Please note revised due dates on Page 5.

Project #3

Computer Science 2334
Fall 2008

User Request:

"Create a Fast Word Search System
with Import/Export and Load/Save Features and a Graphical Display."

Milestones:

1.	Export word information from the dictionary to a text file.	*5 points*
2.	Implement **Serializable** for the word and dictionary classes.	*10 points*
3.	Utilize Object Serialization to save the dictionary to a binary file. Utilize Object Serialization to load the dictionary from a binary file.	*15 points*
4.	Implement a simple graphical display for showing the number of words in the dictionary with various scores using a histogram.	*25 points*
5.	Use the **HashMap** class to save to and retrieve from the dictionary in memory.	*15 points*

- ► Develop and use a proper design. *15 points*
- ► Use proper documentation and formatting. *15 points*

Description:

An import skill in software design is extending the work you have done in a previous project. For this project you will rework Project 2, using the Java **HashMap** class and add a graphical display. For this project, as with Projects 1 and 2, you will put together several techniques and concepts learned in CS 1323 and some new techniques to make an application that searches a large database of words which we will call a "dictionary" even though it is only a list of words, without definitions. Note that much of the code you write for this program could be reused in more complex applications, such as a spell-checker or an actual dictionary (with definitions).

Importing/Exporting and Object Serialization:
Your software will read in a comma-delimited data file and store the information into the dictionary. The name of the import file that contains the dictionary data will be supplied using Eclipse program arguments by the user. Review Project 1 for how to use program arguments.

This process is known as "importing" the dictionary since this file is not a file of dictionary entry objects but a text file that needs to be parsed to construct dictionary entry objects. Each line in the data file contains the word itself, the language from which the word comes, and a numerical value which we will call "score." The meaning of the score will become apparent in later assignments. For now you may simply treat it as an integer.

Once the file has been read in (imported), your program will enter a loop in which it asks the user for the name of four more files, one at a time as it does its work. This should be done using the technique described in the section below on reading input from the keyboard.

The first of these four additional files will contain a list of queries for which your program should

Please note revised due dates on Page 5.

display information. Each query will be a pair consisting of a word and a language. For each query, your program will find the corresponding score. The result will then be written to the screen and written out to a user-specified "score output" file. The score output file will be the second of the four additional files the user specifies.

Once the results of running the program have been written to the screen and the score output file, the user will be asked for the name of the file to which he or she wishes to export the dictionary. This will be the third of the four additional files the user will specify. Once the user has specified this file name, he or she will be asked for a score on which to filter the exported dictionary. If the user enters 0 , the dictionary will not be filtered – the entire dictionary will be exported to the specified file. If the user enters a positive integer, your program should only export the dictionary entries with a score great than or equal to the specified score. If the user enters a negative integer, your program should export only those dictionary entries with a score less than or equal to the absolute value of the specified score. The format for the export file will be exactly the same as the format of the import file.

After the file has been exported, the user will be asked for the name of a file to which the dictionary should be saved. Saving here means to use object serialization to write the data to the file. This is the fourth and final additional filename for which the user will be asked. Once the user has specified this file name, he or she will be asked for a score on which to filter the saved dictionary. If the user enters 0 , the dictionary will not be filtered – the entire dictionary will be saved to the specified file. If the user enters a positive integer, your program should only save the dictionary entries with a score greater than or equal to the specified score. If the user enters a negative integer, your program should save only those dictionary entries with a score less than or equal to the absolute value of the specified score.

After saving the filtered data, the user will be asked if he or she wants to continue with the program or exit. If the user asks to continue, your program will use object serialization to read from the saved dictionary file (the one created using object serialization), replacing the current dictionary in memory, and return to the point in the loop at which it asks the user for the name of four more files.

Graphical Display:

Producing graphical displays of information can be very useful to users. Therefore, in addition to reading and writing the files as described above, each time the dictionary is loaded (the first import and the subsequent replacements), your program will produce a graphical display of the distribution of the scores in the dictionary. In particular, your program will produce what is known as a histogram. The details of the histogram are up to you (whether vertical or horizontal, width of the bars, etc.). However, to make things easy, you may assume that the scores will only range from 1 to 100.

HashMaps, Sorting, and Searching:

Sorting information can be useful to users because the output may be organized in a way that makes it easier to use. It can also be useful to software developers because it can improve the efficiency of their software. Consider finding dictionary entries based on their word and language. If the data structure holding the word data is unsorted, you need to do a linear search through it to find a word. However, if the data structure is sorted based on word and language, you can do a binary search instead. A binary search will, in most cases, take far fewer comparisons to find the desired entry than a linear search.

Hash tables are an alternate way to quickly store and retrieve data. Java provides the **HashMap** class (among others) which provides this functionality.

To observe this efficiency gain, write three versions of the project. In the first version, leave the data structure holding the dictionary entries unsorted and search through it linearly when the user provides

Please note revised due dates on Page 5.

the query and output file names. In the second, sort this data structure based on the word and language, then do a binary search on it. In the third, you must first add the dictionary entries to a **HashMap** and then retrieve them from the **HashMap**. You will measure the amount of time the system uses in each case for internally handling dictionary data. Internally handling dictionary data includes adding the data to the data structure (list or **HashMap**) and retrieving it. **Do not count the time the system uses for reading in the dictionary data or carrying out other activities, like waiting for the user to provide input.** Run each version of your program 5 times using only the sample provided query file one time for each run, record the values for times used for (1) sorting and (2) searching in each version, and present them in a simple table that includes totals and averages. An example of the table format is shown below.

