SPRING 2002: **COP 3530** Data Structures
[Programming Assignment 3; Due April 9 in my office.]
Hashing

Problem Description

Your task is to write a program to read the first Act of Shakespeare's play *Hamlet* and create a hash table of all the words in the text. After the text is processed, your program should print out the ten most frequent words that appeared in the text.

In the second part of this assignment, use the datafile called `SearchList1` and search in the hash table for each of the words mentioned in that datafile. For each word, report its hash value, the number of times it appeared in the text, the number of entries of the hashtable that were accessed for each search, the load factor, the size of the largest cluster, and the average size of clusters.

In the third part of this assignment (do this only after you get the first two parts working), repeat both the above steps using instead the first Act of Shakespeare's *Macbeth*. Your program should use Java's `URL` class. Your program cannot use a local copy of the data file. Instead it must use the following data file from the world wide web:

`httt://www.cs.fiu.edu/~giri/teach/3503/s02/Prog4/MacbethActI.txt`

The data files `HamletActI.txt` and `SearchList1` will be available on your course homepage soon. Use the `hashCode()` method available for Strings in Java. All words in upper case letters are not to be considered as part of the text. For instance names of the speakers such as `BERNARDO` or `HORATIO` are not to be put into the hash table. Words should be stripped of punctuations. For instance the tokenizer might give you a token such as "`castle.`" or "`king!`". The punctuations should be removed before inserting the word into the hash table, leaving the words as "`castle`" or "`king`".

Use the class `HashSet` (or `HashMap`, if you wish) from your text (not the one provided by Java). You can download the code from Dr. Weiss' website. Your code should use *quadratic probing* for collision detection (as given in the text). You may add extra methods (such as one for computing the average size of the clusters) as needed by this assignment.

There are somewhere between 6000 to 7000 words in the given Hamlet text, including the repeats. For now I suggest that you use table sizes of 8191 (Hamlet) and 4093 (Macbeth), both of which are prime. (Any changes to this table size will be announced in class or on the class website.)

Challenges for the bored

For extra credit, you could try the following problems:

Easy When searching for a word, also output the line numbers where it appears.

Easy Output the ten most frequent words that are not prepositions or conjunctions or names.

Medium Draw pictures of occupied cells of the type shown in Figure 20.5 (page 690) in your text.

Hard Animate the picture of occupied cells as the program reads through the text.

Hard Output the ten pairs of words that appear most frequently on the same line.

Hard Output the ten most frequent 3-word phrases in the text.