SI671 Final Project: Deep learning language model on Bible Analysis

Fangzhe Li
University of Michigan School of Information

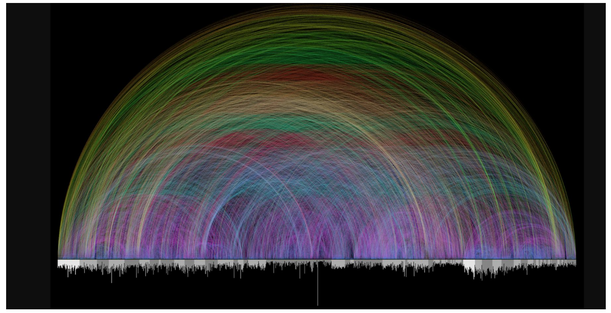
Background

The Bible

The Bible is one of the best selling books in the whole history of human. It is a collections of scriptures containing 66 books, 39 for The Old Testament and 27 for The New Testament.

Cross Reference

There are a lot of cross references in the Bible, which means one verse mentions another verse



Data and Method

The Bible Dataset

King James Version Bible

Cross Reference Dataset

A crowd-sourcing dataset with all the cross references in the Bible and people's vote on if this cross reference between two verses are valid

Sentence Transformer

A sentence transformer sentence-transformers/all-MiniLM-L6-v2based on a BERT family model and fine-tuned on sentence similarities tasks. It is used to generate the sentence embeddings for verse pairs and calculate the cosine similarity between them.

Direct Quotation

```
[29]: fcrs[fcrs.tb != fcrs.fb].sort_values('similarity').tail(20)
```

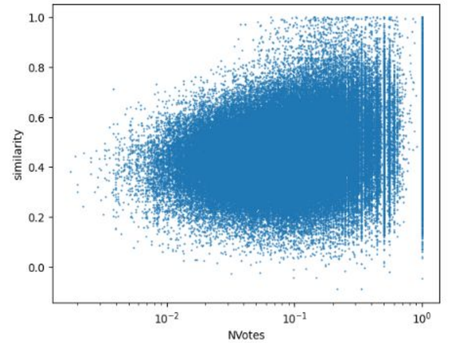
[29]:

	From Verse	To Verse	Votes	Indicator	tt	tb	fb	ft	similarity	NVotes
18223	Exodus.20.7	['Deuteronomy.5.11']	26	False	Thou shalt not take the name of the LORD thy G...	Deuteronomy	Exodus	Thou shalt not take the name of the LORD thy G...	0.995554	0.209677
39366	Deuteronomy.5.11	['Exodus.20.7']	28	False	Thou shalt not take the name of the LORD thy G...	Exodus	Deuteronomy	Thou shalt not take the name of the LORD thy G...	0.995554	0.736842
231071	Matthew.3.10	['Luke.3.9']	14	False	And now also the axe is laid unto the root of ...	Luke	Matthew	And now also the axe is laid unto the root of ...	0.996810	0.104478
252116	Luke.3.9	['Matthew.3.10']	5	False	And now also the axe is laid unto the root of ...	Matthew	Luke	And now also the axe is laid unto the root of ...	0.996810	0.151515
192403	Jeremiah.52.1	['2 Kings.24.18']	4	False	Zedekiah was twenty and one years old when he ...	2 Kings	Jeremiah	Zedekiah was one and twenty years old when he ...	0.997585	0.500000
244245	Mark.1.24	['Luke.4.34']	5	False	Saying, Let us alone; what have we to do with ...	Luke	Mark	Saying, Let us alone; what have we to do with ...	0.998869	0.185185
252702	Luke.4.34	['Mark.1.24']	7	False	Saying, Let us alone; what have we to do with ...	Mark	Luke	Saying, Let us alone; what have we to do with ...	0.998869	0.259259
96401	Nehemiah.7.65	['Ezra.2.63']	4	False	And the Tirshatha said unto them, that they sh...	Ezra	Nehemiah	And the Tirshatha said unto them, that they sh...	0.998995	0.181818
241604	Matthew.24.46	['Luke.12.43']	4	False	Blessed is that servant, whom his lord when he...	Luke	Matthew	Blessed is that servant, whom his lord when he...	1.000000	0.266667
261204	Luke.21.33	['Mark.13.31']	24	False	Heaven and earth shall pass away: but my words...	Mark	Luke	Heaven and earth shall pass away: but my words...	1.000000	0.169014
344766	Revelation.22.21	['2 Thessalonians.3.18']	7	False	The grace of our Lord Jesus Christ be with you...	2 Thessalonians	Revelation	The grace of our Lord Jesus Christ be with you...	1.000000	0.411765
257280	Luke.12.34	['Matthew.6.21']	11	False	For where your treasure is, there will your he...	Matthew	Luke	For where your treasure is, there will your he...	1.000000	0.366667
83442	1 Chronicles.16.22	['Psalms.105.15']	7	False	Saying, Touch not mine anointed, and do my pro...	Psalms	1 Chronicles	Saying, Touch not mine anointed, and do my pro...	1.000000	0.291667
249131	Mark.13.31	['Luke.21.33']	12	False	Heaven and earth shall pass away: but my words...	Luke	Mark	Heaven and earth shall pass away: but my words...	1.000000	0.122449
232977	Matthew.6.21	['Luke.12.34']	27	False	For where your treasure is, there will your he...	Luke	Matthew	For where your treasure is, there will your he...	1.000000	0.141361
18298	Exodus.20.13	['Deuteronomy.5.17']	20	False	Thou shalt not kill.	Deuteronomy	Exodus	Thou shalt not kill.	1.000000	0.180180
39407	Deuteronomy.5.17	['Exodus.20.13']	31	False	Thou shalt not kill.	Exodus	Deuteronomy	Thou shalt not kill.	1.000000	0.632653
164696	Isaiah.37.32	['2 Kings.19.31']	4	False	For out of Jerusalem shall go forth a remnant,...	2 Kings	Isaiah	For out of Jerusalem shall go forth a remnant,...	1.000000	0.500000
192608	Jeremiah.52.24	['2 Kings.25.18']	7	False	And the captain of the guard took Seraiah the ...	2 Kings	Jeremiah	And the captain of the guard took Seraiah the ...	1.000000	0.259259
249477	Mark.14.26	['Matthew.26.30']	5	False	And when they had sung an hymn, they went out ...	Matthew	Mark	And when they had sung an hymn, they went out ...	1.000000	0.555556

