HIST3814O - Final Project Open Notebook/Fail-Log

Aims for this Final Project

- Transcribe the first 7 pages from the 14th Canadian General Hospital war diaries in .xml formats
    - http://collectionscanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayItem&lang=eng&rec_nbr=2005110&rec_nbr_list=3366167,3203123,2005097,2005100,2005101,2005099,2005096,2005110,2005108,2005106
    - These 7 documents present the history of the No. 10 Canadian Stationary Hospital Unit
        - From the Unit's authorization and establishment at Western University (London, Otario, Canada) to its eventual arrival at Canvas Camp (St. Martin's Plain, Shorncliffe, England)
            - I chose this point in the diary to stop transcribing as it neatly ends off the Unit's establishment, training, and transport across the Atlantic before the Unit begins its intended duties of caring for the sick and injured victims of WWI
    - These files shall henceforth be referred to with the abbreviation CWD
- Encode the transcribed files
    - So as to make color-coded versions which provide easy access to available relevant sources, should the reader wish to further his/her knowledge on the topic
- Run a CSV file of all the transcribed information through Open Refine (http://openrefine.org/)
- Run the CSV file though Voyant Tools (http://voyant-tools.org/) to notice any word frequency patterns
    - Document and theorise upon these patterns

Created a generalized *FinalProject_blanktemplate.xml* for the CWD transcriptions

- Opened the *blanktemplate.txt* file from the *module3-wranglingdata-master* repository
    - https://github.com/craftingdigitalhistory/module3-wranglingdata
    - Checked for, and made a word document reminder, all the various sections which had originally been designed for the *Negro Slavery* page transcriptions of Module 2 – Exercise 3
        - I made notes of what the content of the sections should look like, then deleted them
        - Apart from the <biblScope> section, the sections all contain the same information → this information was acquired from the war diary's storage location in the Library of Archives Canada website
            - I must remember to change any "&" symbols in the links to &amp;
            - Both <title> sections
                - First
                    - "War diaries - 14th Canadian General Hospital = Journal de guerre - 14e Hôpital général canadien."
                - Second
                    - "Confidential War Diary of No. 10 Canadian Stationary Hospital, from May 10th/1916 to July 31st/1917, Volume No. 1."
                        - This was ascertained from page 1 in the folder, the war diary's title page

- Both <authority> sections
    - First
        - "Transcribed from digital copy available from Library and Archives Canada at http://collectionscanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayItem&amp;lang=eng&amp;rec_nbr=2005110&amp;rec_nbr_list=3366167,3203123,2005097,2005100,2005101,2005099,2005096,2005110,2005108,2005106"
    - Second
        - "Transcribed by an employee of the Library and Archives Canada"
- The <publisher> section
    - "The Government of Canada"
- The <distributor> section
    - "Open access through the Library and Archives Canada website, http://www.bac-lac.gc.ca/eng/Pages/home.aspx"
- The <date> section
    - "Unknown"
- The <settlement> section
    - "Ottawa, ON"
        - The city in which the Library of Archives Canada (LAC) office building is located
- The <biblScope> section
    - Determinate upon the individual file

Created a Word document reminder concerning the various commands required for the encoding process

- All those interested can find a breakdown of the encoding requirements, along with 3 helpful YouTube tutorial videos, under the subheading *Encoding Your Transcription* at the following URL:
    - http://workbook.craftingdigitalhistory.ca/supporting%20materials/tei/
- For persons, I must surround my text with the following:
    - <persName key="Last, First" **from**="YYYY" to="YYYY" role="Occupation" **ref**="http://www.website.com/webpage.html"> </persName>
- For places, I must surround my text with the following:
    - <placeName key="Sheffield, United Kingdom" **ref**="http://tools.wmflabs.org/geohack/geohack.php?pagename=Sheffield&params=53_23_01_N_1_28_01_W_type:city_region:GB"> </placeName>
- For claims or arguements, I must surround my text with the following:
    - <interp key="reason" n="citation" cert="high" **ref**="http://www.website.com/webpage.html"> </interp>
- I must also remember to save the transcribed CWD files in .xml formats

Began transcribing the CWD files into the *FinalProject_blanktemplate.xml*

- I must remember to transcribe the content of the file *exactly* as they are presented in the images
    - *E.g.* fully capitalized headings

