**ECE 3822: Engineering Computation II**

**Homework No. 1: Command Line Programming**

**Goal:** The goal of this homework is to familiarize you with some basic command line tools to manipulate large repositories of text files. All assignments will be executed on the large text database provided on the course web site.

**Description:** The tasks are:

1. Customize your .profile and .bashrc. Create a directory in your home directory called “bin.” Place an executable in that directory, such a shellscript or a binary file (C program) that prints “hello world.” Modify your path in your .bashrc so that your personal bin directory is included in your standard path. Demonstrate that you can execute the binary in your bin directory by typing it from any directory on your system. Define an alias for a convenient command, “ece\_3822\_d”, that provides a custom directory listing (e.g., “ls -la”). Demonstrate that this is available from any shell in your login session.
2. On the AWS server, you will find data in this directory:

/data/courses/ece\_3822/current/eeg\_reports

The filenames have this structure:

eeg\_reports/01\_tcp\_ar/053/00005391/s003\_2013\_09\_24/00005391\_s003.txt

where “01\_tcp\_ar” and “053” are simply indices used to organize the data, “00005391” is a patient number, and “s003\_2013\_09\_24” indicates session 003 collected on 09/24/2013.

Count the total number of directories in the database by counting the number of sessions (e.g., the last directory in the pathname). Then count the number of sessions that occurred between 2012 and 2015 and have the number “27” somewhere in the filename. Do this with a single command line command.

First demonstrate that your commands work using a small, ten directory subset of the data. Compute the result manually and show that your command gives the same result. Then present your results for the entire database.

1. Each “.txt” file contains text. Find the top 10 words that occur in the entire database. Similarly, find 10 words that occurs exactly 5 times in the database. Show both of these results as a listing where you list the word and the number of times it occurs. For the top 10 words, also show the percentage of the overall data that they account for (the number of times the word occurs divided by the total number of words in the database). Do this ONLY using command line commands.

Do this on a 10-file subset of the database to show that your code works. Show that what you computed manually matches the result of your code. Then present your results on the entire database.