Exploring Philosophical Texts Across Gender & Time with Text Analytics

Haley Egan / vkb6bn
Exploratory Text Analytics - DS 5001

Project Overview

Text Analytics is a powerful tool to help reveal and extract cultural patterns from large quantities of texts. The goal of this project is to use text analytics to explore cultural patterns in philosophical texts. Main topics and cultural themes are analyzed among the texts as a whole, as well as through time and across genders.

Key Questions:

- What are the most important topics philosophers are concerned with?
- Do male and female philosophers explore different or similar topics?
- Do topics change or vary across different eras?

Project Data

For this project, 20 philosophical texts were collected. Thirteen texts are by male philosophers, and span from the 400th century B.C. to the 21st century A.D. Ten of these philosophical texts were written in the Western world, and three by men in South America and Asia. Seven texts are written by female Western philosophers, spanning from the 19th century through 21st century AD. These texts were collected from Project Gutenberg, and other online document archives. The texts were obtained in Plain-Text format, for easier use in data cleaning, parsing, and manipulation. Due to the necessity of the Plain-Text format, the available texts were limited, especially for the female philosophers. Most (known) published female philosophers are from the modern era, with limited free access to their works in the necessary format.

Due to these limitations, the distribution of philosophical texts is not even across era and gender. The imbalance likely affects the results, and is taken into account when analyzing the texts. Further exploration with more texts would provide greater understanding into the cultural themes present in philosophical literature, and should be considered for future studies. However, the twenty texts collected for this project still provides a basic understanding of cultural patterns, and addresses the key questions this project aims to explore.

Methodology

Several Text Analytics, Natural Language Toolkit (NLTK), and Natural Language Processing (NLP) tools were used for this project. Methods used for text exploration and analysis include: Bag-of-words, TF-IDF, Principal Component Analysis (PCA), Latent Dirichlet Analysis (LDA), Topic Modeling, Word Embeddings, word2vec, and Sentiment Analysis. Visualizations used to explore analysis results include: cluster diagrams, t-distributed stochastic neighbor embedding (t-SNE), dispersion plots, and correlations heatmaps. The project code is divided into five project notebooks.

The first project notebook, **'CleanTexts_BuildTables.ipynb'**, imports and cleans the 20 philosophical texts, and then builds the dataframes LIB, DOC, TOKEN, and VOCAB. LIB contains basic metadata about each text, DOC contains preserved paragraphs of each text and the appropriate OHCO index, TOKEN contains an OHCO index and parts-of-speech tags derived from NLTK, and VOCAB contains NLTK to extract stopwords, porter stems, and 'pos_max' that contains the most frequent parts-of-speech tags from the TOKEN table. These are saved as csv files to use for text exploration and analysis in the other project notebooks.

The second notebook, **'TFIDF_comparingAuhtors.ipynb'**, builds a TF-IDF matrix, creates Bag-of-Words, and performs PCA to see if Harriet Taylor Mill wrote *Utilitarianism*, a text attributed to her husband, John Stuart Mill.

The third project notebook, **'TopicModeling.ipynb'**, contains code for LDA and Topic Modeling. The notebook contains an exploration of the overall topics in the texts, as well as the topics in texts grouped by gender and texts grouped by time period.

The fourth notebook, **'WordEmbeddings.ipynb'**, uses word2vec, t-SNE plots, and semantic algebra to explore word use, word similarities and differences, and word evolution between texts, genders, and over time.

The fifth project notebook, **'SentimentAnalysis.ipynb'**, used the NCR lexicon to explore and analyze the top emotions per text, per gender grouping, and per time period. Plots are used to visualize the results.

Gender & Era

While the distribution of texts between male and female authors is not equitable in this project due to many external factors, and the study of gender may in itself be problematic, topics between genders are still explored in this project, in hopes that some trends may be revealed.

For this project, topics are explored over time (era) and by gender. The philosophical texts are broken into three era categories, 'ancient', 'classical', and 'modern'. Ancient texts are considered to be texts written in the B.C. time period. These authors include Aristotle, Plato, Cicero, and Loa Tzu. For this study, classical philosophical texts are considered to be texts written between 1700-1900, which include the authors Hume, Mill, Kant, and Marx. Modern texts are considered texts written after 1900, including the authors Foucault, Arendt, Freire, and hooks.

Topic models are also explored by gender, breaking the philosophical texts into two author gender categories 'male' and 'female'. This classification can help explore the subject between genders, and what the similarities and differences might be. This category has some problems however, due to cultural restrictions. Most of the female authors are in the modern era, because it was largely less acceptable or not permitted for women to study, write, and publish. So while there may be differences between male and female authors, some of these differences may also be influenced by era. Additionally, since women were not permitted to study, write, and publish, many wrote under pseudonyms. One example is the author Harriet Taylor Mill. While she did publish some of her own work, historians believe that the more famous work of her husband, John Stuart Mill, was written as a collaboration between the two. So while John Stuart Mill's publications are not attributed to a female author, Harriet Taylor Mill may have co-authored the texts, or at the very least influenced them in some way.

