Text Analysis of Bible

Victor Cardeno

What is the Bible?

- Christian religious text detailing history and sacred teachings of Christianity
  - Also contains and/or appropriates elements of scripture from other religions, such as Judaism and Islam
- Structure of Bible
  I. Old Testament
     I. Pentateuch – law and original Hebrew Bible
     II. Prophets
     III. Writings
  II. New Testament
     I. Gospels – tell story of Jesus' ministry on Earth
     II. Acts of Apostles – tells story of early Catholic church post-Jesus
     III. Letters – contain Paul's letters to early Catholic communities
     IV. Revelations -

Questions to Answer

- What are most common words, bigrams, and trigrams in the Bible?
- Which books are written with the most positive and negative sentiment?
- Are there notable differences between books of the same section?
  - i.e. Are there differences in sentiment or word usage in Paul's letter or prophets?
- Can we identify different topics?

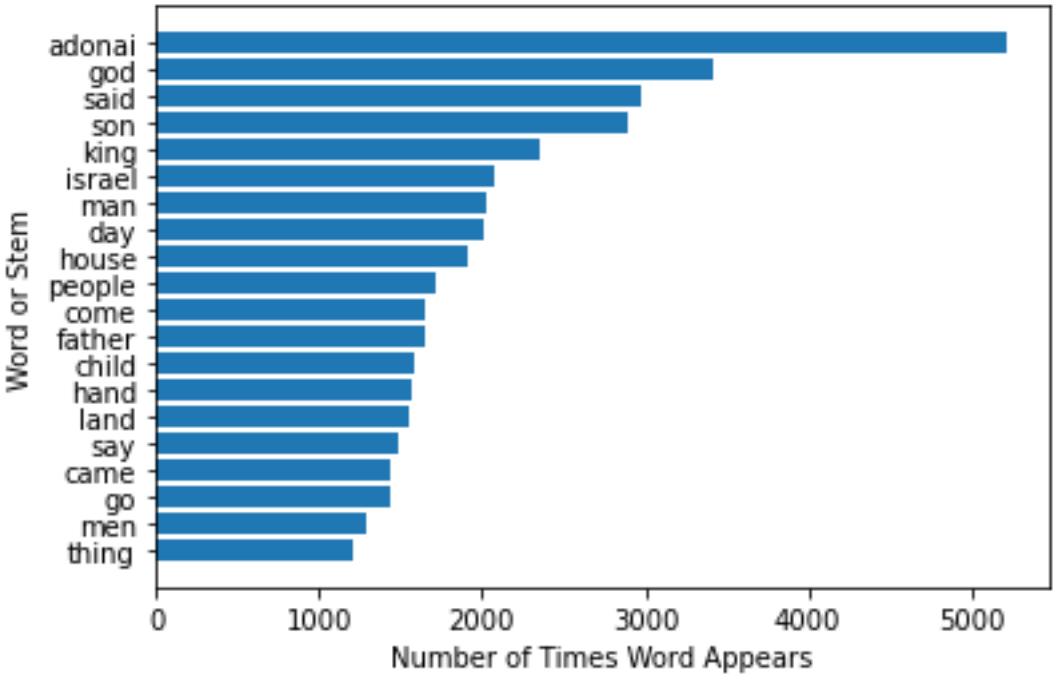
Potential Difficulties

- The Bible, almost by definition, portrays very similar themes and attempts to convey similar messages across books -> difficult to extract different topics
- The Bible has been synthesized and, in some ways, standardized to make it easier to read, so much of the language and syntax is common across different books and chapters
- Language in Bible can be archaic and outdated -> can be difficult for modern NLP algorithms and libraries to decipher
  - Tried to mitigate this by using World English Bible version

Process

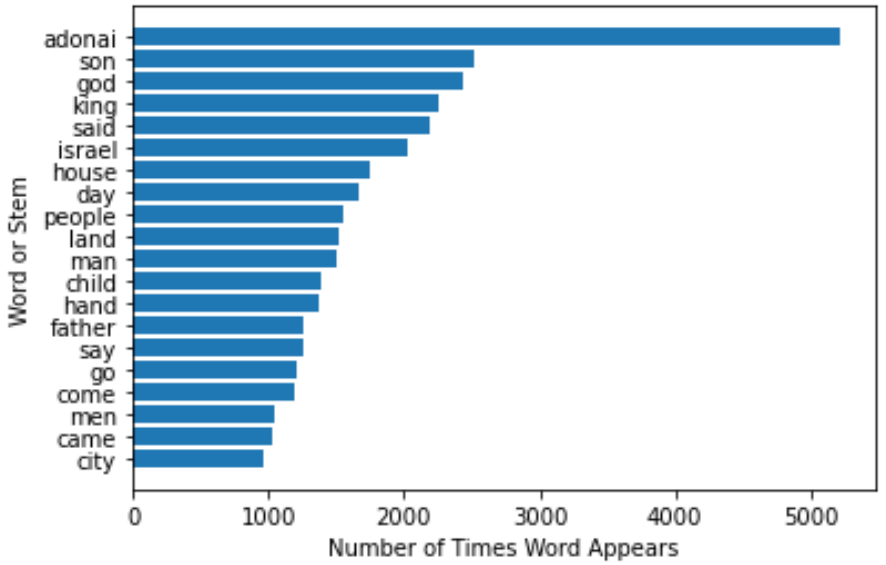
- Data from World English Bible downloaded from Kaggle - https://www.kaggle.com/datasets/oswinrh/bible
- Preprocessing:
  1. Make all words lowercase, remove punctuation, remove footnotes
  2. Lemmatization: looking for root of word to group together words with similar meanings
  3. Remove stopwords – do this to get rid of noise and only keep words that have meaning

Which words are most common in Bible?

