Programming Project 4
75 Points
References: Referenced Textbook and Week 2, 3, 4, 5, 6, 7, 8, 9, 10 & 11 handouts
Due date: Monday, Apr 17 2017, by 11:59 pm

Skills Required:

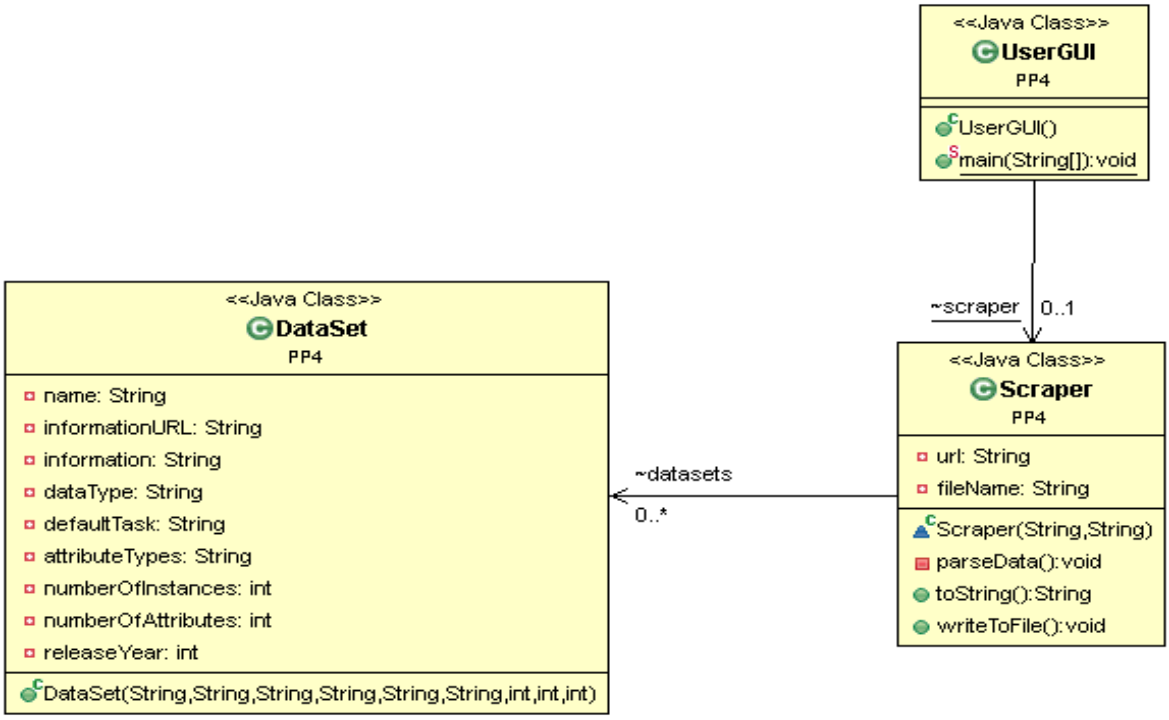
1. Advanced String manipulation techniques
2. Regex, Pattern & Matcher Classes
3. String tokenizer
4. Understanding exceptions, Try and Catch
5. Reading data from the web sources and writing data to files

Description:

1. The program finds specific text strings of statistics in a website using regex
2. A URL link of the web source HTML file for the UCI Machine Learning Data Set Repository is provided

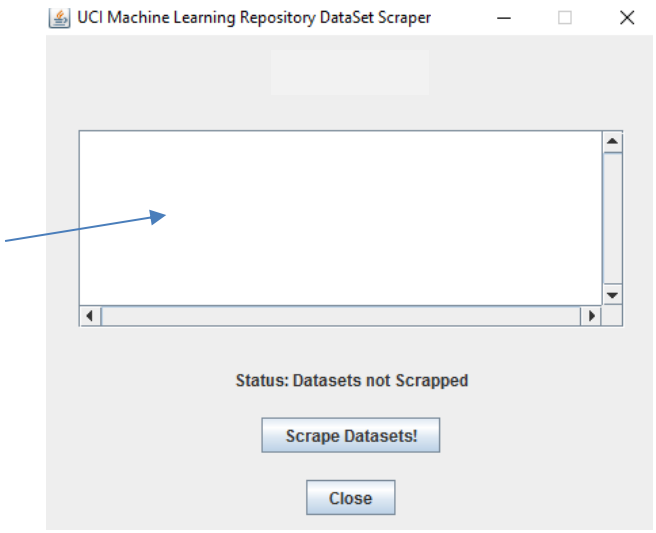
Task Specifications:

1. Create a Java project PP4 in your Eclipse workspace
2. The project has three Java classes as shown above, `DataSet.java`, `Scraper.java`, and `UserGUI.java`
3. The program must be implemented following the provided class diagram model below:



4. All class attributes must be private, that means you need public methods to access class attributes
5. `Scraper.java` has three methods to read and display the web source data
6. The *parseData()* method reads the data from a web page (staring from a given URL) and searches for the match with your defined patterns, using the Java `Matcher` and `Pattern` classes
7. For every dataset, you need to follow another link to acquire the dataset information.
8. Once you have fetched the dataset information, create a DataSet object and add it to the array of DataSet objects as shown in the UML class diagram, `DataSet`
9. The *toString()* method returns the string representation of the DataSet objects in the datasets array in order to display the DataSet information in a GUI JTextArea component in the `UserGUI` class. Note that, each DataSet object must be represented by a single line in the JTextArea component
10. The *writeToFile()* method invokes the *toString()* method in order to store the datasets in a text file. The text file name must be formatted as follows: "output " + *current date and time* + ".txt "
11. The `UserGUI.java` class is the main class that implements the user interface as shown below:

Display songs information in this JTextArea component



12. When clicking on the **Scrape Datasets!** button, your program fetches all the datasets from the Machine Learning Repository and display it in the JTextArea component
13. A user must be able to scroll up/down within the text-area component to read the displayed dataset information
14. The Close button terminates the program
15. `DataSet.java` defines a DataSet object as shown in the UML class diagram; you must follow the same DataSet class contract shown in the UML class diagram

16. Your program reads HTML source page content from the UCI Machine Learning Data Set Repository website using the Java `URL` and `HttpURLConnection` classes
17. Your program opens the URL for the UCI Machine Learning Data Set Repository:
    UCI Machine Learning Repository
18. Capture the HTML source for the page into a String object

19. Find and extract the following dataset information from this dataset website.
    a. Name
    b. URL for the data set information
    c. Data Types
    d. Default Task
    e. Attribute Types
    f. Number of Instances
    g. Number of Attributes
    h. Year of Release
    i. The **first 15 characters** of the dataset information
20. Display the data in a Java application
    a. Enable easy viewing of the list of datasets in the JTextArea component
    b. Use a JTextArea object in your JFrame Windows application
21. Note that acquiring of the data may require you to find and follow another link, such as the dataset information for the current dataset that you will be scraping. (e.g. "`link rel="next" href=`")
22. Each dataset in the repository may contain some missing data. Your program should handle all the missing data attributes and replace it with the default value for an attribute.
23. For example, some of the dataset information might not have their release year mentioned in the website. Your program should not crash while reading such inputs. Instead, the default value of 0 should be set as a release year for those datasets. A similar task has to be performed for all other attributes.
24. Once all the datasets are fetched, the **names of the datasets** will be populated in the combobox.
25. If a user selects a dataset, all the information pertaining to one dataset object will displayed on the JTextArea.
26. Write all the datasets information into a single file, one line per a DataSet object, columns are delimited by commas.

Evaluation Criteria:

1. You must use the class template in your program classes
2. All tasks must be completed to receive credit for this assignment
3. The output should report the correct/expected data values
4. The output file should be formatted for easy viewing

5. The program should not crash while scraping data from the HTML page from NFL.com URL

CIS 611 –

Project IV - Grade Sheet

Group Name(s): _____

Evaluation Criteria	**Comments**	**Max Points Allowed**	**Points Earned**
Program compilation and running (Program should run on your machine during the video call/demonstration): Proper use of the directory structure and the class templates e.g. Scraper, UserGui and Dataset classes with the expected methods (6) Successful compilation and running, meaning no unexpected runtime error, or crash while running the program (2) The program should not crash if the data were scraped instead from another URL, say NFL.com URL (2)		**10**	
Reading and writing data (Demonstrate that it works on your machine during the video call): Reading data from the URL and the use HttpURLConnection class (3) Writing data to a file, use of writeToFile and toString methods (3) Output file is properly formatted for easy reading (2) Writing each dataset information in a single line and		**10**	

delimit columns with commas for easy reading (2)			
General code structure (Walk-through of the code during the video call): Proper use of GUI components., e.g. Text Area to display song information, Scrape dataset button, Close button (3) Proper use of String manipulation, RegEx, Pattern, Matcher, StringTokenizer (3) Proper use of exception handling with Try and catch (3) Proper use of objects and the association, zero to 1, zero to many as shown in the UML (3) Proper scope (private, public etc.) of class attributes and the methods, proper method signature and return type, proper calling of the methods from the other classes (3)		**15**	
Functionality (Demonstrate that the functions work during the video call): Proper fetching of all the datasets and display in the JTextArea component (5) Handling the missing data in the dataset (5) Ability to scroll up/down within the text-area (5) The proper finding of the datasets based on its name (10) Display the first 15 characters of the dataset information in the text description (5) Proper implementation of href, ordering of the datasets (5) Proper functioning of the Close button (5)		**40**	
Total		**75**	