OHCO Model & Structure Decisions

OHCO stands for Ordered Hierarchy of Content Objects. Breaking text elements down to the OHCO hierarchical levels allows for the creation of a database of text elements for data exploration.

For this project, the OHCO is as follows: 'text_id' - the unique ID given to each text to help distinguish it quickly from other texts, 'para_num' - the unique count index given to each paragraph per text, 'sent_num', the unique count index given to each sentence per text, and 'token_num' - the unique count index given for each individual token, or word. Many OHCO models include chapter numbers so that the texts can be examined on a chapter level. This is often less computationally costly than examining a text at the paragraph or sentence level, especially for very long texts. However, for this project chapters are not included, because most of the philosophical texts do not follow a traditional chapter structure. For example, the two texts by Wollstonecraft follow a letter format, and the texts by bell hooks are in essay format. Due to the inconsistencies in document structure, the documents are broken down to the paragraph, sentence, and token level.

Bag-of-Words, TF-IDF, and PCA

Bag-of-Words (BoW) places all the words of each text into a 'bucket' or 'bag'. Through this method, information about the structure of the sentence is lost. TF-IDF stands for Term Frequency - Inverse Document Frequency. TF-IDF scores the relative importance of the words, in order to gain an understanding of the texts as a whole.

The Term Frequency (TF) is the number of times a word appears in the document, divided by the total number of words in the document. Inverse Document Frequency (IDF) is the log of the number of documents divided by the

number of documents that contain a particular word. The IDF finds the weight of rare words across all the documents in the corpus. TF-IDF is TF multiplied by IDF.

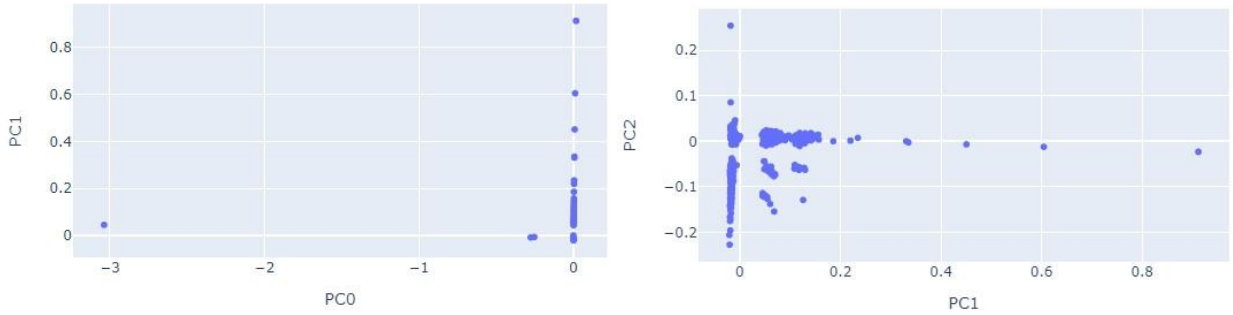
TF-IDF & PCA to Find Similarities Between Texts

Most ancient and classical philosophical texts are attributed to male authors. This is largely because it was not culturally acceptable for women to be educated, let alone become writers and publish their work. However, there may be some instances where a text is attributed to a male author, but a woman may have influenced or written the content. Throughout history, it is known that some men published for female authors, or male pen names were used to disguise the authors' gender.

For this project, there may be an example of this cultural event that is worth exploring. There are two texts that will be examined using TF-IDF, *Utilitarianism* (1861) by John Stuart Mill, and *The Enfranchisement of Women* (1852) by Harriet Taylor Mill. John Stuart Mill and Harriet Taylor Mill were married, and historians believe the couple worked together in their philosophical explorations and publications. *Utilitarianism* is a famous text that is attributed solely to John Stuart Mill. One question that IF-IDF can help explore, is whether or not the writing styles in Utilitarianism are similar to those in *The Enfranchisement of Women*, a text known to be written by Harriet Taylor Mill. If there are similarities in writing styles, then it may be arguable that Harriet Taylor Mill also wrote *Utilitarianism*, or at the very least greatly influenced it.

Principal Component Analysis (PCA)

PCA is a technique for feature extraction that allows for the most important variables to be determined. PCA creates a new matrix with a reduced set of features. The matrix is then projected onto a reduced subspace. The axes of maximum variance are identified, and contain the most information.