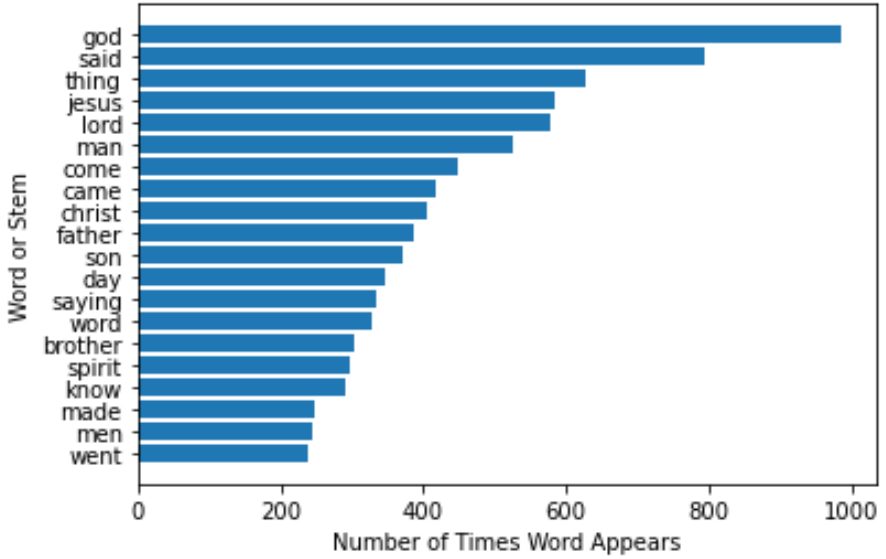


- Most common words in Bible are adonai, god, said, son, king and Israel all appearing > ~2k times
- Adonai is not said “once” in NT, instead preferring to use “god”

