NLP Analysis of Religious Texts

By: Keith Kwong

DataHacks 2021 Submission

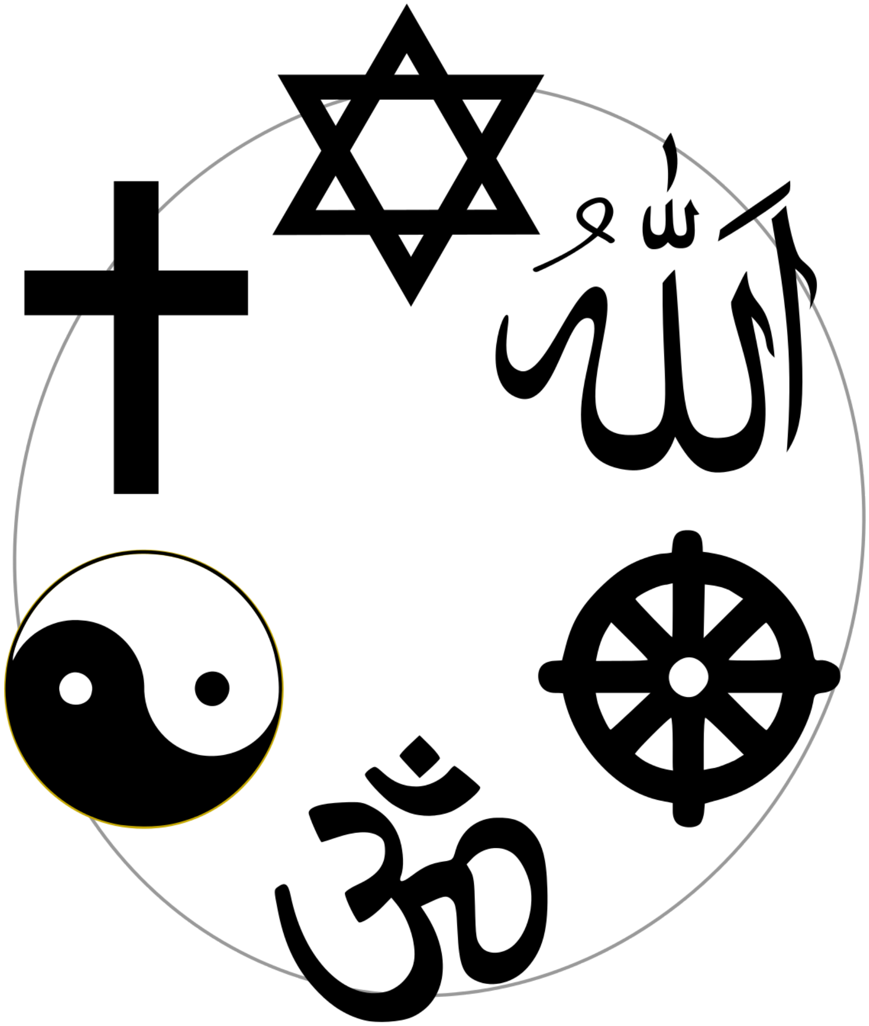


Table of Contents

Introduction

For my project submission for DataHacks 2021, I will be tackling the beginner track's topic of using NLP analysis to draw conclusions on how different religious texts differ in word usage. I was given a dataset containing the chapters of eight books, being the books of Buddha, Ecclesiastes, Ecclesiasticus, Proverb, Wisdom, TaoTeChing, Upanishad, and YogaSutra. The data within the dataset was a count on how many times a certain word was used in a certain chapter, with a total of 8266 unique words accounted for in the dataset. With this sort of data, there are many interesting questions that can be asked and answered, but the most interesting (mandatory) ones are how the wording for the three Old Testament books differ as they were written at separate times and how much Buddhism and Taoism influence on each other can be seen in their texts and practices. So, with these mission statements in hand, let's embark on a journey of how I decided to Data Science my way through this datathon.

