Module 4 – Open Notebook/Fail-Log

Exercise 2: Topic Modeling Tool

- This Exercise went according to plan right up until I opened the *output_html* and *output_csv* folders
    - At that point nothing made sense to me in the least bit, simply because the text was so very messy
    - Overall, The Topic Modeling Tool (TMT) Latent Dirichlet Allocation (LDA) was all very confusing and frustrating
- I followed the instructions to download the LDA for Windows from senderle's GitHub repo *topic-modeling-tool*
    - https://github.com/senderle/topic-modeling-tool
- Ensured *war-diary-text* directory was transferred from my old DH Box account to the new one after the old had expired
    - `$ ls`
- Downloaded a zipped copy of the directory to my computer
    - `$ zip -r wardiaryfiles.zip war-diary-text`
    - Unzipped the folder
- Opened the Topic Modeling Tool
    - Set the *Input Dir…* in the LDA to open the unzipped folder
    - Set the *Output Dir…*
    - Left the number of topics to be modeled at 10
    - Clicked *Learn Topics*
- Clicked on the *all_topics.html* in the new *output_html* folder
    - This opened a browser-based method of navigating my topics and documents
- Analysis
    - I found this tool to be very confusing, and frustrating as a result
        - Mostly because the text was so messy
            - I simply didn't understand what I was looking at, it all looked like gibberish
            - It most certainly challenged my understanding of the material
        - Not knowing what the terms being used meant, I read Marijn Koolen's *Topic Modeling With Newspaper Archives* presentation
            - https://web.archive.org/web/20161025200154/http://humanities.uva.nl:80/~mkoolen1/materials/KB_Mallet_2015/KB_Mallet.html#0
            - Topic models
                - "[Represent] topics in [a] collection of documents"
                - "Use statistics to find topics represented by groups of words"
            - Topic
                - "a mixture of words"
        - In an attempt to learn how to read the results, I Googled "Latent Dirichlet Allocation tutorial"
            - Edwin Chen's blog *Introduction to Latent Dirichlet Allocation* provided some explanation

- http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/
- Overall, the author's explanation does make sense, and eased my frustration a great deal
  - However, his tutorial examples were all based on the assumption that the text in question was perfectly clean to begin with

- I attempted to explore the *all_topics.html*, but the messiness of the text just confused me more and more
  - file:///C:/Users/Danny/Desktop/Carleton%20University/Third%20Year/Third%20Semester/DIGH%203814O/Module%204/output_html/all_topics.html
  - Almost every word is spelt with incorrect symbols
    - I'm assuming this is in computer readable language?
    - Some are so thoroughly jumbled that I can't even guess at the correct form even after reading the file in context
      - E.g. "agittgngggzm"
  - Furthermore, I am not sure how to clean it properly
    - As far as I can assume, there seems to be only two methods, both of which would take considerable time and educated interpretation and guesswork
      - Manually altering every word in context, then reuploading the edited file to the LDA
      - Using regex commands to globally change all incorrect spellings of a word simultaneously
        - However, this assumes that the word in question is misspelled in the same way throughout the file
        - Furthermore, it also assumes the reader is able to interpret the correct spelling before he/she change it
- I also attempted to interpret the material via combining it with the opened Excel documents in the *output_csv* folder which was downloaded at the same time as the *output_html* folder
  - However, this did not help me
    - The results appeared to be the same misspelled topics, as well as lists of numbers and file names, all categorized into rows and columns
      - Truth be told, I did not understand what I was looking at in the slightest

Exercise 6: Voyant Tools

- Apart from some initial confusion, and a sense of being overwhelmed at the start of the analysis section, the Exercise went smoothly throughout
  - This method of comparing and contrasting the data results is **MUCH** more user friendly and easier to understand than the LDA
    - I thus intend to use Voyant Tools for my Final Project
      - At least until I find an alternative method, from among the 7 other Exercises in this Module, which I prefer more
- Opened Voyant Tools
  - http://voyant-tools.org/

- In the CSV of the CND database, I added all of Melodee Beals' content, found here:
    - https://raw.githubusercontent.com/shawngraham/exercise/gh-pages/CND.csv
    - Clicked *Reveal*
- Analysis
    - I played around with results from the upload of Meldoee Beals's colonial newspaper CSV
        - the experimentation process to test everything, including the relations between tool types took a very long time
            - Simply due to amount of differing tools and the ways they can be combined with others
    - Initially it was all very confusing, and a little overwhelming
        - But I eventually got the hang of it, through experimentation, and realized my own personal preference of tool combinations
            - The customization aspect allows the reader to pick and choose between numerous data illustration methods
            - Not only does this allow the data to be seen, and compared, in a variety of ways
                - It makes the reader's life easier by allowing him/her to study the material in ways which the individual finds to be the easiest
    - The system allows the user to use either one tool for the whole screen, or five tools simultaneously in small windows
        - This is changed via the Blue tab → Windows icon on right side → Choose desired option
            - *Corpus View* → the 5 tool option
            - Any of the remaining options, and linking tool types → the 1 tool option
        - After experimenting with Beals' CSV file, I eventually realized my own personal preference of tool usage and combination
        - I preferred the *Corpus View* style, and combined the following tools:
            - *Terms* or *Summary* in one window → Provides lists of the most frequently used words along with word-counts
                - the two are very similar and I was unable to decide which I preferred
            - *Reader* → Provides the entire text in question in context
            - *Trends* → Provides various types of graphs with which to view the word frequency
            - *Phrases* → Provides a list of multiple word frequencies
                - *i.e.* how many times particular words are used in the text in conjunction with one another
            - *Collocates* → Shows how words go together or form fixed relationships
    - Next, I uploaded 2 more files to Voyant Tools in order to further experiment with my preferred tools combination
        - Module 2 - Exercise 1 - Excel Copy - Commonwealth War Graves Commission, Find War Dead
            - It was very interesting to see colour coded and graphical data results for all those Jooste men who fought in WWI and WWII
            - Exported the URL of one of my graph tools, and pasted it into my blog (in a temporary Test Page)

      - I had to update the page twice, as the first resulted in a 404 Error Message
      - The image could then be seen in the blog itself
    - Module 3 – Exercise 2 - OpenRefinedTexasIndex

- After my blog was completed and ready to be updated, I quickly redid this Exercise with the *Commonwealth War Graves Commission, Find War Dead* file as I wanted to export some of the tools for demonstration purposes