

## **Analyzing Works of Early Modern British Philosophers by Digital Tools**

In this final project, I conduct a few tasks of textual analysis on the texts of early modern British philosophers. Specifically, these tasks include finding keywords in contexts, document clustering and classification based on keyword frequencies, and sentiment analysis. In this report, I briefly walk through the methods I have used, and I also report interesting findings from my results.

### **1. Data Collection and Preprocessing**

The object of this research is work by British Philosophers in modern era (1600-1900). These philosophers include Thomas Hobbes, John Locke, George Berkeley, David Hume, Jeremy Bentham, John Stuart Mill, Henry Sidgwick, and others. As I have mentioned in my project proposal, one reason I choose to study these philosophers is their works are all out of copyrights. There are a lot of free texts online. In particular, there are two resources for me to find these books. The first one is Project Gutenberg. (<http://www.gutenberg.org/>) According to its website, Project Gutenberg contains over 54,000 free eBooks that are out of copyrights, and these books covers a variety of categories. These books also come conveniently in well-edited .txt format, which is easy to be processed by many programming languages. But a problem for Project Gutenberg is that it only provides a limited number of books in philosophy. To build a more comprehensive corpus for my project, I need to find another place where I can get more texts in philosophy.

The second and (the main) source of data I use is Intelix Past Master (<http://www.nlx.com/>). It is a website that provides important texts and translations in the history of philosophy and thought. According to their website, “Past Masters series encompasses the largest collection of primary source full-text electronic editions in philosophy in the world.”(<http://www.nlx.com/home>) But a problem for getting texts from Intelix Past Master is that users do not

have direct access to these books in .txt format. Thus, in order to do this project, I have to use a software package to scrape texts from different webpages and put them together.<sup>1</sup>

To fulfill my scraping goal, I use the R software package to do the work. I first manually collect the links of all webpages I scrape texts from, and then, I use the `rvest` library. For each of the webpage, there are only two things I need to scrape – book title and the main text. Given the friendly-organized XML codes of Intelix webpages, all book titles are under the scope of `<.volume_title>` tag and all the main texts are under the scope of `<p>` tag. So scraping contents under these two tags become the statements in the main function. When my program scrapes the main texts, I also make it remove the newline tag `"\n"`. For detailed codes, please see my attachment. I have separate code files for each of the seven philosophers, and outputs are seven .txt files with names of these philosophers. I put the files under the `/authorCorpus` folder.

## 2. Searching and Plotting Keywords

Now we have all the text files for processing. The first (and also the main) part of my project is Keywords in Context (KWIC). That is, to find key words in all books by each authors. There are software programs like AntConc that can directly generate a keyword list for a given text, but I am writing my own program, so I decide not to use any out-of-box software packages. Not directly using such software packages also give me the freedom of being able to control more parameters.

The details of my implementation of searching for keywords is as follows. First, the programs removes stopwords. Stopwords are words that commonly appear across different texts and do not contain too much significance in identifying unique patterns of individual authors. Depending on what kind of corpus one is dealing with, the scope of stopwords could vary. In my project, I use the stopwords list provided by NLTK (a Python library for Natural Language

---

<sup>1</sup> Another problem I have encountered with Intelix Past Master is that the website needs institutional subscription but the University of Georgia has not subscribed to it yet. So I end up using my library account of Rice University, which provides access to all the books.

Processing).<sup>2</sup> I also provide an option of adding more stop words in my program. If one needs an extended list of stop words, she can just change the global Boolean variable MORE\_SW from False to True. The extended list contains stopwords from the Brown Corpus.<sup>3</sup> The next step is to tokenize texts. Tokenizing is the processing of dividing a text into a word list, which is an essential step to count the word frequencies in a text. I again use functions provided by NLTK to accomplish this step. First, I use `sent_tokenize` function to divide a text into each sentence, and then I use `word_tokenize` function to divide sentences into different words. The reasons for converting a text into a collection of sentences first is that we do not want to separate words solely based on space between strings. Punctuations and other sentence dividers also matter, which is why the program has its first step.

Although these two functions seem good, more processes are still needed in order to get totally tokenized texts for counting keywords. The results from the two NLTK functions above also distinguish between words in upper and lower cases. Also, these two functions return numbers and other special characters that are not sentence dividers, which we also need to remove. So I use a regular expression function to search for all characters that only contain English letters, and change them into lower cases. With the filtered word list, we can build a dictionary with all word strings and their corresponding frequencies in the corpus.<sup>4</sup> In the `word_count` function. It put non-stopwords as keys in a dictionary and their counts as corresponding values, and then the program sorts the dictionary based on the counts. The first item in the dictionary is the word with the highest counts.

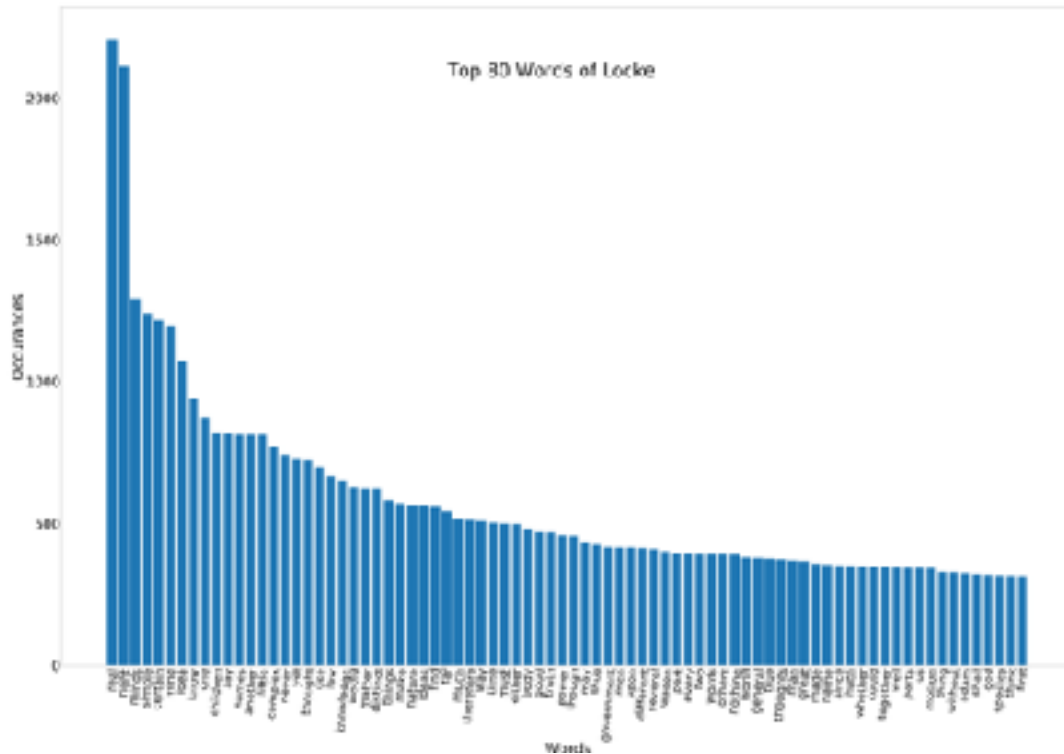
So now we have a keyword frequency ranking for each philosopher. Next step is to plot these results in a graph and see the distributions of these keywords and their counts. Matplotlib, a

---

<sup>2</sup> NLTK stands for “Natural Language Toolkit”, for more documentation, see <http://www.nltk.org/>.

