Nima Beheshti
Nb9pp
DS5001

Analyzing Historical Corpus of Political Texts

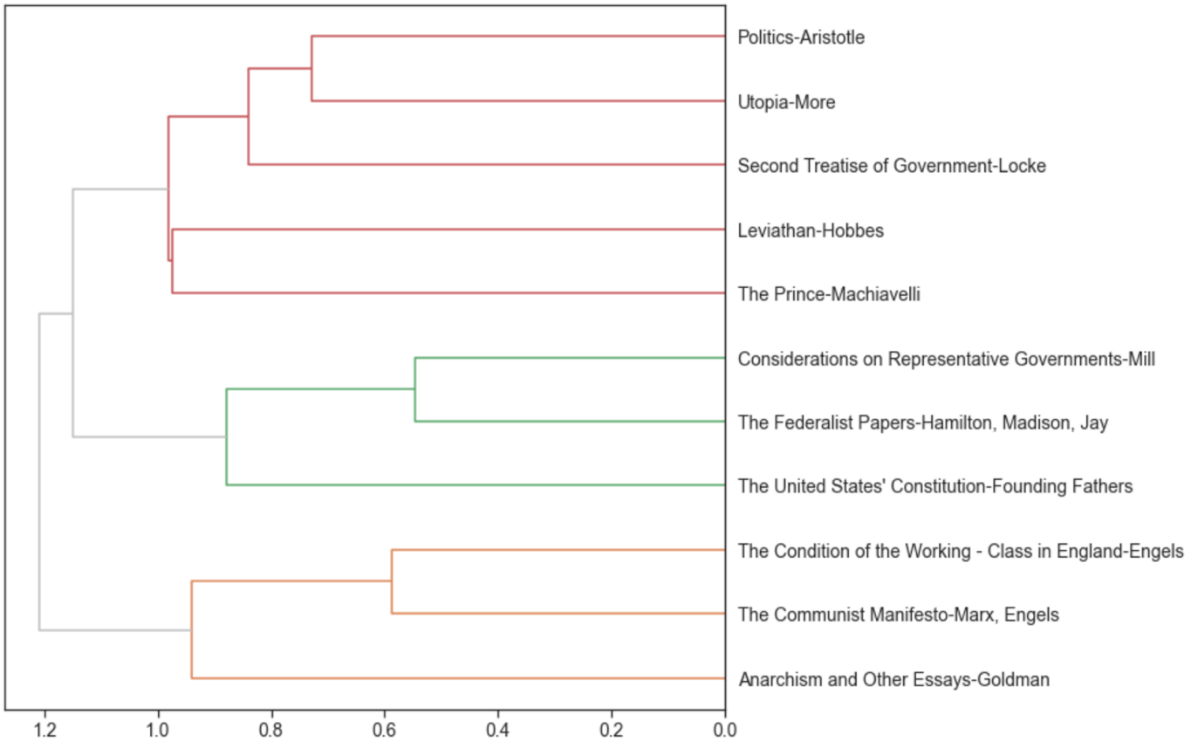
Throughout the course of human civilization, governments have been formed to govern the population of these civilizations based on the values expressed by a combination of governing and governed entities. The world has seen various forms of government during these time periods such as monarchies, republics, democracies, oligarchies, and various others. Oftentimes the fundamental ideas of a respective government can be viewed based on the cultural content published within their ideology. This project looks to identify the proper number of clusters based on a corpus of political texts through testing of simple clustering, principal component analysis, topic modeling, word embedding, and sentiment analysis.

The corpus of text used for this project involves 11 works by various authors and across many centuries in history. These texts include works from Hobbs, Aristotle, Locke, various U.S. founding fathers, Marx & Engles, as well as other individuals scattered throughout these time periods. The initial hypothesis for this corpus is that it will identify two main groups of text, those of the western more democratically/republic minded philosophies, and the communist driven philosophies. Through the course of this project, I will explore this hypothesis and see what modifications may or may not be needed.

