Week 9 assignment

Part 1. Working with sets.

The file genelist.txt contains a long list of gene names for 14 species. Each list is headed by ‘Species\_#’ where # ranges from 0 to 13. All the genes for species 0 are between Species\_0 and Species\_1.

Write a program that creates a text file that:

* Has the count of the total number of gene names for each species
* Has the number of unique gene names for each species.
* Gets the number of gene names that are unique for species i that are also among the unique names for species j. Do this for all pairs and print out values in a 14x14 table. Then have the program generate a matrix, with 14 rows and 14 columns with all pairs and write this to the file.

Turn in the program and the output file

Part 2. Hash tables

Write a program that implements a kind of dictionary using a hash function (that you will write) that returns indices of a list, which in turn holds the data.

The hash function must match the length of your list, so that every call to the hash function returns a valid index. For this assignment this number can be much larger than the amount of data, but not larger than 1 million.

The data values will be species names made by reading in Myotis\_aligned.fa. This has 75 sequences so your list must have a length of at least 75.

The keys will be the accession numbers obtained from reading in the fasta file.

e.g. for

>AF376831.1\_Scotophilus\_heathi

The key will be AF376831 and the data will be Scotophilus\_heathi (ignore the actual sequences)

If your list is myotisseqs and your hash function is called myhash() then the location of the data for AF376831 will be found at myotisseqs[myhash(AF376831)]

Your program will:

1. Read in the data file and store the sequences in your list using your hash function
2. Also put the key values (your accession numbers) in a separate list
3. Loop over your list of accession numbers and for each accession number print to the screen and write to a file the accession number and the species named returned from the list using your hash function.

Turn in your program file and the file generated by running the program