<sup>3</sup> Brown Corpus is a corpus on American English from 1960s.

<sup>4</sup> Also, we can do stemming. I also write a stemming function in it, but considering the difference between modern English and contemporary and for the purpose of looking at the words in a better way, I do not use the stemming function here.



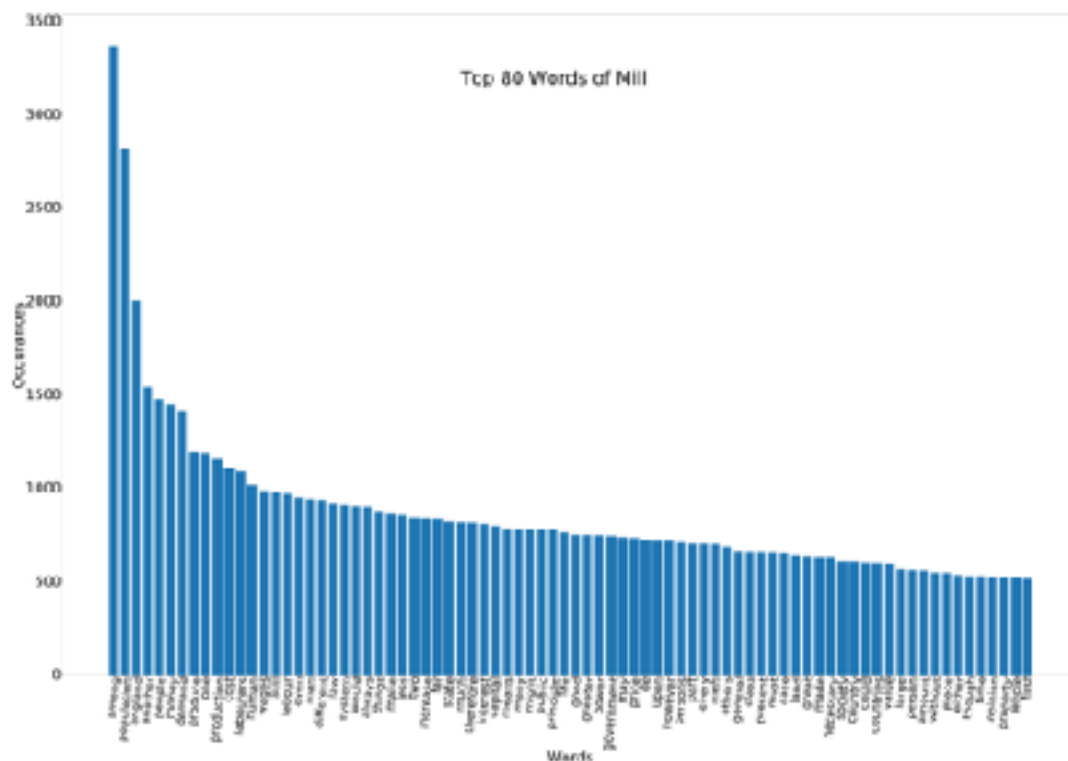
Graph 1. Top 80 words of Locke

Python library, provides a convenient way of plotting the results. I also provide users an option of choosing how many words they would like to see in the plot. I choose 80. In the plot, the x-axis indicates different words and the y-axis indicates the numbers of counts. All the results go into under the /keywordResults folder. Depending on how big the size of an author's file is, the running time for plotting ranges from 10 seconds to 30 seconds, which is not too bad. Graph 1 is an example of what a plot looks like.

Graph 1 shows the top 80 words in the works of Locke. As one of the most comprehensive philosophers of all time, Locke's writings range from epistemology, philosophy of mind, to ethics and political philosophy. The results above also shows it is the case. Some words, like "mind", "body", and "knowledge", concern about epistemology and philosophy of mind, whereas some other words like "right", "law", and "governments" belong to the field of political philosophy. The relation between a philosopher's fields of expertise and frequent words in his writings can also be

found in their corresponding graphs as well, which means that by looking at the keywords of an author, we can roughly guess which field(s) of philosophy his writings are in.

More importantly, based on these plots of word frequencies, we can compare the topics of different philosophers. In history of philosophy, the most important contribution of Mill (together with Bentham) is utilitarianism, which roughly claims that an action is ethically good if and only if it