The first step after downloading the corpus of text was to develop the 'Library', 'Token', and 'Vocab' tables that will be used in the different aspects of this project. These tables were created through a combination of regex, natural language processing, and pandas matrix operations with each highlighting a different use. The Library table will contain the titles, book_id's, authors, and file names for the corpus of text. The Token table will follow various OHCO indexes that break the text down into chapters, paragraphs, sentences, and individual words. In addition, the Token table will be used to create the Vocab table and through matrix operations, the two will be used to display the parts of speech tagging for the respective words and the id of those words. Once the tables have been gathered, the analysis can commence to clustering. Some other tables were created as well such as the TFIDF matrix, which shows the term frequency of words within a text, as well as a reduced version of this table.

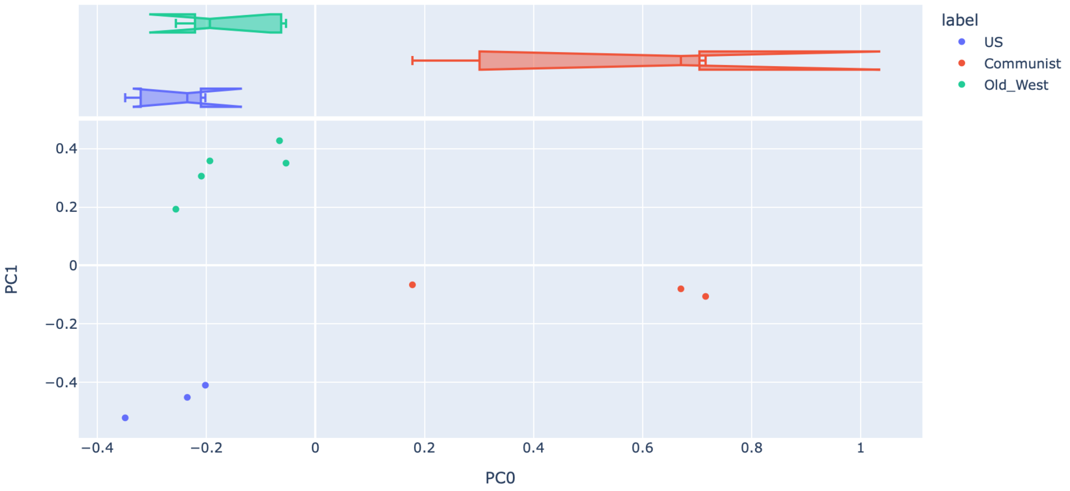
In order to test this hypothesis, the initial step was to create simple clusters for the text to see which works were similar to others. This step required working with the reduced TFIDF table by normalizing it and feeding it into a doc pair table to compare each document with the others to determine which texts most closely resembled each other. Different measurements between the normalized table of term frequencies were used to determine the optimal clusters. The following measurements were used for this test: cityblock, euclidean, cosine, jaccard, dice, jensenshannon, and correlation. The results for some of these text models returned phenomenal clusters. The cosine distance measurement method returned the best results, and those results were used and observed. The cosine method returns three main clusters and can be seen in **Figure 1** below. This figure successfully identifies three clusters. These clusters can be lumped into group names I call Old_West for older western political philosophy with the works of Aristotle, More, Locke, Hobbs, and Machiavelli; US for works that more closely identify with the creation of the philosophy behind the American political system with works by Mill, Hamilton, Madison, Jay, and other founding fathers; and lastly, the third group found within the clusters can be identified as Communist works that resemble ideology later used by communist political leaders and can be seen as the backbone of the creation of communism. These works include works of the authors Marx, Engels, and Goldman. This new three cluster hypothesis will be tested to determine whether the cosine clustering of the texts was accurate.