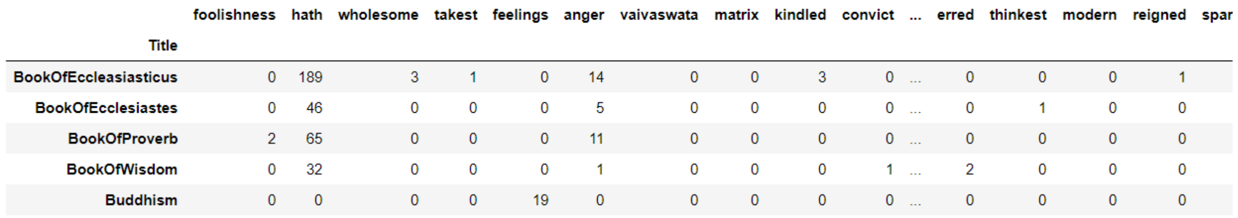
Data Cleaning/Pre-Processing

I. Technologies Used

For this endeavour, I have decided to make use of Jupyter Notebook, Python, the pandas library, Github, and sci-kit learn as my machine-learning library. The entire notebook with all the code will be linked at the end as well as the repository I used to version control this project. Maybe consider looking at the code alongside the report?

II. Making the data palatable

To start working with the data, I first had to turn the data into something that I liked looking at. The original dataset was pretty messy in terms of naming conventions, so I first renamed the column that held the book chapters to simply “Title”. This would allow me to later reference the column without having to type the mess that was the original “Unnamed: 0”. After this, I decided that having the data be separated by book title_chapter was also not that useful for the two problems I needed to solve, so I created a function that would get rid of the chapter moniker and combined all the data for the individual chapters into rows that represented the data for the entire book. Great job me! I (and hopefully you as well) can now look at the data without mentally booming.

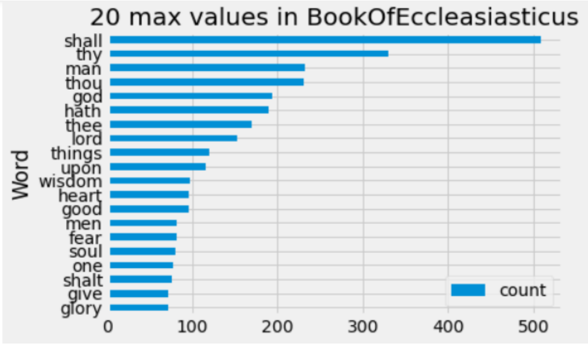


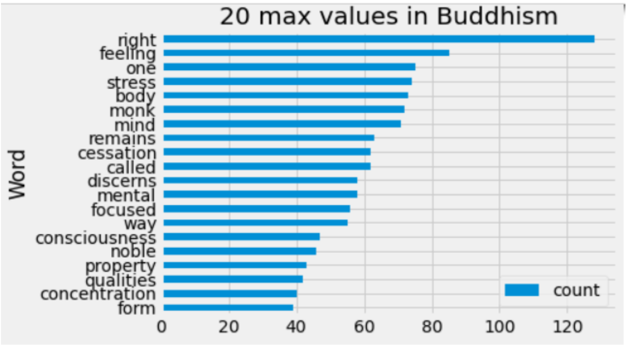
	foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	convict	...	erred	thinkest	modern	reigned	spar
Title																
BookOfEccleasiasticus	0	189	3	1	0	14	0	0	3	0	...	0	0	0	1	
BookOfEcclesiastes	0	46	0	0	0	5	0	0	0	0	...	0	1	0	0	
BookOfProverb	2	65	0	0	0	11	0	0	0	0	...	0	0	0	0	
BookOfWisdom	0	32	0	0	0	1	0	0	0	1	...	2	0	0	0	
Buddhism	0	0	0	0	19	0	0	0	0	0	...	0	0	0	0	

