COMPUTER SCIENCES

CS 368-1 (2011 Summer) — Day 4 Homework

Due Monday, July 18th, at the start of class.

Description

Write a Perl script that processes an input file, and counts the frequency of the lines (one word per line) it contains. This assignment is designed to exercise your knowlege of Perl's strings, collections, and file handling capabilities.

Details

Your script will need two values, defined at the top of the script: an input file name, and a threshold (number). Something like this:

```
my $filename = '...';
my $threshold = 1000;
```

The input file contains a list of words, one word per line. Your script will read this file, and count the frequency of each of the words. When done, it should then print out all words with a frequency count equal to or greater than the threshold from the command line.

I provide an input file: homework-04.txt for you to experiment with. Your script should run successfully on this file (and ones like it).

The script should be case insensitive — that is, it should treat "the" and "The" as the same word. To do this in Perl, just use the lc function (stands for "lowercase"):

```
my $string = "Mixed Case";
$string = lc($string);
print "$string\n";
```

A good threshold number to use is around 1000. Feel free to experiment with numbers higher and lower than that.

Sample output is below. As usual, the exact format of the output does not matter. **Note:** In the sample output, lines beginning with "%" are meant to show the command line prompt and what you type; they are **not** part of the output.

```
'of': 6568
'with': 1908
'at': 1754
'he': 3128
'a': 6602
'on': 2250
'was': 4220
'and': 6984
```

```
'she': 1727
'in': 4747
'from': 1117
'her': 1559
'be': 1294
'had': 2300
'that': 2640
'have': 1094
'they': 1101
'for': 2021
'it': 2551
'i': 3342
'as': 1663
'his': 1983
'said': 1430
'the': 15700
'by': 1213
'you': 1147
'were': 1194
'but': 1495
'is': 1180
'to': 6981
'not': 1286
```

The key to this is the proper use of Perl's collection types. With proper selection, the problem becomes very straightforward.

Also, the script may be easier to test and debug with the use of a smaller dataset. The dataset provided in "homework-04.txt" is over 250,000 lines. Ultimately, however, once you script works on a small dataset, it should work equally well on the larger dataset. In a Linux or Mac OS X shell, you can create a smaller dataset like this:

```
head -n 1000 homework-04.txt > homework-04-small.txt
```

Replace the *1000* with the number of lines/words that you want.

Reminders

Do the work yourself, consulting reasonable reference materials as needed; any reference material that gives you a complete or nearly complete solution to this problem or a similar one is not OK to use. Asking the instructors for help is OK, asking other students for help is not.

Hand In

A **printout** of your code on a **single sheet of paper** (if at all possible). Be sure to put your own name in the initial comment block of the code. Identifying your work is important, or you may not receive appropriate credit.