Beginner Track

Team Name: The Balds

Team Members: Jiani Song, Tiffany Yu, Weiyue Li, and Yi Li

11 April 2021

Introduction

There are two main types of religions in the world: Eastern Religions and Western Religions. Western religions are typically strictly following one supreme God, such as the Holy Mary of Catholics or the Jesus of Christ of Christian. However, Eastern religions believe in more than one God or the presence of some unnatural power, such as the 33 Million Gods of Hinduism and the self-development of Taoism. Besides that, the main points of these two categories of religions are different: the western religions believe that God is above all creatures, whereas the eastern religions believe that all animals are equally created.

Although these religions are different, their word choices are an interesting topic to explore. From a linguistic perspective, we can categorize every word in religious text into positive connotation and negative connotation. Therefore, we are interested in finding the individual trend of the proportion of positive words of each chapter for each book, as well as the general trend of the proportion of positive words for all books in the dataset. The nltk package , which includes the Sentiment Intensity Analyzer, helped us define categorize words into positive and negative categories. While using the package, Python will assign positive and negative indices to each word. If the positive index is significantly greater than the negative index, then the word is positive. If the positive index is considerably smaller than the negative index, then

the word is negative. If the positive index and the negative index are roughly the same, then the word is neutral.

Prompt part 1: Data Observations

In our dataset, we categorized all 8 of our books into Western (The Book of Wisdom, The Book of Proverbs, The Book of Ecclesiastes, and The Book of Ecclesiasticus) and Eastern (The Book of Buddhism, Tao Te Ching, The Yoga Sutra, and The Upanishads).

Data Cleaning

To use groupby by book, we first use the built-in function split to create a new column containing only the name of the books. We then groupby by the book names and sum up all the word counts for each word so that each row represents the total number of times that each particular word appears in a book. We also dropped the none type in our DataFrame so that we could run analysis more efficiently.

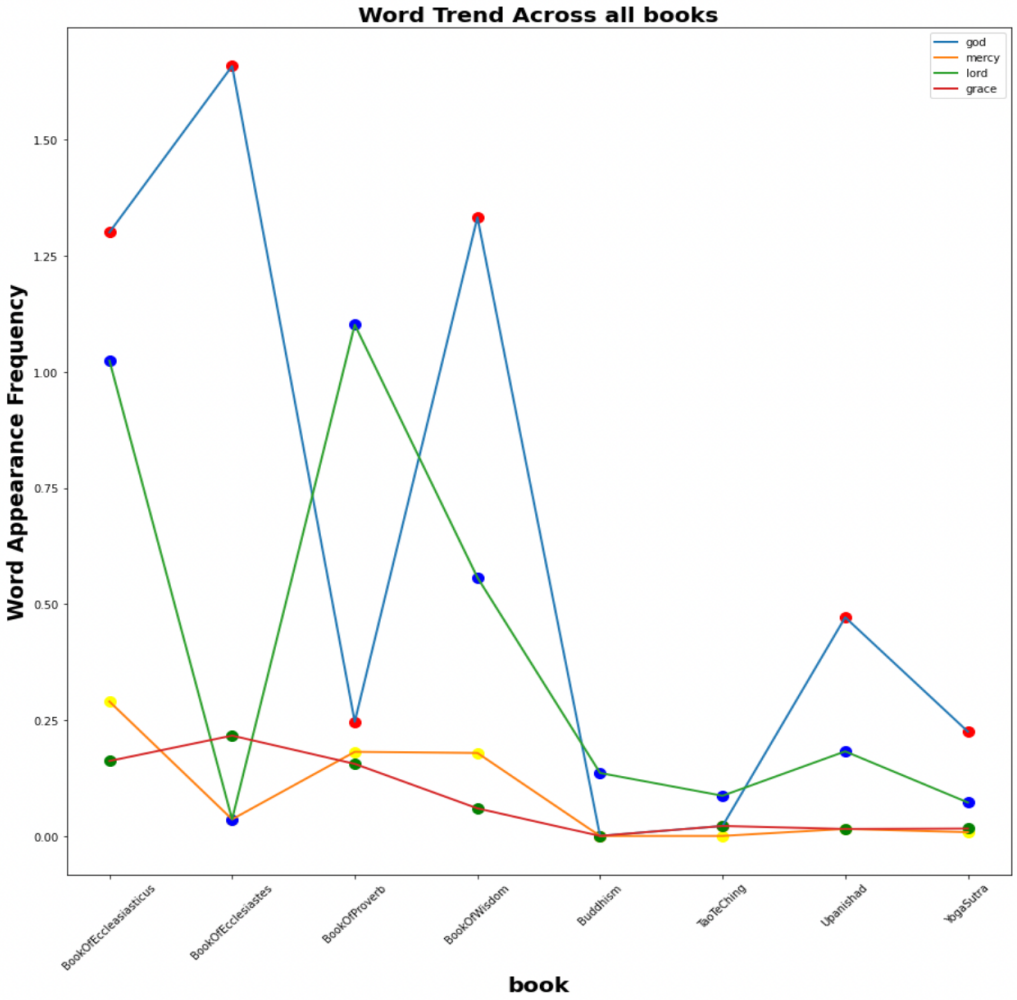
Data Exploration

There are two main types of religions in the world: Eastern Religions and Western Religions. Western religions are typically strictly following one and only supreme god, such as the Holy Mary or the Jesus of Christ; however, Eastern religions believe in more than one god or the presence of some unnatural power, such as the 33 Million Gods of Hinduism and the self development of Taoism. Besides that, the main point of these two types of religions are different: the western religions believe that God is above all creatures, whereas the eastern religions believe that all creatures are created equally. As a result, we predict that words such as "god",

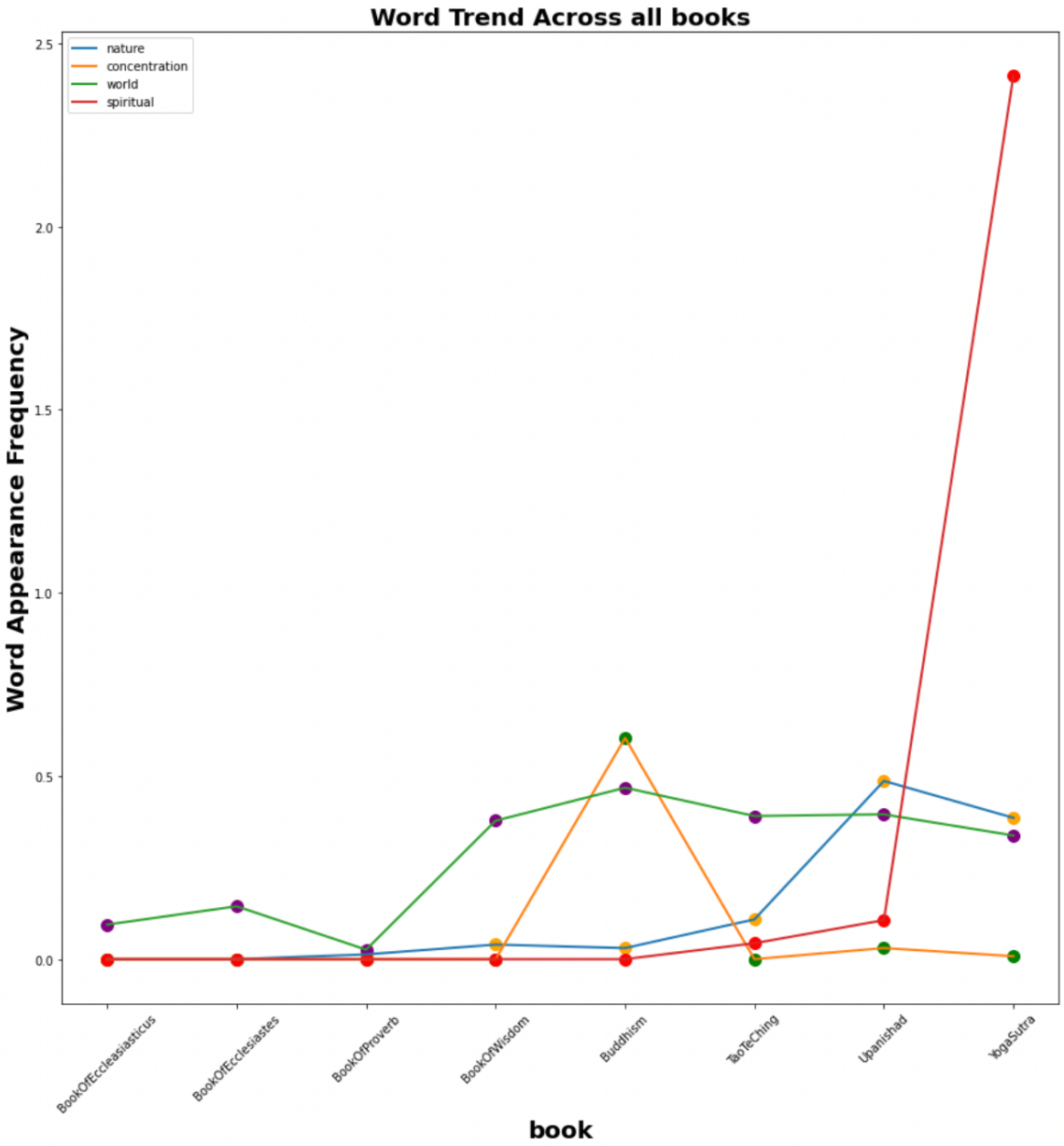
"mercy", "lord", "grace" are more related to Western religion and words such as "nature", "concentration", "world", "spiritual" are more related to Eastern religion.

