Assignment 5

Name: Erin Ledford

Assignment # : Assignment 5

Due date: February 26th, 2021

Honor Pledge: On my honor as a student of the University of Nebraska at Omaha, I have neither given nor received unauthorized help on this programming assignment.

NAME: Erin Ledford

EMAIL: [eledford@unomaha.edu](mailto:eledford@unomaha.edu)

Partners: NONE

**Objectives**: This assignment will give you the opportunity to practice using the Unix commands and tools discussed in class.

**Exercises**: Using only what you learned in Unix Commands and Tools unit of the course, do the following exercises.

Record your answer to each question along with the commands that you used and the order that you used them in a document (preferably PDF). Name this file netID\_Assign5 where netID is your UNO NetID. Be sure to include your name in the file.

For exercises 1 – 10, copy the file surnames.csv (contains data from the U.S. Census1) from Canvas Files/Assignments/Assignment05/surnames.csv directory. After examining the header line (the first line in the file), erase it—your answers to some of the exercises will be incorrect if this line is not erased.

1. Make a copy of surnames.csv called surnames.tab. Use vim to convert surnames.tab from comma-delimited to tab-delimited; i.e., using search/replace, replace all the comma characters (,) with tab characters (\t). How many lines are there in this file (use wc)?
   1. cp surnames.csv surnames.tab
   2. vim surnames.tab
   3. :%s/,/\t/g
   4. wc -l surnames.tab
   5. There are 151672 lines in surname.tab
2. Erase surnames.csv. Open surnames.tab in vim and convert it from tab-delimited to comma-delimited; i.e., using search/replace, replace all the tab characters (\t) with comma characters (,). From within vim save the resulting file as surnames.csv. How many lines are there in this file (use wc)?
   1. rm surnames.csv
   2. vim surnames.tab
   3. :%s/\t/,/g
   4. :w surnames.csv
   5. wc -l surnames.csv
   6. There are 151672 lines in surnames.csv
3. Open surnames.tab at line 100 using less (will involve using a command-line option; note that the same option works with vim). What is the surname on this line?
   1. less +100 -N surnames.tab
   2. The surname on line 100 is BARNES
4. Pipe (i.e., use a |) the first 1000 lines of surnames.tab into less with line numbers displayed (will involve using a command-line option). Use the arrow keys to go to line 500 of the result. What is the surname on this line?
   1. head -n 1000 | less -N surnames.tab
   2. The surname on line 500 is LANG
5. Pipe (i.e., use a |) the last 1000 lines of surnames.tab into less, jumping directly to line 500 of the result with line numbers displayed. What is the surname on this line?
   1. head -n 1000 | less -N surnames.tab
   2. :500g
   3. The surname on line 500 is LANG
6. Use head, tail and a pipe (|) to display only line 2000 of surnames.tab. What is the surname on this line?
   1. Head -n 2000 surnames.tab | tail -n 1
   2. The surname on line 2000 is STRINGER
7. Use cut, grep, a pipe (|) and output redirection to create a new file, unpopularNames.txt, that contains the surnames and the count columns only of entries with a count equal to 150436. How many lines are there in this file (use wc)?
   1. Grep ‘150436’ | cut -f 150436 > unpopularNames.txt
   2. There are 1236 lines in this file
8. Use sort, head, and a pipe (|) to sort the surnames alphabetically and identify the fourth from the top.
   1. Sort surnames.tab | head -n 4
   2. The fourth name is AABY
9. Use sort, head, and a pipe (|) to sort the surnames alphabetically and identify the fourth from the bottom.
   1. Sort surnames.tab | tail -n 4 | head -1
   2. The fourth from the bottom is ZYSKOWSKI
10. Use cut, sort, a pipe (|) and output redirection to create a new file, popularNames.txt, that contains a reverse alphabetized list of the 20 most common surnames (only). What is the first surname in this file?
    1. cut -d $‘\n’ -f1-20 surnames.tab| sort -r > popularNames.txt
    2. The first name in this file is WILSON

For exercises 11 – 15, copy the file bnc-wordfreq.csv, the frequency of words in the British National Corpus (BNC), from Canvas. (The BNC is a 100-million-word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century; bnc-wordfreq.csv is a partial summary of the BNC.) After examining the header line (the first line in the file), erase it—your answers to some of the exercises will be incorrect if this line is not erased. Note that some questions may involve modifying the file in some way, may take more than one step, and/or may require a manual calculation.

1. How many total characters are in the file?
   1. Wc -m bnc-wordfreq.csv
   2. There are 125339 characters in the file
2. How many nouns (part of speech equal to ‘n’) are there in the BNC?
   1. Grep -n ‘n’ bnc-wordfreq.csv
   2. There are 6318 nouns
3. Use a single command to display the most commonly used word (only) in the BNC. What is this word? Use a single command to display the least commonly used word (only) in the BNC. What is this word?
   1. Head -n 1 bnc-wordfreq.csv
   2. The most frequently used word is a
   3. Tail -n 1 bnc-wordfreq.csv
   4. The least frequently used word is zone
4. There are 428 words that start with the letter a in the BNC. Taking into account this fact, use a single command to determine what the first word (alphabetically) starting with the letter b is.
   1. grep -m1‘\bb’ bnc-wordfreq.csv
   2. The first word alphabetically starting with the letter b is baby
5. There are 306 words that start with the letter b in the BNC. Taking into account this fact, use a single command to display what the most-commonly-used “b-word” is (only).
   1. Sort -n -r bnc-wordfreq.csv | grep -m1 ‘\bb’
   2. The most commonly used word is bill

For exercises 16 – 20, copy the file MTgenome.fna, the sequence of human mitochondrial genome, from Canvas. Note that some questions may involve modifying the file in some way, may take more than one step, and/or may require a manual calculation.

1. What is the length in bp (only) of the human mitochondrial genome? (Check your answer here: https://en.wikipedia.org/wiki/Human\_mitochondrial\_genetics). Note that wc includes newlines in the count...
   1. Grep -v ‘>’ | wc | awk ‘{print $3-$1}’
   2. The length in bp is 16569
2. Without using a text editor (e.g., vim), how many nucleotides are in the shortest line in the MTgenome.fna file?
   1. Awk ‘{print length}’ MTgenome.fna |sort -n | head -n 1
   2. The shortest line is made of 49 nucleotides.
3. Using vim, determine how many A nucleotides does this sequence of the human mitochondrial genome contain. (Hint: when doing a search/replace, vim, reports the number of substitutions made).
   1. :%s/A/B/g
   2. There are 5124 A nucleotides
4. How many start codons (ATG) does MTgenome.fna contain? You do not have to worry about a codon being in frame or those on the complementary strand. (Hint: using vim, put the sequence all on one line, then see the hint for question #18).
   1. :%s/ATG/BCD/g
   2. There are 156 ATG start codons
5. One sequence line of the file doesn’t contain any Gs. What is that sequence line? (Hint: you will need to use grep, twice).
   1. Grep -n -v ‘G’ MTgenome.fna | grep -v ‘>’
   2. The sequence line is 122

**How to Submit**: Upload your netID\_Assign5 file to Canvas.