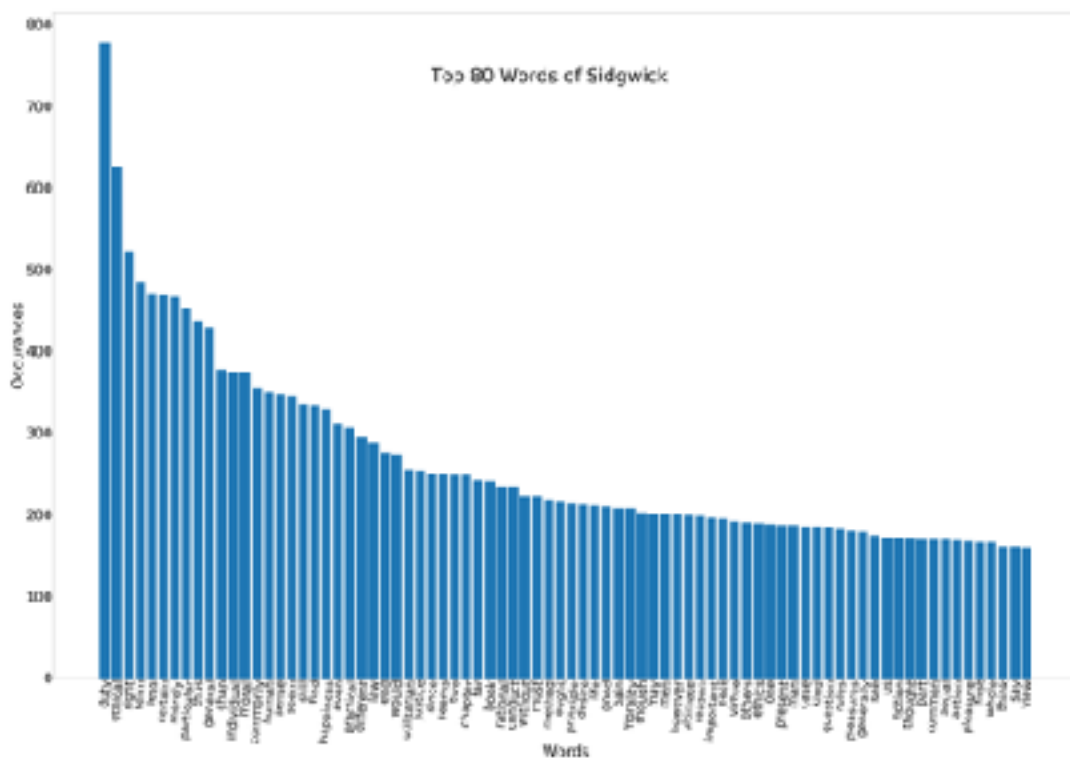


Graph 2. Top 80 words of Mill

maximizes the its utilities (or, "happiness", in layman's term). This doctrine is also somewhat shown in Graph 2 for the top 80 words in Mill's work.

As we can see, words like "price", "money", "wages" are ranked very high in this graph, which is not the case in previous philosophers' work. There are some similarities between Mill and Locke -- they both concern about law, person, government, etc. However, Locke does not take utilitarianism as his principle. He concerns more about liberty and the form of government (as what

he shows in his *Treatise of Government*). Words like "Adam", "God" also appear in top 80 words of Locke, but it is not the case in the work of Mill. This pattern also coheres with a secular turn in the history of philosophy -- in the earlier period of modern philosophy, philosophers like Descartes, Hobbes, Locke related their work more frequently to religion than later philosophers like Bentham and Mill. Philosophy had become more independent with religion. Although these are all theists, the foci of their philosophy are different.



Graph 3. Top 80 words of Sidgwick

Based on keyword lists, we can also conduct analysis on classification of these philosophers' works. For instance, Sidgwick is arguably the most important utilitarian philosopher after Bentham and Mill. Top words of his work are more similar with those of Bentham and Mill than other non-utilitarian philosophers. Here is the graph of the top 80 words of Sidgwick (see Graph 3).

We can find keywords "happiness", "reason", "pleasures" etc. in the graph. These words can also usually be found in the plots of other utilitarian philosophers like Bentham and Mill. The fact

that Sidgwick uses these terms frequently shows that the focus of his writings is mainly ethics and utilitarianism, which is the same with Bentham and Mill. Thus if we want to categorize different philosophers' works together based on the topics in these works, Sidgwick's should be in the same category with Bentham's and Mill's. Similarly, Locke and Berkeley both put heavy focus on epistemology and philosophy of mind, so some books of them should be grouped together. In the next part, I use more contemporary methods to test these hypotheses.

### 3. Classification

In this part of my project, I conduct more detailed analyses on classification of philosophy books by early modern British philosophers. Instead of comparing at the level of different authors, I go deeper and examine at the level of different books by these authors. The main reason for doing so is that I only have a handful of authors and doing classification between them is not as meaningful as doing classification based on different books. A philosopher like Locke might write on many different fields of philosophy and take different inspirations from philosophers with different specialities. Doing classification on their books can help reveal more connections between these philosophers and their works.<sup>5</sup>

From the results of online text scraping, I put all the .txt files of these books into the /authorBooks folder. Since my main focus are seven philosophers (Hobbes, Locke, Berkeley, Hume, Bentham, Mill, and Sidgwick), I choose their classical works first. A fine classification should at least cluster philosophers or books from the same school together. As I have stated earlier, Bentham, Mill and Sidgwick are all proponents of utilitarianism. Thus, their books in ethics should be put into the same category. In order to test this hypothesis, I leave out the work of Sidgwick for testing. To get

---

<sup>5</sup> Readers might have a question here -- why not just have plots for these different books? I think there are two reasons. First, there are a lot of books and plotting all of the them might take a while. Second, I definitely want to have an author-level comparison first before going deeper into book classifications. Plotting keyword frequencies based on different authors serve that goal. After all, if readers want to get the keyword frequencies from particular books, it is definitely doable. She can just tweak the files in the corpus. (Detailed instructions are provided in another file.)

more diversity in my classification, I also added the book *An Inquiry into the Human Mind* from Thomas Reid, who is an empiricist and heavily influenced by Locke and Hume in epistemology and philosophy of mind.

### 3.1. Unsupervised Classification

My classification method is an unsupervised method. That is, I do not tell the program anything about the relations between these philosophers and where their specialities lie. Instead, my program would look at the keywords with different frequencies in these books and identity the unique high-frequent non-stop words in each book. Books with similar highly-frequent non-stop words should be clustered together.

The first step is to tokenize these books. My program has already got a `tokenize` function, which can be conveniently used again. After tokenizing all the texts, I use the function of `tf-idf`, which stands for “term frequency-inverse document frequency”, to get a new representation of these words. Tf-idf could be represented as follows:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$  = number of occurrences of  $i$  in  $j$   
 $df_i$  = number of documents containing  $i$   
 $N$  = total number of documents

In this formula,  $i$  stands for a certain word and  $j$  stands for a certain book. After processing through the `tf-idf` formula, every word becomes a vectorized representation of its frequencies in all processed books. It is possible that the above formula may not be able to distinguish some words with that have the same frequencies in all processed books. And it is also possible that the above



formula fails to account for more nuanced things in language like syntax and grammar. However, if we only look at keywords in context and compare with them, tf-idf is still worth trying. Similar method is also used in Hoover (2013). It takes less than a minute to finish the processing.

Next step is to use an unsupervised algorithm to measure the distances between these vectors. I use KMeans algorithm provided by sklearn,<sup>6</sup> which randomly selects K numbers of centroid points and adjust the locations of the points such that similar documents are close to the same centroid point. Users still need to specify how many clusters there are. In my case, I choose 6, because I have seven philosophers but I only have one book from Reid and it is very similar with some of Berkeley's and Locke's. The results of KMeans clustering are as follows (Table 1). Books are sorted based on which cluster each belongs to.<sup>7</sup>

Some interesting results can be drawn from the result. Readers should note first that the number of each cluster do not really convey any significant meaning. It is just a result from a random process that KMeans algorithm puts centroids in the beginning.