Prompt part 2a: Data Visualization

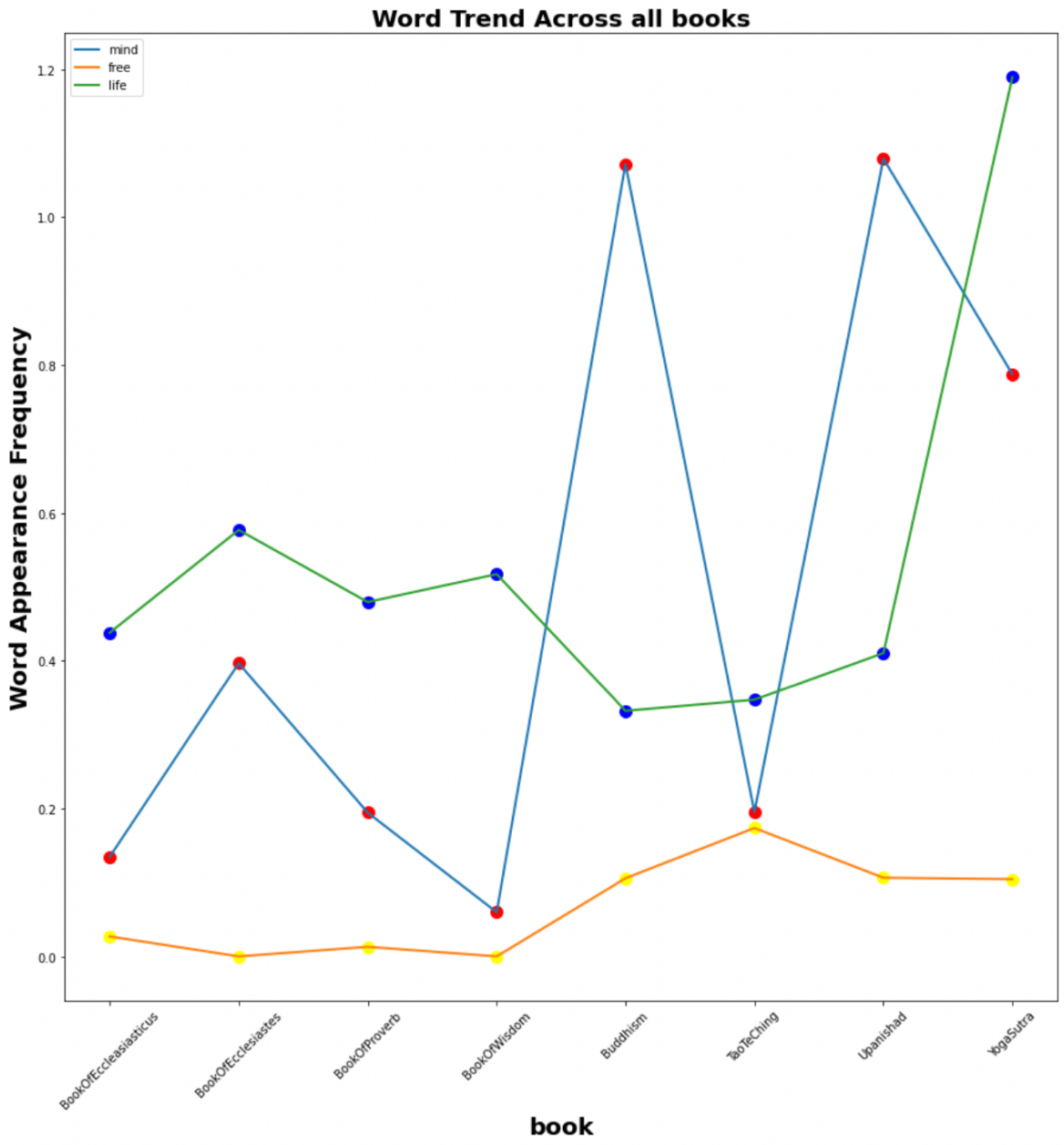
In our dataset, we categorized all 8 of our books into Western (left of the chart) and Eastern (right of the chart). We have picked some specific words to test their trend of proportion in each book. In this particular chart, the words we have picked are: “god”, “mercy”, “lord”, and “grace”. It is clear that these words that are known for their association with the Christianity are appearing more in the Western books than in the Eastern ones. There is also an interesting pattern to note that the Ecclesiates (Old Testament) love to use “god”, whereas the Proverbs (Old Testament) love to use “lord” to express the same meaning.



In this second graph, we picked some words that we think might appear more in the Eastern books instead of the Western ones. These words are: “nature”, “concentration”, “world”, and “spiritual”. In the graph, it is clear that these words match our expectation. And there are two words that are kind of unique for two Eastern religions, which I think can represent the characteristics of those two religions. For example, Buddhists like to use “concentration” and YogaSutra tend to use “spiritual.”



In this third graph, we can clearly see that different word choices have different possibilities in each text. For example, the word “mind” is highly likely to appear in Buddhism and Upanishad; “free” is extremely likely to be in Tao Te Ching, and “life” is extremely likely to occur in Yoga Sutra.



Prompt part 2b: Hypothesis

Background / Justification:

With one of our group members coming from a Catholic high school, he noticed the tremendous amount of the word "sin" appear in his religion and how Jesus cleaned sins for us. Therefore, we believe that Western religious books may contain a large number of negative words such like sin, sinful, sinner, etc. On the other hand, coming from an Eastern country, our group has little understanding of the Eastern religions; instead, our religion focus more on personal life and spirit.

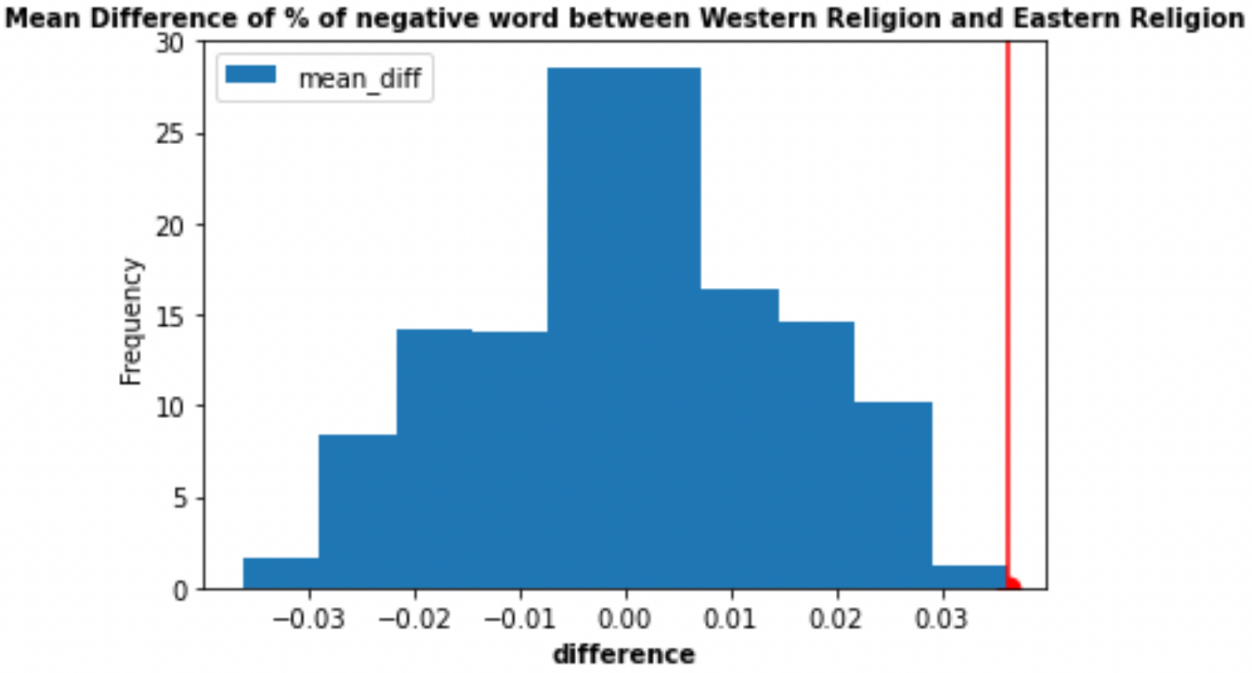
Null and Alternative hypothesis:

As a result, we propose the following hypothesis: "Null: Books of Western religions tend to contain more negative words than Books of Eastern religions do; Alternative: Books of Western religions will have approximately equal amounts of negative words comparing to the books of Eastern religion.

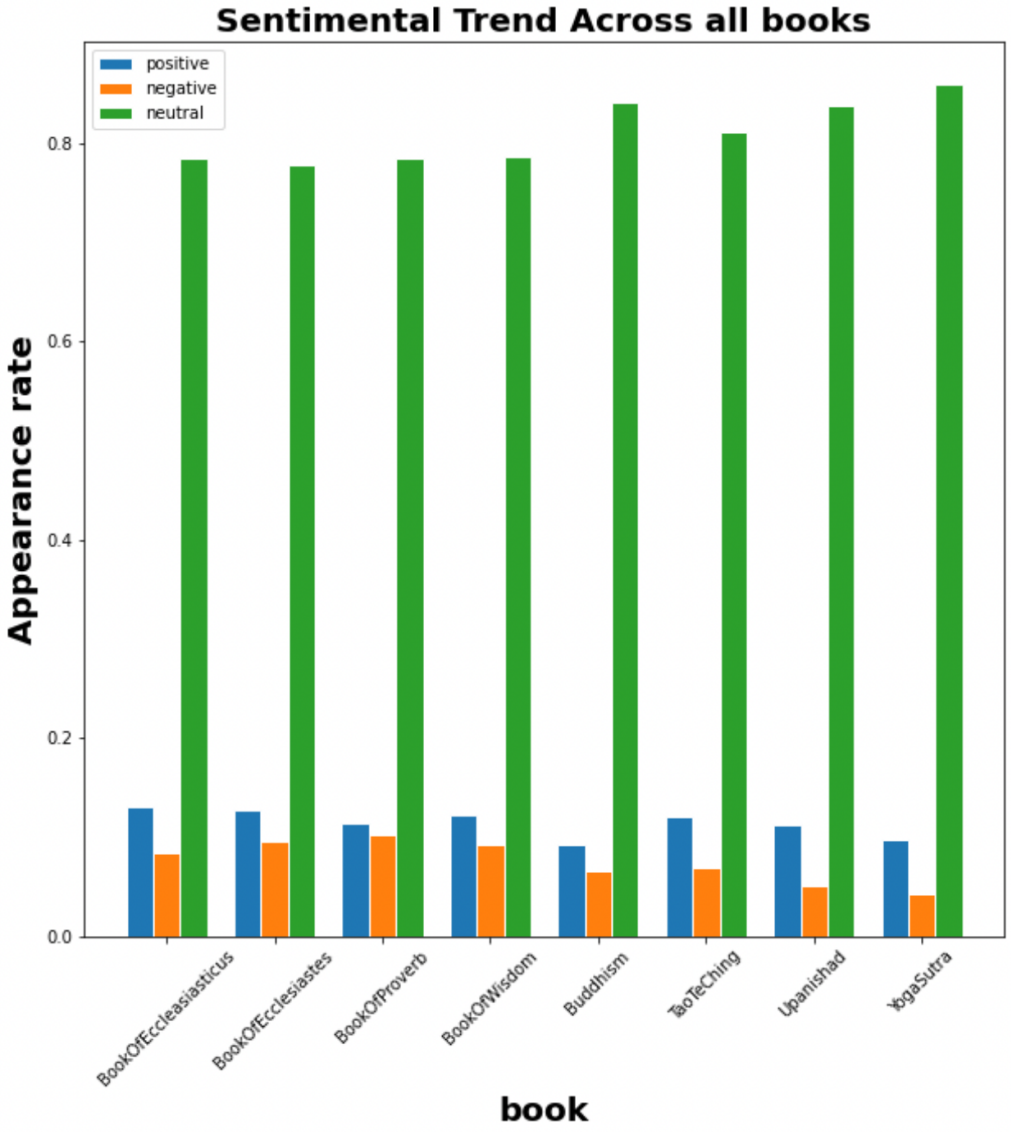
Results of our experiment:

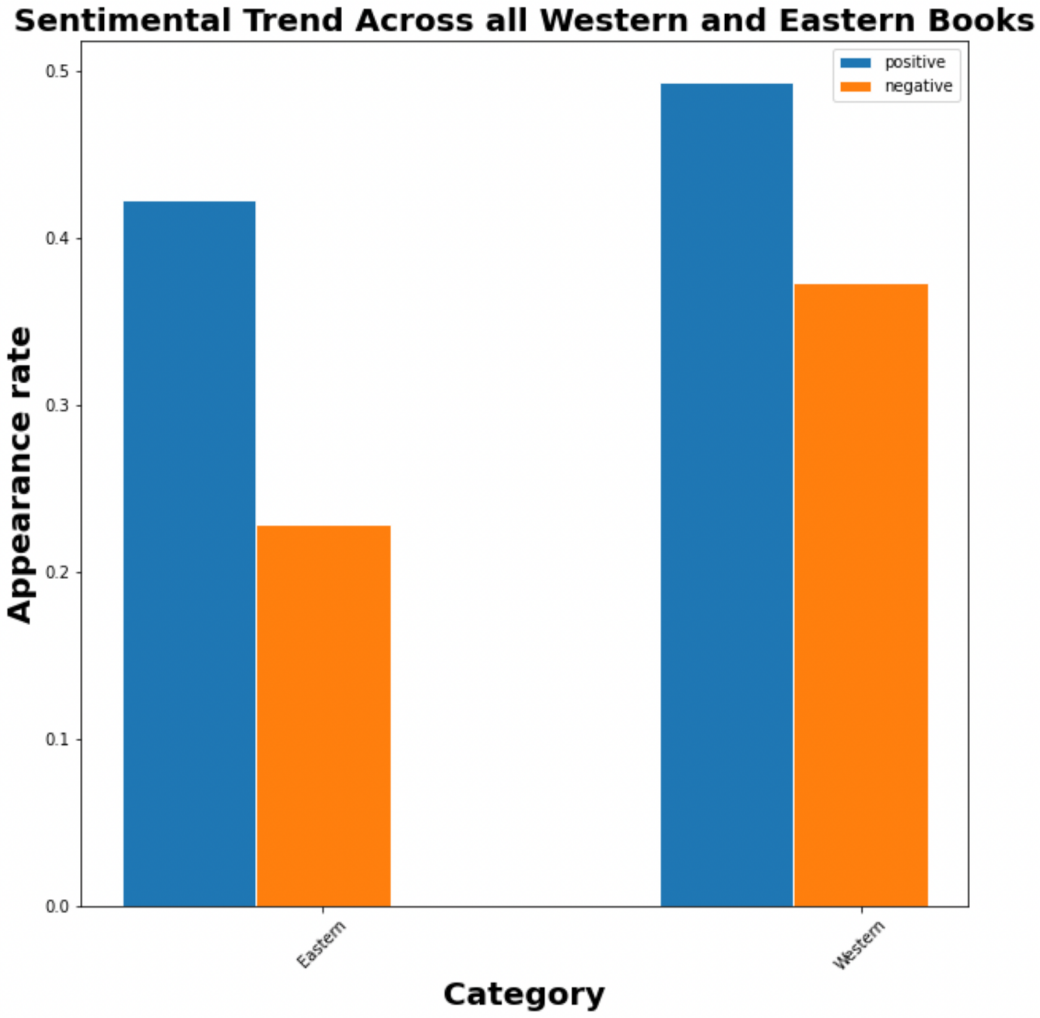
After running the A|B test, we found that the p-value is less than 0.05. This means that the result we got is statistically significant. Our data shows that the Western religious books have higher proportions of negative words compared to the Eastern religious books.

