WE1S “comparing” module

Included in the WE1S Workspace (see S-2), the “comparing” module allows you to compare two corpora (collections of texts and term frequency data about them) to one another to discover how, and how much, they differ. These two corpora may be texts from two different projects you are working on in the WE1S Workspace, or two sets of texts from within the same project (e.g., those classified as being about the humanities vs. those classified as being about science).

The module performs this comparison using the Wilcoxon rank sum test, a statistical test that determines if two samples (i.e., the relative frequencies of shared words in two different datasets) are taken from populations that are significantly different from one another (meaning, if they have different distributions). When used to compare word frequency data from two different corpora, it helps you to determine what the most "significant" words in each corpus are. For more on the Wilcoxon rank sum test in the context of corpus analytics, see these articles: a, b.

To run the code in this module, you must have previously used the “topic_modeling” module (see S-13) to create a bag of words plain-text file of all of your project data.

Module structure

- data
- results
- scripts
- compare_word_frequencies.py
- compare_word-frequencies.ipynb
- README.md

Further Information:
* M-15 (about Jupyter notebooks)

Main Jupyter notebooks in this module:
* compare_word_frequencies.ipynb
* must previously have run the WE1S “topic_modeling” module (see S-13)

Code source: [TBD] (MIT License)

13 Sept. 2020; rev. 13 Sept. 2020 (Lindsay Thomas)