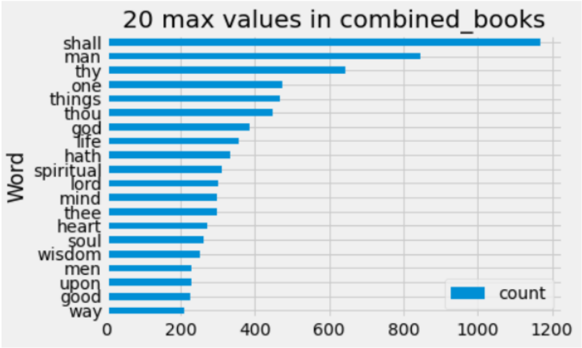
(Picture for reference)

III. Extracting Useful Information

Now that the data is ready to be used, I set out to sharpen my sense of the data by extracting and graphing some useful information, specifically the twenty most used words of each book and the twenty most used words in all the books combined. For this, I created a function that would take in the title of the book and the dataset from which it should find the max values for and then created another function that could handle bulk finding of the twenty most used words in multiple books at once. Running this function allowed me to get a better grasp of how the data looked and how the different books differed in their word usage. An interesting surface-level observation I made here was that books of Catholic origins such as Eccleasiasticus (left) have many top words that are used to point toward a relation to bodily actions, such as “shall” and “thy” whereas the books coming from Asian beliefs such as the book of Budha (right) have many word relating more to mental processes, such as “consciousness” and “concentration”. The rest of the graphs for the other books can be seen in the Notebook.







IV. Pre-Processing For First Prompt

Before I could start working on answering the two prompts given, I did a bit more specific pre-processing of the data that I would be working with. For the first prompt, since it only looks at the books from the Old Testament, I first extracted only the three books of Proverb, Wisdom, and Ecclesiastes from my modified data set. I then got rid of any word columns where all three values were zero, indicating that none of the books had that word.

V. Pre-Processing For Second Prompt

Much like the first prompt, I did some more specific preprocessing before starting my analysis for the second prompt. I extracted only the Buddhism book and the Taoism book from my modified data set and got rid of any word data where both the Buddhism and Taoist book had zeros.

Proposal/Hypothesis

I. Prompt 1: A proposal

The first prompt tasked me with defining a model that shows how the three Old Testament books have changed in wording since they were published at different times. To do this, I propose that the best way to show how their wording has changed is by building a model that compares the frequency of their words, using sci-kit learn to vectorize the word count data and drawing conclusions from seeing how different the individual word usages are.

II. Prompt 2: My Hypothesis

The second prompt tasked me with experimenting on the Buddhist and Taoist books and seeing if the influence both religions have on each other can be seen through the text of both books. My hypothesis is as follows: Knowing that Buddhism and Taoism are influenced by each other, then the word composition of both these books will be similar, with a difference of less than 5% in each word category.

Analysis

Prompt 1 Analysis: Difference between three Old Testament books

I. Background Information

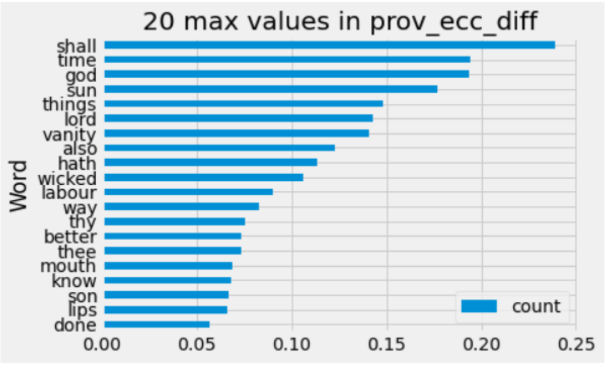
To start off with this analysis, I first did a bit of background research to find basic information about each of the three books. The Book of Proverbs is the oldest of the three books, being written around 700 BC and mainly focusing on morality and the meaning of human life. The book that follows is the Book of Ecclesiastes, being written around 350 BC and focusing on the meaning of life and God’s punishment of sin. Finally the Book of Wisdom is the most recent, being written during roughly 50 BC and focusing on the pursuit and benefits of wisdom, as the name implies.