A | B Testing Visualization:



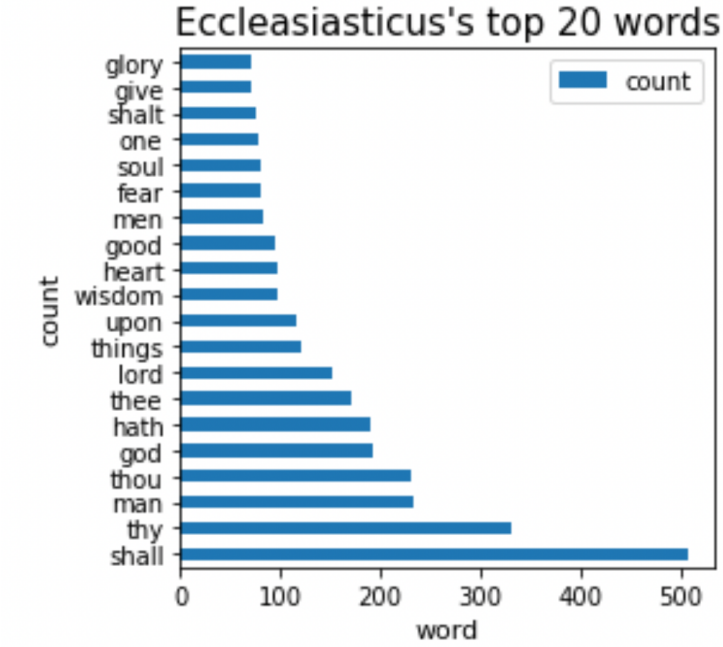
<u>Prompt part 3:</u> Visualizing our findings

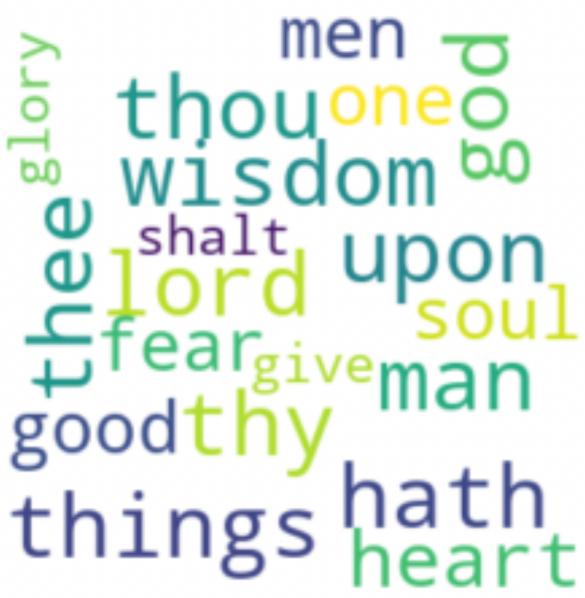



Sentimental Trend Across all Western and Eastern Books
positive
negative
Appearance rate
0.0
0.1
0.2
0.3
0.4
0.5
Eastern
Western
Category

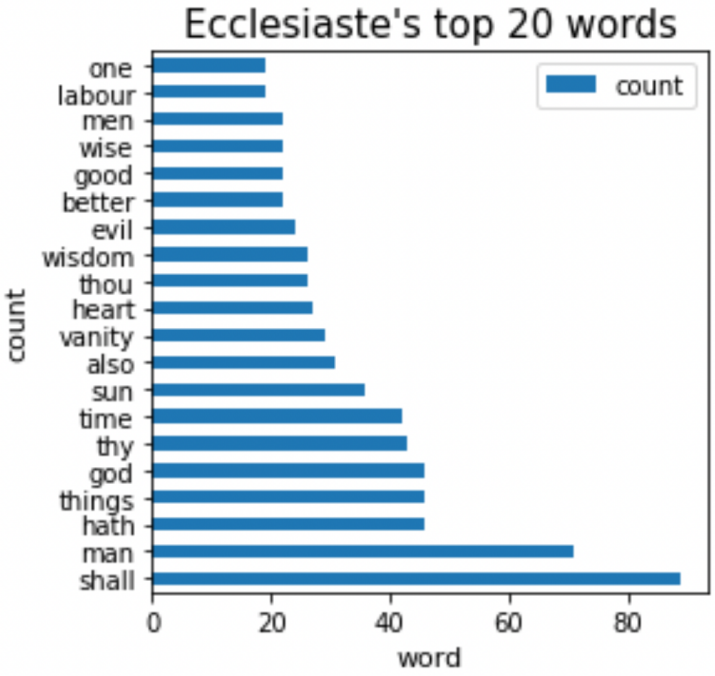
Prompt part 4: Finding Top 20 Words

Eccleasiasticus:

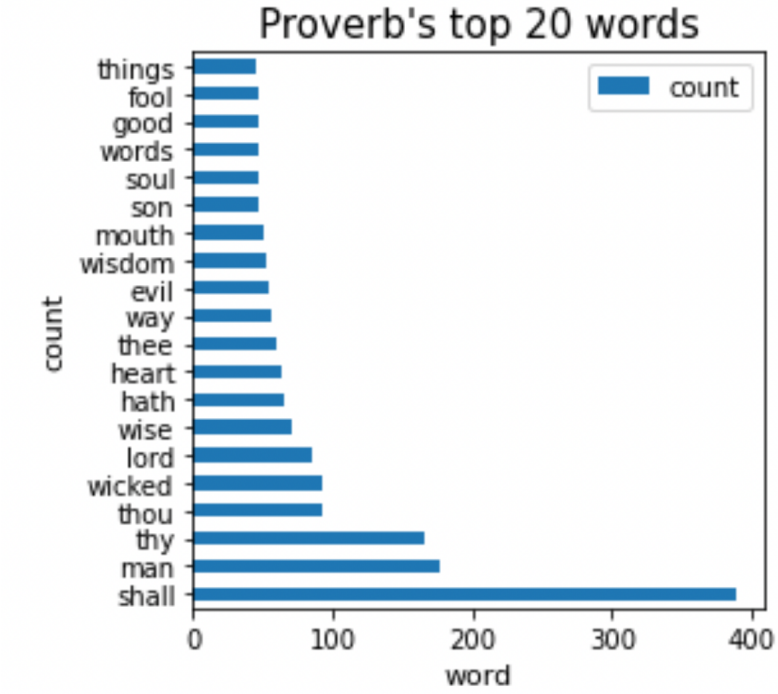




Ecclesiastes:

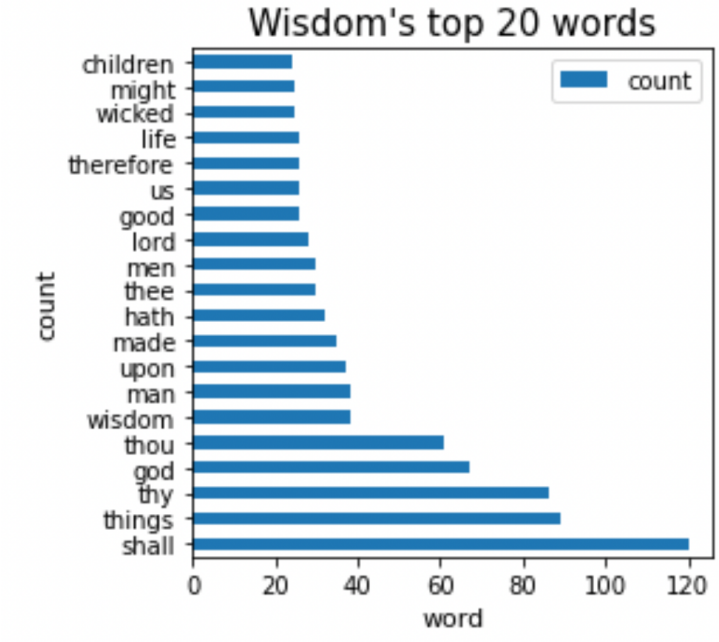


Proverb:


