Stylometric Analysis in Modern Philosophical Novels

Wan Zhi Jun
Singapore University of Technology & Design

ABSTRACT

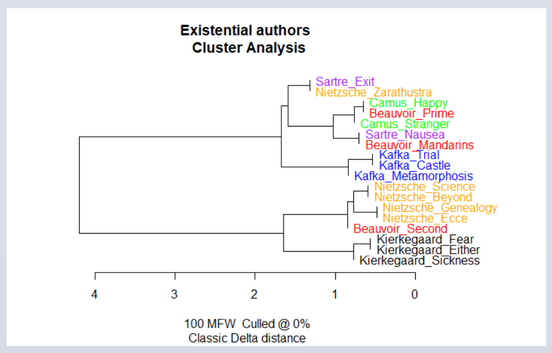
Algorithms excel at finding surface level patterns, because they are able to process a large amount of text at once, unlike people who are limited by time and energy. With algorithms we are able to process so many instances of textual patterns like word frequency and word distances, and to a small extent genre and style. As part of research in Digital Humanities, text mining techniques are used in literary analysis and humanities interpretation over large number of texts. This project aims to apply techniques defining Digital Humanities in analyzing philosophical fiction, a genre of works that which predominantly addresses questions discussed in discursive philosophy while depicting a fictional account. In particular, most of the authors included in this analysis are notable Existentialist philosophers, namely, Søren Kierkegaard, Friedrich Nietzsche, Jean-Paul Sartre, Simone de Beauvoir, Albert Camus, and Franz Kafka.

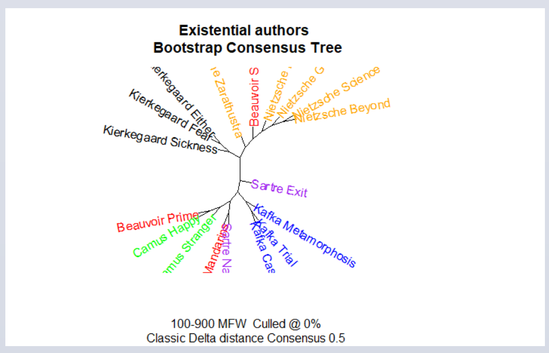
Existentialism is a subset of modern philosophy spanning across 19th and 20th century, during a period of time where technological advancements, social hierarchy, and cultural exchange, all simultaneously happen at the same time with rapid pace, especially technological advancements. In arts and culture, it is also a time when artists start presenting concepts in more abstract manners. Likewise in philosophy, philosophers tend to explore more abstract ideas and build upon them, elaborating more on observations in offbeat manners that could be contrary to those of older philosophers. Existential philosophers also tend to build or branch their ideas onto previous existential philosopher's work. Therefore the data set covers philosophers spanning across the time period. Since different authors will have different writing styles, it is interesting to observe the progression which occurs in this field.

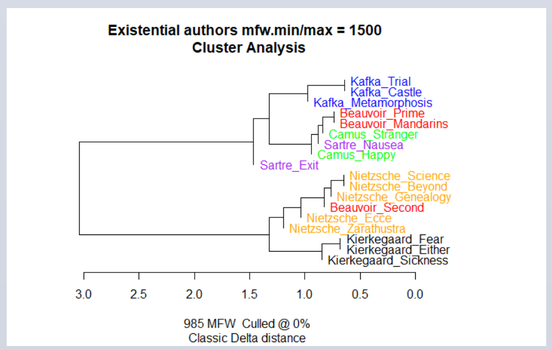
METHODOLOGY AND RESULTS

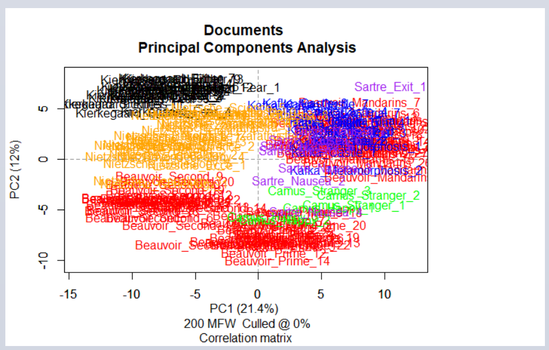
R programming language and Stylo package are used for text mining and visually analyzing 18 novels by 6 different authors. The experiment flow is to collate and strip prefaces, footnotes and translator's notes within text files version of the novels both manually and using scripts. Once collated, the preprocessing is conducted in RStudio using the Stylo package. In RStudio, any leftover XML markups were deleted, the collection of text files, now known as a corpus, were tokenized and stop words were also deleted. Stylometric analysis was done using feature frequency. In this experiment culling level of 80% was used. Cluster analysis, Principal Component Analysis, and Bootstrap Consensus Tree were used to visualize the data for style analysis in the existentialism philosophy genre. Data consists of corpus with stop words removed and the same corpus with stop words unremoved.

