Ananda Francis February 27, 2022

DS 3500: Advanced Programming with Data Homework 2

Natural Language Processing and Analysis Applied to the Bible

BACKGROUND:

As someone who recently started getting re-invested in spirituality and religion I wanted to focus my project on books in the Bible. Sticking with mostly popular and well known books I chose 4 Old Testament books (Genesis, Psalms, Proverbs and Isaiah) and 4 New Testament books (Matthews, John, 1 Corinthians and Revelation). I used this website link (Kings James Version): http://www.stewartonbibleschool.org/bible/text/index.html and saved each book as an .txt file. Here's a brief summary of what each book focuses on for context (this context will help understand insight provided in analysis after seeing the visualizations):

1. **Genesis**: first book in the Bible known for some of its many popular stories: Story of Creation, Adam & Eve, Cain & Abel, Noah's Arc, Story of Abraham and Story of Jacob.

2. **Psalms**: 150 "songs, poems and prayers" expressing praise and thanks to God as well as recounting stories of His believers

3. **Proverbs**: important teachings and lessons about morality and goodness; one of the most quoted books for its wisdom and guidance

4. **Isaiah**: consists of prophecies, Isaiah becoming prophet and the deliverance of Jewish peoples from foreign rule and captivity of Babylonians

5. **Matthews**: written by disciple and apostle Matthew, first recountation in the Bible of Jesus' conception, birth, life, ministry, crucifixion, resurrection and stories such as The Last Supper; considered the most important and well known of the 4 Gospels

6. **John**: written by disciple and apostle John, focusing on the miracles of Jesus (like the infamous water to wine and giving a blind man sight), the experience of his divinity and ministry in the flesh, baptisms Jesus performed and ending in Judas' betrayal leading to Jesus' crucifixion and resurrection

7. **1 Corinthians**: one of the 2 letters from Paul writing about the problems of immorality (focusing on division, sex, food, gathering and ressurection) plaguing the church and people of Corinth, a city he brought to the faith, and how the Gospel can fix these issues

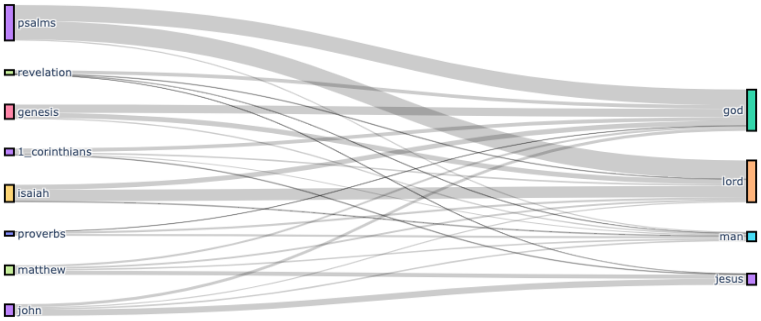
8. **Revelation**: known most as a prophecy for the future and the end of the world focusing heavily on mystical symbolism and numerological importance, many have connected it to events in ancient and modern history

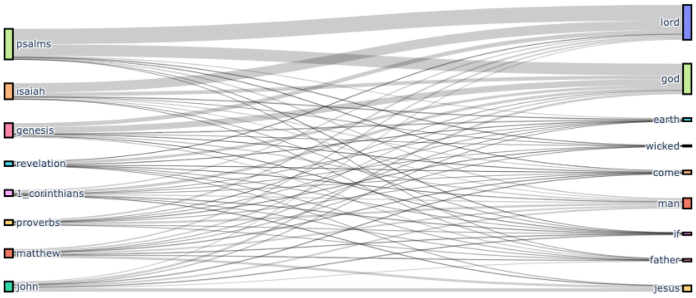
VISUALIZATIONS:

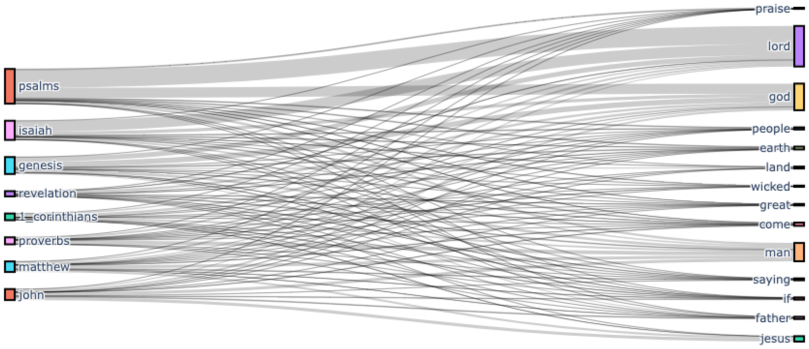
For this project I created 6 visualizations. While we were only supposed to do 3, I felt my investigation required more, although I still stayed consistent with creating 3 types of visualizations. The first type of visualization is Sankey Diagrams. The figure in the top left (look below) is the product of a Sankey Diagram when I take the top word from each Bible book and map the occurrence of each top word to each text (thickness of the line representing how many times a specific word appeared in a specific book). The second diagram (top right) is the resulting Sankey Diagram when I took the top 2 most common words and then the bottom Sankey Diagram is the result when I took the 3 most common words. The reason you see only 5 words instead of 8 in the first image, 9 words instead of 16 in the second image and 15 words instead of 24 in the third image is because some texts have overlapping top words.

If you look under the diagrams (to improve readability of the visualizations) I have created a list of each Biblical book used in this investigation as well as how many times the 15 common words appeared in each book. Taking a closer look at this you will see that "Lord" was a top 3 word for all of the Old Testament books (Genesis, Isaiah, Psalms and Proverbs). You'll also see "God" in the top 3 for Genesis, Psalms, 1 Corinthians and Revelations. I find this interesting as to why: "God" appearing in the top 3 for Genesis makes sense as the book focuses heavily on God's relationship with the first generations of mankind and Psalms consist of at least 100 prayers to God so it makes sense for God to appear very often in these books. 1 Corinthians having "God" appear more than "Jesus" is particularly interesting especially considering that Paul, a disciple of Jesus, was talking about the Gospel (while Matthew and John clearly are more focused on Jesus, his seems more focused on God).

While I can dive deeper into many other trends in this data I'd like to note the significance of "Jesus" being the top word in Matthew and John and having a count of 0 in all of the Old Testament books. While I grew up Christian and know little of Judaism I do know that Jesus being the Messiah, 1 of the Holy Trinity and the son of God is a belief not shared by both religious systems. The Hebrew Bible is also known to be the Old Testament (from the diagrams you see there is no mention of Jesus at all and it isn't until the first Book of the New Testament, Matthew, that we even see Jesus come up). The Hebrew Bible is composed of the Torah (which has the Book of Genesis), Nevi'im (which has the Book of Isaiah) and Ketuvim (which has the Book of Psalms and the Book of Proverbs). NLP is useful here, as, if we were historians who knew little of these religious systems or their texts, the data helps provide some context for the difference of significance of Jesus Christ in these two Abrahamic religions.







genesis:

1. god: 230
2. lord 206
3. land 187
4. father 169
5. earth 121
6. man 107
7. saying 79
8. come 76
9. if 53
10. people 36
11. great 35
12. things: 16
13. wicked 5
14. praise 2
15. jesus 0

psalms:

1. lord 779
2. god 439
3. praise 158
4. earth 141
5. people 130
6. man 99
7. wicked 90
8. great 69
9. come 60
10. land 43
11. if 34
12. things: 33
13. saying 8
14. father 4
15. jesus 0

proverbs:

1. man 157
2. wicked 89
3. lord 86
4. if 29
5. things: 23
6. come 22
7. earth 16
8. great 12
9. father 18
10. people 11
11. god 8
12. land 6
13. praise 4
14. saying 1
15. jesus 0

isaiah:

1. lord 476
2. come 149
3. people 138
4. god 136
5. earth 102
6. land 81
7. man 68
8. things: 50
9. saying 36
10. great 28
11. if 23
12. praise 15
13. father 13
14. wicked 12
15. jesus 0

matthew:

1. jesus 172
2. man 125
3. saying 120
4. come 89
5. lord 76
6. if 74
7. father 60
8. things: 57
9. god 55
10. great 37
11. earth 27
12. people 24
13. land 11
14. wicked 9
15. praise 1

john:

1. jesus 256
2. father 126
3. man 125
4. come 88
5. if 82
6. god 82
7. things: 70
8. lord 46
9. saying 40
10. people 20
11. great 6
12. land 5
13. earth 4
14. praise 3
15. wicked 0

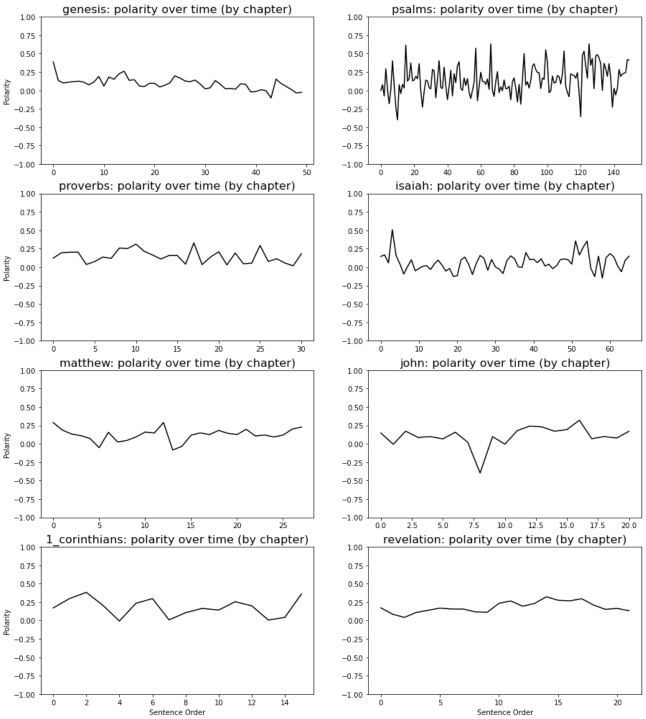
1_corinthians:

1. god 103
2. if 80
3. man 78
4. things: 75
5. lord 62
6. come 31
7. jesus 27
8. praise 5
9. earth 4
10. father 3
11. great 2
12. saying 2
13. people 2
14. wicked 1
15. land 0

revelation:

1. god 99
2. earth 81
3. things 75
4. great 72
5. come 50
6. saying 46
7. man 33
8. lord 22
9. jesus 14
10. if 9
11. people 7
12. father 4
13. praise 1
14. land 0
15. wicked

My second visualization is a group of subplots showing the changes in polarity between each chapter of all the Biblical books used in this investigation. If you look under the graph you will see a table with each Biblical book's minimum polarity score and maximum polarity score (by chapter) and the chapter in the book that the min/max polarity score occurred in. For context, the closer the polarity is to -1 the more "negatively" connotated a text of speech is and the closer the polarity is to 1 the more "positively" connotated a text of speech is.

