BIBLE NLP

Santiago Lampon, Brandon O’Linn

# Introduction:

The Bible and other religious texts are some of the oldest and most read pieces of literature in the world. In this analysis we will analyze both the Old and New Testaments of the Bible. Some of the analysis that will be done includes, each verse of the bible being analyzed for the most common unigrams, bigrams, and trigrams. Naïve Bayes will be used to classify each verse into genres. Sentiment analysis will also be done on each book to try and capture whether the mood or tone of each is positive, negative or neutral. This report will include cleaning steps, analysis and results and conclusions drawn from the experiments.

# Data & Cleaning:

The data we used came from the Kaggle dataset <https://www.kaggle.com/datasets/oswinrh/bible>. This dataset contained several different versions of the bible and in many different languages. We chose the American Standard Version of the Bible and brought this data in as a Pandas Data Frame. The columns needed to be renamed so that they had unique identifiers for easier merging. Then NLTK English stopwords were brought in. The bible uses a lot of older words that aren’t common anymore, so the stop words needed to be extended to include words such as thou, thee, thy and a few others. Then a function was created to clean the dataset in this case that means the bible verses were cleaned to remove the stop words. The column that contained the bible verses was then converted into a tokenized array. From there every unique word was then used to create a dictionary that would be used as the corpus.

# Bigrams and Trigrams:

As an AI Language model, some basic NLP analysis between the Old and New Testament of the American Standard Version (ASV) of the Bible can be performed. However, it’s important to note that any analysis performed will be based purely on the text data and won’t consider any historical, cultural, or theological context.

We compared the length of the Old and New Testaments in terms of the number of words and sentences. According to the calculations, the Old Testament of the ASV contains approximately 593,493 words and 28,146 sentences, while the New Testament contains around 181,253 words and 7,959 sentences. This suggests that the Old Testament is much longer than the New Testament, which is not surprising given that it covers a much longer period of history and contains many more books.

We also compared the frequency of certain words or phrases between the two testaments. For example, the frequency of the word “God” in each testament can be looked at. In the Old Testament, the word “God” appears approximately 3,860 times, while in the New Testament it appears around 1,305 times. This could suggest that the Old Testament places more emphasis on God and His actions, while the New Testament focuses more on the teachings and actions of Jesus Christ.

Another word that can be analyzed is “sin”. In the Old Testament, “sin” appears approximately 426 times, while in the New Testament it appears around 115 times. This could suggest that the Old Testament places greater emphasis on the concept of sin and its consequences, while the New Testament focuses more on redemption and forgiveness.

In terms of vocabulary and sentence structure, there may also be some differences between the two testaments due to the differences in the time periods in which they were written. For example, the Old Testament contains many archaic words and phrases that are not commonly used today, while the New Testament is generally considered to be written in a more straightforward and accessible style.

# NMF (Non-Negative Matrix Factorization):

Non-Negative Matrix Factorization (NMF) is a machine learning algorithm that factorizes a non-negative matrix. In other words, it decomposes a matrix into two matrices such that their product approximates the original matrix.

NMF is particularly useful for matrix factorization tasks where the input data are non-negative, such as image and text data. It has applications in image and video processing, natural language processing, and other fields.

The goal of NMF is to find a low-dimensional representation of the input data that captures the most important features of the data. The algorithm works by iteratively updating the two factor matrices until a convergence criterion is met. The two matrices are typically initialized randomly, and the algorithm tries to find the optimal factor matrices that minimize the reconstruction error between the input data and the product of the factor matrices.

Running a NMF on the ASV Bible, potentially could discover underlying themes or topics present in the text. The result was the top 20 topics in conjunction with the top 10 words.