Proverb's top 20 words
count
things
fool
good
words
soul
son
mouth
wisdom
evil
way
thee
heart
hath
wise
lord
wicked
thou
thy
man
shall
0
100
200
300
400
word

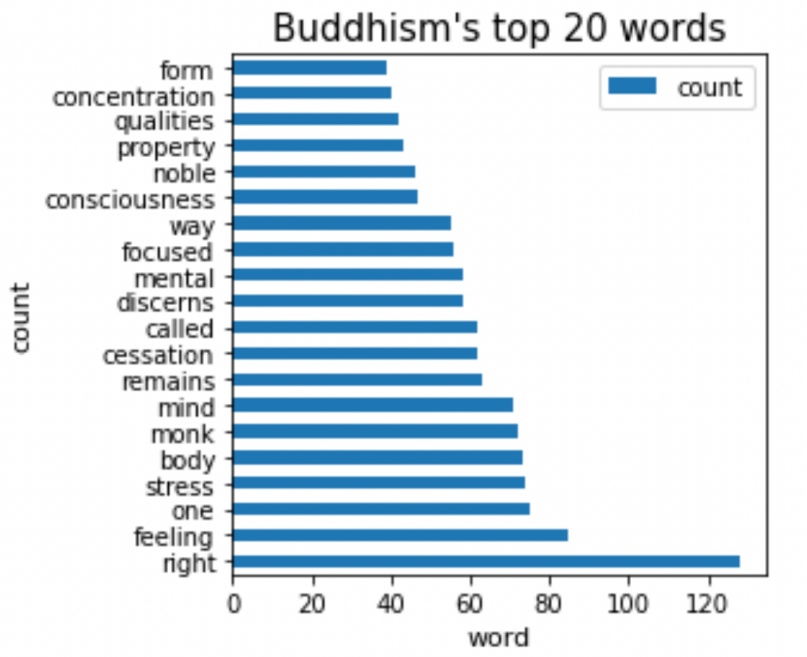

heart
fool
wisdom
hath
soul
lord
thee
things
good
wicked
thy
son
evil
way
mouth
man
thou
wise
words

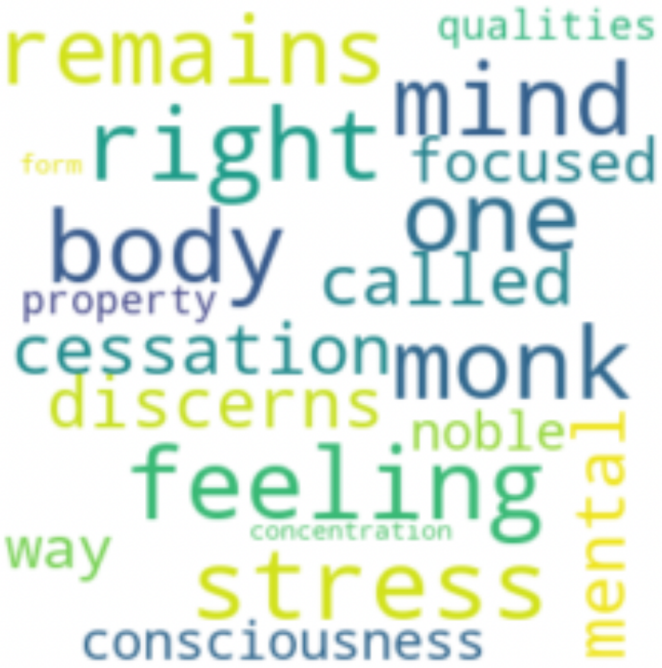
Wisdom:


Wisdom's top 20 words
count
children
might
wicked
life
therefore
us
good
lord
men
thee
hath
made
upon
man
wisdom
thou
god
thy
things
shall
0
20
40
60
80
100
120
word

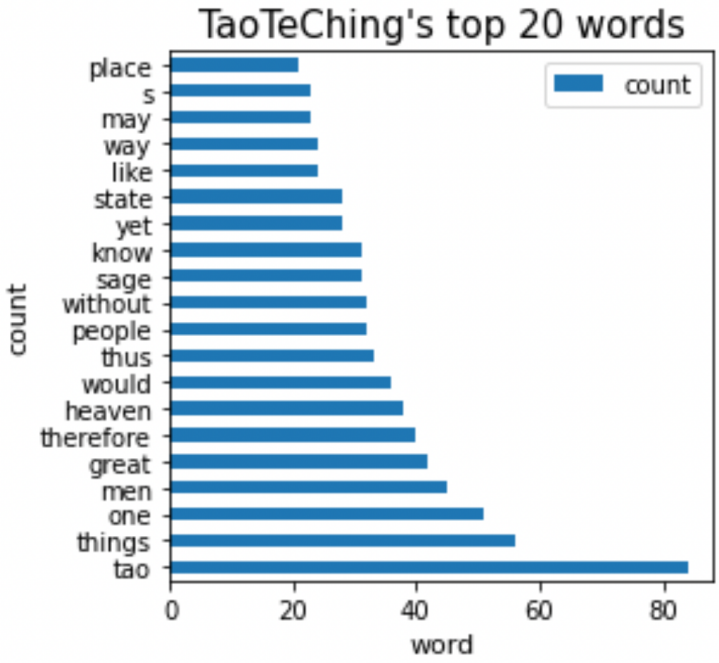

god
wicked
wisdom
life
man
thou
men
hath
things
good
might
us
upon
thy
lord
thee
made
children

Buddhism:


Buddhism's top 20 words
count
form
concentration
qualities
property
noble
consciousness
way
focused
mental
discerns
called
cessation
remains
mind
monk
body
stress
one
feeling
right
0
20
40
60
80
100
120
word


remains
qualities
mind
right
form
focused
one
body
called
property
cessation
monk
discerns
noble
feeling
mental
concentration
way
stress
consciousness

TaoTeChing:


TaoTeChing's top 20 words
count
place
s
may
way
like
state
yet
know
sage
without
people
thus
would
heaven
therefore
great
men
one
things
tao
0
20
40
60
80
word

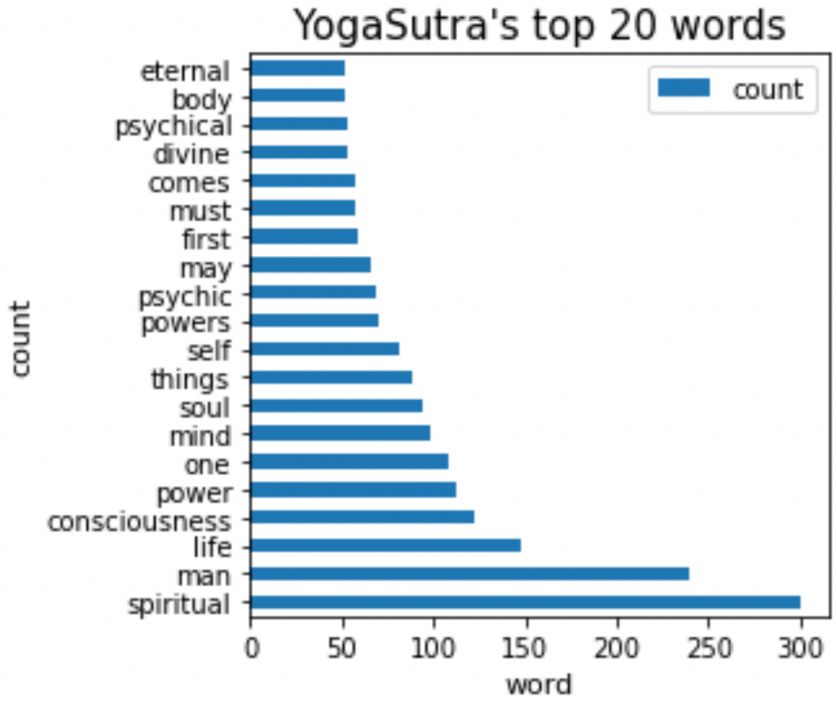

way
know
great
without
s
tao
yet
place
thus
things
heaven
sage
may
men
one
people
state