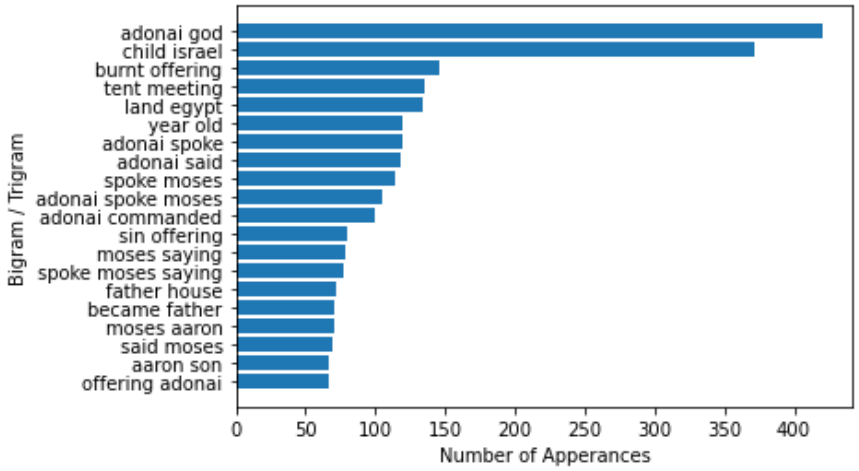
Old Testament



New Testament

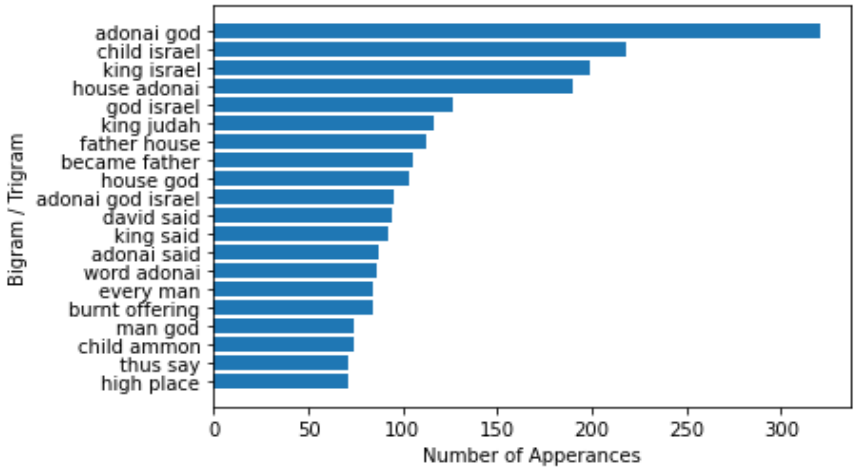


Bigrams and Trigrams in Old Testament

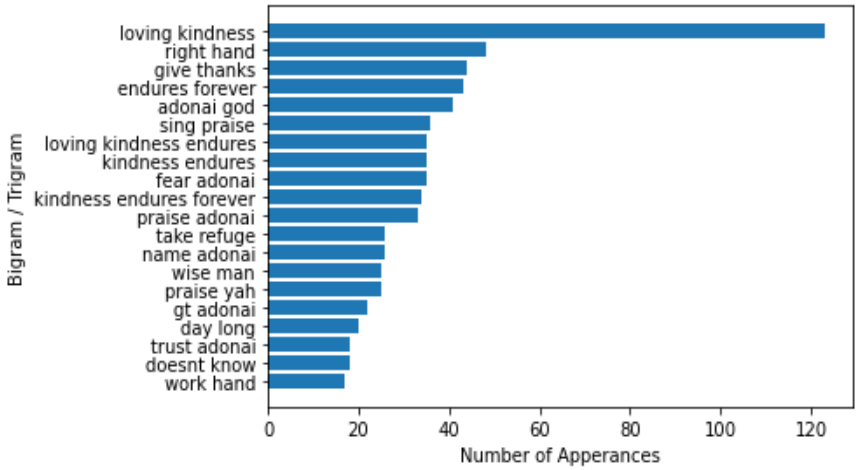


Pentateuch

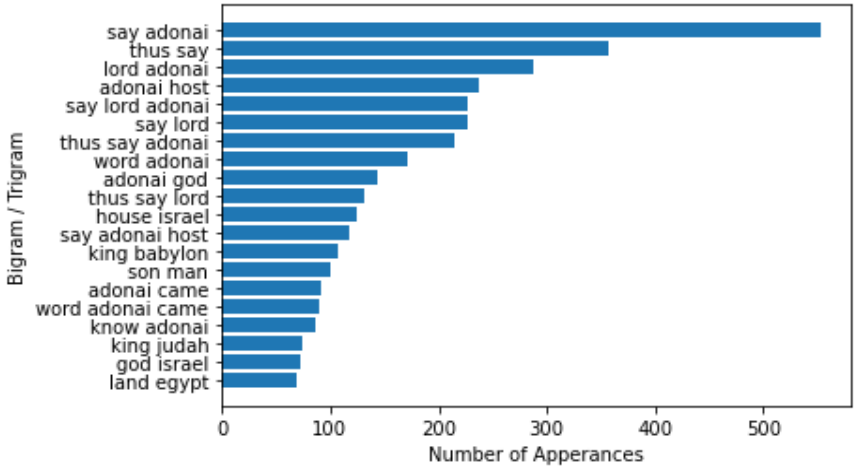
- You can see how different the Writings (Psalms, Proverbs, Songs, etc.) are from the other three Old Testament books.
- It appears the Pentateuch is defined more by its main characters (Moses and Aaron) than the prophetic books.
- The books of the Former Prophets were more historical in nature, while the Latter Prophets were more “prophetic”, describing God’s judgement more.