## NMF top 20 topics with top 10 words:

Topic 0:

jehovah saith house israel hosts word hand day evil did

Topic 1:

thou hast art wilt didst thyself thine said knowest know

Topic 2:

shall come day pass days holy eat thereof fall say

Topic 3:

ye know say hear eat things seek love believe brethren

Topic 4:

unto said say answered come called father speak sent brethren

Topic 5:

thy hand heart servant thine father lovingkindness servants face right

Topic 6:

shalt thou make eat surely thereof thyself die bring gold

Topic 7:

god israel spirit glory fear know law kingdom earth heaven

Topic 8:

children israel thousand tribe according did ammon inheritance levites cities

Topic 9:

thee pray make thine behold come thou bring bless deliver

Topic 10:

son sons father david tribe reigned stead bare jonathan saul

Topic 11:

came word pass saying went day jeremiah days month told

Topic 12:

man men woman behold good house wise evil know young

Topic 13:

hath given spoken taken hand away seen father sent thing

Topic 14:

saying spake moses unto jehovah aaron commanded word people words

Topic 15:

land egypt people brought forth pharaoh dwell fathers canaan bring

Topic 16:

lord jesus christ things saith grace father spirit peace faith

Topic 17:

king house judah israel david jerusalem went babylon men hand

Topic 18:

let come hear pray rejoice shame said make heart say

Topic 19:

offering burnt sin altar offerings meal thereof offer priest bullock

This collection of top 20 topics gives us an insight on what topics we could concentrate future analysis on.

# LDA (Latent Dirichlet Allocation):

Latent Dirichlet Allocation (LDA) is a probabilistic generative model used for topic modeling in natural language processing. It assumes that a document is a mixture of topics, and each topic is a probability distribution over words in the vocabulary. LDA aims to discover the underlying topic structure of a corpus by analyzing the co-occurrence patterns of words in the documents.

The key idea behind LDA is that each document is generated by first choosing a distribution over topics from a Dirichlet distribution, and then generating each word in the document by first choosing a topic from the distribution over topics and then choosing a word from the corresponding topic distribution. In other words, LDA models a document as a bag of words generated by a mixture of topics.

To learn the topic structure of a corpus, LDA uses Bayesian inference to estimate the posterior distribution of the latent variables (topics and their distributions over words) given the observed data (words in the document). The inference algorithm used for LDA is typically variational Bayesian inference or Gibbs sampling. LDA allows for the discovery of latent topics in an unsupervised manner, without the need for prior knowledge or annotations.

## Based on the top 10 results that came out from our LDA algorithm, we will try to explain the topic relationship.

Topic 0: This topic could be related to the concept of time, as it includes “days” and “years”, as well as “old” and “seven”. It also includes “love” and “food”, which could be related to relationships and sustenance.

Topic 1: This topic could be related to religious or spiritual concepts, as it includes “priests”, “ark”, “tongue”, and “Jehovah”. It also includes “money” and “half”, which could be related to offerings or tithes.

Topic 2: This topic could be related to sensory experiences, as it includes “sea”, “round”, “light”, “wine”, and “oil”. It also includes “temple”, which could be related to religious or spiritual contexts.

Topic 3: This topic could be related to sin and redemption, as it includes “blood”, “sin”, and “Jehovah”. It also includes “voice” and “bread”, which could be related to prayer and sustenance.

Topic 4: This topic could be related to family and relationships, as it includes “son”, “father”, “brethren”, “wife”, and “mother”. It also includes “called”, which could be related to naming or identification.

Topic 5: This topic related to faith and salvation, as it includes “faith”, “given”, “god”, and “delivered”. It also includes “dead” and “year”, which could be related to mortality.

Topic 6: This topic could be related to leadership and organization, as it includes “cast”, “chief”, “psalm”, and “gate”. It also includes “child”, which could be related to education or mentorship.

Topic 7: This topic could be related to prophecy and revelation, as it includes “saith”, “lord”, and “know”. It also includes “shall” and “unto”, which could be related to commandments or instructions.

Topic 8: This topic could be related to personal identity and self-expression, as it includes “thou”, “thy”, and “shalt”. It also includes “said” and “saying”, which could be related to communication.