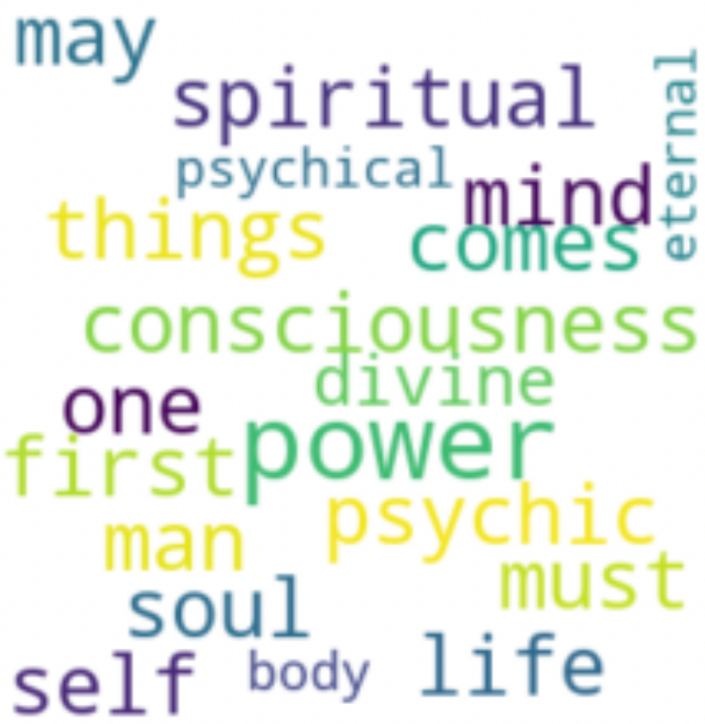
Upanishad:


Upanishad's top 20 words
count
body
god
heart
nature
therefore
knows
atman
beyond
senses
said
nachiketas
must
know
knowledge
death
man
brahman
mind
self
one
0
20
40
60
80
100
word

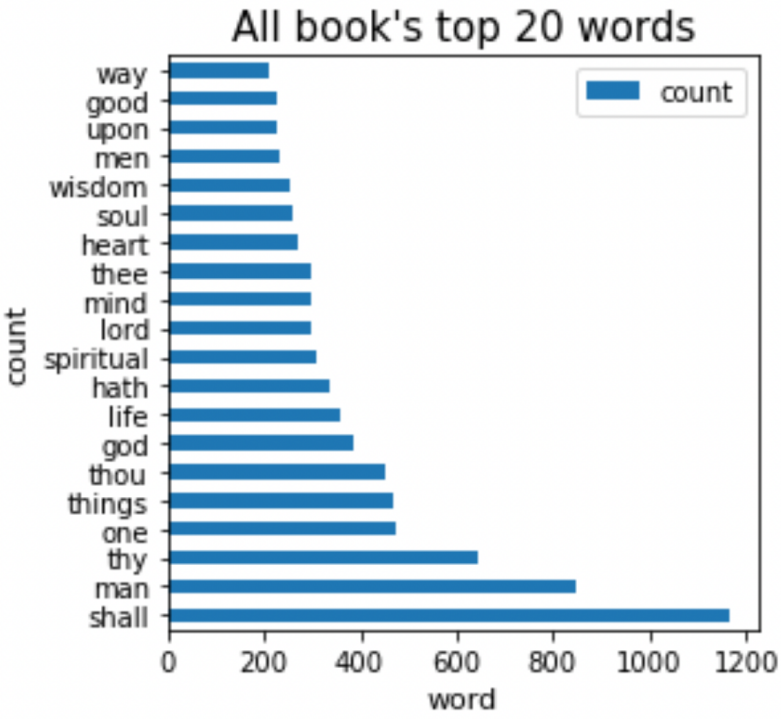

death
body
said
nachiketas
senses
knowledge
mind
nature
atman
man
god
heart
self
beyond
know
one
brahman
must

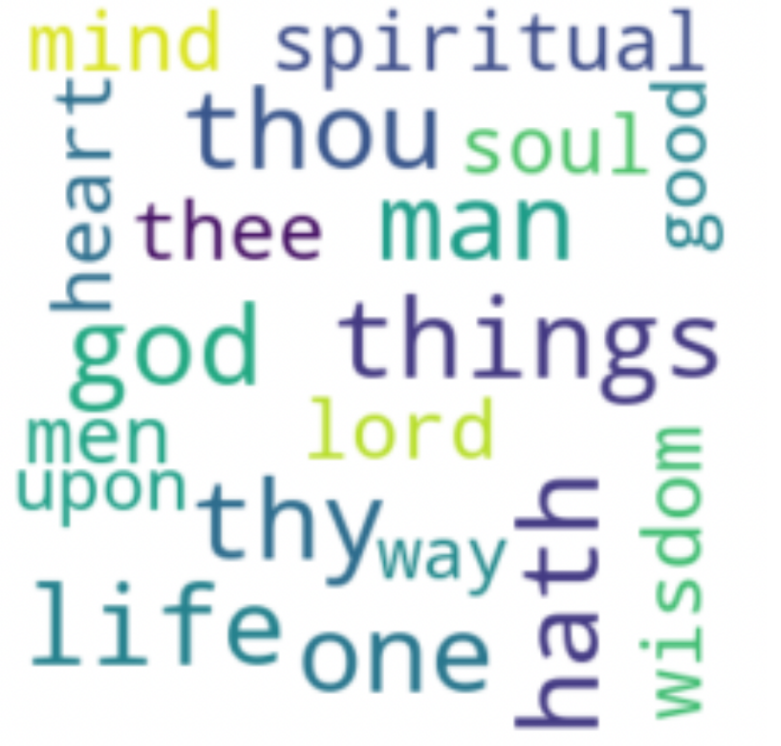
YogaSutra:


YogaSutra's top 20 words
count
eternal
body
psychical
divine
comes
must
first
may
psychic
powers
self
things
soul
mind
one
power
consciousness
life
man
spiritual
0
50
100
150
200
250
300
word


may
spiritual
psychical
mind
eternal
things
comes
consciousness
divine
one
first
power
psychic
man
must
soul
self
body
life

Across all books:





Prompt part 5: Wording Differences among books "Proverbs", "Ecclesiastes", and "Wisdom"

F1-score / precision score from DecisionTreeClassifier model:

	precision	recall	f1-score	support
Ecclesiastes	1.00	1.00	1.00	4
Proverb	0.25	0.33	0.29	3
Wisdom	0.33	0.25	0.29	4
accuracy			0.55	11
macro avg	0.53	0.53	0.52	11
weighted avg	0.55	0.55	0.55	11