We can see that this unsupervised algorithm sometimes classifies books based on authors, but more often it classifies books based on topics. For instance, Locke's *Treatise of Government* is about political philosophy, which is also the main topic for the famous *Leviathan* and the less famous *Elements of Law* by Hobbes. And Locke and Hobbes also live in similar periods in history, so it is no surprise that they have many similar keywords, including “sovereign”, “government”, “commonwealth” (which sounds very British), etc. Interestingly, the most frequent term is “hath”, which is not a common word in contemporary English at all. In fact, “hath” is equivalent to “has” in

---

<sup>6</sup> For more documentation, see <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>.

<sup>7</sup> Here are what some of these names stands for — hobbes-law: *Elements of Law*; locke-treatise: *Treatise of Government*; hobbes-element: *Elements of Philosophy*; mill-representative: *On Representative Government*; bentham-FOG: *Form of Government*; mill-liberty: *On Liberty*; mill-util: *Utilitarianism*; bentham-POM: *Principles of Morals*; hume-enquiry: *An Enquiry Concerning Human Understanding*; hume-treatise: *Treatise of Human Nature*; hume-morals: *An Enquiry Concerning the Principles of Morals*; hume-religion: *Dialogues on Natural Religion*; berkeley-TOK: *Theory of Knowledge*; berkeley-TD: *Three Dialogues between Hylas and Philonous*; locks-understanding: *An Essay Concerning Human Understanding*; hobbes-liberty: *On Liberty and Necessity*

	book_title	cluster
0	hobbes-leviathan	0
3	hobbes-law	0
7	locke-treatise	0
15	mill-representative	1
14	bentham-FOG	1
4	mill-liberty	1
5	mill-util	1
13	bentham-POM	1
10	hume-enquiry	2
8	hume-treatise	2
9	hume-morals	2
17	hume-religion	2
16	reid-mind	3
11	berkeley-TOK	3
12	berkeley-TD	3
6	locke-understanding	3
2	hobbes-elements	4
1	hobbes-liberty	5

Table 1. Unsupervised Classification Result

Top terms per cluster:

Cluster 0 words: hath, sovereign, adae, christ, government, commonwealth, covenant, father, dominion, children, king, unto, civil, sovereignty, war, every, saviour, law, moses, amongst,

Cluster 1 words: government, society, public, utilitarian, happiness, moral, offences, conduct, interest, political, utility, social, duty, however, class, character, punishment, morality, interests, institutions,

Cluster 2 words: ideas, ideas, cleanthes, sentiment, causes, reasoning, society, philo, passions, sentiments, entirel y, philosophy, moral, species, existence, influence, connexion, project, gutenbergm, passion,

Cluster 3 words: ideas, phil, idea, exist, motion, existence, sensation, perceived, bodies, extension, sensations, ha th, without, substance, sensible, colour, visible, distance, shape, human,

Cluster 4 words: b, c, f, e, strait, q, h, b, motion, line, k, angle, arch, lines, w, n, wherefore, moved, superficies s, b,

Cluster 5 words: lordship, hath, election, sin, maketh, causes, necessitated, prayer, lord, deliberation, thou, lord, god, spontaneous, good, compulsion, saith, rain, doth, necessary,

Figure 1. Screenshot of the Top Terms per cluster from classification

contemporary English. Words like “has” or “have” are usually considered to be stop words. The inclusion of “hath” shows that contemporary stop words in some cases may not work so well in modern English, for instance, the identification of “has”. Next group is utilitarianism, which puts all the works of Bentham and Mill together, and it is marked by the words like “happiness”, “utilitarian”, etc. So I consider it to be a good result.<sup>8</sup> Cluster 2 is also a good — it puts together all the works of Hume in the list. Although these books of Hume covers a few different fields, like ethics, epistemology, and philosophy of religion, the algorithm still puts all these works together. It might be the case that compared to other modern British philosophers, Hume has a distinctive writing style so it is easy to put these books together. Group 3 is more about perception, epistemology, and philosophy of mind. Locke, Berkeley, and Reid are all seminal figures in British Empiricism. All these books in group 3 touch on similar issues, though they might come with different writing styles (like Berkeley’s *Three Dialogues between Hylas and Philonous*). The last two clusters are not as clear as the previous four. Hobbes’s *Element of Philosophy* has its own cluster. In the beginning, I was not sure why. And the keywords are also very weird compared to other. Since I had never read that book before, I decided to look into that book. My interesting finding from that book is, although it is titled “Element of Philosophy”, it has actually got a lot of physics into it. No wonder why it is not widely read by philosophers! The result also corresponds with the top words in Cluster 4, which includes a lot of single letters (they are probably used in graphs in physics) and terms in physics like “angles”, “arch”, etc. The last cluster only contains Hobbes’s *On Liberty and Necessity*, I haven’t read that book either. But as far as I can tell, it concerns a lot about religion.