Version 1	Insert Time	Sort Time	Search Time	Total Time
Run 1				
Run 2				
Run 3				
Run 4				
Run 5				
Average				
Version 2	Insert Time	Sort Time	Search Time	Total Time
Run 1				
Run 2				
Run 3				
Run 4				
Run 5				
Average				
Version 3	Insert Time	Sort Time	Search Time	Total Time
Run 1				
Run 2				
Run 3				
Run 4				
Run 5				
Average				

If you only use the sample query file, does the first, second, or third version of your code spend less time internally handling zip code data? **Why?** If you use many additional query files, do you expect the first, second, or third version of your code to spend less time internally handling zip code data? Why? Rather than creating additional query files, try having your system repeatedly process the same query file. Is there some number of repetitions at which the less efficient version becomes the more efficient version? Is so, approximately what number is that?

Put the table of data and your answers to these questions into MILESTONES.txt under milestone 5.

Please note revised due dates on Page 5.

File Formats:

Dictionary Data File:
Each line of the file contains a single word, followed by a comma, followed by a space, followed by the language from which the word comes, followed by a comma, followed by a space, followed by an integer value. For example:

```
valedictorian, English, 27
```

Score Output File Format:
The text written to the screen and the score output file for each word matching the search criteria must follow the following output format.

Line 1: Name of the word and language searched for.

Line 2: Score of dictionary entry found

Sample Output:

```
Word and language searched for: valedictorian, English
Score: 27
```

Implementation Issues:

File I/O:

To export to a file, use the **FileWriter** class with the **BufferedWriter** class as follows.

```
FileWriter outfile = new FileWriter("output.txt");
BufferedWriter bw = new BufferedWriter(outfile);
bw.write("This is a test -- beep.");
bw.newLine();
bw.close();
```

When you have finished writing to a file, you must remember to close it, or the file won't be saved. If you fail to close the file, it will be empty!

Remember to add '`throws IOException`' to the signature of any method that uses a **FileWriter** or **BufferedWriter** or that directly or indirectly calls a method that performs File I/O.

Object Serialization:

In lab, you did an exercise where you stored a list of objects to a data file using **ObjectOutputStream** and then read them back into the program using **ObjectInputStream**. You are to use this method for saving objects to a file to save the dictionary and to read them back in to the dictionary in memory.

Reading Input from the Keyboard:

In order to get the word information from the user, you need to read input from the Keyboard. This can be done using the **InputStream** member of the **System** class, that is named "`in`". When this input stream is wrapped with a **BufferedReader** object, the `readLine()` method of the **BufferedReader** class can be used to read and store all of the characters typed by the user into a **String**. Note that `readLine()` will block until the user presses the Enter key, i.e., the method call to `readLine()` will not return until the user presses the Enter key.

Please note revised due dates on Page 5.

The following code shows how to wrap and read strings from `System.in` using an **InputStreamReader** and a **BufferedReader**.

```
BufferedReader inputReader = new BufferedReader(
                                    new InputStreamReader( System.in ) );
System.out.print( "Type some input here: " );
String input = inputReader.readLine();
System.out.println( "You typed: " +  input );
```

You need to add '`throws IOException`' to the signature of any method that uses or that directly or indirectly calls a method that uses a **BufferedReader** or **InputStreamReader**.

Due Dates and Notes:

Your revised design and detailed Javadoc documentation are due on **Wednesday, October 15th**. Submit the project archive following the steps given in the submission instructions **by 9:00pm**. Submit your revised UML design on *engineering paper* or a hardcopy using UML layout software, a hardcopy of the index page of your Javadoc documentation, and a hardcopy of the stubbed source code at the **beginning of lab on Thursday, October 16th**.

The final version of the project is due on **Wednesday, October 29th**. Submit the project archive following the steps given in the submission instructions **by 9:00pm**. Submit your final UML design on *engineering paper* or a hardcopy using UML layout software, a hardcopy of the index page of your Javadoc documentation and a hardcopy of the source code at the **beginning of lab on Thursday, October 30th**.

You are not allowed to use the **StringTokenizer** class. Instead you must use `String.split()` and a regular expression that specifies the delimiters you wish to use to "tokenize" or split each line of the file.

You may write your program from scratch or may start from programs for which the source code is freely available on the web or through other sources (such as friends or student organizations). If you do not start from scratch, you must give a complete and accurate accounting of where all of your code came from and indicate which parts are original or changed, and which you got from which other source. Failure to give credit where credit is due is academic fraud and will be dealt with accordingly.

As noted in the syllabus, you are required to work on this programming assignment in a group of at least two people. It is your responsibility to find other group members and work with them. The group should turn in only one (1) hard copy and one (1) electronic copy of the assignment. Both the electronic and hard copies should contain the names and student ID numbers of **all** group members. If your group composition changes during the course of working on this assignment (for example, a group of five splits into a group of two and a separate group of three), this must be clearly indicated in your write-up, including the names and student ID numbers of everyone involved and details of when the change occurred and who accomplished what before and after the change.

Each group member is required to contribute equally to each project, as far as is possible. You must thoroughly document which group members were involved in each part of the project. For example, if you have three functions in your program and one function was written by group member one, the second was written by group member two, and the third was written jointly and equally by group members three and four, both your write-up and the comments in your code must clearly indicate this division of labor. Giving improper credit to group members is academic misconduct and grounds for penalties in accordance with school policies.

This project originally said to use **HashSet**. This was a mistake. Please use **HashMap** for this project.