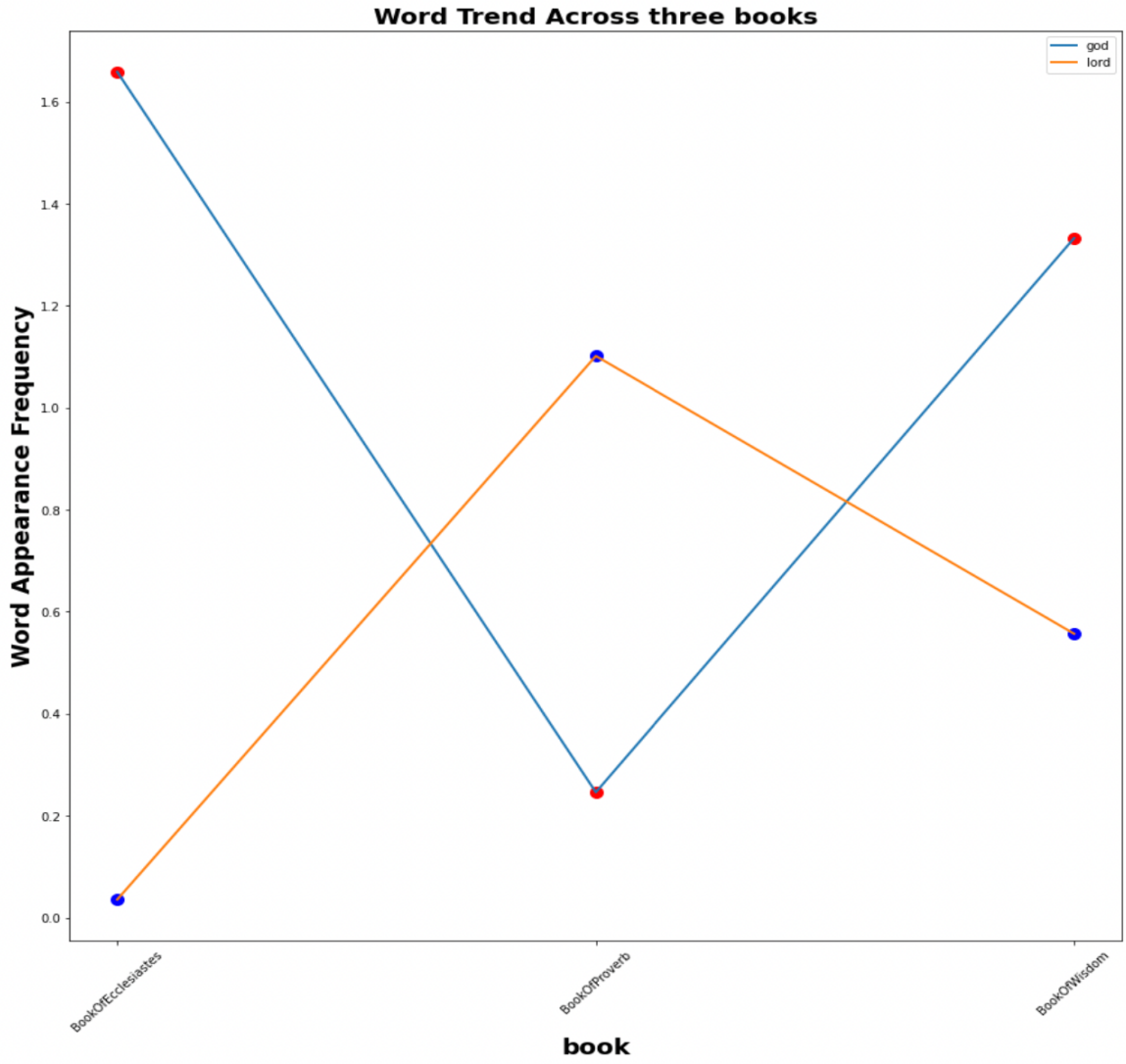
Interpreting tables and conclusions:

Through observation, we noticed that even though Proverbs, Ecclesiastes, and Wisdom are all three books from the Old Testament, the different periods of time when these three books were written have made some of their word choices different. For example, when trying to express the meaning of God, or deity, the book Proverbs tends to use more “lord” instead of “God”; however, the book Ecclesiastes tends to use more “God” instead of “lord”. We think it is a pretty interesting pattern to notice, and thus, we have defined a ML model to help determine which book the given words distribution belong to. This ML model collected data of the word counts in each book, and found the trend in each book.

Based on the results of machine learning predictions from the DecisionTreeClassifier model. We can see that the predictions accuracy for books “Ecclesiastes” is much higher than the rest two books, which indicates that the word distribution between "Ecclesiastes" and the rest

two books are significantly different because the accuracy score will be super low if the word distribution is similar since the machine cannot distinguish between "Ecclesiastes" and the other two books. Therefore, we conclude that the word usage is different between "Ecclesiastes" and the other two books though they all come from the old testament. One obvious example will be "God" and "Lord" that we mentioned before.

Visualization that will prove the result from machine learning model:



Prompt part 6: Buddhism and Taoism

Background research:

Based on our research, Taoism and Buddhism shared many similar practices and beliefs, the most common of which are meditation and reincarnation. The followers tend to practice with harmony and balance throughout their lives. In contrast with Western religion, Taoism and Buddhism do not contain any ideas of worshiping the creator god; instead, they both focus on personal and spiritual development, which lead followers to better understand nature and reality.

Hypothesis about word usage in books "TaoTeChing" and "Buddhism":

Null Hypothesis: Words from two lists (“likely” and “unlikely”) that we have defined above are equally distributed in both “TaoTeChing” and “Buddhism”.

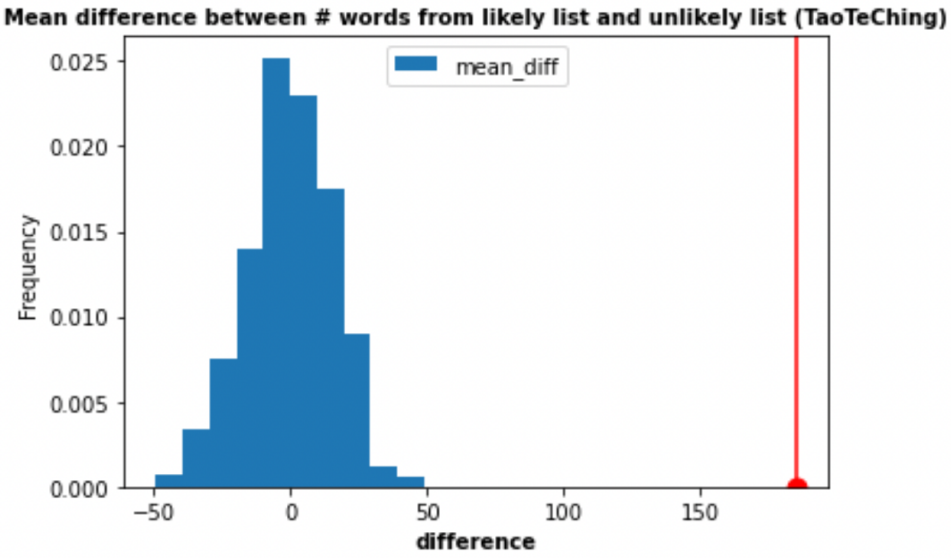
Alternate Hypothesis: Both “TaoTeChing” and “Buddhism” are more likely to contain words from the “likely” list rather than “unlikely” list.

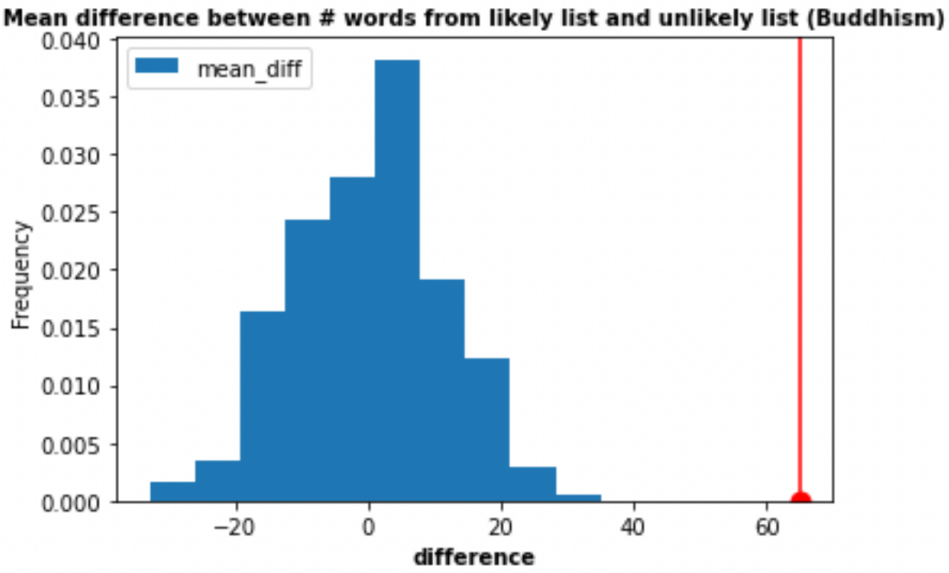
Interpretation of the results:

By using for loop, we found out that the number of words from the “likely” list is greater than the words from the “unlikely” list for both “TaoTeChing” and “Buddhism”. Then, we produced bootstrap testing and found out that p-values are approximately equal to zero for both “TaoTeChing” and “Buddhism” results, which proves that our findings are significant. Furthermore, we produce a line graph that shows the trend of word distribution for both “TaoTeChing” and “Buddhism” and we surprisingly see that the trend is pretty similar between

