

Corpus Analysis Report

1 Dataset

For this assignment, I chose to analyze the King James Version (KJV) of the Bible, source from project Gutenberg. The Bible serves as an excellent corpus for linguistic analysis because it is naturally divided into two distinct categories: the Old Testament and the New Testament. These two sections were written in different time periods, originally in different languages (Hebrew vs. Greek), and cover distinct theological themes (Law/History vs. Gospel/Epistles). Analysing the linguistic shift between these two sections offers a clear test case for the text classification and topic modelling techniques covered in class.

1.1 Data Collection and Processing

I collected the dataset programmatically using Python's requests library to fetch the raw text file from Project Gutenberg. The raw text was then split into the two testaments.

To meet the requirement of at least 100 documents per category, I treated individual chapters as documents rather than whole books. Using Python's re (regular expression) library, I parsed the text by splitting on chapter headers. This resulted in a robust dataset well exceeding the minimum requirements.

The following table summarizes the size of the corpus after processing:

Category	Number of Documents
New Testament	260
Old Testament	929

2 Methodology

The analysis was performed using Python 3.12. The pipeline consisted of three main stages: preprocessing, distinctive word analysis (Naive Bayes), and topic modelling (LDA).

2.1 Preprocessing

I implemented a custom cleaning function in `preprocessor.py` to provide custom preprocessing steps to the historical text. The steps included:

1. **Noise Removal:** Regex substitution was used to strip verse numbers (e.g., “1:1”, “12:1”) and stray digits, ensuring that topic models would not be dominated by numerical markers.
2. **Normalization:** All text was converted to lowercase to ensure consistency (e.g., “God” and “god” are treated as the same token).
3. **Stopword Removal:** I used the NLTK English stopwords list but augmented it with a custom list of archaic Biblical pronouns and verbs (e.g., thou, thee, hath, shalt, ye, etc.). Without this step, these high-frequency archaic terms would have obscured the meaningful content words in the analysis.

2.2 Analysis Techniques

1. **Naive Bayes:** To identify words unique to each category, I utilized the Log-Likelihood (LLR) derived from a Naive Bayes model. Instead of using the classifier for prediction, I used the feature log probabilities to calculate

$$LLR(w, c) = \log(P(w|c)) - \log(P(w|C_o)).$$

This highlights words that are statistically over-represented in one testament to the other.

2. **Topic Modelling (LDA):** I used the `gensim` library to perform Latent Dirichlet Allocation (LDA). I chose to model 15 topics to capture high-level themes across the entire Bible. The model was trained using 15 passes over the corpus. The resulting topics were then mapped back to the documents to calculate the average topic distribution for each category.

3 Results and Analysis

3.1 Naive Bayes

The Log-Likelihood Ratio successfully identified the vocabulary shift between the two testaments.

Category	Top Distinctive Words (Word, LLR)
New Testament	(jesus, 8.16), (christ, 7.59), (paul, 6.41), (peter, 6.35), (john, 6.19), (disciples, 6.08), (gospel, 5.93), (pharisees, 5.77), (faith, 5.68), (church, 5.63)
Old Testament	(judah, 4.73), (hosts, 4.43), (philistines, 4.27), (joshua, 4.11), (congregation, 3.93), (families, 3.89), (moab, 3.86), (zion, 3.76), (jeremiah, 3.73), (manasseh, 3.70)

The results strongly align with the theological divisions of the text. The Old Testament is characterized by geopolitical terms (e.g., Judah, Philistines, Moab, Zion), reflecting its focus on the history of the Israelite nation. In contrast, the New Testament is defined by evangelical terms (e.g., Jesus, Christ, faith, gospel, church). The high LLR scores indicate that these words are not just frequent, but specifically discriminatory between the two classes.

3.2 Topic Modelling (LDA)

The LDA model discovered several latent themes running through the corpus. Below are 6 of the 15 topics found, along with manually assigned labels based on the top 25 terms.

Topic	Top 25 Terms
Wisdom and Morality (0)	earth (0.011), like (0.009), heart (0.007), wicked (0.007), great (0.006), away (0.006), behold (0.006), thereof (0.006), forth (0.006), every (0.005), mouth (0.005), make (0.005), evil (0.005), way (0.004), neither (0.004), let (0.004), eyes (0.004), yea (0.004), voice (0.004), righteous (0.004), waters (0.004), words (0.004), days (0.004), good (0.004), wise (0.004)
Psalms of Praise and Lament (4)	let (0.021), hast (0.020), praise (0.013), mine (0.013), ever (0.012), earth (0.011), name (0.011), soul (0.010), mercy (0.009), enemies (0.008), salvation (0.007), heart (0.007), strength (0.006), sing (0.006), righteousness (0.006), make (0.006), hear (0.006), like (0.005), give

