**ECE 3822: Engineering Computation II**

**Homework No. 2: Text Processing**

**Goal:** The goal of this homework is to teach you how to do simple text processing using a combination of command line tools available in Linux. Before you attempt this assignment, please study this tutorial:

*https://www.isip.piconepress.com/courses/temple/ece\_3822/resources/tutorials/unix\_for\_poets.pdf*

**Description:** The tasks are:

1. Using only standard Linux commands, generate a histogram of all three-word sequences in the EEG Report database provided (see /data/courses/ece\_3822/current/eeg\_reports). We refer to these sequences as a trigram. Your output should list these sequences in decreasing order of occurrence. Compute the number of occurrences (essentially a histogram), the percentage of time a trigram occurs (the number of occurrences / the total number of trigrams) and a cumulative distribution (which is a useful representation because it shows how many trigrams are needed to cover 80% of the data).

Note that your trigram counter should be case insensitive and ignore punctuation. For example, suppose you have two text files, file1.txt and file2.txt. These files contain the following text:

|  |  |  |  |
| --- | --- | --- | --- |
| file1.txt: | See Jane run. SeeJohn run. | file2.txt | See jane r%un. Se-e jan.E run. |

The trigrams present in this data are:

see jane run

jane run see

run see john

see john run

see jane run

jane run see

run see jane

see jane run

The output of your command line should be:

|  |  |
| --- | --- |
| Trigram | Frequency |
| No. | Percentage | Cumulative |
| see jane run | 3 | 37.5000% | 37.5000% |
| jane run see | 2 | 25.0000% | 62.5000% |
| run see john | 1 | 12.5000% | 75.0000% |
| see john run | 1 | 12.5000% | 87.5000% |
| run see jane | 1 | 12.5000% | 100.0000% |

Trigrams should be counted even when text is split across lines. However, you do not need to deal with beginning or end of file boundaries (edge effects).

Since the list of trigrams you compute for the entire database will be very long, abbreviate your list to show: (1) the 10 most frequently occurring trigrams, (2) the trigrams that occur at the 25%, 50% and 75% percentiles, and the 10 least frequently occurring trigrams.

The output of your code **MUST** contain the columns above but does not need to contain the vertical or horizontal lines. In your document, you can insert the data into an MS Word table.

1. Demonstrate that you can run the command you construct for task no. 1 from within a shellscript. Create a shellscript called compute\_trigrams.sh, insert your command into the file, set the permissions and other properties correctly, run it and demonstrate that it gives the proper output.

This shellscript be run as shown below, must take a root directory as input, search all files below that directory and produce the same output as in task no. 1. It should be run as follows:

compute\_trigrams.sh /data/courses/ece\_3822/current/eeg\_reports

Note that your code should run on any popular version of the Linux operating system and on any machine (I should be able to copy your script to our local Linux cluster and run it).