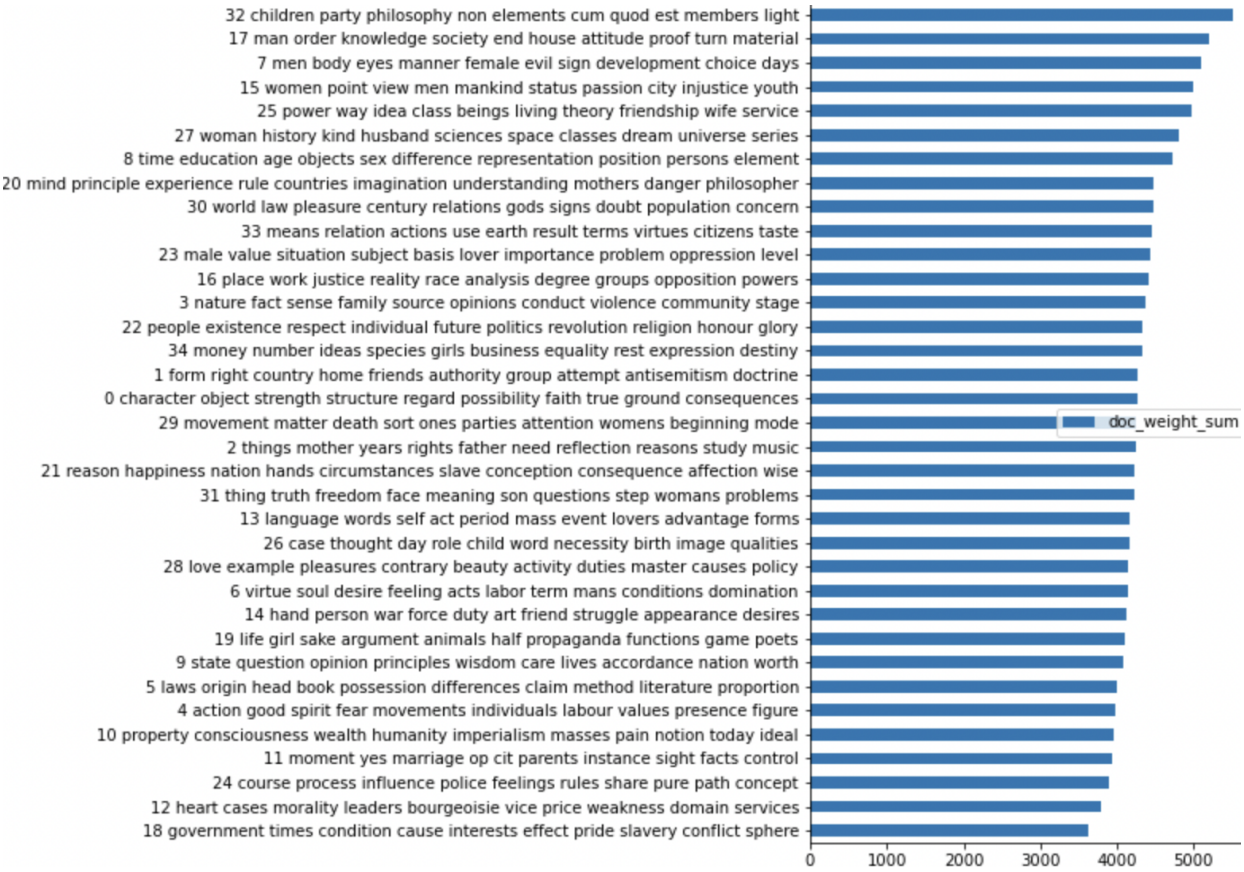


The Vector Space Model helps find similarities between documents and words, by grouping similar documents. The word vectors are used to find synonyms and word networks. Looking at the positive and negative loadings for the top components gives some insight into the texts, and the most important components, providing insight into whether or not *The Enfranchisement of Women* and *Utilitarianism* were written by the same person (or people).

The top components suggest that the two texts share similarities in the following words: amiable, apparent, accessible, attaches, attributes, appears, ascertained, arises, accidentally, affairs. The scatter plots of the top components per text show a close relationship between words. This may suggest that the two texts may indeed be both written by Harriet Taylor Mill, or at the very least, she collaborated closely in the writing of Utilitarianism.

Topic Modeling & LDA

Topic Modeling is helpful for the study of culture and history by uncovering themes and patterns in documents. With Topic Modeling, topics can be extracted over time and examined by groups, such as gender and era.



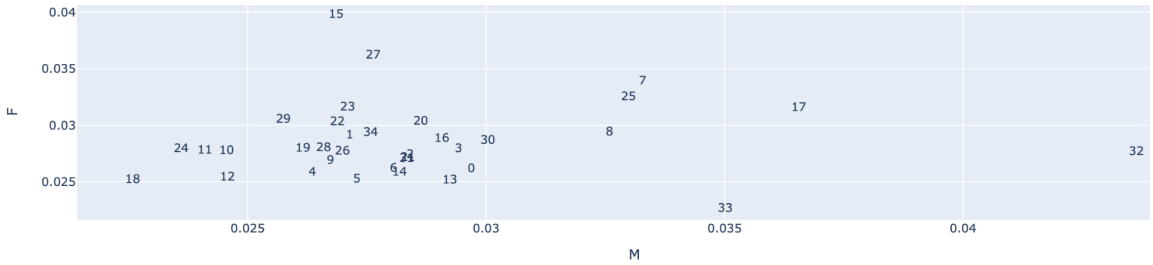
The top topics when looking at all the philosophical texts are all pretty similar at the sum of the weights level. There are about 7 sets of topics that have the largest weights. Some of the most significant topics (from topic IDs 32, 17, 7, 15, 25, 27) include: children, philosophy, man, order, knowledge, society, female, evil, development, choice, women, status, passion, injustice, time, education, representation, and position.