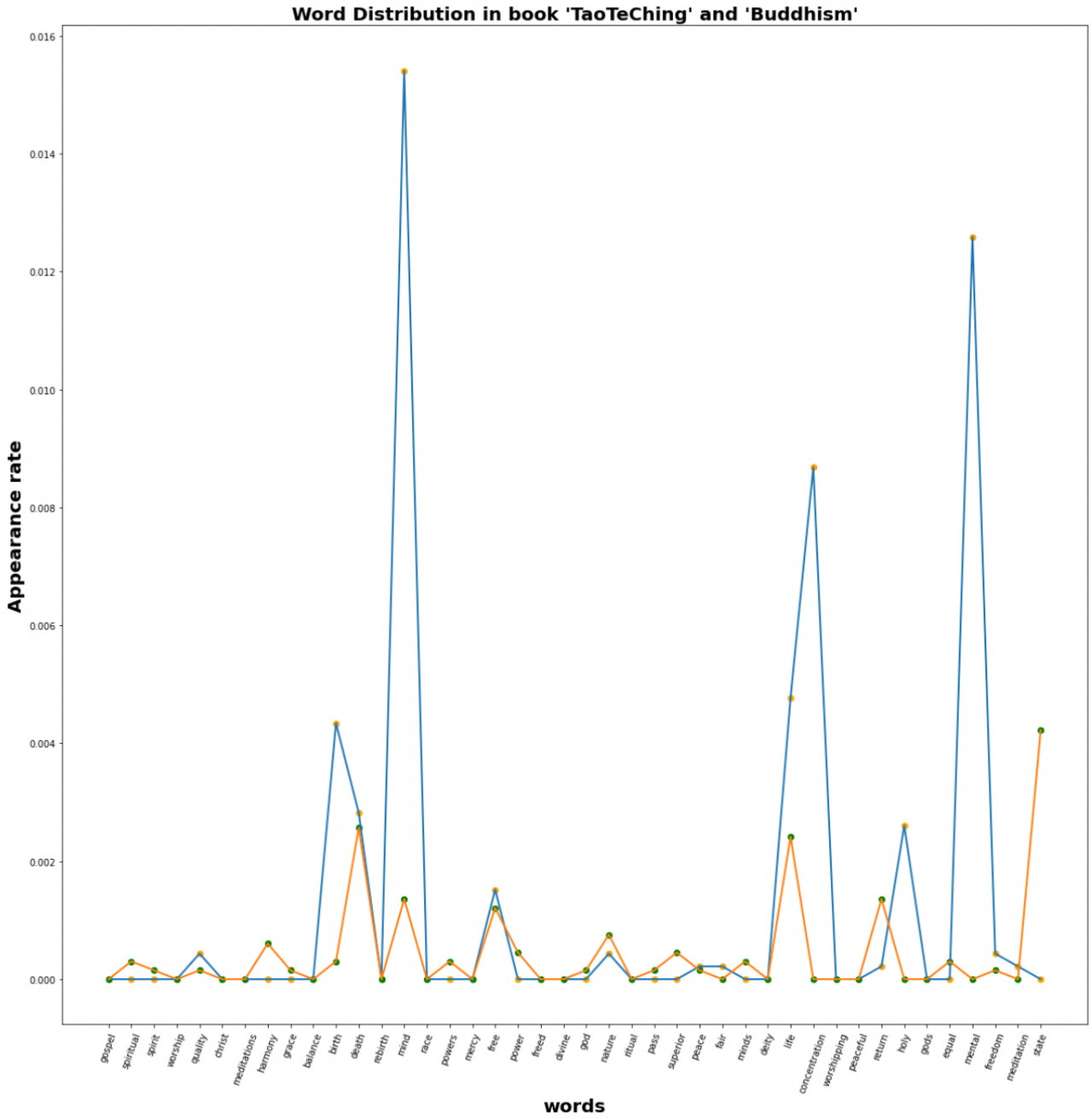
two books. Therefore, we conclude that the word usage for both “TaoTeChing” and “Buddhism” are similar.

Belows are the histograms of our hypothesis tests:





Visualization that further supports our hypothesis:



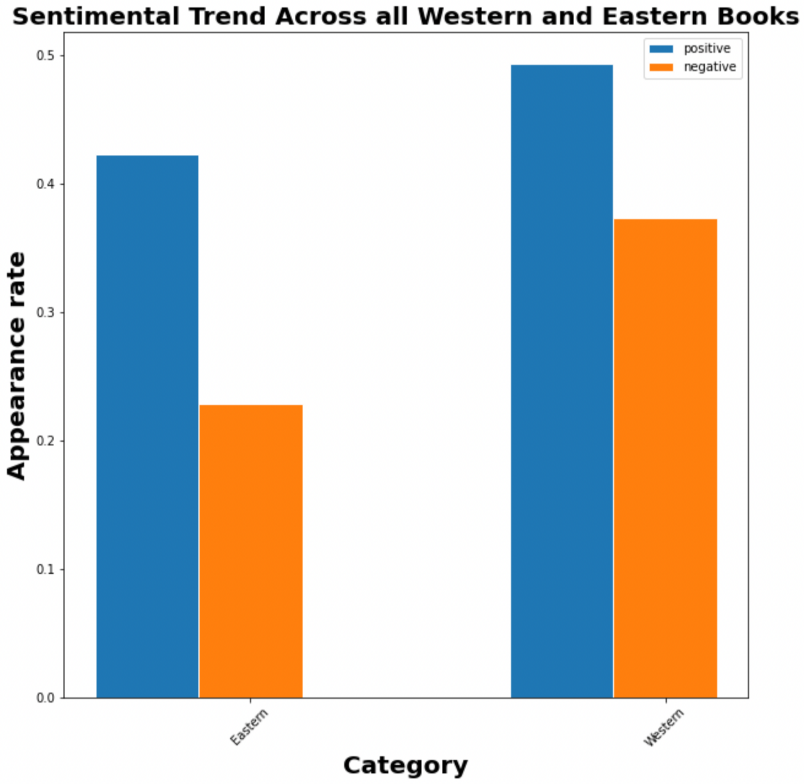
Analysis the result for Buddhism and Taoism:

Both religions advocate equality among all living things. Men and women have no distinctions and animals are considered equal to humans. According to diffen.com, there are many similarities between Buddhism and Taosim. They both value the equality between all creatures, especially they both propose to treat animals equally. Their goals are similar: Buddhism wants to eliminate mental suffering, whereas Taoism wants to find a balance in life, both in the sense of self-development. They even have a similar practice: meditation. Their similarities can be supported by investigating the word usage in the scriptures, according to our results in Part Four, which is the extensive use of words like “mind”, “mental”, “concentration”, and “state”.

Prompt 7: Proposal

With one of our group members coming from a Catholic high school, he noticed the tremendous amount of the word sin appeared in his religion class and the fact that God is cleansing all of us from our sins. Therefore, we believe that the Western religious books may contain a large number of negative words like sin. On the other hand, coming from an Eastern country, our group has a little understanding of the Eastern religions that focus more about personal spiritual development. As a result, we propose the following hypothesis:

- Null hypothesis: the proportion of negative words in the Western religious books is larger than that in the Eastern religious books.
- Alternative hypothesis: the proportion of negative words in the Western religious books is smaller than or equal to that in the Eastern religious books.



Conclusion:

From our two hypotheses, we tested the difference of the number of the word usage in the “likely” list and “unlikely” list by finding the p-value. The p-values we found for are both approximately equal to zero, which rejected the null hypothesis. There is significant evidence that both “TaoTeChing” and “Buddhism” are more likely to contain words from the “likely” list rather than “unlikely” list.

From our exploration of these eight religious books and their word usages, we have found some interesting patterns that support the claim that the Western religions and the Eastern religions are different in terms of their focuses. The Western religious books are more likely to contain words related to god, mercy, pray, etc. because of their characteristics of monotheistic. In contrast, the Eastern religious books are more likely to contain words related to nature, reality, spiritual, etc. because of their focus on personal spiritual development. Also, the word usage may also vary within the Western Religions. For example, Proverbs are more likely to represent God as “Lord” and Ecclesiastes are more likely to represent God as “god.”

Indeed, there are many hidden patterns in this powerful dataset that we were trying to work on but failed to reach a conclusion. Therefore, if we had a more abundant amount of time, we would like to spend more time learning different packages and models to help us analyze our data.

Thank you for organizing this amazing event! It helped us, a group of first-year Data Science students, experience what a real-world project is like, allowing us to obtain a wonderful experience!