Module 3

FailLog

The hard and frustrating parts of module 3 came from the messy data rather than the commands. Like they say practice makes better - when my commands don't work I know now to either leave if, i've tried a bunch of times to fix it or closely examine the codes for minor details; a full stop here, colon there might just be the problem.

Exercise 1:

To start off, we had to extract the data needed for the exercise. To do this, I used the code **curl** instead of **wget**. The two codes are quite similar but also slightly different. I learned that the main difference is that curl has a library whereas wget is just a command line, no library. It's also stated that it's harder to create a library than it is to make a 'mere' command line. So here the curl code helps to pull the data and store it in the texas.txt file. This part was simple enough and easy to follow. Now the next instruction got me confused, cause I was reading the instructions and following the video tutorial; only to realize the insturctions was more detailed than the video.

I opened up the text using nano in dhbox and followed the steps by highlighting the unnecessary texts **ctrl+shift+6** but I kept getting confused. At first i thought i wasn't doing it right and the cutting part should've gone much faster. Took a while for it to click in my head that there's just a ton of the text to cut out the texts.

I opened the file on BBEdit to see if that would make it go faster but either

I wasn’t doing it right or it was just still slow to clean out. Finally finished cutting out the unnecessary text with **ctrl+k**, resaved it and re-uploaded it into the dhbox. The text was a mess still and there were alot of things we had to be rid of.
Using the command
$ grep ‘\bto\b’ texas.txt
we were able to find all the lines that contained the word ‘to’ Went ahead to copy the last 10 lines of the texas.txt on dhBox into RegExr. It showed all the lines (highlighted in red).
$ sed -r -i.bak ‘s/(.+\bto\b.+)/~\1/g’ texas.txt into dhBox which also helped to create a backup file in case a mistake occured! Thankfully none did and I was able to navigate my way to the end of the exercise and I got results; seperated senders list from receiers and transformed my file into CSV format.

I looked through the file on Numbers and I found a few errors like:
James Webb [April 7 1839] in my data and I realized i made a mistake somewhere but for the most part, my data was better organized and understandable.

Excercise 2
Downloaded OpenRefine on my laptop. Thought it would come up like BBEdit did, then realized it actually opens up on the browser.
Using the Texan file from Exercise 1, i played around with the drop down menu and saved my file in github as module2exercise2.csv

I attempted the optional exercise: **Exploring other Named Entity**

Extraction tools and I find it so confusing. I couldn’t find the save icon for some strange reason and i’m still in the process of figuring out what I’m not doing right, which is my biggest **fail** for now and it’s honestly really frustrating to me cause it’s something as simple as a *save* icon. So for now, that exercise is incomplete but I am able to see graphs and certain word analysis on there.