BIOI 3500, *Advanced Bioinformatics Programming*
Assignment 5

Assignment 5 :: 50 points :: Due: 11:59 p.m. Friday, 26 February 2021

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. Census[1]) 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**)?

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**)?

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?

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?

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?

6. Use **head**, **tail** and a pipe (**|**) to display only line 2000 of **surnames.tab**. What is the surname on this line?

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**)?

8. Use **sort,head,** and a pipe (**|**) to sort the surnames alphabetically and identify the fourth from the top.

9. Use **sort,head,** and a pipe (**|**) to sort the surnames alphabetically and identify the fourth from the bottom.

10. Use **cut**, **sort**, **head,** and 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?

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.

11. How many total characters are in the file?

12. How many **nouns** (**part of speech equal to 'n'**) are there in the BNC?

13. Use a single command (i.e., multiple Unix tools in a pipeline) 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]Obtained from *http://introcs.cs.princeton.edu/java/data*

14. There are 428 words that start with the letter **a** in the BNC. Taking into account this fact, use a single command (i.e, multiple Unix tools in a pipeline) to determine what the first word (alphabetically) starting with the letter **b** is.

15. There are 306 words that start with the letter b in the BNC. Taking into account this fact, use a single command (i.e., multiple Unix tools in a pipeline) to display what the most-commonly-used "b-word" is (only).

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.

16. 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...

17. Without using a text editor (e.g., **vim**), how many nucleotides are in the shortest line in the **MTgenome.fna** file?

18. 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).

19. 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).

20. One sequence line of the file doesn't contain any **G**s. What is that sequence line? (Hint: you will need to use **grep**, twice).

How to Submit

Upload your **netID_Assign5** file to Canvas.