Module 3 – Open Notebook/Fail-Log

Exercise 1: Regular Expressions

- This Exercise taught the usefulness of Regex Search and Replace management, as it is used through DH Box and our Text Editors
- I completed this Exercise without any error messages etc., to my great surprise and appreciation
    - However, I was rather confused in a number of places
        - The most difficult part was attempting to combine numerous regexes into a single search/replace function line
            - I had completely read, and created a personal summary, of the Exercise's introduction, so…
            - For experimentation sake, I attempted to calculate the required regexes on my own, and tested them in *RegExr: Learn: Build & Test RegEx* (https://regexr.com/)
            - Unfortunately, I failed in that regard
            - I could not figure out exactly which regexes were required, nor in what order the various regexes needed to go
        - Thus, the additional YouTube videos in the Workbook provided me with any and all help that I may have required in the various sections for this Exercise
            - They were invaluable in this regard
- Grabbed the index from the correspondence of the Republic of Texas
    - OCR'd text found at http://archive.org/stream/diplomaticcorre33statgoog/diplomaticcorre33statgoog_djvu.txt
    - Downloaded file using "curl" at the Command Line, then pushed it into a new file "texas.txt"
        - $ curl http://archive.org/stream/diplomaticcorre33statgoog/diplomaticcorre33statgoog_djvu.txt > texas.txt
        - Downloaded the file, opened it in Sublime Text, and deleted everything before and after the Index of Letters
        - Reuploaded the file back into File Manager
- Editing the text via the Command Line
    - Selected every line with the word "to" in the "texas.txt" file
        - $ grep '\bto\b' texas.txt
    - The following code was implemmented
        - $ sed -r -i.bak 's/(.+\bto\b.+)/~\1/g' texas.txt
            - This code created a parenthetical group named "\1" for each of the lines
            - It also added a tilde ("~") to the start of each group/line
            - New file "texas.txt.bak" created
            - The individual pieces themselves…
                - -r → The "extended regex" saves us from having to 'escape' certain characters
                - -i.bak → created a backup of the original input file
                - -'s/old-pattern/newpattern/g' → how we find and switch what we're looking for

        - The final "g" → 'globally', in the file
    - texas.txt → file name looking to be changed
- Deleting all lines without a tilde
    - grep found all the lines that had a tilde in them, and wrote them into a new file called "index.txt", entered new file to confirm
        - $ grep '~' texas.txt > index.txt
        - $ nano index.txt
- Used the following command to remove the page numbers, found after the years on each line, as well as the commas between the years and the months-dates
    - $ sed -r -i.bak 's/(,)([0-9]{4})(.+)/\2/g' index.txt
        - The comma, as well as the space following, captures the comma prior to the year in the line search
        - [0-9] → searches for all numbers between 0-9
        - {4} → searches for numbers containing 4 digits
        - .+ → captures the entire rest of the line
        - Parenthetical groups are created with the various "()"
            - comma as the first group → "\1"
            - the space and the year as the second → "\2"
            - rest of the line as the third → "\3"

- Could not continue work for the day
    - Created a .html file for my history of commands of everything conducted today
        - File name "Module 3 – Exercise 1 – commands1.html"
        - Uploaded to GitHub repo
- Editing the text via the Command Line continued…
    - Replaced the tildes with nothing to delete them
        - $ sed -r -i.bak 's/~//g' index.txt
    - Separated the Sender and Receiver by a comma
        - Replaced all the words "to" with a comma
            - $ sed -r -i.bak 's/(\b to\b)/,/g' index.txt
                - Space before "to" → searches for "to", specifically with a space before it
                - \b → start/end of the word
                - Comma → symbol intended to replace parenthetical group
    - Downloaded file via the File Manager, then deleted the File Manager's copy
        - Opened file in Sublime Text then copied text to *RegExr: Learn: Build & Test RegEx*
            - Added ".+,.+,.+" to the Expression Line
            - Sifted through the text to the appropriate lines
                - deleted all unnecessary commas and excess information
            - Copy/replaced Sublime Text version with edited text
            - Added new header line: "Sender, Recipient, Date"
            - Uploaded file back to File Manager
            - Checked nano to ensure success
    - Made a copy of the file in a CSV format, titled "cleaned-correspondence.csv"
        - $ cp index.txt cleaned-correspondence.csv
    - Downloaded CSV file to my computer