Figure 1



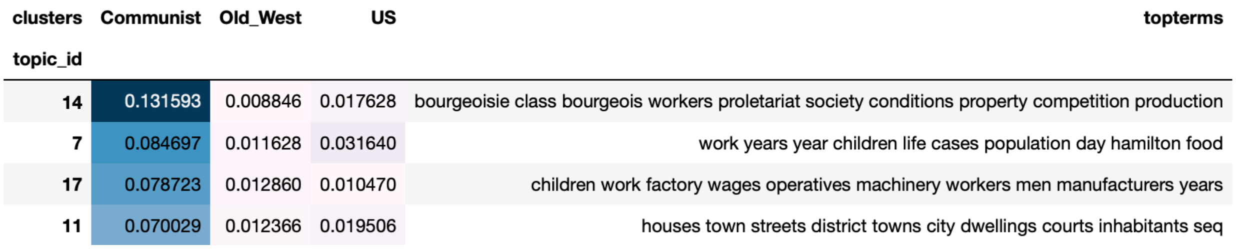
Two fundamental issues with clustering based on similarity and difference metrics are that texts similar in content may use different words and so appear orthogonal, and simple clustering does not capture subject matter of texts. A better method of finding clusters is by identifying the axes of maximum variance, which can be done through principal component analysis. To test the three-cluster hypothesis through the PCA method, it required computing a covariance matrix, decomposing that matrix into eigenvectors and eigenvalues, and creating eigen pairs with explained variance. With these steps completed, I was able to compute PC0 and PC1 loadings that showed PC0+, PC1+, and PC1- words that resembled a mix of words from clusters Old_West, and US, while PC0- contained words from the Communist cluster. Through further exploration, I was able to create a reduced document matrix that showed the top 10 PC components of each text, then visualized these components on a scatter plot to see if these three clusters could still be identified. The scatter plot for PC1 can be seen in **Figure 2** below. As can be seen from the figure, all three previously identified clusters remained as separate clusters with no overlap. This gives further evidence to the three-cluster hypothesis found just using simple clusters. Next, I will look to topic models to see how our new hypothesis fairs against topic modeling.

Figure 2

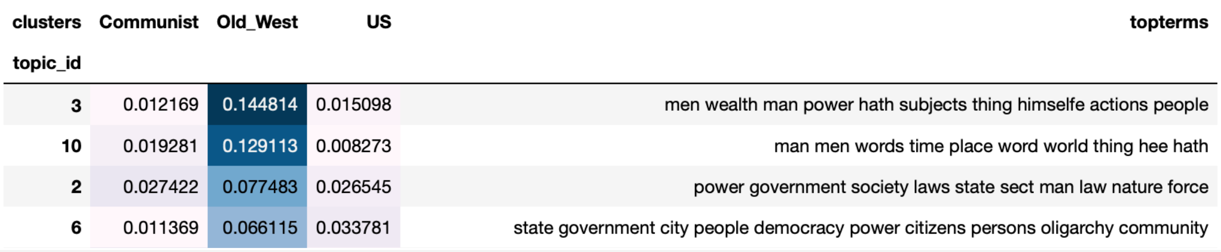


To further test the three-cluster hypothesis, proposed by the simple clustering method, I looked to use topic modeling to find the specific topics popular in each of the clusters. Topic models reveal themes as well as patterns in structure and process. I used the Latent Dirichlet Allocation (LDA) model on the corpus of texts to test. This topic model used 25 different topics and displayed the top results for each cluster. The model **Figure 3** shows the results of the topics most common in each of the clusters, along with some of the specific words that make up each topic. As can be seen in the figures, each of the clusters had a unique set of top topics. The Old_West and US clusters had some small similarities but overall, still had different main topics. The topics and language models made up of the words in the topics seem to match each of what we would expect to see for the respective clusters. The Communist cluster's top topics and words have to do with economic class divides seen with the words such as "bourgeoisie", "proletariat", "working conditions", "wages", and "time periods" indicating long hours and years of work. The Old_West cluster seems to talk more about man in nature which we would expect from the works by Locke and Hobbs who wrote about the nature of mankind with words such as "man", "nature", "law", and "power". The US cluster seems to form around the ideas that make up the foundations of the republic system used by the United States with topics and words such as "people", "government", "executive", "legislative", "courts", "laws", and "liberty". These different topics add further evidence of the three cluster conclusions discussed earlier.