    - *E.g.* punctuation oddities
    - *E.g.* Spelling mistakes
- I must remember to indicate paragraphs correctly
    - <p> → signals the start of a paragraph
    - </p> → signals the end of a paragraph
- I must remember the change any “&” symbols in the links to &amp;
- Transcribed file *e001518029* → page 1
    - Process was done manually
        - The OCR’d version from Module 2 was simply too messy to bother with
    - Saved as *FinalProject-e001518029.xml*
- Transcribed file *e001518030* → page 2
    - I decided to copy the previously OCR’d text (Module 2), and paste it into the <body> section of the file
        - While most of the text had been OCR’d, the heading sections had not
            - These were then included manually
        - Cleaned the pasted OCR’d text
            - Used Sublime Text’s *Find* + *Replace* function to make global corrections every time a word was misspelled
            - The intended regexes were first tested in *RegExr: Learn: Build & Test RegEx* before being applied to the file
                - https://regexr.com/
    - Saved as *FinalProject-e001518030.xml*
- It was at this point that Lauren Rollit very graciously gave free access to the class to download her already transcribed copies of the first 58 CWD files
    - https://github.com/laurenrollit/hist3814o-final-project
    - I opened her GitHub page and downloaded the repo as a zipped folder
    - I extracted all the files
- I copied the content from each of Lauren’s files into the <body> section of my *FinalProject_blanktemplate.xml*
    - I ensured that each file displayed the correct page number in the <biblScope> section
- Read through and corrected any mistakes which Lauren accidentally made
    - *E.g.* Not capitalizing various headings or missed commas *etc*.
- Thus, the following files were transcribed:
    - *FinalProject-e001518031.xml*
    - *FinalProject-e001518032.xml*
    - *FinalProject-e001518033.xml*
    - *FinalProject-e001518034.xml*
    - *FinalProject-e001518035.xml*
- **NB**: Columns exist in the multiple CWD files, which cannot be transferred to Sublime Text
    - Therefore, I added Vertical Bars (|) to demonstrate the column breaks
    - Differing rows are shown simply as new paragraphs
    - Additionally, in order to represent blank columns, I added double asterixis (**)
    - The CWD images use single hyphens (-) to represent information being the same as the cell directly above it
    - In the final column, the initials “ES” are presented 77 times
        - I assumed these to be Edwin Seaborn’s signature initials
            - The Commanding Officer of the No. 10 Canadian Stationary Hospital
            - This hypothesis is based off the existence of his full signature, not merely the initials, in various CWD files

Began encoding the transcribed CWD files

- I realized that I had forgotten what a few of the required topics in the encoding meant, and thus how to fill them in accurately
    - *E.g.* "citation" when attempting to encode an argument of claim
    - Watching the tutorial YouTube videos from Module 2 helped to remind me
- I must remember to change any "&" symbols in the links to &amp;
- Regarding the various claims and arguments
    - Websites were found for some of the claims
        - Their URLs were added as references in the encoding
    - However, various other claims did not yield any further results when I searched their keywords/phrases in Google
        - Thus, the individual image's URL, from the LAC database, is used as the reference citation
- During the process, I realized that numerous codes were very identical and I did not want to waste time by typing out each one individually
    - Created a Word document
        - Copy/pasted the repeat codes and resulting information
        - This provided much easier access to the codes, rather than searching through already completed encoded files
- Encoded *FinalProject-e001518029.xml*
- Encoded *FinalProject-e001518030.xml*
- Encoded *FinalProject-e001518031.xml*
- Encoded *FinalProject-e001518032.xml*

Possible CSV file creation

- I wanted to create a CSV Excel file containing all the transcribed information, neatly categorized into their appropriate tables
    - Manually doing this will take far too long
    - Unfortunately, DH Box isn't working properly
        - It's giving me a *Internal Server Error* message
        - Dr. Graham has sent an email to the application's maintenance team in the hopes of rectifying the issue
        - This problem prevents me from creating a CSV file of the info through the Command Line
    - While I wait for DH Box to be fixed
        - I copied all the transcribed information from each file into a blank word document
        - I did this so that I may continue with the encoding process while still maintaining a fully encoded version to later attempt the CSV file construction with
- Unfortunately, the issue with DH Box was never rectified
    - I manually copy/pasted all the information into a 5 column Excel document, one cell at a time
        - Saved the document in a .CSV format → *FinalProject-WordFrequencyCheckSheet.csv*
        - This file will later be cleaned further in Open Refine