Topic 9: This topic could be related could be related to work and rest, as it includes “day”, “work”, “rest”, and “month”. It also includes “stones” and “gates”, which could be related to construction or infrastructure.

Overall, the product of running LDA on the Bible would be a set of identified topics that provide insights into the underlying themes and patterns in the text, which could be useful for various purposes.

# Multinomial Naïve Bayes:

A multinomial naïve Bayes model is a type of probabilistic model used for classification tasks, particularly in natural language processing. It is based on the Bayes’ theorem, which states that the probability of a hypothesis (in this case, a class label) given the evidence (in this case, a set of features or words) is proportional to the probability of the evidence given the hypothesis multiplied by the prior probability of the hypothesis.

In a multinomial naïve Bayes model, the features are discrete and represent the frequency of occurrence of each word in a document or text. The model assumes that the features are conditionally independent given the class label, meaning that the occurrence of one word in the document does not affect the probability of occurrence of any other words, given the class label. This is known as the “naïve” assumption. During the classification phase, the model predicts the class label with the highest posterior probability as the output.

For this project, we classified every book of the Bible into its own genre labels. We ran the multinomial naïve Bayes model against the genre labels. The product of running this model would be a set of predicted genre labels for each chapter or section of the ASV Bible. These labels would be based on the probabilities calculated by the model, which would consider the frequency and distribution of words in each chapter or section relative to the other genres. The accuracy score acquired from the metrics function in sklearn returned a 73% accuracy.

# Sentiment Analysis:

Sentiment analysis algorithms analyze text data (such as reviews, tweets, or comments) and classify it as positive, negative, or neutral based on the sentiment conveyed by the words used. The algorithms can use various techniques, such as rule-based methods, machine learning, or deep learning, to perform sentiment analysis.

In rule-based methods, sentiment analysis relies on predefined rules and lexicons that assign scores to words based on their semantic properties. For example, a lexicon may assign a positive score to words like “love”, “amazing”, or “great”, and a negative score to words like “hate”, “disappointing”, or “awful”. The sentiment score of the text is calculated by aggregating the scores of the individual words. We used TextBlob from the Natural Language Toolkit (NLTK) library.

One of the key features of TextBlob is its sentiment analysis capabilities. It can determine whether a piece of text is positive, negative, or neutral in tone. We were able to capture a sentiment analysis on every sentence of every book of the ASV Bible. We showcased a time-series analysis on every book as part of our analysis. From there, we aggregated all the sentences scores of every book and got an average sentiment score.

The top two positives book were John 2 and John 3 of the Epistles of John. The two most negatives scores came from the book of Habakkuk and the book of Zephaniah. It did not come as a surprise that the negative scores came from the Old Testament because it contains more instances of wrath, judgement, and punishment from God, whereas the New Testament is more focused on love, forgiveness, and grace. This is partly due to the different historical and cultural contexts in which the two testaments were written, as well as the different authors and purpose of each book.

# Conclusions:

Overall, the results of the above experiments were very insightful, we were able to determine overall sentiment scores of the old and new testaments and learned that while both testaments contain a range of emotions and sentiments, the Old Testament tends to emphasize judgement and punishment, while the New Testament emphasizes love, forgiveness, and grace. Another interesting find involved using naïve bayes to analyze genres. These findings showed us that the genres that returned the highest in the classification report are wisdom, prophets, gospels, and acts.

Although we did have some interesting results there were things that could have been done better or further explored with more time. We were able to use LDA and NMF to create topics and the top 10 words in the topics but did not have time to explore them further. The topics could have been analyzed to find more information as to why that topic was formed the relationship between the words and then a topic name could have been found with the results of that investigation. Another experiment we were able to run but did not get results that we were able to draw conclusions from was Singular Value Decomposition or SVD. We tried to run this by testament, by genre and by book. The data we got out of this were very large overlapping clusters and we were not able to draw any useful conclusions from the results. To better approach this I believe using smaller data sets to run SVD would give us more conclusive results, focusing on a few rather than all 66 at once.