Build Project” which helps IntelliJ to detect the new JDK.
- Step 6: Go to src -> sample -> main and run the program.

III. Features completed:

1. Include 10 categories:

The application successfully places all the news into 10 categories that are related to *Newest, Sports, Technology, Covid, Health, Business, Politics, Entertainment, World,* and *Others*.

2. Each category has 50 news:

For each category, there will be 50 news. This news is scraped from five sources (VnExpress, ZingNews, TuoiTre, ThanhNien, and NhanDan) in the controller and stored in an array list. Then the news will be sorted by time and 50 latest news will be kept and display.

3. News slot content (Title, cover image, description, source, time duration):

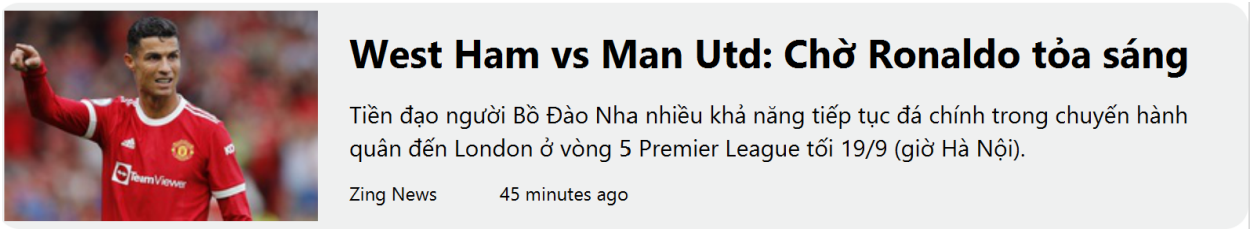

Figure 1: A news slot contains content from Zing News[1].

After the news is stored. It will be distributed into 5 pages (10 news each page). The content will be displayed through FXML elements using set methods. For each set method, the method will take an array list of FXML elements (labels or image views), an integer, and the list of news in the controller as parameters. The method will create the loop, where it will take the news' attributes, starting from the integer (parameter), and set to the FXML elements.

```
<fx:define>
    <ArrayList fx:id="descriptionList">
        <fx:reference source="description1" />
        <fx:reference source="description2" />
        <fx:reference source="description3" />
        <fx:reference source="description4" />
        <fx:reference source="description5" />
        <fx:reference source="description6" />
        <fx:reference source="description7" />
        <fx:reference source="description8" />
        <fx:reference source="description9" />
        <fx:reference source="description10" />
    </ArrayList>
```

Figure 2: FXML elements stored in an array list.

```java
public void setDescriptionList(ArrayList<Label> labelList, int begin, NewsManagement newsList){
    int count = begin;
    for (Label description: labelList) {
        description.setFont(Font.font( family: "Time New Roman", FontWeight.NORMAL,  size: 15));
        description.setAlignment(Pos.TOP_LEFT);
        description.setWrapText(true);
        description.setText(newsList.getNews(count).getDescription());
        count++;
    }
}
```

Figure 3: An example of a set method.

4. The news content (Title, description, texts, images, authors, outlets, time duration):

News content has all the important and necessary information. This can be exemplified by the following figures.

Tuchel tặng nhà cho giúp việc người Philippines

HLV Thomas Tuchel trả toàn bộ chi phí mổ tim cho con của người giúp việc và tặng một căn nhà tại Philippines.

HLV Thomas Tuchel trả toàn bộ chi phí mổ tim cho con của người giúp việc và tặng một căn nhà tại Philippines.

Figure 4: Images taken in a post are displayed by our program, title and description is displayed in the certain order. Besides that, image and image's description is also showed by the program. Image'source: VNExpress [2]

Figure 5: Text and images are shown interlaced that look like the origin. Image's source: Tuoi Tre [3]

Additionally, the **article's source**, **author,** and **time** will be shown at the foot of the article.

5. Intuitive GUI:

There is a menu including 10 buttons for 10 categories at the top of the application. The news is displayed below by rows. Each row has a separated background to be distinguished from another. The navigation for the page (5 pages) is placed under the news. Every section (menu, news, page navigation) will not overlay on another when the resolution changes.

6. Web Scraper:

To adapt to the requirements of the project, the system needs to scrape the articles' content from five outlets which were mentioned above. Thus, the external tool which is applied to the systems to get the content is Jsoup[4]. It is a tool which is using java's syntax and the ability to access the HTML path to get the articles' content. However, each news outlet has a different HTML structure, so the system needs to be divided into five

different scraping methods for five outlets. In each scraping method, there would be five different algorithms to access the HTML content of each outlet into a news list as a class object. Especially, every category in 1 outlet has the same HTML structure, so the system only needs to access the category's links to get the article's URL and article's brief content such as title, image, and description. To separate each category link from the other, the system has 10 classes represent 10 categories of the article.