gender	F	M	topterms
topic_id			
15	0.039739	0.026856	women point view men mankind status passion city injustice youth
27	0.036153	0.027631	woman history kind husband sciences space classes dream universe series
7	0.033860	0.033282	men body eyes manner female evil sign development choice days
25	0.032483	0.032985	power way idea class beings living theory friendship wife service
23	0.031560	0.027098	male value situation subject basis lover importance problem oppression level
17	0.031538	0.036556	man order knowledge society end house attitude proof turn material

gender	F	M	topterms
topic_id			
32	0.027646	0.043634	children party philosophy non elements cum quod est members light
17	0.031538	0.036556	man order knowledge society end house attitude proof turn material
33	0.022592	0.035015	means relation actions use earth result terms virtues citizens taste
7	0.033860	0.033282	men body eyes manner female evil sign development choice days
25	0.032483	0.032985	power way idea class beings living theory friendship wife service
8	0.029377	0.032588	time education age objects sex difference representation position persons element

Some of the main topics in texts by female authors include: (15) women, point, view, men, mankind, status, passion, city, injustice, youth, (27) woman, history, kind, husband, sciences, space, classes, dream, universe, series, (7) men, body, eyes, manner, female, evil, sign, development, choice, days.

Some of the main tops in texts by male authors include: (32) children, party, philosophy, members, light, (17) man, order, knowledge, society, end, house, attitude, proof, turn, material, (33) means, relation, actions, use, earth, result, terms, virtues, citizens, taste.



The scatterplot of texts grouped by male and female authors reinforce the information provided in the gradient plots. The group of topics on the edges of the scatterplot that stand apart from the other topics, are the same as the top topics in texts by female authors, and by male authors. This graph helps visualize how the majority of topics in texts by male and female authors are more similar than different, based on how closely they are grouped together. Some of the most similar topics, that are most closely grouped together in the graph, include: (6) virtue, soul, desire, feeling, acts, labor, term, mans, conditions, domination, (14) hand, person, war, force, duty, art, friend, struggle, appearance, desires, (2) things, mother, years, rights, father, need, reflection, reasons, study, music, (21) happiness, nation, hands, circumstances, slave, conception, consequence, affection, wise, (31) truth, freedom, face, meaning, son, questions, step, womans, problems.

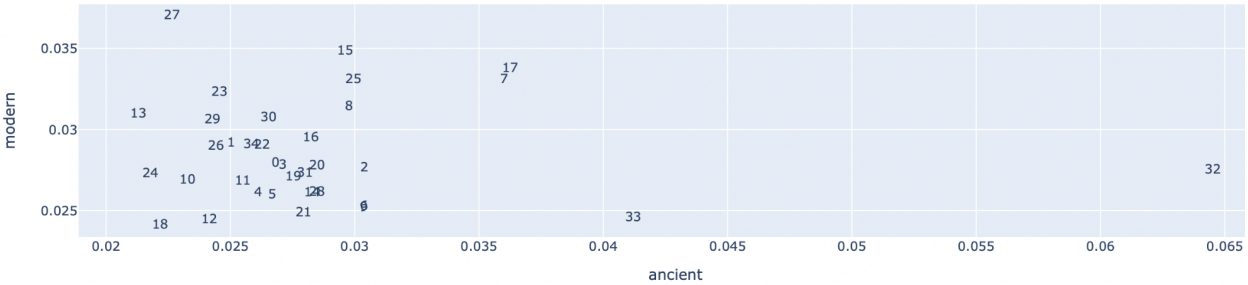
era	ancient	classical	modern	topterms
topic_id				
27	0.022648	0.025807	0.037000	woman history kind husband sciences space classes dream universe series
15	0.029621	0.030669	0.034846	women point view men mankind status passion city injustice youth

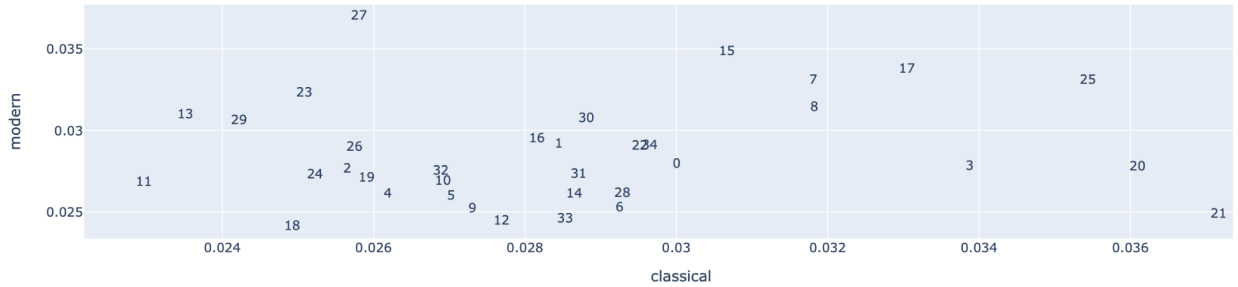
era	ancient	classical	modern	topterms
topic_id				
21	0.027947	0.037166	0.024857	reason happiness nation hands circumstances slave conception consequence affection wise
20	0.028489	0.036096	0.027756	mind principle experience rule countries imagination understanding mothers danger philosopher

era	ancient	classical	modern	topterms
topic_id				
32	0.064518	0.026887	0.027509	children party philosophy non elements cum quod est members light
33	0.041204	0.028530	0.024573	means relation actions use earth result terms virtues citizens taste

The top topics for modern texts include: (27) woman, history, kind, husband, sciences, space, classes, dream, universe, series.