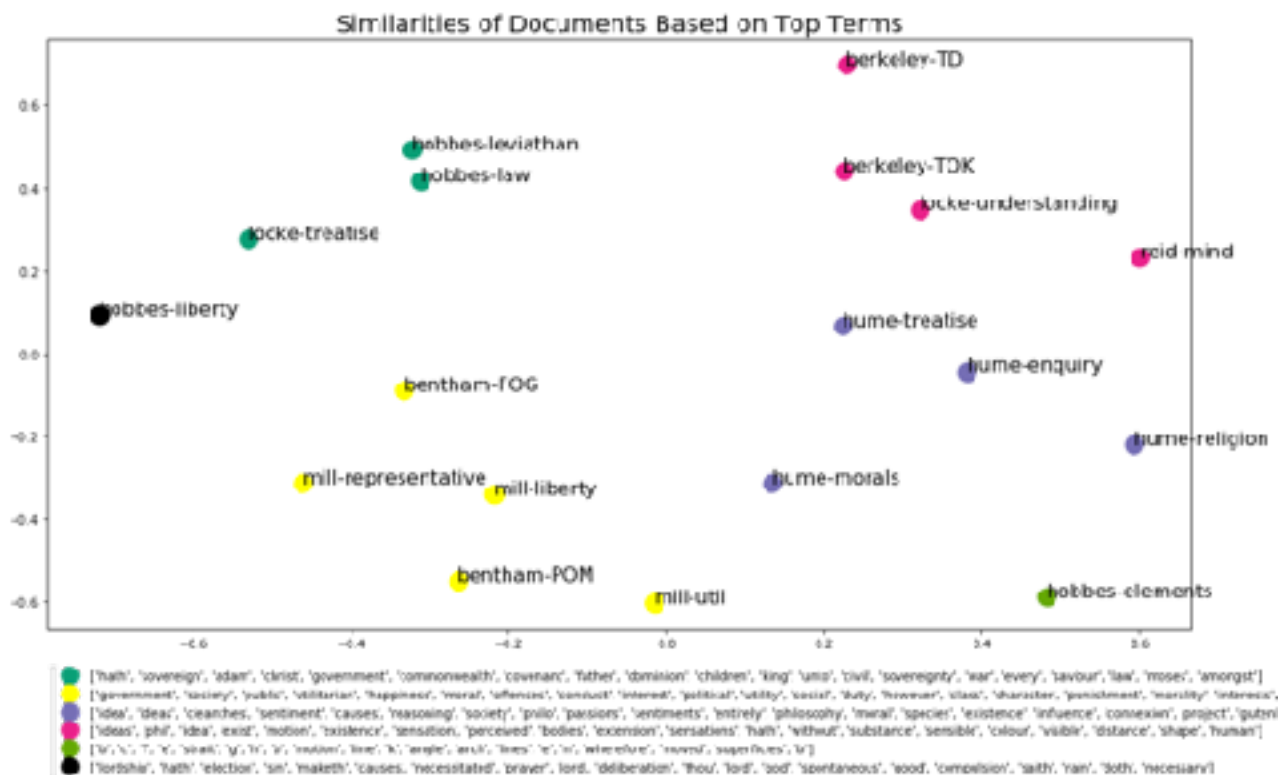
### 3.2. Similarities of Documents

However, categorization based on KMeans clustering is not enough. Depending on where the algorithm puts the initial centroids, the results might fluctuates. I have tried my best to reduce

---

<sup>8</sup> In an additional experiment I have done later in the project, I also found that after my model has learned the classification of these books, it will also categorize Sidgwick’s book into the same category with that of Bentham’s and Mill’s, which means the model does have some predictive power.

the randomness in the process by increasing maximum running time and and, but in rare occasions, the results might still differ a little. In order to get a more comprehensive picture on the similarities and differences between these documents, I look into the metric of cosine similarity and use Ward hierarchical clustering to draw a dendrogram based on the results of cosine similarities between these books.<sup>9</sup> After the processing, we have the following graph and dendrogram (Graph 4 and graph 5).



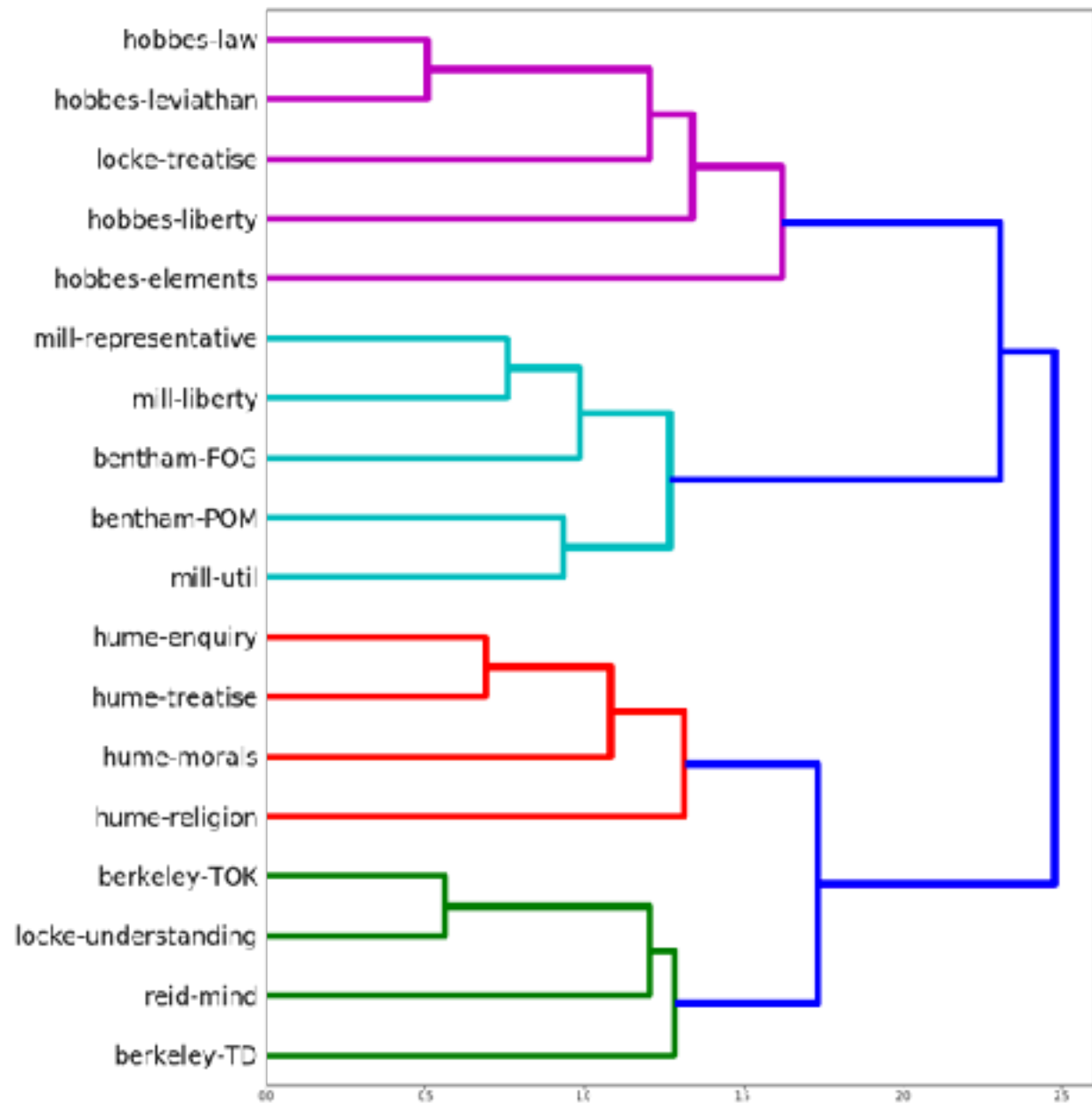
Graph 4. Similarities of Documents Based on Top Terms

In my opinion, cosine similarity provides more detailed information than KMeans clustering on classifying these documents, because even if some documents belong to different clusters and are shown in different colors in the graph, by looking at distances and their relative locations in the graph, we can still have a rough idea of how they are related. For instance, Hobbes's *On Liberty and*

<sup>9</sup> Detailed documentation for the algorithms of cosine similarities and Ward hierarchical clustering can be found at [http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine\\_similarity.html](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html) and <http://scikit-learn.org/0.15/modules/generated/sklearn.cluster.Ward.html>, respectively. Basically both methods are built upon the results from tf-idf.