Former Prophets

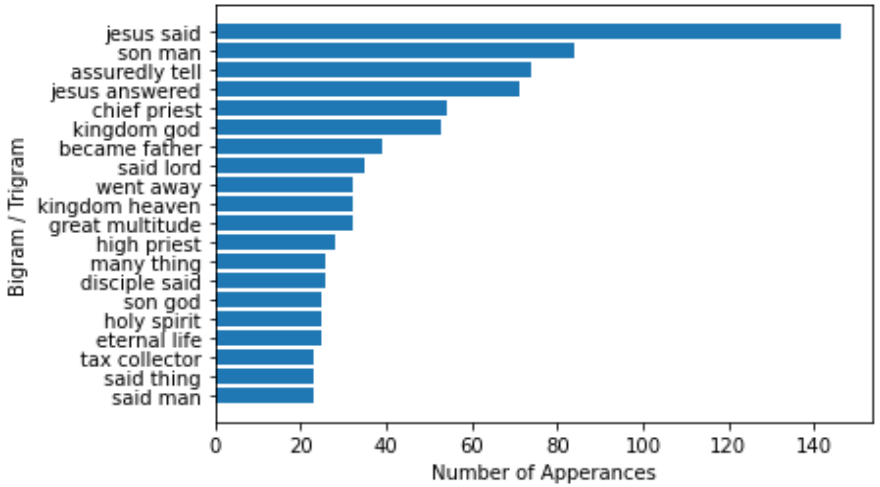


Writings

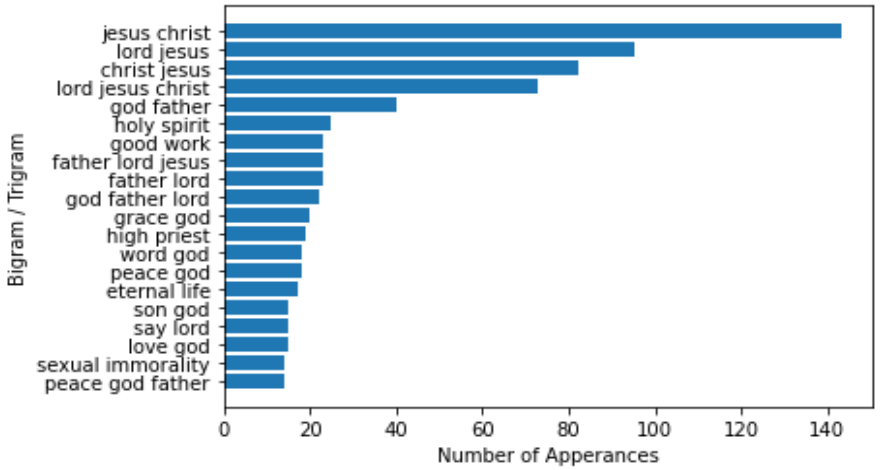


Latter Prophets

Bigrams and Trigrams in New Testament

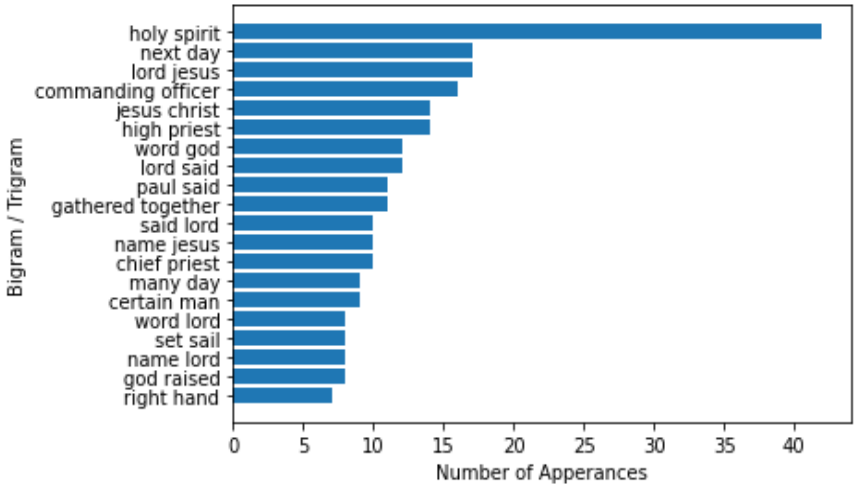


Gospels

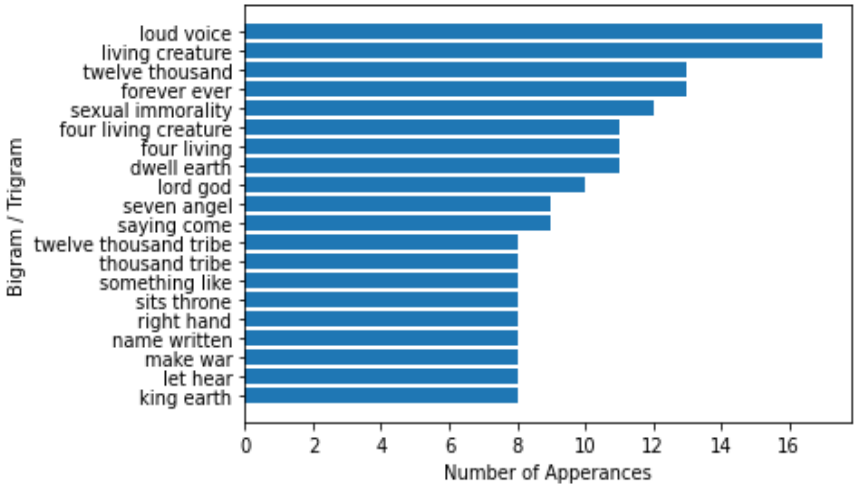


Letters

- Three different syntactical groups emerge:
  - Gospels: detailing Jesus' mission -> discussing him and what he said more
  - Acts and Letters: talking a lot about God's word and Jesus
  - Revelation: discussing judgement and numbers
- It's interesting that sexual immorality appears as a relevant bigram in 2 different sections.



Acts of the Apostles



Revelations

Topic Modeling with Latent Dirichlet Allocation (LDA)

- LDA is a method to extract topics from a series of documents by attempting to combine words in your documents into undiscovered topics and then representing how each document fits into each new topic
  - Output:
    - a series of topics that have weights assigned to each word, from which the coder themself interprets and defines the topic
    - Weights of each topic assigned to each document in your original set of documents, allowing you to identify which documents are most related to which topics
  - Conclusion:
    - It seems that LDA on the whole Bible appears to separate books that have similar themes

LDA on the Whole Bible – Topic Definition

- I'm not super familiar with the intricacies of LDA, so I selected number of topics based on intuition, interpretability of topics, and assignment of topics across the documents.