The top topics for classical texts include: (21) reason, happiness, nation, hands, circumstances, slave, conception, consequence, affection, wise, (20) mind, principle, experience, rule, countries, imagination, understanding, mothers, danger, philosopher. The top topics for ancient texts include: (32) children, party, philosophy, elements, members, light.





There is a close relationship between the topics in ancient and classical texts. There are a few outliers (32, 33), but as a whole, classical and ancient texts have a lot of overlap in topics. These similarities were also seen in the distance cluster diagrams. However, there is a very wide distribution between the topics seen in modern and classical texts. This may be surprising, since modern and classical texts are closer in eras, and would perhaps have more in common than with ancient texts. Modern and ancient text topics appear to have the most similarities, excluding a few outliers (32, 33). This is an example of how patterns in culture can be revealed and discovered through text analytics.

There are more similarities between ancient philosophical and modern texts than classical texts. This could be because when philosophers were writing in ancient Greece and Rome, it was a period of stability, which could also be said about modern times. But for classical times, between 1700 and 1900, there was great societal upheaval. Many of the philosophers in the classical era were taking a close look at themselves, at society, and at human nature. These authors were critiquing government systems, monarchy, and the rapid changes that were happening within the world.

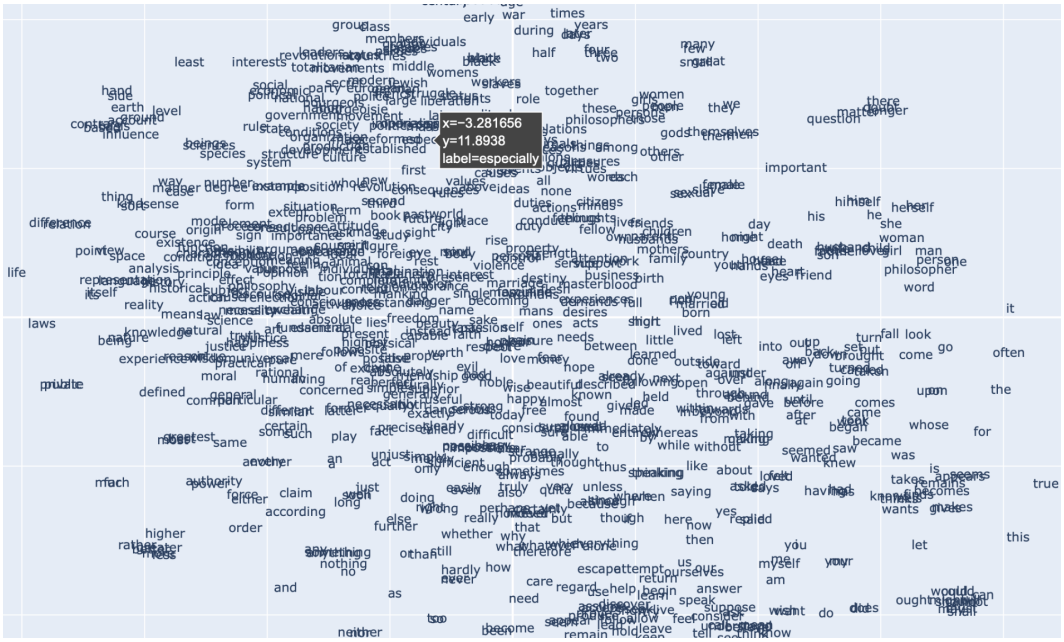
Word Embeddings

Word Embeddings use unsupervised Machine Learning algorithms for learning the meanings of words. The word contexts are mapped to vector space, where the contexts are the 'embeddings'. The distributional hypothesis states that words that occur in similar contexts tend to have similar meanings. Similar row vectors tend to have similar meanings.

With word embeddings, semantic algebra can be used to add, subtract, and multiply word vectors to extract additional insights. According to the laws of semantic change, over time, words that are used more frequently change less, and words that have the same meanings change at a faster rate.

Word2vec is a simple two layer neural network that produces word embeddings. Word2vec uses one-hot encodings for words and their contexts, factorizing the word-context matrix of pointwise mutual information (PMI), as a measure of the association between words. PMI compares the probability of two words occurring together, versus them being independent of each other.

