Fundamentals of Text Mining:
Analysis Activity
1: How do female and male authors' prose differ?

Summary: In this activity we will explore different ways to examine and visualize your datasets using Voyant.

About Voyant
Voyant is a web-based reading and analysis environment for digital texts. It is freely available here: https://voyant-tools.org

The tool is open source, and is widely used by Digital Humanists using text analysis and visualization. It is an ideal tool to experiment with in order to familiarize yourself with the process of importing textual data, analyzing it, then creating and interpreting the results as visualizations. One drawback of the tool is that it is hosted on the McGill University servers, and so its ability to process very large datasets is limited. It is possible, however, to install it on a home or local server.

Key features include:

- Importing documents in various formats (plain text, HTML, XML, PDF, RTF, MS Word, ODF, etc.)
- Several tools for studying term frequencies and distributions within documents and within a collection of documents (a corpus)
- A full-text reader that supports very large texts and includes interactive features
- Interaction between the tools that facilitates navigation and exploration at different scales (from "close reading" to "distant reading")
- A mechanism for bookmarking and sharing instances of Voyant Tools (specific texts and tools) through persistent URLs

---

Instructions:

- **Objective**
  - In this activity, we will compare the works of British female and male authors from the 1880s using Voyant.

- **Upload to Voyant**
  - To upload a group of documents, you must first create a zip file of your dataset folder:
    - You can convert the female and male corpus data sets into zip files by right-clicking the respective folders and selecting the "Download" option from the drop-down menu.

        - Once the folders and zipped and downloaded, navigate to your "Downloads" folder or wherever your browser sends your downloads on your computer.

    - Upload the zip file to Voyant:
        - We will need to open two instances of Voyant in our browser since we are comparing two different data sets.
        - From the landing page, select the zip folder you have just created.
        - Click 'upload'.
        - Voyant will do the work of expanding the archive and processing all of the documents in your dataset.
        - Replicate the steps above for the second data set.
        - ***Keep both tabs open for the rest of the session!***

- **Understanding the Dashboard View**
    - Familiarize yourself with the dashboard in one of your open tabs.
    - List three pieces of information about your data set that you can see at a glance from the dashboard view.
    - What are your overall impressions of the Voyant dashboard? Do you find it intuitive and user friendly? If not, what do you find unclear or challenging?
    - What is a stopword? [hint: read the 'help' documentation]
    - How would you add a tool to the dashboard that is not included in this default dashboard view?
- **Voyant Suite of Tools**
    - Voyant provides a range of tools and options for text analysis. What information can you learn from the following tools and visualizations? [hint: Voyant help documentation is useful].
        - Cirrus
        - Document Terms
        - DreamScape
        - Contexts
        - Choose your own tool

- **Explore your DataSet using Voyant**
    - An opportunity to explore your data sets using the tools embedded in Voyant. The goal is for you to experiment with your data, to customize tool options and to create a visualization or two.

    - **Most Frequent Words comparison**

        - Let's take a moment to look at the most frequency words in each data set now that you have had an opportunity to explore the dashboard.[1]
        - Compare the Cirrus word clouds in the two tabs:
            - Are the most frequently used words in the female corpus the same ones that appear most frequently in the male corpus?
            - Describe any differences you observe with your group.
    - **Topic Comparison**
        - Now let's take a look at the Topics in each data set.
        - You can switch over to this tool by clicking the window pane icon in the upper right hand corner of the panels and navigating to Corpus tools > Topics.
        - Compare the list of Topics in each data set:
            - Are there any trends in the female corpus topics that jump out? What about the male corpus?
            - Do you notice any similarities between the two topic lists? Differences?
            - Discuss with your observations with your group.

If you want to continue exploring, find the four-box icon for the upper right panel to select a different visualization tool for that panel. The index of available tools describes what each option can do.

[1] Credit for text file compilation: http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244469/Data%20Collections%20and%20Datasets#demo-corpora