WE1S uses the Wilcoxon rank sum test to compare groups of documents to one another.

While WE1S uses methods like topic modeling (see M-2) and keyword extraction (see M-14) to explore broad topics or themes in large collections of data, it has also used a test of statistical significance, the Wilcoxon rank sum test, to compare groups of documents to one another directly. The Wilcoxon rank sum test identifies specific words that appear significantly more in one group of documents as compared to another, thus providing researchers with an understanding of what words are "distinctive" to each group.

The Wilcoxon rank sum test is a test of statistical significance. Like other commonly used tests of statistical significance such as the two-sample t-test and the log-likelihood ratio test, it tests whether and how two distributions of values (in this case, two different sets of word frequencies) are significantly different from one another. "Significantly" in this context refers to a statistical measure of confidence; if a word is shown to occur significantly more in one group than another, this means that it occurs more than we would expect given the distribution of other word frequency values in that group as compared to the other group. Unlike the two-sample t-test or the log-likelihood ratio test, however, the Wilcoxon rank sum test does not assume that the values being tested are independent of one another (i.e., that one does not determine another). Since language use is not random, and therefore how often a word appears in a document is at least partly determined by the presence of other words, this assumption means the Wilcoxon rank sum test is more accurate than the two-sample t-test or the log-likelihood ratio test when used to compare two sets of word frequencies (see Ford, "The Wilcoxon Rank Sum Test" for a more in-depth discussion of the mathematics, and Lijffijt et al., "Significance testing of word frequencies in corpora" for a comparison of the Wilcoxon rank sum test to other tests of statistical significance in the context of corpus linguistics).

Conducting Wilcoxon rank sum tests allows WE1S researchers to answer questions such as, "What words occur more frequently in news articles about the humanities as compared to articles about the sciences?" These answers allow WE1S to better understand its data at the level of language use. Using Wilcoxon rank sum tests, researchers can understand what specific words characterize particular groupings of documents.

Resources

Recommended readings:

* Clay Ford, "The Wilcoxon Rank Sum Test," University of Virginia Library, https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/.

* Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, et al, "Significance testing of word frequencies in corpora," Digital Scholarship in the Humanities 31, no. 2 (2016): 375-397.

* Andrew Piper and Eva Portelace, "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading," Post45 (May 10, 2016), http://post45.org/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/#identifier_27_7012.

12 July 2021 (Lindsay Thomas)