7. Ability to display news in ascending order

After scraping articles' URLs of a category into a list, the program will execute the sorting method which is a private method inside the category class to short the list. First, it will connect to the articles through their URL to access the HTML structure. Then, it scrapes out the published date and time of the article through the metadata. After that, there will be an algorithm to split the date and time out of the date published format of the article and add to the Time object. However, this method has 2 purposes which are getting the date and time published and check if the article is an error or not. From the article category link, it will contain some error link, which requires the system to detect and eliminate it out of the list. For that reason, in the sorting method, it will have a time list which contains time's component such as day, month, hour and minute, especially the available news matched the time. When the system is scraping the time, if Jsoup cannot connect to the article's link or time published invalid, the method will return null for the Time object, and it will not be added to the Time list. Significantly, the project requirements need to display 50 news per category, so if the method scraped until 50 categories it will interrupt the loop and move to the sorting process. In the sorted project, the Time list will be sort by the primary priority of the month in descending order, if the month is equal, it will get the greater day, otherwise, it will be comparing the earliest article depending on the date and time published. After all, the original news list will be cleared and reorganize by the time object by adding the news object which is the Time class's attribute.

8. Added Features:

At the starting time, the program needs to load the newest news from the main website of the outlets and set the scene components which waste a little time, even reduce the reader's experience on the application. Thus, we add the loading page at the beginning

using the multithreading method. While the system is scraping the content of the outlets, there would be a splash screen showing to get the application to become more impressive. Until the system has finished, the splash screen will close right away and get the user to start suffering on the application.

IV. Known bugs & debug:

1. **Bug:** Scene get bigger after using back button in article window.

Debug: by using stage.sizetoScene(), then stage.setWidth() and stage.setHeight().

2. **Bug:** For the scraping tool, before reading the article's URL, the program needs to scrape it from the article's category URL and place it in a new list. However, while the system is scraping the article's URL, it may contain some error URL which can get the system throws Exception which can get the system to become at risk when it tries to connect to that URL.

Debug: There is a filter that was combined with the sorting method. It will surround the connect method into the try-catch block and when the system through any Exception, it will consider that URL is illegal and return null for that news.

3. **Bug:** News from TuoiTre technology category display blur images in category scene but still display proper images in the article scene.

Debug: Go to the article and get the first image to set as the cover image.

4. **Bug:** Paragraphs in ThanhNien's news can be stored in two different elements <p> or . So sometimes, there will be ThanhNien news that displays no content.

Debug: We fix this by comparing the size of the two elements, and we will scrape the paragraphs from the elements that have a bigger size because the other elements (smaller in size) will contain content that is irrelevant to the article.

5. **Bug:** The author of the articles in Zingnews is encrypted or locked information, thereby cannot use Jsoup to scrape.

Debug: None

V. Project demo video:

Link: https://youtu.be/DrXySD4curk

0:00: Intro to the demo

0:09: Installation guide – **Presenter:** Thai Thuan

5:02: UI design - **Presenter:** Tan Phong

7:46: Features - **Presenter:** Hung Anh

9:51: UML diagram - **Presenter:** Hoang Phuc

VI. Acknowledgment:

[4] Jsoup, "Jsoup: Java HTML parser", Jsoup, 2021. [Online]: Available: https://jsoup.org. [Accessed: 19-Sep-2021].

[5]"Download OpenJDK builds of Liberica JDK, Java 8, 11, Java 17 Linux, Windows, macOS", *BellSoft LTD*, 2021. [Online]. Available: https://bell-sw.com/pages/downloads/?version=java-16. [Accessed: 19-Sep- 2021].

Link:

http://tutorials.jenkov.com/javafx/index.html

https://www.tutorialspoint.com/java/index.htm

https://www.youtube.com/watch?v=9XJicRt_FaI&t=5536s

https://youtu.be/f06uUtkmtDE

https://youtu.be/o-lAsVuskKI

V. References:

[1]*Zingnews.vn*, 2021. [Online]. Available: https://zingnews.vn/west-ham-vs-man-utd-cho-ronaldo-toa-sang-post1264484.html. [Accessed: 19- Sep- 2021].

[2]"Tuchel tặng nhà cho giúp việc người Philippines", *vnexpress.net*, 2021. [Online]. Available: https://vnexpress.net/tuchel-tang-nha-cho-giup-viec-nguoi-philippines-4358376.html. [Accessed: 19- Sep- 2021].

[3]T. ONLINE, "11 tháng 5 ngày: Tuổi trẻ trưởng thành với bao nước mắt, nụ cười", *TUOI TRE ONLINE*, 2021. [Online]. Available: https://tuoitre.vn/11-thang-5-ngay-tuoi-tre-truong-thanh-voi-bao-nuoc-mat-nu-cuoi-20210915181029269.htm. [Accessed: 19- Sep- 2021].