T-SNE, or T-distributed Stochastic Neighbor Embedding, is used to visualize high-dimensional word vector data. The t-SNE plot for all the philosophical texts shows a wide distribution of word clusters in the vector space. One of the biggest clusters near the middle, at (-5, -1) of the plot include the words 'totalitarianism', 'religion', 'expression', 'ideal'. Another cluster at (-10, 10) includes words of Western identity and nationality, such as 'jewish', 'french', 'german', and 'european'. Another cluster around (-19, 0) includes the words 'beings', 'species', 'virtues', 'sciences', and 'principles'. These clusters make sense, and show the effectiveness of similar words plotted in vector space.



When looking at the word embeddings for all the texts written by women, the words are distributed evenly around the center of the plot, and interesting word clusterings can be seen. At coordinates (-12, 4) grouped words include 'theory', 'character', 'idea', 'possibility', 'condition'. In the center of the plot, at (-0.4, -1) the cluster includes 'desire', 'respect', 'passion', and 'love'.

For texts written by male authors, the word embeddings are skewed more toward the right of the t-SNE plot. A cluster near (0, 6) includes the words 'expression', 'consequence', 'attitude', 'importance', 'ideal'. Near the coordinates (8, 11) include the words 'wisdom', 'justice', 'injustice', 'happiness', and 'virtue'.

These words are clearly related, and could even be synonyms. There are more words that are less clustered, than there are clearly clustered words. This follows the distributional hypothesis, that similar words have the same or similar meanings.

```
#FEMALE
complete_analogy_female('religion', 'education', 'freedom')
```

```
[('free', 0.4744895100593567), ('assume', 0.46264272928237915)]
```

```
#FEMALE
complete_analogy_female('leaders', 'revolutionary', 'totalitarian')
```

```
[('european', 0.6137496829032898), ('political', 0.5984175205230713)]
```

```
#MALE
complete_analogy_male('religion', 'education', 'freedom')
```

```
[('husbands', 0.49874797463417053), ('thus', 0.4820350408554077)]
```

```
#MALE
complete_analogy_male('leaders', 'revolutionary', 'totalitarian')
```

```
[('political', 0.6747063994407654), ('european', 0.5981352925300598)]
```

```
#ANCIENT
complete_analogy_ancient('people', 'classes', 'groups')
```

```
[('bourgeois', 0.7378024458885193), ('parties', 0.6962724924087524)]
```

```
#ANCIENT
complete_analogy_ancient('life', 'death', 'happiness')
```

```
[('pain', 0.5637077689170837), ('sake', 0.5506163239479065)]
```

```
#CLASSICAL
complete_analogy_classical('people', 'classes', 'groups')
```

```
[('bourgeois', 0.7102118134498596), ('parts', 0.706534206867218)]
```

```
#CLASSICAL
complete_analogy_classical('life', 'death', 'happiness')
```

```
[('sake', 0.5197551846504211), ('pain', 0.5091485977172852)]
```

```
#MODERN
complete_analogy_modern('people', 'classes', 'groups')
```

```
[('parties', 0.7044128775596619), ('countries', 0.6912365555763245)]
```

```
#MODERN
complete_analogy_modern('life', 'death', 'happiness')
```

```
[('pain', 0.5679351091384888), ('pleasure', 0.5310956835746765)]
```

The analogies for male and female authors, as well as across the eras, align well with the topic models for the respective categories. There are similar analogies as topics, such as virtue, justice, passion, and society. While it is harder to get a sense of what the texts are about from the analogies, there are enough similarities that text meanings and author explorations can be extracted. The word analogies also provide insights into the change, or

lack thereof, of words between genders and across eras. There are many similarities between the texts by male and female authors. For example, when looking at the words 'leaders', 'revolutionary', and 'totalitarian', in a strong cluster in t-SNE plots, the resulting analogies for male and female authors were the same, 'european' and 'political'.

While there are some differences over the eras, there appear to be more similarities than differences. Many of the words remain the same from ancient times to modern times. This may suggest that core ideas and language have not changed very much. However, there are often subtle changes in words that may not be core ideas. For example, in the last analogy comparison across eras, 'life', 'death' and 'happiness' all result in the analogy 'pain'. But the other analogy changes over time, from 'sake' in ancient and classical times, to 'pleasure' in modern times. This change is rather interesting, because it may give a small glimpse into a cultural shift over time. In another example, 'bourgeois' and 'parties' is language used in ancient and classical times, but in the modern era, this language shifts to 'countries' and 'parties'. 'Bourgeois' is not a word that is used in modern times, but there are indications that the meaning has been kept in time, by similar words in the cluster that have endured, such as 'class.'