Allusion


```
[72]: {'result_codes': ["['Numbers.21.7', 'Numbers.21.8', 'Numbers.21.9']",
  "['John.12.32', 'John.12.33', 'John.12.34']",
  "['Luke.18.31', 'Luke.18.32', 'Luke.18.33']",
  "['2 Kings.18.4']",
  "['John.8.28']",
  "['Luke.24.44', 'Luke.24.45', 'Luke.24.46']",
  "['Luke.24.26', 'Luke.24.27']",
  "['Acts.2.23']",
  "['Luke.24.20']",
  "['Acts.4.27', 'Acts.4.28']",
  "['Matthew.26.54']",
  "['Psalms.22.16']"],
 'result_texts': ['Therefore the people came to Moses, and said, We have sinned, for we have spoken against the LORD, and against thee; pray unto the LORD, that he take away the serpents from us. And Moses p
rayed for the people.And the LORD said unto Moses, Make thee a fiery serpent, and set it upon a pole: and it shall come to pass, that every one that is bitten, when he looketh upon it, shall live.And Moses m
ade a serpent of brass, and put it upon a pole, and it came to pass, that if a serpent had bitten any man, when he beheld the serpent of brass, he lived.',
  'And I, if I be lifted up from the earth, will draw all men unto me.This he said, signifying what death he should die.The people answered him, We have heard out of the law that Christ abideth for ever: and
how sayest thou, The Son of man must be lifted up? who is this Son of man?',
  'Then he took unto him the twelve, and said unto them, Behold, we go up to Jerusalem, and all things that are written by the prophets concerning the Son of man shall be accomplished.For he shall be deliver
ed unto the Gentiles, and shall be mocked, and spitefully entreated, and spitted on:And they shall scourge him, and put him to death: and the third day he shall rise again.',
  'He removed the high places, and brake the images, and cut down the groves, and brake in pieces the brazen serpent that Moses had made: for unto those days the children of Israel did burn incense to it: an
d he called it Nehushtan.',
  'Then said Jesus unto them, When ye have lifted up the Son of man, then shall ye know that I am he, and that I do nothing of myself; but as my Father hath taught me, I speak these things.',
  'And he said unto them, These are the words which I spake unto you, while I was yet with you, that all things must be fulfilled, which were written in the law of Moses, and in the prophets, and in the psal
ms, concerning me.Then opened he their understanding, that they might understand the scriptures,And said unto them, Thus it is written, and thus it behoved Christ to suffer, and to rise from the dead the thi
rd day:',
  'Ought not Christ to have suffered these things, and to enter into his glory?And beginning at Moses and all the prophets, he expounded unto them in all the scriptures the things concerning himself.',
  'Him, being delivered by the determinate counsel and foreknowledge of God, ye have taken, and by wicked hands have crucified and slain:',
  'And how the chief priests and our rulers delivered him to be condemned to death, and have crucified him.',
  'For of a truth against thy holy child Jesus, whom thou hast anointed, both Herod, and Pontius Pilate, with the Gentiles, and the people of Israel, were gathered together,For to do whatsoever thy hand and
thy counsel determined before to be done.',
  'But how then shall the scriptures be fulfilled, that thus it must be?',
  'For dogs have compassed me: the assembly of the wicked have inclosed me: they pierced my hands and my feet.'],
 'error': 'N/A'}
```


Important Result



Findings

- There are no statistically important relationship between votes and similarity of the text.
- There are some verse pairs with relatively high similarity but negative votes.
- The Bible is consistent among different books even though they are written by different people in different periods.
- The allusions between different books are caught by similarities between the embeddings

Future Works

- Calculate the all the 9.6 million verse pairs and use the correlation coefficent as an indicator of how well our sentence transformer understands the semantic of the Bible
- Combine other information retrieval technique like BM25 to build a better search engine.

SI671 Project Proposal

Fangzhe Li
fangzhel@umich.edu

December 13, 2022

Abstract

In this project, we will use a sentence transformer model to generate the embeddings of King James Version Bible verses, try to dig the similarity between different verses and make recommendations according to the input verses, which may be very helpful to both Christians and non-believers who want to know more about the Christianity to read the context within the Bible. Some annotated data from crowd sourcing will be used as the ground truth. Several discoveries are made on the characteristics of the Bible. We also use the similarities calculated by the embeddings generated by the sentence transformer model to rerank the verse recommendations to let it catch allusions relation.

1 Objectives

Build a sentence transformer model on The Bible corpus to explore the cross reference between The Bible's verses. Try to figure out some insight on how Bible is organzied.

2 Introduction

The Bible is one of the best selling books in the whole history of human. It is a collections of scriptures containing 66 books, 39 for The Old Testament and 27 for The New Testament. And another good characteristic of it is that it will never change so we can have the same content even after 10 years. However, it is usually hard to get into the world of the Bible because of the deep cross references between different verses from different books. This project aimed at showing some insights on how the cross references are related to each other and how all scripture is given by inspiration of God, and is profitable for doctrine, for reproof, for correction, for instruction in righteousness.

Figure 1: Bible Sturcture

There are a lot of cross reference from the Bible but some of them are controversial to different readers. Also, these connections are limited in the perspective of studying the Bible. Here is a version of cross reference visualized by Chris Harrison.