- The cleaned version will be used to study word frequency throughout the final 5 files transcribed via Voyant Tools

- **NB:** Only the content from the final five transcribed CWD files was added to the .csv file
  - The first file is simply the War Diary's cover page and provides no information on names, locations, dates, intentions *etc*.
  - The second file is an introductory passage, written by Sgt. V.A. James, and is not in a row-column format
    - I felt this would have confused the overall Excel format, and decided to leave it out of the word frequency check

Checked that each file was successfully encoded

- Ensured all the encoded files were the same folder as the *000style.xsl* file
  - XSL file found in the same *module3-wranglingdata-master* repository as the *blanktemplate.txt* file
- Opened Firefox
  - Dragged each individual file into a new tab
    - Encoded file *FinalProject-e001518032.xml* returned an error message
      - I had accidentally ended each paragraph with a capital P in the code
        - I used Sublime Text's Find and Replace All function to quickly change each error simultaneously
- I created a full bibliographical list of all the sources I discovered and linked into the encoding
  - A pdf version can be found in the *HIST-3814O-FinalProject* GitHub repo

Used Open Refine to double check the cleanliness of the *FinalProject-WordFrequencyCheckSheet.csv*

- Created a New Project with the *FinalProject-WordFrequencyCheckSheet.csv*
- Clicked the arrow to the left of each column's title
  - Selected "Facet" → "Text Facet"
- Clicked on "Cluster" within each of the resulting facet boxes
  - "Merge Selected & Re-Clustered" any clusters which were clearly meant to be the same, but which were accidentally typed incorrectly during the transcribing process
    - However, this was only done for a very few examples as the transcriptions were mostly done accurately
      - In the "Place" column the addition of an extra space between "LONDON, ONTARIO" in two rows was made
      - In the "Remarks and references to Appendices" column the addition of a period (".") was made
    - No changes were made to any of the remaining "Date", "Hour", or "Summary of Events and Information" columns
- Clicked the arrow to the left of each column's tiltle
  - Selected "Edit Cells" → "Common transforms" → "Trim leading and trailing whitespace"
- Clicked "Export" → "Custom tabular exporter"
  - Changed 'Tab-separated values (TSV)' to 'Comma-separated values (CSV)' in the Download tab
  - Downloaded the file under the save name *FinalProject-WordFrequencyCheckSheet-OpenRefine.csv*

- This file was meant to be uploaded to the *HIST-3814O-Final-Project* GitHub repo
    - However, when I opened the file to check the success of the Open Refine process, I found that all the dates had been transformed to "########" and refused to be altered back again, even after formatting each cell individually
        - The same error occurred with the original *FinalProject-WordFrequencyCheckSheet.csv*
    - For this reason, I chose to copy/paste the content as a whole into a Word document (*WordFrequencyCheckSheet.txt*), in table format
        - I then manually corrected all the changed dates
        - This file was then exported as a pdf before being uploaded to GitHub

Human error is one of the most fatalistic issues when it comes to recognizing patterns within text

- I chose to use Voyant Tools point out any possible patterns of word and phrase frequency
- I uploaded the *FinalProject-WordFrequencyCheckSheet-OpenRefine.csv*
    - **NB**: Since the file in question had created the dating error, mentioned above, I specifically set all the resulting dates as "Stopwords" so that Voyant would ignore any pattern(s) they make-up
        - Said patterns would be relatively easy to recognize manually anyway
            - *E.g*. June 1st, 1916 presents the highest amount of individual submissions for a single day → 6 in total
                - *FinalProject-e001518033*
    - In the resulting corpus, I clicked on one window's "Define options for this tool."
        - Set "Stopwords" to "English"
        - Added a few conjunction words to the "Edit List"
        - Selected the "apply globally" check-box
- Exported URLs from various Voyant tools and embedded html snippets into the blog for this project, so as to better demonstrate the results
    - Exported copies of both *Terms* and *Summary* after customizing the tools to portray the highest to lowest word frequencies
    - Exported two different display styles of the *Trends* tool as it portrays the frequency usages of the 6 most counted words in the text
        - *Stacked Bar* format
        - *Line + Stacked Bar* format
    - Exported the *Collocate* list to demonstrate the usage frequency of particular words when they are used in conjunction with one another