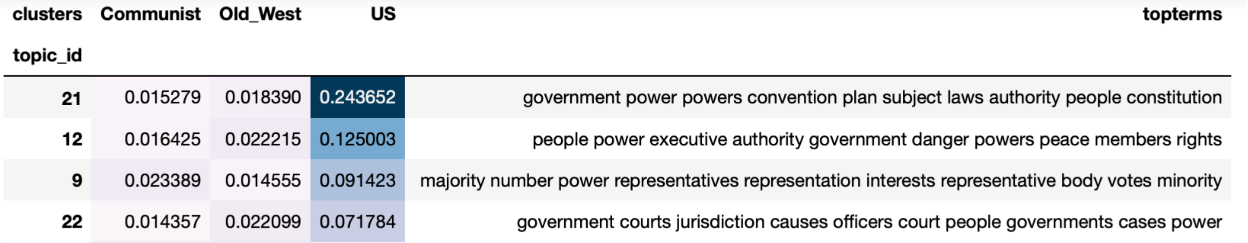
Figure 3



clusters	Communist	Old_West	US	topterms
topic_id				
14	0.131593	0.008846	0.017628	bourgeoisie class bourgeois workers proletariat society conditions property competition production
7	0.084697	0.011628	0.031640	work years year children life cases population day hamilton food
17	0.078723	0.012860	0.010470	children work factory wages operatives machinery workers men manufacturers years
11	0.070029	0.012366	0.019506	houses town streets district towns city dwellings courts inhabitants seq



clusters	Communist	Old_West	US	topterms
topic_id				
3	0.012169	0.144814	0.015098	men wealth man power hath subjects thing himselfe actions people
10	0.019281	0.129113	0.008273	man men words time place word world thing hee hath
2	0.027422	0.077483	0.026545	power government society laws state sect man law nature force
6	0.011369	0.066115	0.033781	state government city people democracy power citizens persons oligarchy community

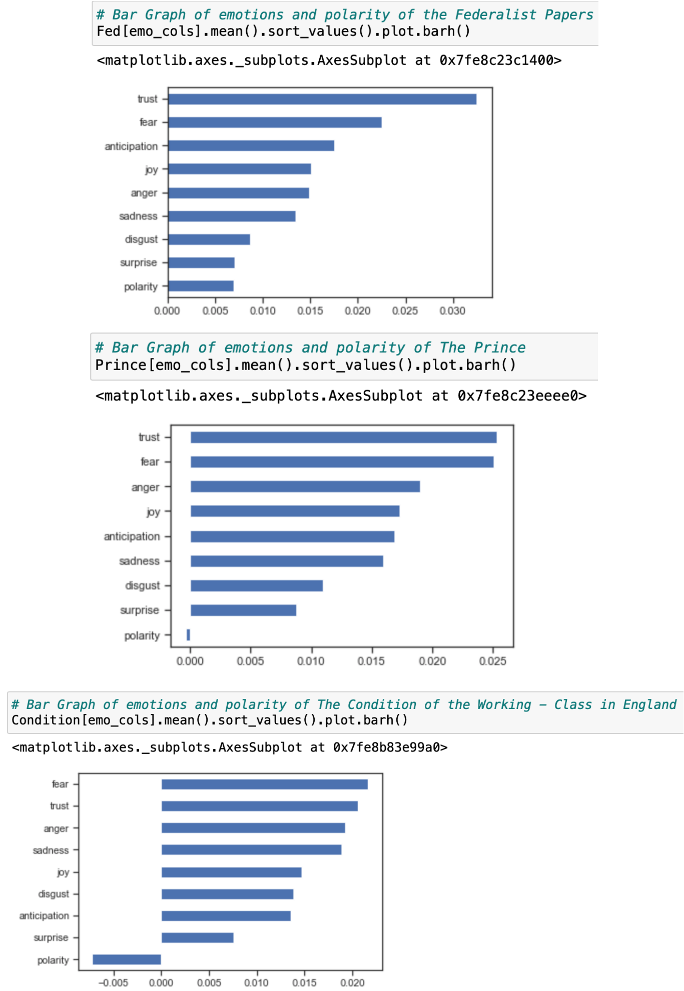


clusters	Communist	Old_West	US	topterms
topic_id				
21	0.015279	0.018390	0.243652	government power powers convention plan subject laws authority people constitution
12	0.016425	0.022215	0.125003	people power executive authority government danger powers peace members rights
9	0.023389	0.014555	0.091423	majority number power representatives representation interests representative body votes minority
22	0.014357	0.022099	0.071784	government courts jurisdiction causes officers court people governments cases power

