Assignment 5 – Large Data Sets

Maximum Points = 50

The purpose of this lab is to focus on the reading of data from a large dataset. This assignment also requires the use of search algorithms.

The U.S. Census Bureau maintains a dataset of the most frequently used names (http://www.census.gov/genealogy/names/names_files.html).

Each of the three files, (dist.all.last), (dist. male.first), and (dist female.first) contain four items of data. The four items are:

1. A "Name"
2. Frequency in percent
3. Cumulative Frequency in percent
4. Rank

In the file (dist.all.last) one entry appears as:

```
MOORE        0.312        5.312        9
```

In our Search Area sample, MOORE ranks 9th in terms of frequency. 5.312 percent of the sample population is covered by MOORE and the 8 names occurring more frequently than MOORE. The surname, MOORE, is possessed by 0.312 percent of our population sample.

BASIC ASSIGNMENT

a) Design and implement a program that asks the user to select one of the three datasets of names (last name, male first name, or female first name) one at a time.

b) You can read the contents of a web page with the following sequence of commands:

```
String address = "http://csc.colstate.edu/summers/security.htm";
URL url = new URL (address);
Scanner in = new Scanner(url.openStream());
```

c) Once you have read in the dataset,
   a. allow the user to select a "name" and then display the frequency of the name, the cumulative frequency in percent, and/or the rank. Be sure to provide a response if the name is not in the list.
   b. Find the "name" based on rank
   c. Find the "name" based on frequency

d) Make sure to include necessary constructors, accessors & mutators (gets/sets), and toString methods for all classes.

e) Use a GUI to interface with the user.

   1. NOTE: Some of these methods may throw exceptions—check out the API documentation.
   2. Throw an exception if you find a malformed link (e.g. missing a protocol).

EXTRA CREDIT: Find another large dataset to query.

(Due before class on Wednesday, April 6, 2011) Submit a .doc file containing the UML class diagram showing inheritance for all the classes used in your program. [10 pts]

(Due before class on Wednesday, April 13, 2011) Submit your .java files containing your program to the dropbox in WebCT. [50 pts]

Grades are determined using the following scale:

- Runs correctly..…………………:___/10
- Correct output……..……………:___/10
- Design of output..………………:___/8
- Design of logic…………………:___/10
- Standards……………………….:___/7
- Documentation.………………...:___/5

Grading Rubric (Word document)