*Necessity* was clustered into its own category, but it is in general still about ethics and political philosophy, which also corresponds to its close distance to the Locke's *Treatise on Government*. What is more, Hume's works covers different fields of philosophy, and as we can see from the above graph, his works in epistemology and moral psychology are close to the Locke and Berkeley's works in these fields. On the other hand, Hume's work in ethics is close to the works by Bentham and Mill, even though Hume is not considered to be a utilitarian in any sense.

The dendrogram (Graph 5) also contains very helpful information. There are five colors in the dendrogram. And the blue cluster is at the highest level of these clusters. There are the four different colored clusters below— purple, light green, red, and dark green. Purple shows all the works by Hobbes and Locke's *Treatise*, which, as we have already analyzed, is similar with Hobbes's work in political philosophy. As Hobbes's *Elements of Philosophy* contains its unique content in physics, it does not seem to be too close to other books by Hobbes in the hierarchical clustering. The cluster of light green shows the works by utilitarians, including Mill and Bentham. There are also basically two types of books by Bentham and Mill — one in ethics, and the other in political philosophy. *Principles of Morals* and *Utilitarianism* are both seminal works in ethics, but the other three books, *On Representative Government*, *Form of Government*, and *On Liberty* all concern more about political philosophy. Thus they are clustered together in lower hierarchies. The red cluster indicates books by Hume, which is the same with the cluster we have classified before. The last cluster contains books from Berkeley, Locke, and Reid, which confirms our hypothesis — Reid's work is indeed clustered together with Berkeley's and Locke's books in epistemology and philosophy of mind. But we can also see that Reid's book is still not clustered together with Berkeley's and Locke's books at the lowest level. That is probably because Reid is chronically later than Berkeley and Locke, and also partly influenced by Hume. Another thing that is worth mentioning is Berkeley's book *Three Dialogues between Hylas and Philolous*. This book has a couple of distinctive features. First, compared to other books on the list, it is a much shorter book. Second, it is written in the form of a dialogue, just



Graph 5. Dendrogram of Clustering Early Modern Philosophy Books

like Socrates's dialogues. In my clustering result, *Three Dialogues between Hylas and Philolous* is still categorized into the dark green category with other books in epistemology and philosophy of mind together, but it is only at a higher cluster. That is probably because of its distinctive style and a lack of variety in its word use because of the short length.

## 4. Sentiment Analysis

### 4.1. Introduction

In my project, I have also conducted sentiment analysis on these books. As I have stated in my project prospectus, my hypothesis is that sentiment analysis does not really work for books in philosophy. There are two reasons for my hypothesis. First, sentiment analysis might be easy to conduct and usually helpful if the object of study is a movie or online shopping review, where we can just look at keywords, see how many emotional words there are, and analyze whether these words indicate positive or negative sentiments. However, it might not be simply the case for books in philosophy. First, different from movie reviews, philosophy sometimes could be very esoteric or sarcastic. More importantly, philosophical discussions are largely argumentative, so there should be less emotion involved in these texts. So sentiment analysis will not really help understanding these texts. But given I have already collected my corpus and I have different books in .txt format, I think I should just give a try and test my hypothesis. It should also be fun to look at the results.

Instead of using the 18 books for classification earlier, I add more books into sentiment analysis this time. These books include not only philosophy books by modern British philosophers from 1600 to 1900, but also books by authors from continental Europe during the same period (including Baruch Spinoza, Immanuel Kant, Karl Marx, etc.), books by American authors (*The Federalist Papers*, which is partly influenced by the political philosophy of Locke.), books by Bertrand Russell (which are in the 20th century and usually out of copyrights also), and even some of my papers in philosophy.

## 4.2 Implementation and Results

To implement this part of my project, I use the Python library of Textblob.<sup>10</sup> According to its website, Textblob is a high-level library for processing textual data, and it is compatible with both Python 2 and 3. It is a high-level library because it is built on NLTK and Pattern<sup>11</sup>, and it uses some functions from them to facilitate the speed and convenience of textual processing. Textblob does a

---

<sup>10</sup> For more documentation, see <https://textblob.readthedocs.io/en/dev/>.

<sup>11</sup> See <https://www.clips.uantwerpen.be/pages/pattern-en>.





After getting the numbers of subjectivity and polarity for each book or paper, the last step I need to do is to plot these results. Here is a scattered plot (Graph 6) I have for the subjectivity and polarity for each book on my test list.

In this scattered plot, each blue dot represents a book or paper on my list. Here are some interesting findings from these results. First, most texts do not have strong emotional leanings, as we can see from their polarity scores, most texts (especially ones in the field of political philosophy) are around 0.1, which means it is barely above the threshold of being positive,<sup>13</sup> and even if we take some texts to have positive emotions, the emotions are not really very strong. Thus, if we assume the methods provided from Textblob, NLTK, and Pattern are valid, then the results above confirms our hypothesis that philosophical texts are usually argumentative unemotional. And, on the other hand, if we are certain that these texts in philosophy are unemotional (by other means of checking, for instance, manually reading these texts and making judgments), then it validates the use of Textblob, NLTK, and Pattern on sentiment analysis on texts in philosophy. Some outliers from the plot above are also interesting to examine. For instance, Hobbes's *Elements of Philosophy* has the lowest score of polarity and subjectivity, which is probably due to its large portion of contents in physics. Marx's *Communist Manifesto* and Engels's *The Condition of the Working Class in England* (represented as "marx-CM" and "engels-condition in the plot respectively) do not carry high scores of polarity and subjectivity either, which surprises me. I would expect their books to be more emotionally loaded. On the contrary, Thomas More's *Utopia* has the highest score of polarity and the second highest score of subjectivity, which probably reflects the book's sarcastic feature and its difference with Marx and Engels's works, although they are all pertinent in the history of Communism. Last but not least, based on these scores, it seems that some authors have consistent

---

<sup>13</sup> In fact, when interpreting these scores, the documentation of Pattern suggests that 0.1 is usually a threshold that delivers the best results. So we use 0.1 as the threshold for being emotionally positive, then most texts are around the threshold, which confirms my hypothesis that texts in philosophy are largely argumentative, so their sentiments are not strong.

writing styles (evidenced by consistent subjectivity and polarity in their books) whereas some others do not. I do not think consistent scores of polarity and subjectivity should be a metric for us to judge whether the an author is good or bad, but these scores may still nevertheless reflect their distinctive and usually excellent writing styles (e.g. Russell, Hume) or sometimes the broadness of an author's specialities (e.g. Hobbes).

## 5. Conclusion

In this project, I have conducted textual analyses on works by early modern British philosophers. The main parts of my analyses are keyword frequencies and classification based on these frequencies. In addition, sentiment analysis is also applied on these texts. I have also shown interesting findings in each part of my analyses. The results from these analyses are largely coherent with my previous understanding of these texts. But for the texts that I was previously less familiar with or texts that I have never read before, these results do open a window for me to explore these texts and gain more general understandings on what texts are about. If these results and findings are all plausible, then it also shows that textual analysis in the field of philosophy, which is currently under-studied in Digital Humanities (as I have stated in my previous papers), is worth more attention and endeavors.