Topic # / Label	Word Distribution	Interpretation
1 / God	0.021*"**adonai**" + 0.014*"**god**" + 0.010*"**like**" + 0.010*"**man**" + 0.008*"**let**" + 0.007*"**heart**" + 0.007*"**hand**" + 0.007*"**earth**" + 0.006*"**come**" + 0.006*"**word**"	Discussing God and his relationship with humans and the earth
2 / Jesus	0.018*"**god**" + 0.016*"**said**" + 0.014*"**jesus**" + 0.011*"**thing**" + 0.010*"**lord**" + 0.009*"**man**" + 0.008*"**christ**" + 0.008*"**come**" + 0.007*"**came**" + 0.007*"**father**"	Discussing Jesus and his relationship with God
3 / Covenant	0.031*"**adonai**" + 0.011*"**land**" + 0.011*"**god**" + 0.010*"**said**" + 0.010*"**day**" + 0.010*"**son**" + 0.009*"**israel**" + 0.009*"**say**" + 0.008*"**people**" + 0.007*"**child**"	Discussing God's convenant with the people of Israel? Emphasis on adonai means likely Old Testament
4 / Kings	0.023*"**adonai**" + 0.022*"**king**" + 0.020*"**son**" + 0.017*"**said**" + 0.015*"**israel**" + 0.012*"**house**" + 0.012*"**god**" + 0.011*"**david**" + 0.010*"**child**" + 0.009*"**people**"	Discussing kings of Israel, including God's appointment of David as king of Israel

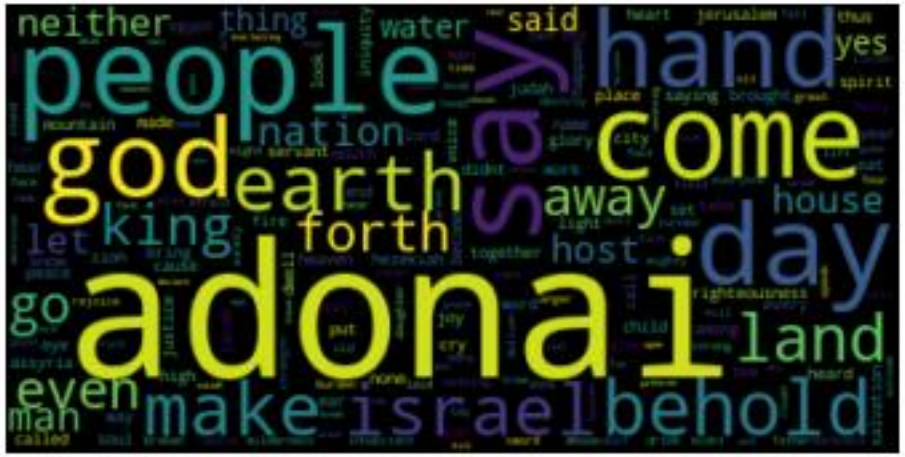
LDA on the Whole Bible – Document Assignment

Section	Topic Distribution	Comment
Pentateuch	3 / Covenant - 0.9999255	Makes sense as Pentateuch told story of formation of covenant
Former Prophets	4 / Kings - 0.9955973	Make sense as Former Prophets told story of Kingdoms of Israel and Judah before conquest
Writings	1 / God - 0.9993451	Makes sense as writings talk about how to form and maintain a relationship with God
Latter Prophets	1 / God - 0.28362772 3 / Covenant - 0.71159524	Prophets were largely predicting judgement of those who persecuted God's people, so maybe I'd expect a bit more of Topic 1, but these topics make sense
Gospels	2 / Jesus - 0.9995394	Makes sense as Gospels discussed Jesus' ministry
Acts of the Apostles	2 / Jesus - 0.99441904	Makes sense as Acts follow up Jesus' time on Earth
Letters	1 / God - 0.027808066 2 / Jesus - 0.9721701	Mostly discussing Jesus, but with a bit more of a personal touch of guidance like in the Writings
Revelations	1 / God - 0.44376022 2 / Jesus - 0.4186992 3 / Covenant - 0.13689654	Revelations stands as pretty different than other books in Bible, so makes sense there'd be less distinction here

Later Prophets Deep Dive

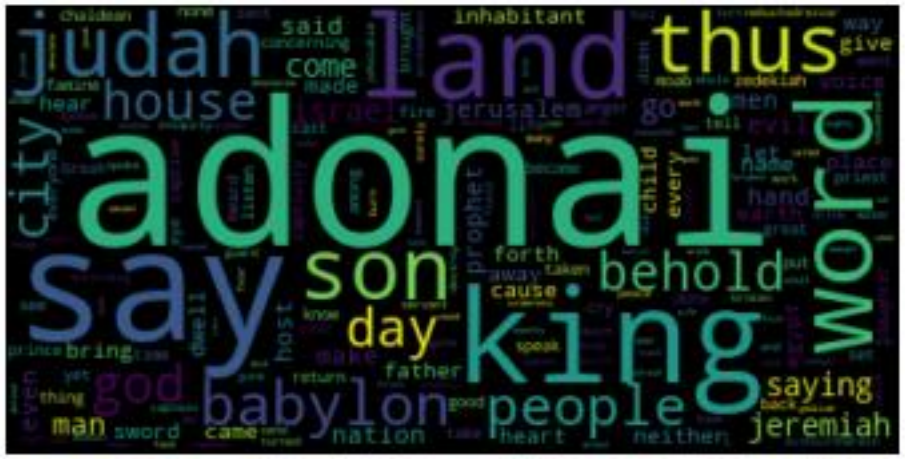
- Later prophets likely had similar overall message but may have presented it in different ways based on audience or perspective, so we can do a bit of a deeper dive into sentiment, syntax, and topics among all the later prophets.
- Later Prophets
  - Isaiah
  - Jeremiah
  - Lamentations
  - Ezekiel
  - Daniel
  - Hosea
  - Joel
  - Amos
  - Obadiah
  - Jonah
  - Micah
  - Nahum
  - Habakkuk
  - Zephaniah
  - Haggai
  - Zechariah

What words do major Latter Prophets use?