The distributional hypothesis, from linguist Zellig Harris, revolves around the central idea that words that occur in similar contexts tend to have similar meanings. This idea is the fundamental thought behind word embeddings, which I will use next to map out and group the specific words used in the corpus of the three clusters. Using this word embedding model, we can further examine the words that make up the topics discussed earlier and see if similar words are grouped accordingly. To conduct word embeddings, I had to use a word2vec model with a corpus of words from each cluster. This model allows for coordinate creation and mapping using a TSNE model. The result of this mapping is a scatter plot of words from each cluster that groups similar words. Each cluster shows very tight groupings of related words that express similar ideas to the words in respective topics. For example, one of the topics for the US cluster revolved around the US government style of branches of government. In the scatter plot, we can find similar word groupings such as “legislative”, “executive”, and “department” near each other. In addition to seeing these groupings, I tested an analogy function, but the results did not turn out that conclusive and didn’t work as was expected. One additional test I ran with the word embeddings was a similarity test, which produces the words found to be similar a given word. The word I used was “power” and the main response from the Old_West cluster were “father”, “authority”, and “master”. The US clusters were “government”, “authority”, “constitution”, and “legislature”. The response from the Communist cluster were “modern”, and “movement”. These differences aren’t surprising but lend some additional arguments in favor of the three-cluster hypothesis, with some added similarities between the Old_West, and US clusters.

The final test I ran on the corpus of text was sentiment analysis. I didn’t expect sentiment analysis to help in our quest to find the proper number of clusters and the similarities between the texts, but it did lead to interesting results. To build this sentiment analysis model, I joined the Token table created with a lexicon of words and their emotions. I then chose one work from each of the clusters and viewed the sentiments associated with them. From the Old_West cluster I chose “*The Prince*”, from the US cluster I chose “*The Federalist Papers*”, and from the Communist cluster I chose “*The Condition of the Working - Class in England*”. Based on the sentiment bar graphs seen in **Figure 4**, it can be seen that “*The Federalist Papers*” had an overall positive sentiment value given that the polarity value was greater than 0.005; “The Prince” had a neutral sentiment polarity value; and “*The Condition of the Working - Class in England*” had a negative sentiment polarity value with a value less than -0.005. These outcomes were not that shocking given my understanding of the text, but still interesting to observe. “*The Federalist Papers*” were created to attempt to influence the people of New York to adopt the U.S. Constitution so it makes sense for the sentiment to be positive given that the authors would speak about why the Constitution was a good idea to adopt. On the other side, "*The Condition of the Working - Class in England*" was created as an observation in attempts to gain followers based on anger for the current system, and so Engles likely used negative words to describe the observations in his work.

Figure 4



```
# Bar Graph of emotions and polarity of the Federalist Papers
Fed[emo_cols].mean().sort_values().plot.barh()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe8c23c1400>
```

```
# Bar Graph of emotions and polarity of The Prince
Prince[emo_cols].mean().sort_values().plot.barh()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe8c23eeee0>
```

```
# Bar Graph of emotions and polarity of The Condition of the Working - Class in England
Condition[emo_cols].mean().sort_values().plot.barh()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe8b83e99a0>
```

The original hypothesis of this corpus was that the texts would fall under two clusters: one for Western political philosophy and one for communist political philosophy. The first contrary result was the simple clustering model where the cosine similarity method gave a three-cluster hypothesis. Based on the results of the principal component analysis, topic modeling, word embedding, and sentiment analysis I conclude that this corpus is made up of three clusters as hypothesized by the cosine similarity method. It can be stated that these three clusters are older Western political philosophy, United States foundational political philosophy, and Communist political philosophy.