## Works Cited (excluding ones cited in my previous papers for this project)

Hoover, David, Textual Analysis, *Literary Studies in the Digital Age*. MLACommons. Accessed November 2017 on: <https://dlsanthology.mla.hcommons.org/textual-analysis/> [DOI: 10.1632/llda.2013.3]

## Appendix I. List of Stopwords

'doesn', 'did', 'then', 'i', 'it', 'do', 'm', 'these', 'such', 'was', 'can', 'hasn', 'your', 'herself', 'a', 'than', 'they', 'been', 'where', 'shan', 'does', 'doing', 'we', 'have', 'up', 'against', 'during', 'yourself', 'haven', 'in', 'too', 'by', 'each', 'whom', 'themselves', 'now', 'through', 'further', 'll', 'had', 'me', 'before', 'after', 'some', 'wouldn', 'he', 'her', 'no', 'most', 't', 'its', 'and', 'ourselves', 'o', 'itself', 'why', 'down', 'the', 'but', 'that', 'being', 'from', 'of', 'isn', 'myself', 'shouldn', 'weren', 'not', 'with', 'am', 'which', 'or', 'when', 'd', 'into', 'over', 'mustn', 'ain', 'you', 'at', 'wasn', 'nor', 'just', 'few', 'other', 'our', 'here', 'under', 'any', 'this', 'is', 'aren', 'same', 's', 'there', 'to', 'my', 'yours', 'should', 'won', 'ours', 'what', 'for', 'couldn', 'hers', 'more', 'having', 'didn', 'his', 'once', 've', 'below', 'about', 'hadn', 're', 'if', 'be', 'out', 'so', 'an', 'only', 'will', 'are', 'mightn', 'needn', 'off', 'has', 'while', 'above', 'him', 'who', 'on', 'were', 'own', 'very', 'himself', 'those', 'how', 'them', 'y', 'ma', 'she', 'between', 'as', 'until', 'their', 'again', 'theirs', 'yourselves', 'don', 'both', 'all', 'because', "s"

## Appendix II. List of Scraped or Downloaded Texts

Jeremy Bentham:

Bentham, Jeremy. *A Comment on the Commentaries and A Fragment on Government*. Edited by J. H. Burns and H. L. A. Hart.

London: University of London, The Athlone Press, 1977.

\_\_\_\_\_. *A Fragment on Government*. London: Printed for T. Payne, at the Mews Gate; P. Elmsley, opposite Southampton Street in the Strand; and E. Brooks, in Bell Yard, Temple Bar, 1776.

\_\_\_\_\_. *A Fragment on Government*. Edited by F. C. Montague. London: Oxford University Press, 1891.

\_\_\_\_\_. *A Fragment on Government and An Introduction to the Principles of Morals and Legislation*. Edited by Wilfrid Harrison. Oxford: Basil Blackwell, 1948.

\_\_\_\_\_. "Historical Preface, intended for the Second Edition." In *The Works of Jeremy Bentham*. Vol. 1, 240-59. Edited by John Bowring. Edinburgh: W. Tait, 1838-43.

\_\_\_\_\_. *An Introduction to the Principles of Morals and Legislation*. 2 vols. London: E. Wilson, Royal Exchange, and W. Pickering, Lincoln's Inn Fields, 1823.

\_\_\_\_\_. *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press, 1907.

\_\_\_\_\_. *An Introduction to the Principles of Morals and Legislation*. Edited by J. H. Burns and H. L. A. Hart. London: University of London, The Athlone Press, 1970.

\_\_\_\_\_. *The Works of Jeremy Bentham*. Edited by John Bowring. Edinburgh: W. Tait, 1838-43.

George Berkeley:

Berkeley, George. *Alciphron: or, the Minute Philosopher*. 3rd ed. London: J. and R. Tonson and S. Draper in the Strand, 1752.

\_\_\_\_\_. *The Works of George Berkeley, Bishop of Cloyne*. Edited by T. E. Jessop. Vols. 2 and 3 of 9. London: Thomas Nelson and Sons Ltd., 1949, 1950.

\_\_\_\_\_. *The Works of George Berkeley, Bishop of Cloyne*. Edited by A. A. Luce. Vol. 1 of 9. London: Thomas Nelson and Sons Ltd., 1948.

\_\_\_\_\_. *The Works of George Berkeley, D. D., Bishop of Cloyne*. Edited by Rev. G. N. Wright. Vol. 1 of 2. London: Printed for Thomas Tegg, 1843.

Thomas Hobbes (original publication records are missing on Intelex Website):

*De Cive*

*The Elements of Law*

*Leviathan*

*Elements of Philosophy* (including *De Corpore*)

*Answer to Dr. Bramhall*

*An Historical Narration Concerning Heresy, And The Punishment Thereof*

*Considerations upon the Reputation, Loyalty, Manners, and Religion of Thomas Hobbes of Malmesbury*

*Answer to Sir William Davenant's Preface before Gondibert*

*Letter to Edward Howard*

*Questions concerning Liberty, Necessity, and Chance*

*Dialogue between a Philosopher and a Student*

*Behemoth*

*The Whole Art of Rhetoric*

*The Art of Rhetoric*

*The Art of Sophistry*

*Seven Philosophical Problems*

*Decameron Physiologicum*

*Proportion of a Straight Line to Half the Arc of a Quadrant*

*Six Lessons to the Professors of the Mathematics, one of Geometry, the other of Astronomy, etc.*

*{STIGMAI}, or Marks of the Absurd Geometry, Rural Language, etc.*

*An Extract of a Letter Concerning the Grammatical Part of the Controversy Between Mr. Hobbes and Dr. Wallis.*

*Three Papers against Dr. Wallis*

*Considerations on the Answer of Dr. Wallis*

Letters and other Pieces

Dave Hume:

Hume, David. *An Enquiry Concerning Human Understanding and Selections from A Treatise of Human Nature*. Chicago: Open Court, 1912.

\_\_\_\_\_. *A Treatise of Human Nature*. Edited by P. H. Nidditch. 2nd ed. Oxford: Clarendon Press; New York: Oxford University Press, 1978.

\_\_\_\_\_. *A Treatise of Human Nature*. Edited by Ernest Rhys. Everyman's Library ed. London: J. M. Dent & Sons Ltd.; New York: E. P. Dutton, 1911.

\_\_\_\_\_. *Dialogues Concerning Natural Religion*. Edited by Norman Kemp Smith. Indianapolis: Bobbs Merrill, 1947.

