Feminist Text Analysis, Spring 2023

Can there be a feminist text analysis? Feminism, text, and analysis in a computational world

HDTMT: Text Mining Oral Histories in Historical Archaeology

Paper Title: Text Mining Oral Histories in Historical Archaeology

Author: Madeline Brown & Paul Shackel

Link: https://link.springer.com/article/10.1007/s10761-022-00680-5

Introduction:

Oral histories help provide important context for the past. It provides firsthand accounts from people who have experienced particular historical events in the course of their own life – a valuable insight for researchers to explore the social context, culture, politics, economics, etc. of the period being studied.

Text mining oral records and interviews is a relatively new method of retrieving important contextual information about the communities being studied in historical archaeology.

While the qualitative interpretation of texts typically conducted in historical archaeology is important and crucial, text mining and NLP can provide valuable insights that enhance and complement these methods.

The authors outline the following benefits:

(1) Rapid and efficient analysis of large volumes of data

(2) Reproducible workflows

(3) Reducing the potential for observer bias

(4) Structured analysis of subgroup differences

The authors also mention that they do not propose that text mining and NLP replace the traditional techniques used but as a supplementary tool.

This project not only shows an implementation of such techniques for historical archaeology, but it also provides a framework that can be reused and provides insights into the areas of improvement that can help further enable the use of text analysis and NLP methodologies for this particular area of study.

Data Source:

26 oral interviews acquired in 1973 and transcribed in 1981 from the anthracite coal mining region of Pennsylvania which were collected by the Pennsylvania Historical and Museum Commission and are now on file at the Pennsylvania State archives.

Link to Data Source:

Available on Github: https://github.com/maddiebrown/OralHistories

How Did They Make That:

– Interview transcripts that were available in PDF were converted to text format for analysis.

– Text was formatted to ensure that everything was standardized i.e., has the same sort of font, text size, etc.

– When PDF is converted to text sometimes extra characters and spaces come up which were removed to make the text tidy.

– Line breaks were added to signify speaker change.

– The text was then imported into R for further cleaning and then analysis.

– In R, metadata was removed. The text was then standardized. All terms were converted into lowercase (which for example avoids counting “coal” and “Coal” as separate terms), and numbers and apostrophes were removed.

– Tokens were then into individual words and bigrams, and stop words were removed “using tidytext’s stop_words lexicons: onix, SMART, and snowball.”

– After this certain slang terms, names, and common words which did not influence any meaning of the text were removed.

– In terms of actual text analysis, in this project, the authors focused on word frequency and n-grams, and they discussed future potential of sentiment analysis and tagged lexicon analysis.

Tools/skills you would need to do the project yourself:

– R Programming Language

– Text mining and NLP concepts (for example what stop words to remove and why)

– R Packages Used: tidyverse, textclean, tidytext, and cleanNLP

Tutorials:
(I used Google search to come up with primary resources, although interested analysts are suggested to search more resources for further understanding)

– R: https https://www.datacamp.com/courses/free-introduction-to-r

– Text mining with R: https://www.tidytextmining.com

– NLP with R: https://s-ai-f.github.io/Natural-Language-Processing/

– tidyverse R package: https://www.tidyverse.org/learn/

– textclean R package: https://cran.r-project.org/web/packages/textclean/index.html

– tidytext R package: https://cran.r-project.org/web/packages/tidytext/index.html

– cleanNLP R package: https://cran.r-project.org/web/packages/cleanNLP/index.html

This entry was posted in Posts on May 1, 2023 [https://femethods2023.commons.gc.cuny.edu/hdtmt-text-mining-oral-histories-in-historical-archaeology/] by Muhammad Rakibul Islam (Rakib).