Project Requirements

This project serves as a means for you to demonstrate the amalgamation of a number of technologies that you have studied (or will learn this semester.) Overall, the project is to obtain data from a set of webpages from the website assigned to you. Each student will have a different website. If you do not want the site assigned to you, then provide three website URLs that are not on the list emailed to you and not one of Amazon, Craigslist, Ebay and Abebooks. In this case, I will then choose one of the three provided as alternatives if it has the capabilities of searching, returning results and its results page can be read in the source html.

Requirement1. Online (Search) function. The user should be able to obtain/search for data online from your approved chosen website. Included in this is the ability to download at least one picture or icon found on the website. We will be discussing the code to accessing the webpages and downloading images in upcoming lectures. Decide within the context of your URL what data (and to what extent) can your user search for/obtain from the website chosen. GUIs should be provided for this purpose.

Requirement2. Storage of Data. All data obtained from/for the user should be stored in a storage structure. The program must provide a hash based storage for this purpose. When data is entered into your storage, the user name and a timestamp should be entered with each item of data obtained in Requirement#1. The user should have the option to delete data that they have requested and have been stored prior. In addition, the user should be able to modify the data provided for the user profile (Requirement#11).

Requirement3. Offline queries. (By Requirement#4 next, the data must be maintained between executions of the program. So, if you shut down the program and start it up again, the data is not lost.) The user should be able to conduct offline queries of your data storage via a GUI. The extent of these queries will depend on what you have stored in your storage structure; i.e., the queries will target the specific information you have stored about each item and/or transactional history in your storage. The user should only be able to retrieve data that they requested and was entered to the data storage on the user's behalf. The admin can retrieve data from a specific user(s) or all users.

Requirement4. Persistence of Data. The notion of persistence states that the data is preserved even if the system crashes. The privileged user/admin only from a GUI should be able to reconstruct the storage from backup data (possibly the Transactional Log from Requirement#9, assuming you have enough information in the log to carry out this task.) There are number of ways that persistence can be implemented in Java. You can use any such method that allows reconstruction of the data system.

Requirement5. Processing individual pages. This requirement dovetails Requirement#1. The data obtained from the website will be embedded within the webpages provided by the website. These pages are in html. You are to temporarily store the html *into a temporary file* and then process it. You must write the code yourself and cannot use a third party html API or parser. We will be discussing an approach using regex in class in upcoming lectures. You will be processing at two different types of webpages from the site. The first page is the "initial results" page (first level). This is the page that their server sends in response to your query. It contains a summary of individual results that matches your query and embedded in this page are links to the *actual* individual results pages. The latter are the second type (second level) of webpages you will be processing. You will be extracting pertinent data (text) to be stored into your data storage and images that will be stored locally and the names of the image files and their locations stored in the database. As to which data should be stored will depend of course on which website you are assigned.

Rx`equirement6. Input/Output from Command Line. While you develop your code using a visual editor/debugger utility, the user cannot be expected to learn that development tool. Typically, the compilation of such code and execution of the program will occur from command line (command.exe/cmd.exe on Windows platforms). You certainly can develop your system on such a tool, but you must make sure that it can be compiled

and run from the command line. Till this end, you will need to download the Oracle/Sun Java JDK, which includes the Java compiler 'javac' and the Java runtime 'java'.
Source: http://www.oracle.com/technetwork/java/javase/downloads/index

When running your program from the command line, you should allow the user to pass information or select options via pre-defined command line flags that you will have defined for parsing. (I emailed the class a website that provides pseudo-code outlining how to accomplish this.) Which flags should you minimally provide?
a) –i input_file (where the input_file has an initial set of queries and/or search requests that the user wants the system to process in a batch);
b) –o output_file (where the output_file will be a printout of the essential data stored in your data store as a result of the queries/requests/searches found in the input_file;
c) –p (this feature directs the program to process the input_file and print the output to the output_file without starting up the GUIs, without necessarily storing the data and then immediately terminates
Feel free to add flags to the command line such as –debug etc. but these extra flags will not be considered an innovation.

Requirement7. GUIs. Your system should include appropriate GUIs to enhance the user experience. You may user JavaFX to enhance the user's experience. Development with GUIs will not be considered an innovation.

Requirement8. Documentation. Every project must contain proper documentation and a reasonable amount of code comments. The protocol followed should be as discussed in class: meaningful variable names, class and method descriptions, purpose of variable by declaration and by parameter list of methods, before each block of code, and when appropriate line by line comments. A user manual (.doc/x) should be provided describing the functionality of your project and showing a picture of each GUI provided with a description below it as how to use it. An installation guide should describe how to compile/install your system, listing system requirements. Feel free to add to this list, but this requirement will not be eligible to count as an innovation.

Requirement9. Transaction Log. Every transaction that interacts with the data storage should be written to a text file (the log) that contains the name of the transaction (e.g. INSERT, DELETE, MODIFY) with its parameters (data values), the user that requested it, along with the date/timestamp for security purposes. If the transaction doesn't make changes to the data storage, it need not be logged in for persistence, but you still need to keep it for security logging purposes. The admin is the only one who can directly interact with this log.

Requirement10. User Registration. Because security is such an important aspect of modern software engineering, you are to implement a User account system, with a special account for Administrator who can have access to the transaction log and rebuild system based on persistence. The User account in addition stores data and interact with User's obtained data that is stored in storage structure.

ADD TWO NEW FEATURES TO IMPLEMENT FOR THE ABOVE SYSTEM ON YOUR OWN.

Requirement11. Innovation#1.
There are typically a number of initial result pages provided by their server in response to your query. Have your system go through all (if requested, but be careful testing live, as their system may log you out) or some (user-chosen how many) or predefined limit (system sets the limit). Then, you process the individual (second level) results pages, as described above.

Requirement12. Innovation#2.
If your program implements a search function for data provided by the website, and there exists an advanced search feature on that site, implementing the advanced search completely via your own GUI would be an acceptable innovation