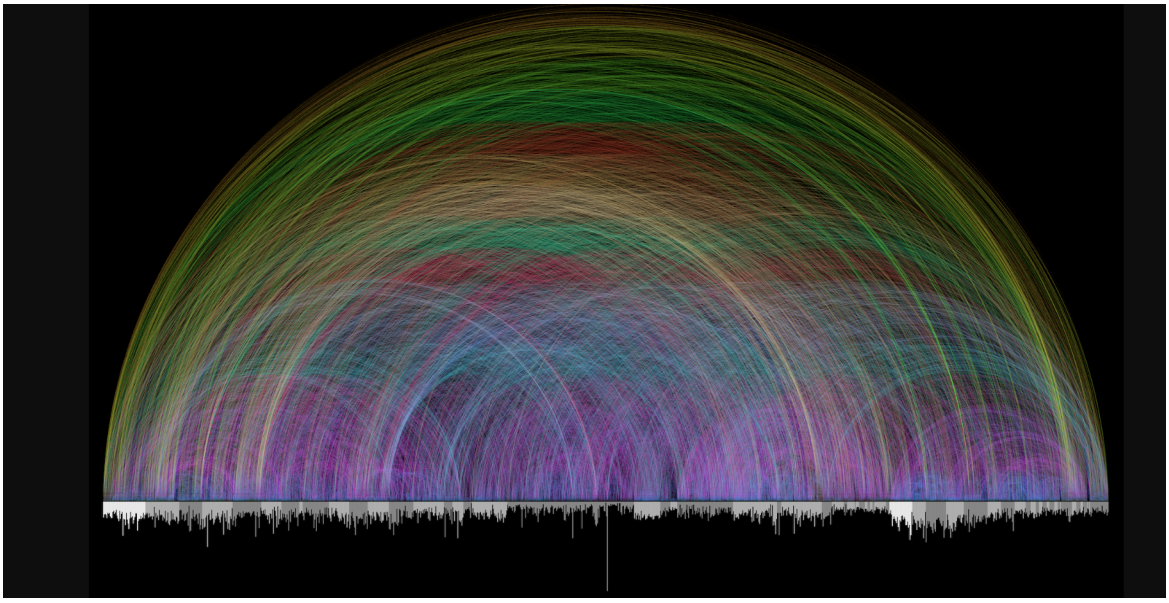


Figure 2: Bible cross reference - Harrison

And here is another version made by openbible.info.

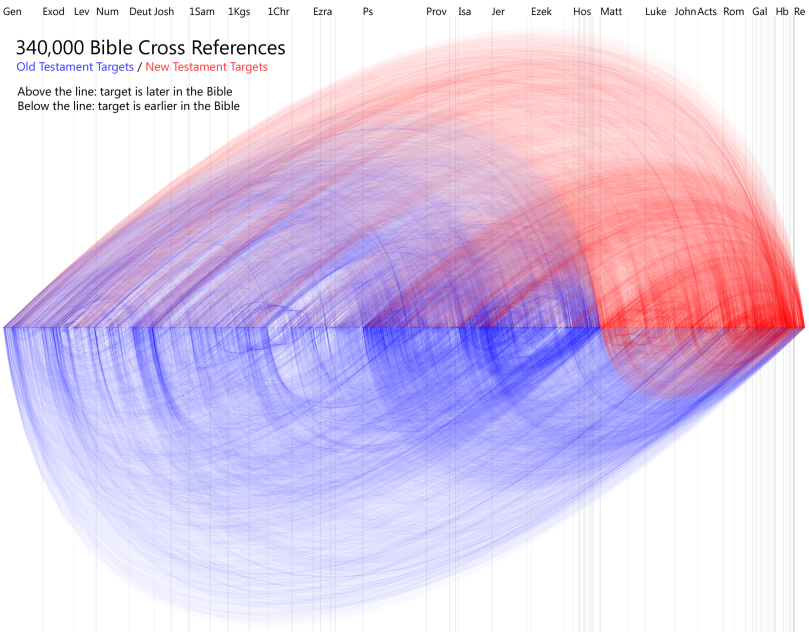


Figure 3: Bible cross reference - Openbible

3 Group members

This is a solo project. Fangzhe Li is the only member in this team.

