Fundamentals of Text Mining:
Analysis Activity
2: How does news coverage differ across the globe?

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:

- **Upload to Voyant**
    - To upload a group of documents, you must first create a zip file of your dataset
        - You can convert the entire dataset or subset of your choosing into a zip file 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:
        - 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.
- **Understanding the Dashboard View**
    - Familiarize yourself with the dashboard.
    - List three pieces of information about your content 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
        - Mandala
        - Contexts
        - Choose your own tool

- **Explore your Dataset using Voyant**
    - An opportunity to explore your datasets 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.
- **Your dataset**
    - In this dataset, you will find the following folders…
        - **All - Watergate Scandal Global Reactions**
            - Contains OCR’d newspaper articles about the Watergate Scandal from the International Herald Tribune, The Sunday Times, The Daily Mail, The Times, and The Telegraph
        - **China (subset of the “All” dataset)**
            - Contains OCR’d newspaper articles from reporters in the field in China regarding the Watergate Scandal.
        - **France (subset of the “All” dataset)**
            - Contains OCR’d newspaper articles from reporters in the field in France regarding the Watergate Scandal.
        - **Russia (subset of the “All” dataset)**
            - Contains OCR’d newspaper articles from reporters in the field in Russia regarding the Watergate Scandal.
        - **United Kingdom (subset of the “All” dataset)**

        - Contains OCR'd newspaper articles from reporters in the field in the United Kingdom regarding the Watergate Scandal.
    - **United States (subset of the "All" dataset)**
        - Contains OCR'd newspaper articles from reporters in the field in the U.S. regarding the Watergate Scandal.

You may work with the entire data set or choose a specific region of interest for this activity.

- **Most Frequent Words comparison**
    - Open two new instances of Voyant
    - There are two different options for this activity:
        - **Option 1:** Upload your entire dataset in one window, and load a single text in the other Voyant window.
        - **Option 2:** Upload two of the subsets of the dataset into separate tabs.
            - **For example:** You could generate a zip file from the United States folder and load it into one window, and load a zip file from the United Kingdom folder to the other window.
    - Compare the word clouds.
        - Are the most frequently used words in the single document (or subset) the same ones that appear most frequently in the larger corpus (or other subset)?
    - Discuss any differences you observe with your group.
- **WordTree**
    - Open two instances of Voyant
    - Load two of the subsets of the dataset into separate tabs.
        - **For example:** You can generate a zip file from the United States folder and load it into one window, and load a zip file from the United Kingdom folder to the other window.
    - Find the four-box icon in the upper right panel and select the WordTree option under Corpus Tools
    - Repeat this process for the subset in the other tab
    - Compare the WordTrees.
        - Explore the options for this visualization
    - Potential questions to consider:
        - What similarities or differences do you see between the two WordTrees?
        - How does the place of publication affect the coverage of this political scandal?
        - What are some of the terms associated with:
            - Watergate
            - Nixon
            - Policy
        - How do these terms differ between the two subsets?

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.