Sophia Nicolas
Prof. Vroman
CS362N
05/10/2025

NLP Final Project Report:
Psalm Genre Detector

Main task:

The main task of this project is to use word embeddings to identify genres of biblical Psalms by detecting reoccurring words or topics within each Psalm.

Steps of Project:

- Psalms will be labelled 0-9 according to genre, read from a labelled Bible corpus, and tokenized by word.
- Initial vector representations of individual Psalms will be built through a TFIDF matrix, where each row is an individual Psalm and each column is single word in the vocabulary of words appearing across all the Psalms.
- Word embeddings are created using SVD to create sub-matrices that further define words or topics.
- For each genre, frequencies of the 10-20 most common words appearing within each Psalm in the genre will be collected, and the word embeddings from these Psalms will be chosen to represent the genre they belong to.
- Plots of PCA values of groups of Psalms by genre will be used to show the most common words in each genre. The purpose of the plots is to visualize word similarity between these common words by how near they are to each other on the plots.
- Cosine similarity between a genre's most common words and all the other words in the Psalm vocabulary will be computed. The top few most similar words will be analyzed to note any word or topic patterns that may provide insight into

how themes across the Psalms are communicated through word use and frequency.

Heart behind project choice:

- Reflect on most common words appearing in types of Psalms to gain a deeper understanding of themes that the Bible focuses on.
- Note what word similarities might attest to God's character, human nature, and the relationship we can have with Him.
- Consider how topic modelling and detecting word relationships through similarity and frequencies can help build Bible study tools such as dictionaries or theme-based concordances.
- Reflect on how God is the creator of language, poetry, and math, and how He communicates His character through both fluently in ways we will never fully understand but will always be in awe of.

Part One:
Labelling, loading, and preprocessing the Psalms.

There are 10 genres that Psalms are labelled as, represented by integers 0 - 9:

- Lament: 0
  - Bringing to God sorrows and requesting for His presence and deliverance.
- Thanksgiving: 1
  - Giving thanks to God for who He is and what He has done.
- Enthronement: 2
  - Ascribing power and glory to God as He exercises His authority as King on the throne.
- Pilgrimage: 3
  - Preparing hearts for an attitude of worship.

- Royal: 4
  - Worshipping God as King of creation, anticipating Christ as King.
- Wisdom: 5
  - Sharing wisdom found through the fear and pursuit of God.
- Imprecatory: 6
  - Requesting God's judgment upon evil and bringing of justice.
- Historical: 7
  - Reflecting on God's faithful guidance of the people of Israel.
- Praise: 8
  - Giving praise to God for who He is and what He has done.
- Confidence: 9
  - Expressing assurance in God's faithfulness, strength, and goodness.

These genres are merely human labels for us to help categorize and remember specific themes that God's Word talks about in each Psalm. A psalm can be in multiple genres at once, but for simplicity each are placed in only one category for this project.

Part Two:
Create vectors that represent each Psalm using TF-IDF.

A TF-IDF matrix representing the Psalms is created, each row vector representing individual Psalms in their entirety.

- The elements in a row correspond to a column in the matrix representing a word in the vocabulary of words that appear across tokenized Psalms.
- An element at index (row, column) is a number representing individual document (row) to word (column) relationship.

Part Three:
Create word embeddings of each word appearing in a generalized vocabulary of the Psalms using SVD.

This part continues the vectorization task started in Part Two, taking the TF-IDF matrix representing the collective Psalms and now zooming in representing vocabulary words that appear across all Psalms.

Embeddings generated will numerically represent each individual word that appears in the vocabulary of the tokenized Psalms. Word embeddings are made with SVD, which is helpful for numerically representing topics as they are used in a corpus.

Create dictionaries that store Psalms and their TI-IDF and SVD vector representations.

A dictionary will be created of vocabulary word-vector mappings for easy access. Another dictionary for Psalm-vector mappings will also be created. The values for this dictionary map a Psalm to its TIF-IDF vector and also a list of the appropriate word embeddings in place of the initial word tokens.

Part Four:

Analyze what words may be representative of a genre.

By detecting the frequency of the most common words appearing in each Psalm within a given genre, perhaps we can find what words may be representative of a genre.

A dictionary for accessing and analyzing words in the Psalms that a genre is associated with will be created.

Then, the most common words in the Psalms of each genre will be collected. These words and their frequencies as (word, frequency) tuples will be stored and printed.

In the next part, these words will have their relatedness levels visualized and analyzed.

Part Five:

Visualize word-relatedness between the 20 most common words of a Psalm genre through PCA value plotting.

PCA takes in the SVD-generated word embeddings and reduces their dimensionality to 2D. These 2D points are then plotted, so that one can see word-relatedness a little more clearly.

This visualization is done in hopes that words or topics representative of a genre can be understood at a deeper level by seeing a model's perception of relatedness between these words and topics.

Part Six:

Cosine similarities will be computed between words in each given genre corpus and the words "lord" and "god".

Something interesting I noticed is that the words "lord" and "god" are almost the most common words used in a group of Psalms associated with a

given genre. I assume these words are primarily references to God Himself, His name and titles often translated as either "Lord" or "God".

The word "praise" is also usually far from the other points, and is the nearest word to "lord" across genres where both of these words are seen as among the most commonly occurring words in a genre.

These thoughts can be confirmed or found otherwise through doing a study on the Psalms that contain these words within them.

Out of curiosity, I will be using what scores cosine similarity computes for each word pair in a given Psalm genre containing the words "lord" or "god".

Part Seven:

The reason why I thought it was cool that "Lord" and "God" are consistently shown as outliers among the word-relatedness graphs is because I am wondering if the Psalms poetry structure uses God's name in such a unique way compared to the rest of the vocabulary, and if so, is the model able to detect that?

This is something I would like to look into, and am curious if this is an artistic choice of God through the psalmists to communicate His attribute of holiness. It was also cool seeing that “Lord” and “God” are almost always the most commonly occurring words in a group of Psalms, since everything in the Bible is ultimately about God. “Lord” and “praise” having the closest relationship among most common words is cool to see too, since to praise the Lord is the purpose of the Psalms and of the life God designed for His people.

If I were to continue moving further with this project, I would be curious about learning more about topic modelling and how to accurately represent texts with it, especially if it is God's Word.

Another thing that I am curious about is how embeddings perform on the original language or different translations, and how they can be used for Bible study, dictionary, or translation tools.