Isaiah

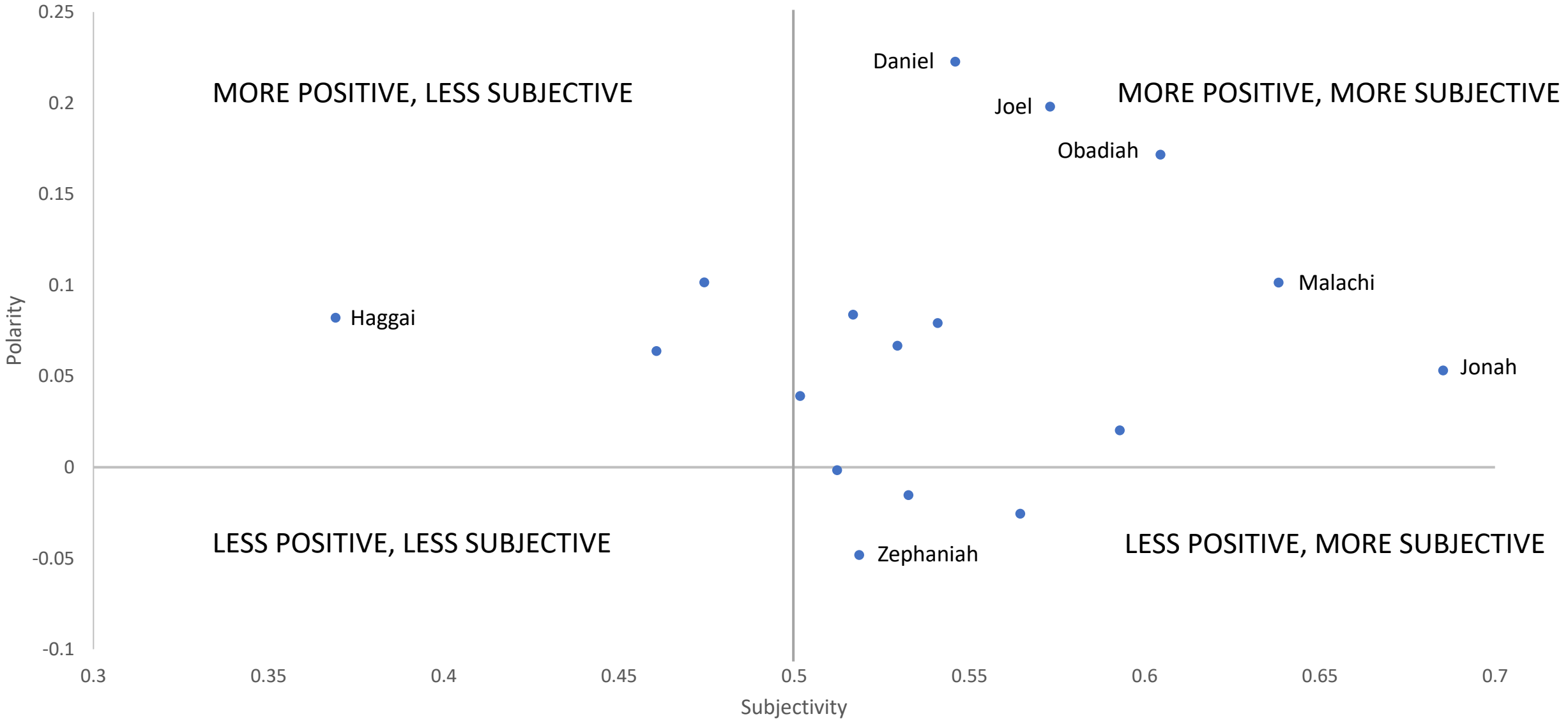
Ezekiel



Jeremiah

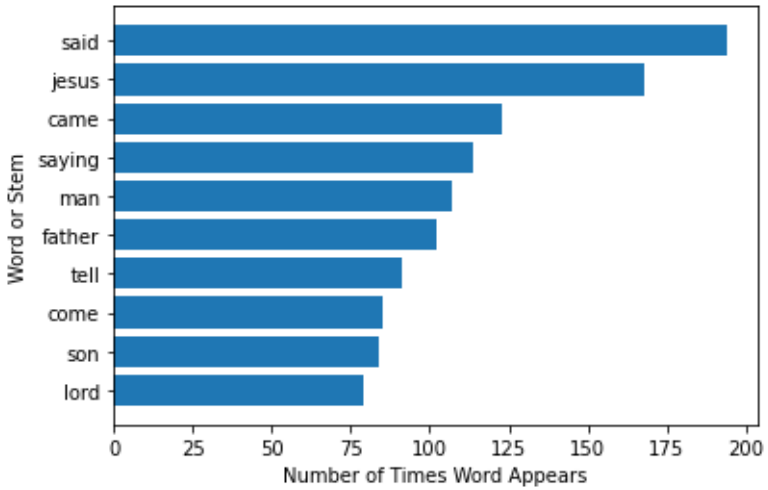
- It's interesting that ***Ezekiel*** seemed to use the title "lord" often while the other two did not seem to use it frequently, if at all.
- ***Isaiah*** appears to be more focused on the people of Israel and the earth.
- ***Jeremiah*** appears to discuss the politics more (Babylon, Judah, land, king)
- ***Ezekiel*** seems to align with Isaiah a bit more and discuss the relationship of God to his people.

How are prophets conveying their message?



How do gospels compare?

