Writing 2100 - Text Mining Intro

Preparation

- This page is a companion to the guest lecture for WRIT2100.
- **Please do bring a laptop!** If you have one, bring your own. If you do not have one, please feel free to check one out at the Olin circulation desk. Having a laptop will allow you to participate in the exercises and get the most out of this exploration.
- **Please bring samples of text you want to explore**. These can be in several formats: plain text, Microsoft Word Document, PDF and URLs to online sources. We will be loading them into various tools; hopefully you will see interesting results. I will encourage sharing.
- **No special software will be needed.** All exercises will be done through a Web browser, without any special plugins.

Agenda

Presentation

There will be a presentation during which your questions and comments are welcome. My aim is to discuss as much as is useful to you. Please feel free to chime in at any time.

Exercises

All exercises will be demonstrated, so no prior knowledge of the tools are required. The room is equipped with a jack to allow easy sharing of your desktop on the screen, so if you discover something that you would like to share and discuss, we can easily do so, and I will encourage that.

Voyant

Voyant is a low barrier text analysis tool that delivers a rich, interactive interface and a variety of visualizations (all of which are explained in the help file). I nput format can be plain text, a PDF (with OCR), a MS Word Document or a URL for HTML analysis. Please feel free to bring your own material for upload to the workshop, understanding that upload of any material will be subject to the Voyant privacy policy. Sample texts and URLs for analysis are listed below for experimentation, in case you run low on ideas.

- Sample texts for upload are below, courtesy of Project Gutenberg. Download **plain text** version to your local machine for upload into the Voyant interface
    - Clarke, Marcus Andrew Hislop, 1846-1881. For the Term of his Natural Life - https://www.gutenberg.org/ebooks/3424
    - Churchill, James Morss, 1796?-1863. A Treatise on Acupuncturation - https://www.gutenberg.org/ebooks/50985
- Sample PDFs
    - Supreme Court of the United States. Website with opinions, dissents, etc.
    - Jones, Adam. 2006. Except of Chapter 1 from Genocide: a comprehensive introduction. London: Routledge.
- Sample URLS: copy and paste into the Voyant upload browser window to get started.
    - Economics of Crisis - http://www.economicsofcrisis.com/indications.html
    - Instructions to major John Sullivan. Washington, George, 1732-1799. The writings of George Washington from the original manuscript sources. Electronic Text Center, University of Virginia Library
    - Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code - http://www.copyright.gov/title17/92preface.html
- Sample Visualization
    - Martin Luther King, Jr's "I Have A Dream..."

Google nGram Viewer

We will also explore Google's nGram Viewer. Google nGrams depict the frequency of a word or word phrase by publication year. Note that many modifications can be made to refine the analysis, so please consider the links below as starting points. Syntax for refinement is found on the About page.

- nGram tool - delete the words and supply your own.
- Guerrilla Theater Groups
- Contrast "acupuncture" as found in publications of USA and Great Britain
- Contrast "convict ship" as found in publications of USA and Great Britain
- When are we "patients" and when are we "healthcare consumers"?
- Paris...of the Nile/along the Nile/of Egypt/etc.
- Words for destroying populations
- Spanish words in English publications
- Terms related to racial integration
- "Ebola" 1975-2008 in various languages (compare with data from WHO on Ebola outbreaks and a map of official languages of African countries)

Immersion

Immersion is a tool for discovering the connections in a corpus of email. It analyzes the flow data (information found in email headers) and represents these as a network of entities. The analysis is done in real time on the flow data for which you provide credentials. The display is rich and interactive. *By design, Immersion collects only header information (From, To, Cc and Timestamp).* However, using the actual flow data from your account may cause concerns regarding privacy - Be sure to read over the FAQs to understand what information you are granting access to, and how it will be used. If you do not like the terms of the tool, you can experience it with their demo data.

- Immersion
- Demo (If you would rather not use your own account)