II. Vectorizing with Sci-kit learn

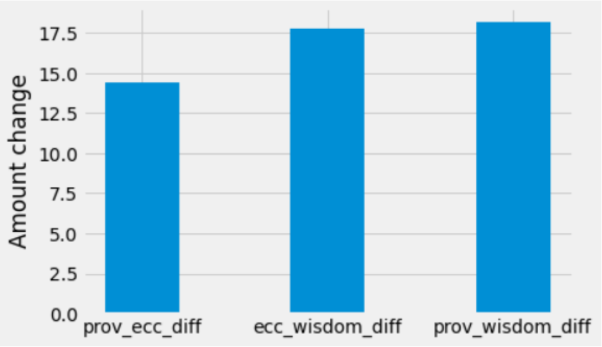
Given that our original data is already tokenized and counts occurrences, I skipped straight to vectorizing the data with the sci-kit learn library. Importing their TfidfTransformer, I fit my data with the estimator before transforming it into a tf-idf representation, obtaining the frequency information of the three books. With this information, I then find the absolute value of the differences in frequencies between each of the three books and turn that information into a dataframe. With the magnitude in differences obtained, I could now begin analyzing the differences between these three books.

III. Analyzing the Differences Based Word Frequency

Before diving into my analysis of the data, let's look at what the data currently looks like shall we? To the right is data that shows the twenty biggest differences in word frequencies



between the books of Proverb and Ecclesiastes. The maximum word frequency difference here is the word “shall”, hovering at around 0.24 and the smallest (of the twenty largest) difference is the word “done” with a frequency difference of around 0.05.



Now that we’ve seen a sample of the data we are working with, I’ll hit you with another graph. This graph showcases all the differences added up together, summing into a number that represents the total difference of word frequency, and therefore the total difference between the two books. The largest difference comes from comparing the frequencies of the Book of Proverb and Wisdom here, a result that makes sense given that they are the two books that were written the farthest apart. The next biggest difference, however, is actually that of Ecclisiastes and Wisdom, which has a smaller time gap than that of Proverbs and Ecclisiastes! Why is this the case? The answer is most likely an increase in effort for the field of philosophy. It is around the latter half of the 4th century when famous philosophical schools such as Plato’s Academy were established, and the sudden increase in IQ must have changed how they wrote the book of Wisdom (this last claim is my own thoughts, I’m too lazy to go looking for the data to back it up). What isn’t my own opinion, though, is that the word usage for these books are different, differing in word frequencies.

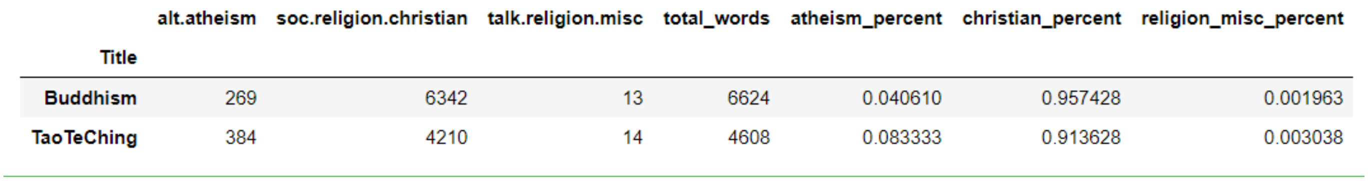
Prompt 2 Analysis: Influences in Buddhism and Taoism

I. Background Information

Similar to the first prompt, I once again started off by looking into some background information for Buddhism and Taoism. The prompt mentioned that Buddhism and Taoism were influenced by each other, and it is pretty clear to see why when looking at their core values. Both believe that there is a higher existence above that can be achieved through following a path of practices and both have a set of heavenly principles that people need to follow.