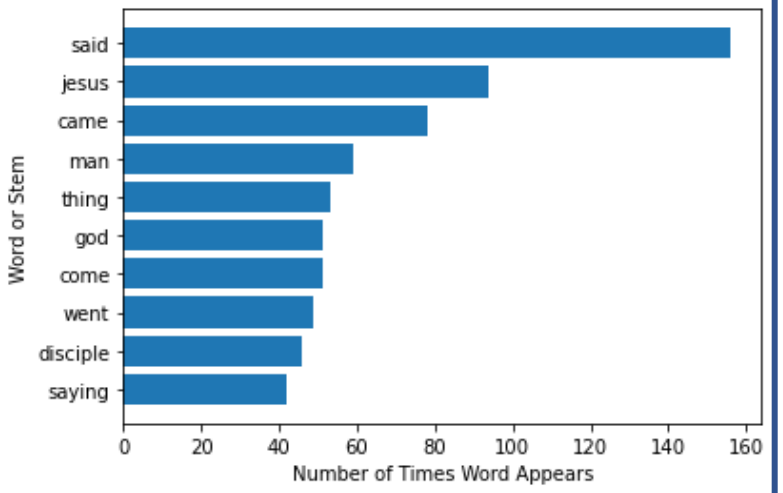
Matthew



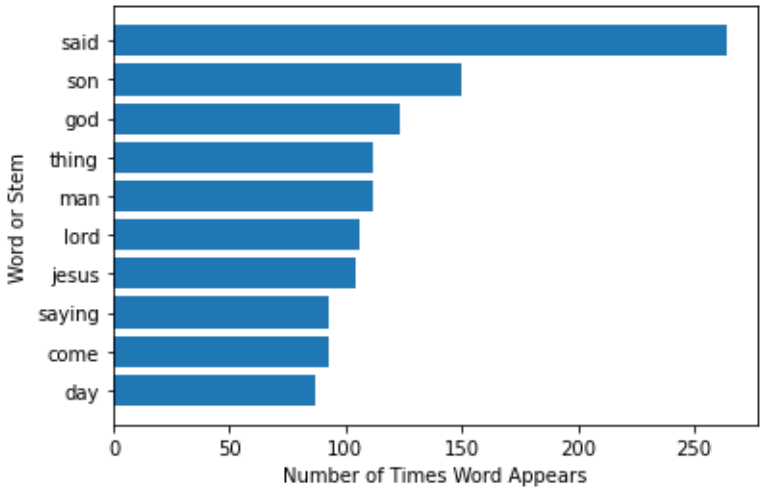
- Polarity: 0.09
- Subjectivity: 0.52

Mark

- Polarity: 0.11
- Subjectivity: 0.48



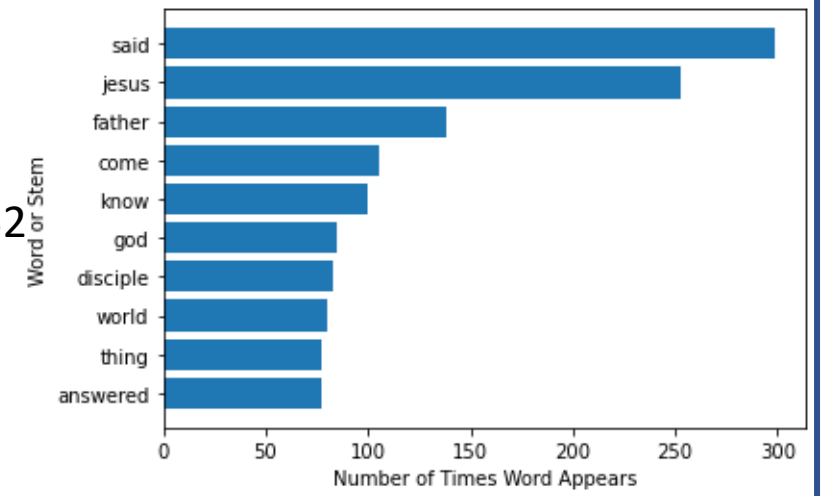
Luke



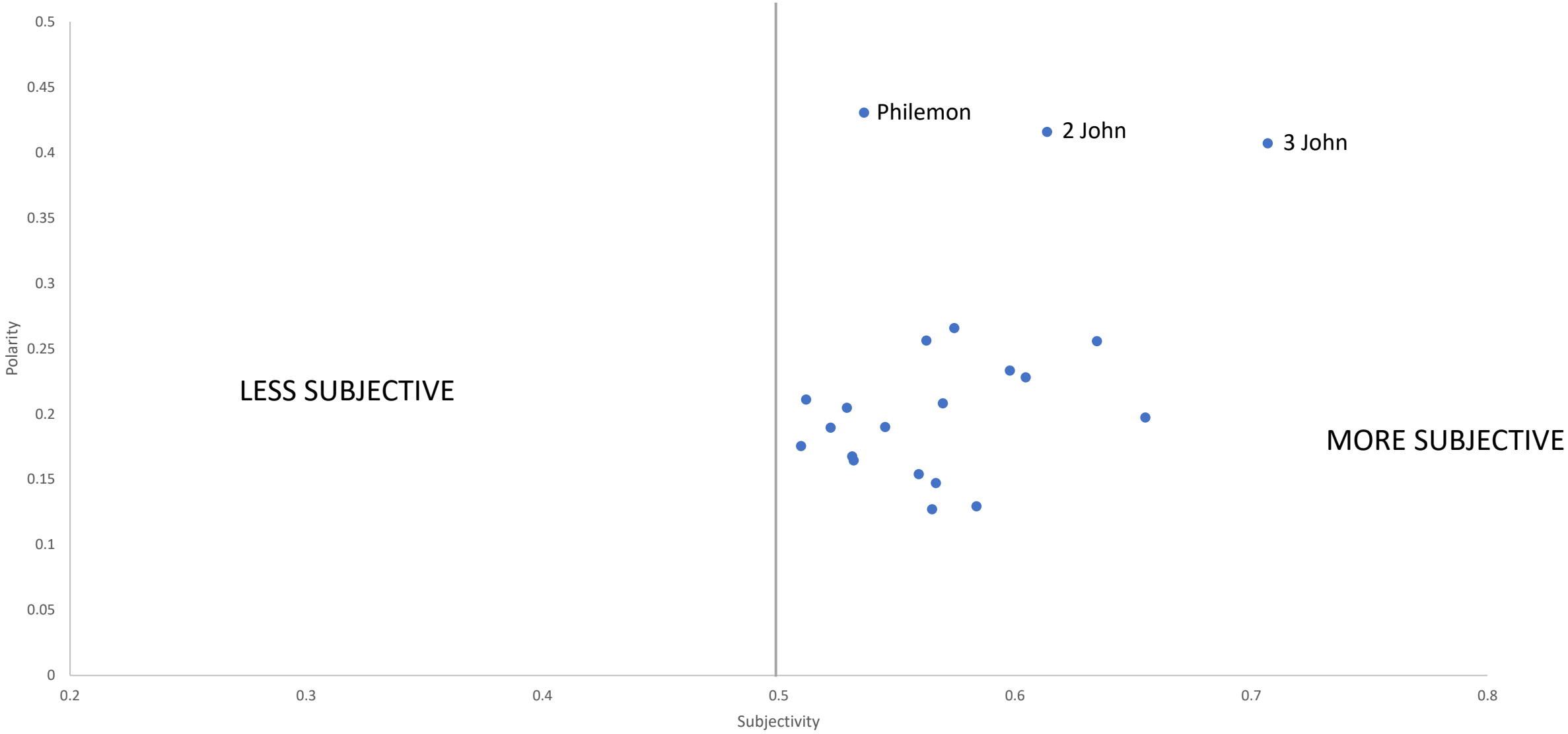
- Polarity: 0.14
- Subjectivity: 0.51

John

- Polarity: 0.12
- Subjectivity: 0.52



Paul's letter to Philemon, as well as John's epistles, appear to be relatively more positive than the others.



Conclusions

- LDA actually worked better than I expected on this dataset; I was able to identify four separate topics, which could be attributed coherently to different books and sections.
- I had to transition in the middle of my analysis to use the World English Bible because modern packages weren't working well with the older language of New American Standard Edition; the Bible is translated and "written" to be consistent for easier read, which made it difficult for certain algorithms like TextBlob to pinpoint differences among them.
- I didn't find anything super interesting, although I did only apply relatively basic text analytics. I suppose the different polarity and subjectivity scores for different prophets and letters are somewhat cool, especially to see the ones that stand out.
- I think some ideas for further analysis may be a closer inspection of the Gospels and of Peter's Letters; the Gospels tell a similar story in four different ways, while the Letters are written largely be the same person to different audiences, meaning there may be more to flesh out across different letters.

Dictionary

- Bigram: consecutive word pair
- Trigram: set of three consecutive words
- Stopword: common grammatical linkage words that have limited or no meaning to text (i.e. “a”, “or”, “the”, “but”, etc.)