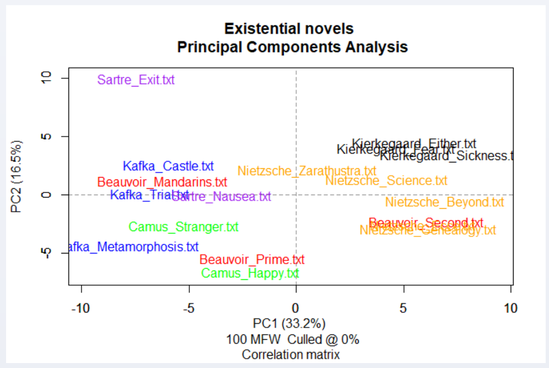
Results shown below are plots using the Stylo package with R programming language. It can be observed that the classification of style is generally accurate with some exceptions. Later works by philosophers like Camus, Sartre, and Beauvoir, tend to be slightly scattered due to having multiple influences for their philosophical ideas, an example being Sartre interpreting Nietzsche, while earlier works by philosophers like Kierkegaard and Nietzsche seem to be more distinctly classified.

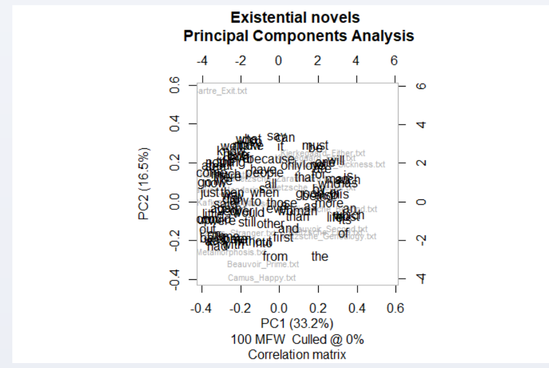


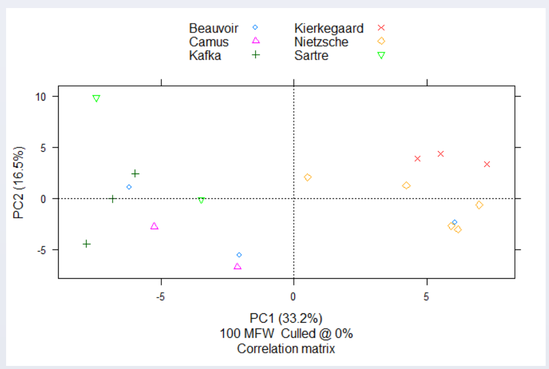


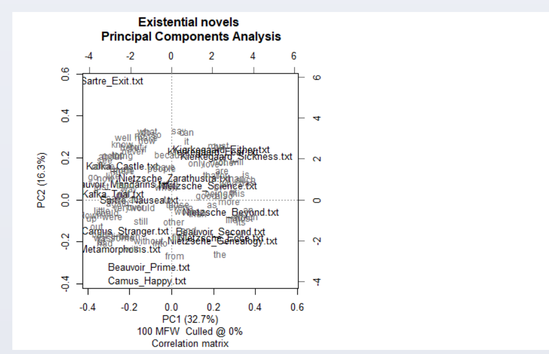












A quick observation on the word cloud reveals that absolute words such as 'must', 'only', 'all', and 'nothing' appear as a whole in the 18 novels corpus. Such words could be associated with Nihilism, a closely related field to Existentialism in modern philosophy. Notable nouns like 'love' and 'people' are also found in the word cloud, suggesting the trend of philosophy to be gradually more self centered and individual-centric, as comparison to earlier philosophy which may gravitate around environmental observations, or observing people as a collective society instead of single independently operating entities.

CONCLUSION

In literary analysis and humanities studies, there exists a "hidden layer", contextual cues that are often lost on the latest machine learning algorithm. We need to use our notion of a deeper hidden layer, the contextual understanding, to further analyze patterns algorithms reveal. Perhaps from the word frequency, we can analyze the relationships of this data with the overall genre of the texts, and gain certain understandings of the overall concept the author tries to convey. Different authors in different genres will have different outlooks towards a single concept, and even under the same genre, different authors will continue to have vastly different ideas of what a word or an idea means to them, and how they want to use this word or idea to convey conceptually. Algorithms can help us identify a linguistic pattern and reduce time spent combing through every book manually, but ultimately we as researchers are in control of the data we discover.

REFERENCES

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Jockers, M. (2013), Macroanalysis: Digital Methods and Literary History (University of Illinois Press)

Rockwell, G. and Sinclair, S. (2016), Hermeneutica: Computer-Assisted Interpretation in the Humanities (MIT Press)

Underwood, T. (2014), Understanding Genre in a Collection of a Million Volumes (University of Illinois, Urbana-Champaign)

Liu, A. 'Where Is Cultural Criticism in the Digital Humanities?' http://liu.english.ucsb.edu/where-is-cultural-criticism-in-the-digital-humanities/ (Retrieved 2 May 2019)