	(0.005), rejoice (0.005), great (0.005), selah (0.005), forth (0.004), servant (0.004), children (0.004)
Kingdom History (5)	king (0.041), house (0.020), israel (0.020), judah (0.018), jerusalem (0.016), son (0.012), kings (0.012), david (0.009), city (0.007), solomon (0.007), brought (0.007), land (0.006), fathers (0.006), saying (0.006), great (0.006), sent (0.006), years (0.005), servants (0.005), year (0.005), priests (0.005), went (0.005), saul (0.005), took (0.005), hezekiah (0.005), days (0.005)
Prophetic Dialogues (8)	say (0.017), father (0.014), things (0.011), son (0.008), know (0.008), give (0.007), behold (0.007), saith (0.007), hast (0.006), neither (0.006), speak (0.006), saying (0.006), heaven (0.006), eat (0.006), house (0.005), verily (0.005), life (0.005), answered (0.005), take (0.005), name (0.005), hear (0.005), yet (0.005), let (0.005), see (0.005), world (0.004)
The Gospels (12)	jesus (0.028), saying (0.014), went (0.013), things (0.009), son (0.009), saw (0.008), disciples (0.007), saith (0.007), heard (0.007), many (0.007), say (0.007), jews (0.007), answered (0.006), called (0.005), john (0.005), great (0.005), sent (0.005), took (0.005), forth (0.005), peter (0.005), certain (0.005), jerusalem (0.005), days (0.005), behold (0.005), city (0.004)
Pauline Epistles (14)	christ (0.019), things (0.017), law (0.010), jesus (0.010), may (0.009), let (0.009), faith (0.008), every (0.008), good (0.007), spirit (0.007), brethren (0.007), love (0.007), know (0.006), might (0.006), according (0.006), glory (0.006), grace (0.006), flesh (0.005), world (0.005), another (0.005), word (0.005), yet (0.005), sin (0.005), without (0.005), body (0.004)

By averaging the topic distributions across all documents in each category, we can see which themes dominate each testament.

Category	1st Topic	2nd Topic	3rd Topic
Old Testament	0 (Prob: 0.2161)	4 (Prob: 0.1242)	5 (Prob: 0.1083)
New Testament	14 (Prob: 0.4256)	12 (Prob: 0.3070)	8 (Prob: 0.1463)

The distribution shows a clear thematic split. The Old Testament is heavily dominated by topics related to Law/Ritual and Kingdom History, while the New Testament is almost exclusively dominated by the Gospel/Epistle topic. This confirms that while the corpus shares a common vocabulary, the latent thematic structure is distinct between the two categories.

3.3 Experimentation

As an experiment, I re-ran the Naïve Bayes analysis using TF-IDF normalization instead of raw counts. Using TF-IDF actually degraded the interpretability of the results for this specific dataset. Because TF-IDF penalizes words that appear in many documents, central religious terms like "God" or "Lord" (which appear in almost every chapter) were down-weighted. While this is usually desirable, in a theological corpus, these frequent words are often the most important. The raw count method provided a clearer picture of the dominant themes (Israel vs. Jesus) than the TF-IDF method, which tended to surface rarer, less significant proper nouns.

Category	Top Distinctive Words (Word, LLR)
New Testament	(jesus, 4.32), (christ, 4.05), (faith, 3.31), (paul, 3.15), (disciples, 3.12), (peter, 2.90), (john, 2.79), (gospel, 2.77), (church, 2.54), (pharisees, 2.45)
Old Testament	(judah, 2.28), (israel, 2.08), (land, 2.08), (king, 1.91), (hosts, 1.88), (offering, 1.72), (sons, 1.71), (congregation, 1.68), (philistines, 1.64), (zion, 1.63)

4 Discussion

4.1 Dataset Insights

If I were to describe these results to a friend with no NLP experience, I would explain that the computer was able to "read" the Bible and figure out the difference between the Old and New Testaments without being told what they were about.

The "Distinctive Words" analysis acted like a highlighter, marking the words that make the Old Testament sound "old" (words like offering and king) versus the words

that make the New Testament distinct (words like faith and church). It effectively quantified the shift from a story about a specific nation (Israel) to a story about a specific person (Jesus). The topic modeling went a step further, grouping words into "themes." It automatically realized that words like priest, blood, and offering belong together, detecting the "Ritual Law" theme without knowing anything about religious history.

4.2 Personal Learning

This assignment highlighted the critical importance of domain-specific preprocessing. My initial attempts at topic modeling produced poor results because the top words were all archaic "stopwords" like thou, thee, and hath. Standard English stopword lists (like NLTK's) do not cover 17th-century English. I learned that data cleaning is not a "one size fits all" process; I had to manually inspect the text and create a custom exclusion list to get meaningful topics.

Additionally, I learned that more complex metrics are not always better. As noted in the experimentation section, the simpler "Bag of Words" count model actually outperformed the more complex TF-IDF model for this specific task. This taught me that the choice of vectorization must align with the nature of the dataset, in a corpus where repetition is a stylistic feature (like the Bible), penalizing frequency can be counterproductive.