4 Computing resources

Normal computers will be used as the computing resources.

5 Method

The sentence transformer sentence-transformers/all-MiniLM-L6-v2 is pre-trained on nreimers/MiniLM-L6-H384-uncased model and fine-tuned by 1 billion sentence pairs dataset. This embedding maps sentences and paragraphs to a 384 dimensional vector space and can be used in downstream tasks. The sentence vectors are used in sentence similarity tasks for our cases.

6 Datasets

The data source for this project is King James verison bible and some annotated cross reference from openbible.info. The King James version bible dataset contains four columns: Book, Chapter, Verse and Text.

- Book: The book name of the Bible verse
- Chapter: The chapter number of the Bible verse
- Verse: The verse number of the Bible verse
- Text: The text content of the Bible verse

	citation	book	chapter	verse	text
0	Genesis 1:1	Genesis	1	1	In the beginning God created the heaven and th...
1	Genesis 1:2	Genesis	1	2	And the earth was without form, and void; and ...
2	Genesis 1:3	Genesis	1	3	And God said, Let there be light: and there wa...
3	Genesis 1:4	Genesis	1	4	And God saw the light, that it was good: and G...
4	Genesis 1:5	Genesis	1	5	And God called the light Day, and the darkness...
...	...	...	...	...	...
31097	Revelation 22:17	Revelation	22	17	And the Spirit and the bride say, Come. And le...
31098	Revelation 22:18	Revelation	22	18	For I testify unto every man that heareth the ...
31099	Revelation 22:19	Revelation	22	19	And if any man shall take away from the words ...
31100	Revelation 22:20	Revelation	22	20	He which testifieth these things saith, Surely...
31101	Revelation 22:21	Revelation	22	21	The grace of our Lord Jesus Christ be with you...

Figure 4: Sample annotated bible data

- From Verse: From Verse in the format of Book.ChapterNo.VerseNo
- To Verse: To Verse in the format of Book.ChapterNo.VerseNo
- Votes: crowd sourcing data until Nov 1st, 2022, the vote of connection between this row

	From Verse	To Verse	Votes
0	Gen.1.1	John.1.1-John.1.3	217
1	Gen.1.1	Ps.104.30	33
2	Gen.1.1	Rev.10.6	37
3	Gen.1.1	Ps.121.2	36
4	Gen.1.1	1Cor.8.6	40
...	...	...	...
344784	Rev.22.21	2Cor.13.14	5
344785	Rev.22.21	2Thess.3.18	7
344786	Rev.22.21	Rom.16.20	5
344787	Rev.22.21	Rev.1.4	-1
344788	Rev.22.21	Rom.16.23	-1

Figure 5: Sample annotated bible cross reference data

7 Data Preprocessing

We firstly observed the the two datasets we have and found to get the sentence similarity of cross reference pairs, we firstly need to unify the citation to verse. So We decided to use the pattern of Bookfullname.ChapterNo.VerseNo to refer any verses in both dataset.