        - Renamed file to “Module 3 – Exercise 1 – cleaned-correspondence.csv” for personal categorization purposes
            - Note: this was only done for the downloaded copy, the DH Box copy retains the original title
        - Uploaded file to hist3814o repo
- Created a new history of commands file for the remainder of the Exercise
    - “Module 3 – Exercise 1 – commands2.html”

Exercise 2: Open Refine

- This Exercise taught me to use Open Refine, a program which allows us to clean up our messy data
    - E.g. of Messy data → “Shawn” and “S4awn” are probably the same person
    - At first everything was rather confusing, and the Workbook even claimed I should have had less unique names in my list than I actually did
        - Was a mistake made in Exercise 1? I don’t think so…
- Watched 3 YouTube videos concerning Open Refine, found at: http://openrefine.org/
    - The first and third videos were much easier to understand and made me think that the program will make life much easier!
    - However, the second video, with all it’s technical aspects was very confusing. At one point, I caught myself physically vocalize: “Wait, what? You lost me lady…”
- Downloaded and opened Open Refine
    - Created a new project via uploading the “cleaned-correspondence.csv” file
        - Named the project “Module 3 – Exercise 2 – cleaned-correspondence”
    - Clicked on the arrow to the left of "Sender"
        - Selected Facet → Text Facet
    - Repeated action for “Recipient”
    - Within the "Sender" facet box
        - Clicked on “Cluster”
            - The Workbook claims we should have 189 unique names, but Open Refine counted 192, reduced to 156 after the merge
                - Not sure if I made a mistake in Exercise 1 or not, but the ussie seems to be rectified since the Workbook states +/- 150 names should exist post-merge
        - Merged and Re-Cluster any “Sender” name clusters that misspelled a name and was attempting to merge it incorrectly
            - One cluster was recommended to be merged as “Joaquin G. Rej6n”, since neither of the choices spelled the full name correctly
                - I tried Googling “Joaquin G. Rej” to find the correct spelling but nothing substantive came up, so I kept it the way Open Refine suggested for now
            - One cluster only had two instances/rows in which a genuine option was presented
                - I thus Googled the first option “Geo. L. Hammeken” to ascertain if it was the correct one

        - The Texas State Historical Association (TSHA) would suggest that this is, and I merged the two accordingly
            - https://tshaonline.org/handbook/online/articles/fha41
    - Repeated action for “Recipient”
        - Like “Sender”, Open Refine counted more unique names within my “Recipient” column than the Workbook says should be found
            - It counted 205 initially, reduced to 178 post-merge
        - Like “Joaquin G. Rej6n”, more clusters here were misspelled in all rows
            - I chose to stick with Open Refine’s suggestion for “Alc^ La Branche” due to a lack of sufficient Google results
            - “Mirabeau B. I^mar” was translated to “Mirabeau B. Lamar” by the TSHA
                - https://tshaonline.org/handbook/online/articles/fla15
            - “Count Mol6” was translated to “Count Molé” after finding mention of the name in
                - https://books.google.ca/books?id=R7FlAAAAMAAJ&pg=PA672&lpg=PA672&dq=Count+Mol6&source=bl&ots=yPEtfaZeP6&sig=2-i5QK_8l4VK7_yXZm8SBlqBuqk&hl=en&sa=X&ved=0ahUKEwick7nLsLDbAhXn24MKHdDJAtgQ6AEIKTAA#v=onepage&q=Count%20Mol6&f=false
- Clicked the arrow next to "Sender"
    - Selected Edit Cells → Common transforms → Trim leading and trailing whitespace
- Repeated for "Recipient"
- Exported the project as a .csv file
- Clicked the arrow next to "Sender"
    - Selected Edit column → Rename this column
        - Renamed it as “source”
        - Renamed “Recipient” as “target”
- Clicked Export → Custom tabular exporter
    - Unchecked “Date”
    - Changed 'Tab-separated values (TSV)' to 'Comma-separated values (CSV)' in the Download tab
    - Downloaded the file
- Dragged the file into a Palladio interface (http://hdlab.stanford.edu/palladio-app/#/upload)
    - I don’t really see any particular pattern, not sure what I’m looking for
- Uploaded the file into DH Box File Manager
    - Under the new name “Module3–Exercise2–OpenRefinedTexasIndex.csv”