\_\_\_\_\_. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Edited by P. H. Nidditch. 3rd ed. New York: Oxford University Press, 1975.

\_\_\_\_\_. *Essays and Treatises on Several Subjects*. London: Millar, Kincaid, Donaldson, 1758.

\_\_\_\_\_. *Essays Moral, Political, and Literary*. Edited by T. H. Green and T. H. Grose. 2 vols. London: Longmans, Green, and Co., 1898.

\_\_\_\_\_. *Essays Moral, Political, and Literary*. Edited by Eugene F. Miller. Revised edition. Indianapolis: Liberty Fund, Inc., 1987.

\_\_\_\_\_. *The Philosophical Works*. Edited by Thomas Hill Green and Thomas Hodge Grose. 4 vols. London: Longman, 1882-86.

\_\_\_\_\_. *The Philosophical Works of David Hume*. 4 vols. Edinburgh: Printed for A. Black and W. Tait, 1854.

John Locke:

Locke, John. *An Essay Concerning Human Understanding*. 4th ed. London: Printed for Awnsham and John Churchill at the Black Swan, and Samuel Manship at the Ship, 1700.

\_\_\_\_\_. *An Essay Concerning Human Understanding*. Edited by Alexander Fraser Campbell. New York: Dover, 1959.

\_\_\_\_\_. *An Essay Concerning Human Understanding*. Edited by P. H. Nidditch. Oxford: Clarendon Press, 1975.

\_\_\_\_\_. *Collected Works*. 1st ed. 3 vols. London: Printed for John Churchill at the Black Swan in Pater-Noster-Row, and Sam. Manship at the Ship in Cornhill, 1714.

\_\_\_\_\_. *Two Treatises of Government*. 5th ed. London: Printed for A. Bettesworth in Pater-Noster Row, J. Pemberton in Fleetstreet, and E. Symon in Cornhill, 1728.

\_\_\_\_\_. *Two Treatises of Government*. Edited by Peter Laslett. Student edition. Cambridge: Cambridge University Press, 1988.

\_\_\_\_\_. *The Works of John Locke*. 12th ed. Vols. 1, 2, and 4 of 9. London: Printed for C. and J. Rivington, etc., 1824.

John Stuart Mill:

*Utilitarianism*. 4th ed. London: Longmans, Green, Reader, and Dyer, 1871.

\_\_\_\_\_. *Utilitarianism*. In *Collected Works of John Stuart Mill*. Vol. 10, *Essays on Ethics, Religion, and Society*. Edited by J. M. Robson, 203-59. Toronto: University of Toronto Press, 1969.

\_\_\_\_\_. *Utilitarianism*. Edited by George Sher. Indianapolis: Hackett Publishing Co., 1979.

\_\_\_\_\_. *On Liberty*. 4th ed. London: Longmans, Green, Reader, and Dyer, 1869.

\_\_\_\_\_. *On Liberty*. In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 18, *Essays on Politics and Society*, 213-310. Toronto: University of Toronto Press, 1977.

\_\_\_\_\_. *On Liberty*. Edited by Elizabeth Rapaport. Indianapolis: Hackett Publishing Co., 1978.

\_\_\_\_\_. *The Subjection of Women*. 1st ed. London: Longmans, Green, Reader, and Dyer, 1869.

\_\_\_\_\_. *The Subjection of Women*. In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 21, *Essays on Equality, Law, and Education*, 259-340. Toronto: University of Toronto Press, 1984.

\_\_\_\_\_. *The Subjection of Women*. Edited by Susan Moller Okin. Indianapolis: Hackett Publishing Co., 1988.

\_\_\_\_\_. *Considerations on Representative Government*. 3rd ed. London: Longman, Green, Longman, Roberts & Green, 1865.

\_\_\_\_\_. *Considerations on Representative Government*. In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 19, *Essays on Politics and Society*. Toronto: University of Toronto Press, 1977.



\_\_\_\_\_. "Chapters on Socialism." *Fortnightly Review* (February 1879): 217-37; (March 1879): 373-382; (April 1879): 513-530.

\_\_\_\_\_. "Chapters on Socialism." In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 5, *Essays on Economics and Society*, 703-53. Toronto: University of Toronto Press, 1967.

\_\_\_\_\_. "Bentham." In *Dissertations and Discussions*. 2nd ed. Vol. 1, 330-92. London: Longmans, Green, Reader, and Dyer, 1867.

\_\_\_\_\_. "Bentham." In *Collected Works of John Stuart Mill*. Vol. 10, *Essays on Ethics, Religion, and Society*. Edited by J. M. Robson, 75-115. Toronto: University of Toronto Press, 1969.

\_\_\_\_\_. "Coleridge." In *Dissertations and Discussions*. 2nd ed. Vol. 1, 393-466. London: Longmans, Green, Reader, and Dyer, 1867.

\_\_\_\_\_. "Coleridge." In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 10, *Essays on Ethics, Religion, and Society*, 117-63. Toronto: University of Toronto Press, 1969.

\_\_\_\_\_. "Dr. Whewell on Moral Philosophy." In *Dissertations and Discussions*. 2nd ed. Vol. 2, 450-509. London: Longmans, Green, Reader, and Dyer, 1867.

\_\_\_\_\_. "Whewell on Moral Philosophy." In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vol. 10, *Essays on Ethics, Religion, and Society*, 165-201. Toronto: University of Toronto Press, 1969.

\_\_\_\_\_. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. 4th ed. London: John W. Parker and Son, 1857.

\_\_\_\_\_. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. 7th ed. London: Longmans, Green, Reader & Dyer, 1871.

\_\_\_\_\_. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. Edited by Sir W. J. Ashley. London: Longmans, Green & Co., 1909.

\_\_\_\_\_. *Principles of Political Economy, with Some of Their Applications to Social Philosophy*. In *Collected Works of John Stuart Mill*. Edited by J. M. Robson. Vols. 2 and 3. Toronto: University of Toronto Press, 1965.

(All the following texts are from the Gutenberg Project website )

Edmund Burke

*Reflections on the Revolution in France*

Friedrich Engels

*The Condition of the Working Class in England*

Alexander Hamilton, John Jay, James Madison

*The Federalist Papers*

Immanuel Kant

*Critique of Pure Reason*

Niccolò Machiavelli

*The Prince*

Karl Marx

*The Communist Manifesto*

Bertrand Russell

*The Problems of Philosophy*

*Mysticism and Logic*

*The Analysis of Mind*

Adam Smith

*Moral Sentiments*

*The Wealth of Nations*

Baruch Spinoza

*Ethica*

*On the Improvement of the Understanding*