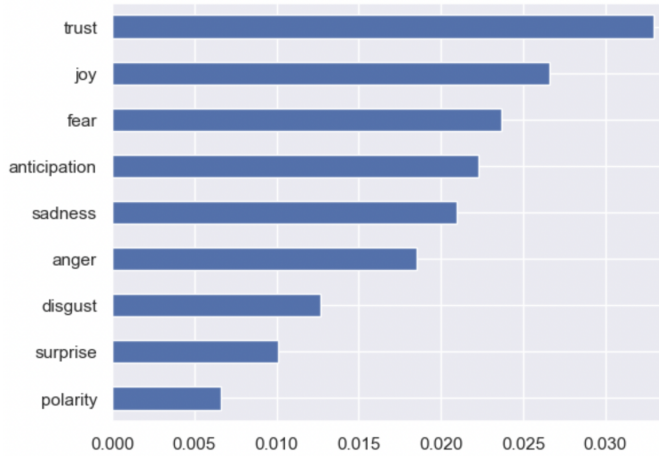
Sentiment Analysis by Gender

Sentiment Analysis is the computational treatment of opinion, sentiment, and subjectivity in text. 'Sentiment' can be defined as values, opinions, emotions, and attitudes. The emotional weight of a text as a whole is gathered, with the assumption that it is tied to the sentiment of the person writing the text. For this sentiment analysis of the philosophical texts, the NRC Word-Emotion lexicon is used. The NRC lexicon is a list of English words and their association with eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The two sentiments are explored, negative and positive, which are used to analyze polarity.

Emotions in Texts by FEMALE Authors

```
female[emo_cols].mean().sort_values().plot.barh()
```

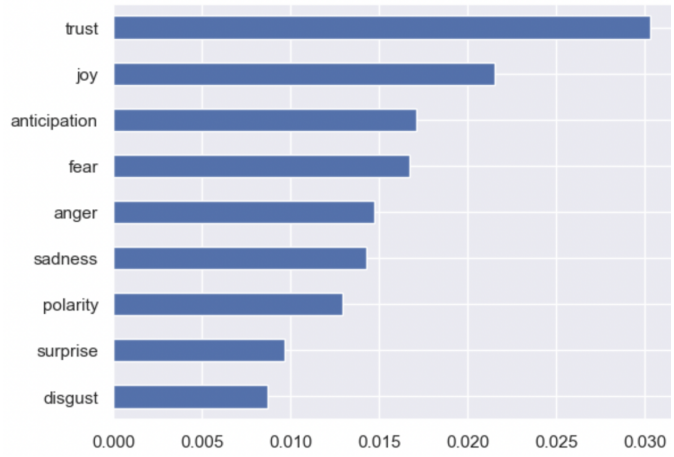
```
<AxesSubplot:>
```



Emotions in Texts by MALE Authors

```
male[emo_cols].mean().sort_values().plot.barh()
```

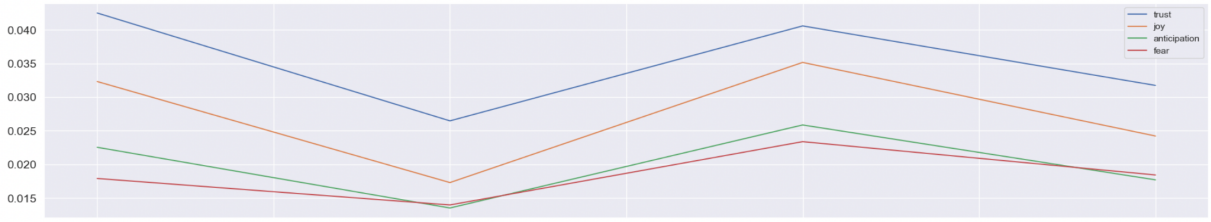
```
<AxesSubplot:>
```



When examining the top emotions in philosophical texts written by female and male authors, the top two emotions are the same, 'trust' and 'joy'. The emotions deviate at the third top emotion. For female authors, the following top emotions are 'fear', 'anticipation', 'sadness', and 'anger.' For male authors, the next top emotions are 'anticipation', 'fear', 'anger', and 'sadness'.

Sentiment for Ancient Texts

```
title_ancient = ancient.groupby(TITLE)[emo_cols].mean()
plot_sentiments(title_ancient, ['trust', 'joy', 'anticipation', 'fear'])
```

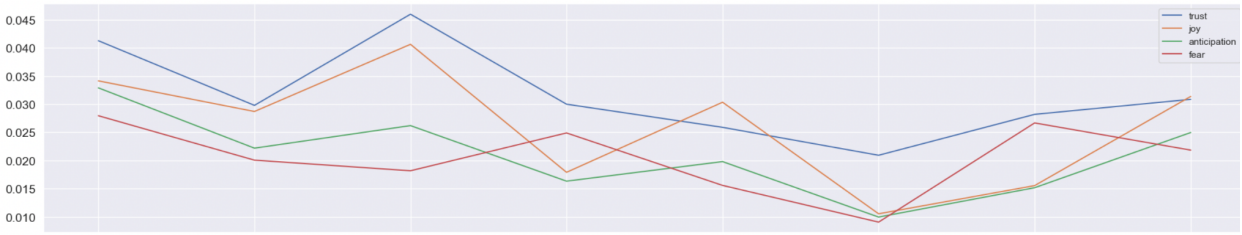


Sentiment for Classical Texts

```
title_classical = classical.groupby(TITLE)[emo_cols].mean()
plot_sentiments(title_classical, ['trust', 'joy', 'anticipation', 'fear'])
```

Sentiment for Modern Texts

```
title_modern = modern.groupby(TITLE)[emo_cols].mean()
plot_sentiments(title_modern, ['trust', 'joy', 'anticipation', 'fear'])
```