II. Implementing sk-learn’s classifier

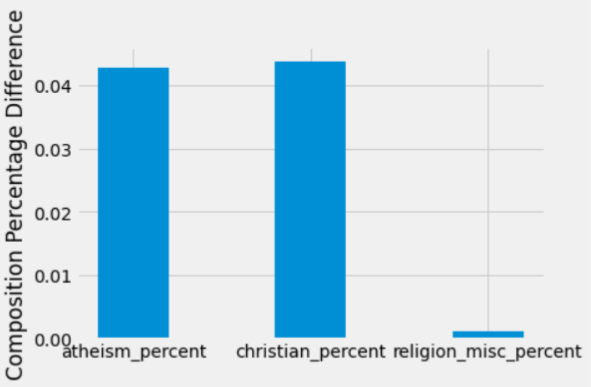
Recall the previous section of the report, and my hypothesis stating that the word composition of the two books would not be too far apart. Well, in order to compare word composition, the words themselves need to be classified. To this end, I once again made use of the sci-kit learn library to build and train a classifier that would allow me to categorize and label the words. Importing the 20 newsgroup dataset from sk-learn, I tokenized the newsgroup data, transformed it into tf-idf format, and trained my classifier on it using a naive Bayes classifier. In order to save time, I only included three categories here, being 'talk.religion.misc', 'alt.atheism', 'soc.religion.christian' as they are all related to religion. I then applied my classifier to “predict” the label of all the words in the Buddhism-Taoism dataset, summing up all the words with the same label in a dataframe. I am then finally able to calculate the word composition percentage for each label, adding them to the dataframe.



	alt.atheism	soc.religion.christian	talk.religion.misc	total_words	atheism_percent	christian_percent	religion_misc_percent
Title							
Buddhism	269	6342	13	6624	0.040610	0.957428	0.001963
TaoTeChing	384	4210	14	4608	0.083333	0.913628	0.003038

III. Analyzing word composition

Recall in my hypothesis that I made mention of a property I liked to call “word composition”, but what is word composition? In this scenario, word composition refers to the percentage of the text that contains a specific type of word. A sentence such as “I like that dog” would have a word composition of 25% “animal”, since, of four words, there is one animal type word. With that explanation done, let's look at the data. Using the previously described dataset, we can see that the Buddhism book contains roughly 4% “atheism” words, 96% “christian” words, and 0.01% miscellanioius words. TaoTeChing likewise contains 8.3% “atheism” words, 91% “christian” words, and 0.03 miscellanious words. While these labels may seem random, what matters here is that certain words are grouped together into a label, not the meaning of the label itself. From here we can graph the differences of these word compositions and see just how similar these two books are. As we can see from the graph on the right, the percent differences between the two books’ word compositions are quite small, and, in line with my hypothesis, are all less than 5%.



Conclusions

The religious texts that were given to me all are unique, but with these analyses of the word data of the texts, I am now able to scientifically explain how these types of text are different. Looking at the top twenty words used by the books give an insight into how the religion and culture as a whole think and utilize a religious text. Diving deeper into our analysis of the three books from the Old Testament showed that religious books that were written in roughly the same era are still capable of being changed drastically and by comparing the different word compositions of the Buddha and Taoist texts it becomes clear that it is not just a mere coincidence that these two religions have very similar practices and beliefs. The model and method I proposed for the first prompt clearly showed the distinct difference in word frequency between the three texts, and even my hypothesis for the second prompt was validated by my method of comparing word compositions. However, there are a couple of experimental errors that I unfortunately have to address here. To begin with, my analysis of my results of the first prompt may not be completely accurate, because it is quite hard to pinpoint the exact year the books were written. Most websites only mentioned vague time eras, such as the “latter half of the third century”. For my experimentation of the second prompt, there may also be some error due to the rather inaccurate labels I used for the classifier. The dataset I had used didn’t have the best labels for the method that I had in mind, so I had to choose the ones best fit for the texts I had. The classifier was also trained on newsgroup data rather than religious texts, so that may have also impacted the labeling process. In conclusion, with the exception of a few errors in accurate data and date labeling, I would consider my experiments here a success.

Bibliography/Github

- https://github.com/khkwong/datahacks2021