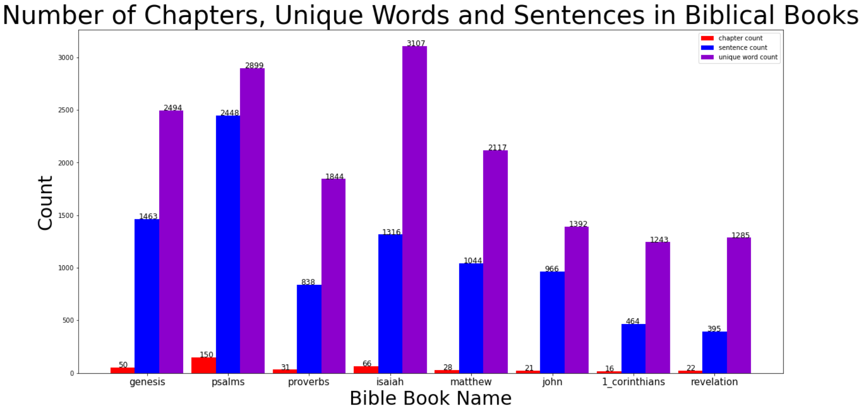


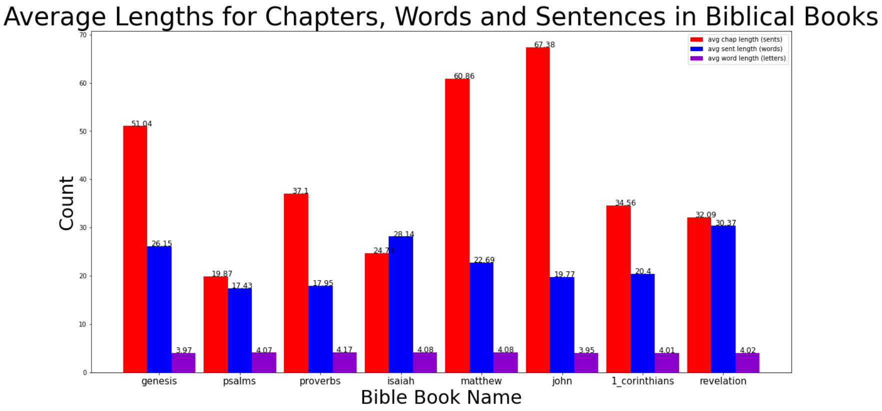
Revelation being of the 2 books who's minimum polarity not negative is interesting especially considering this is the book that talks about the earth being destroyed by fire. Psalms being the book that has the lowest minimum polarity is also interesting considering the book is supposed to be collections of songs praising the Lord (although that specific chapter does talk about the Lord punishing the "wicked"). But Psalms also has the most "positively" connotated chapter(s) of all in the books: chapter 67 being about the blessings from God and giving thanks ("Let all the people praise thee!" in the King James Version, "Let the people thank you, God!" in the Common English Version repeats over and over) and chapter 126 being about freeing and rewarding the people of Israel (or Zion).

book	min_polarity	chapter min_pol occured	max_polarity	chapter(s) max_pol occurred
genesis	-0.102	44	0.387	1
psalms	-0.4	11	0.633	67, 126
proverbs	0.018	30	0.329	18
isaiah	-.150	59	.508	4
matthew	-0.085	14	0.289	13
john	-0.397	9	0.32	17
1 corinthians	-0.008	5	0.382	3
revelations	0.042	3	0.322	15

The final type of visualization are overlaying bar graphs using data from each text and 3 different statistics collected from each text. The first shows the number of chapters and sentences that comprise each book as well as the number of unique words. There appears to be a correlation between the number of unique words and the number of sentences and chapters that comprise each book. A line plot with a unique word count on x axis and sentences/chapters count on the y axis can create another visualization to deny or confirm this observation.

The Old Testament books are a lot longer in chapter length than the New Testament, which makes sense as it has more sentences and unique words as well. A reasoning as to why that could be because of the Old Testament is primarily teachings, stories and the history of Israelites, involving more characters and requiring longer storytelling while the New Testament focuses mostly on Jesus' storyline and the Christian church. Despite total lengths though, the New Testament on average has more sentences per chapter, about the same words per sentence and about the same letters per word (though slightly less).





CONCLUSION:

Natural language processing can be a powerful tool to understand historical texts such as the Bible and how statistics like polarity score , common words and average chapter lengths may affect one's experiences when reading and interpreting such texts. This investigation has many areas of growth, one of which being exploring other books in the Bible as well as other translations and seeing the discrepancy in results, especially with polarity and common words. I greatly enjoyed this investigation and would love to expand it further.

For the code behind these visualizatoins visit my github repo at: https://github.com/anandafrancis/nlp-library