Looking at sentiment over all the texts in each time period may not be the most revealing study of cultural patterns in philosophical texts through time. However, there are a few insights that can be gained. In the ancient texts, the range of emotions is the most steady of all the eras, with the least amount of variance. The texts written in the modern era have the widest range of emotion and the most variability. This likely does not mean that humanity today has more emotions, or even that they are more expressive with emotions. It also does not necessarily mean that the topics explored were different, as seen from the topic modeling. The trends likely point to cultural differences in the understanding of philosophy, and how authors approached the subjects in their writing. In ancient times, philosophy was seen as a science and gave birth to many of the scientific fields we have today, such as physics. Over time, philosophy began to change, particularly during the classical era, as explored in this project, to become a study of society, government, human nature, and emotion. Philosophy in modern times is often a method to examine ourselves.

Conclusion

This project uses text analytics, NLTK, and NLP practices to extract cultural patterns and meaning from twenty philosophical texts that span 2,066 years. Topics and sentiment were examined across all texts, as well as between genders of authors and across time periods. Some discoveries were not surprising, such as the main topics of the philosophical texts including virtue, justice, universe, consequences, reason, pain and happiness. But some surprising trends were revealed, such as there being more similarities between ancient and modern texts than classical and modern texts. Also, while there are some differences in topics between male and female authors, there are more similarities in key topics than differences. There are also many similarities in top emotions between male and female authors.

Additionally, a theory was explored that suggests Harriet Taylor Mill may have written or highly influenced the text *Utilitarianism*, which was published by and attributed solely to John Stuart Mill. Through TF-IDF and PCA,

comparing *The Enfranchisement of Women* published by Harriet Taylor Mill, and *Utilitarianism* by John Stuart Mill, revealed that word usage is very similar between the two texts. This is an exciting discovery, which suggests that historian's theory may indeed be true that Harriet Taylor Mill could have authored or co-authored the famous text, *Utilitarianism*. To further test this argument, a larger sample of texts written by both people should be collected and analyzed. It would also be interesting to explore a wider range of philosophical texts to see if others were actually written by women, even though they are attributed to male authors.

Text Analytics is a powerful tool to explore patterns in texts over time and across groups. This project is an example of what can be gained from this type of exploration. However, this project is just a small example of what is possible, and barely scratches the surface. More meaning and insight can be gained from a larger sample size of philosophical texts, including a larger and more balanced gender distribution. Another interesting area to explore could be examining a large collection of philosophical texts from different regions of the world. It may be easier to uncover cultural patterns when looking at texts from a wide variety of cultures. Many of the similarities in the texts for this project may be due to most texts being by Western authors. It is possible that there are more differences in cultural patterns between different world regions, than across gender and era within the same region. This is an exciting area to explore further, and an apt example of the power of text analytics in uncovering cultural truths.

List of Texts Used In Project

- 10 Western Philosophical Texts by Male Authors:
    - Aristotle: *Nicomachean Ethics* (4th century BC)
    - Plato: *The Republic* (4th century BC)
    - Cicero: *On Moral Duties / De Officiis* (44 BC)
    - David Hume: *An Enquiry Concerning Human Understanding* (1748)
    - Immanuel Kant: *Fundamental Principles of the Metaphysic of Morals* (1785)
    - Karl Marx: *The Communist Manifesto* (1848)
    - John Stuart Mill: *Utilitarianism* (1861)
    - Friedrich Nietzsche: *Beyond Good and Evil* (1886)
    - Søren Kierkegaard: *Selections from the Writings of Kierkegaard* (1923)
    - Michel Foucault: *The Order of Things* (1966)
- 3 Non-Western Philosophical Texts by Male Authors:
    - Laozi (Lao Tzu): *Tao Te Ching* (400 BC)
    - Paulo Freire: *Pedagogy of the Oppressed* (1968)
    - Herman Hesse: *Siddartha* (1922)
- 7 Western Philosophical Texts by Female Authors:
    - Mary Wollstonecraft: *A Vindication of the Rights of Men* (1790)
    - Mary Wollstonecraft: *A Vindication of the Rights of Woman* (1792)
    - Harriet Taylor Mill: *The Enfranchisement of Women* (1852)
    - Simone de Beauvoir: *The Second Sex* (1952)
    - Hannah Arendt: *The Origins of Totalitarianism* (1951)
    - bell hooks: *Ain't I a Woman: Black Women and Feminism* (1981)
    - bell hooks: *Feminist Class Struggle* (2002)

References

Archive-It, 2014, https://archive-it.org/.

Bansal, Shivam. *Beginners Guide to Topic Modeling in Python*, Analytics Vidhya, 2020, analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/.

Gupta, Shashank. *Sentiment Analysis: Concept, Analysis, and Applications*, Towards Data Science, 018, https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

Project Gutenberg, 2022, https://www.gutenberg.org/.

University of Virginia, Course: DS 5001, Sree Mallikarjun, Fall 2022.

Violante, Andre. A*n Introduction to t-SNE with Python Example*, Towards Data Science, 2018, https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1.