7.1 Choose data structure for Bible storage

We also need a few helper function for our bible dataset to have access to the text of a verse through bookname, chapterNo and VerseNo. Here are the code for the Class Bible:

Listing 1: Class Definition for BiBle

```
class Bible():
    def __init__(self):
        self.df = pd.read_csv("bible_data_set.csv")
        self.verse_group = self.df.groupby(['book','chapter','verse'
            ])
        self.chapter_group = self.df.groupby(['book','chapter'])
        self.df['to_reference'] = self.df.apply(lambda x: x.book + '.
            ' + str(x.chapter) + '.' + str(x.verse), axis=1)

    def get_verse(self, book, chapter, verse):
        return self.verse_group.get_group((book, chapter, verse)).
            text.iloc[0].strip('\n')

    def get_verse_num(self, book, chapter):
        return len(self.chapter_group.get_group((book, chapter)))
```

Here through Pandas groupby function, we can have a quick access to each verse through bookname, chapterNo and VerseNo. Even though it is not as fast as Python built-in Dictionary, it is good enough for our case and it is very easy to implement. Also we have another method called "get_verse_num" to get the number of verse in a certain chapter of a book.

7.2 Multiple verses referred in "To Verse" Column

And since in column 'To Verse' in cross reference dataset may have "-" between and may refer to a couple of different verses in a chapter or even different chapter. We need to make a decision for how to deal with this case. The first method I came out is to firstly fill the verses included by the "-" to make sure we can get all the paragraph related to the target verse instead of just the first and last verse. Here are the code for the Class Cross_reference and fill_list():

Listing 2: Class Definition for Cross_reference

```
class Cross_reference():
    def __init__(self):
        self.df = pd.read_csv("cross_reference.txt", delimiter="\t")
        self.bible = Bible()

    def fill_list(self, l):
        # one verse
        if len(l) == 1:
            b,c,v = l[0].split('.')
            book = self.get_whole_book_name(b)
            return [book + '.' + c + '.' + v]
        book = self.get_whole_book_name(l[0].split('.')[0])
        start_c = int(l[0].split('.')[1])
        end_c = int(l[1].split('.')[1])
        start_v = int(l[0].split('.')[2])
        end_v = int(l[1].split('.')[2])
        # print(book, ':', start_c, end_c, start_v, end_v)

        # multiple verses but in the same chapter
        if (start_c == end_c):
            new_l = []
            for i in range(start_v, end_v + 1):
                new_l.append(book + '.' + str(end_c) + '.' + str(i))
            return new_l
        # cross chapter case
        else:
            new_l = []
            # print(start_c, end_c + 1)
            for c in range(start_c, end_c + 1):
                # print(c)
                if c == start_c:
                    sv = start_v
                    ev = bible.get_verse_num(book, c)
                elif (c != start_c and c != end_c):
                    sv = 1
                    ev = bible.get_verse_num(book, c)
                else:
                    sv = 1
                    ev = end_v
```

```
                # print("sv: ", sv, "ev: ", ev)
                for v in range(sv, ev + 1):
                    new_l.append(book + '.' + str(c) + '.' + str(v))
        return new_l
```

After we got all the verses in a list, the next step we decided to treat the paragraph as a whole instead of using dataframe explode() method to create a 1-to-1 map for each verse in the paragraph because sometimes one verse itself doesn't make any sense but the whole paragraph is related to the original verse. Also, either we keep the votes for exploded rows or divide it evenly to each verse makes little sense. Also, we checked that for the related paragraph, most of them are less the 256 words, which means little information will be truncated by our sentence transformer model when trying to get the embedding based on the texts.

7.3 Book name Abbreviation and non-existing verses

In our Cross Reference dataset, the names of book appear in the way of abbreviation, so we need a dictionary to convert it to the full name of book. e.g. 'Gen' will be 'Genesis'. We also checked that if there are any verse that are not existing. So this part is generated by the ChatGPT developed by OpenAI, which is really fancy:

Listing 3: Converting abbreviation of book to full name

```
class Cross_reference():
    def __init__(self):
        self.df = pd.read_csv("cross_reference.txt", delimiter="\t")
        self.bible = Bible()
        self.not_existing_verses = {'2 John.1.14', '2 John.1.15', '3 
            John.1.15'}
        # honorably genreated by ChatGPT
        # Create a dictionary of book abbreviations and their full
            names
        self.bible_abb_books = {
            "Gen": "Genesis",
            "Exod": "Exodus",
            "Lev": "Leviticus",
            "Num": "Numbers",
            "Deut": "Deuteronomy",
            "Josh": "Joshua",
            "Judg": "Judges",
            "Ruth": "Ruth",
            "1Sam": "1 Samuel",
            "2Sam": "2 Samuel",
```

```
"1Kgs": "1 Kings",
"2Kgs": "2 Kings",
"1Chr": "1 Chronicles",
"2Chr": "2 Chronicles",
"Ezra": "Ezra",
"Neh": "Nehemiah",
"Esth": "Esther",
"Job": "Job",
"Ps": "Psalms",
"Prov": "Proverbs",
"Eccl": "Ecclesiastes",
"Song": "Song of Solomon",
"Isa": "Isaiah",
"Jer": "Jeremiah",
"Lam": "Lamentations",
"Ezek": "Ezekiel",
"Dan": "Daniel",
"Hos": "Hosea",
"Joel": "Joel",
"Amos": "Amos",
"Obad": "Obadiah",
"Jonah": "Jonah",
"Mic": "Micah",
"Nah": "Nahum",
"Hab": "Habakkuk",
"Zeph": "Zephaniah",
"Hag": "Haggai",
"Zech": "Zechariah",
"Mal": "Malachi",
"Matt": "Matthew",
"Mark": "Mark",
"Luke": "Luke",
"John": "John",
"Acts": "Acts",
"Rom": "Romans",
"1Cor": "1 Corinthians",
"2Cor": "2 Corinthians",
"Gal": "Galatians",
"Eph": "Ephesians",
```

```
            "Phil": "Philippians",
            "Col": "Colossians",
            "1Thess": "1 Thessalonians",
            "2Thess": "2 Thessalonians",
            "1Tim": "1 Timothy",
            "2Tim": "2 Timothy",
            "Titus": "Titus",
            "Phlm": "Philemon",
            "Heb": "Hebrews",
            "Jas": "James",
            "1Pet": "1 Peter",
            "2Pet": "2 Peter",
            "1John": "1 John",
            "2John": "2 John",
            "3John": "3 John",
            "Jude": "Jude",
            "Rev": "Revelation"
        }
    def get_whole_book_name(self, abb):
        return self.bible_abb_books.get(abb, "")
```

7.4 Preprocessing Pipeline

In this part, we connect the steps mentioned above, constructing data structure for Bible and Cross_reference, dealing with multiple verses referred in "To Verse" Column, mapping abbreviation book names to full book names and getting rid of non-existing verses.

Listing 4: preprocess pipeline

```
    def preprocess_to_verse(self):
        # deal with multiple verses case
        self.df['To Verse'] = self.df['To Verse'].apply(lambda x:x.
            split('-')).apply(self.fill_list)
        # Just ignore all the to verses that cross books because they
            are not that important in our analysis
        self.df = self.df[self.df['To Verse'].map(lambda d: len(d)) >
            0]
        self.df['Indicator'] = self.df['To Verse'].apply(lambda x:
            any(a in self.not_existing_verses for a in x))
        self.df = self.df[self.df['Indicator'] == False]
        self.df = self.df.dropna()
```

```
        self.df['tt'] = self.df['To_Verse'].apply(self.generate_text)
        self.df['tb'] = self.df['To_Verse'].apply(lambda x: x[0].
            split('.')[0])

    def preprocess_from_verse(self):
        self.df['From_Verse'] = self.df['From_Verse'].apply(lambda x
            :[self.get_whole_book_name(x.split(".")[0]) + '.' + x.
            split(".")[1] + '.' + x.split(".")[2]])
        self.df['Indicator'] = self.df['From_Verse'].apply(lambda x:
            any(a in self.not_existing_verses for a in x))
        self.df = self.df[self.df['Indicator'] == False]
        self.df['fb'] = self.df['From_Verse'].apply(lambda x: x[0].
            split('.')[0])
        self.df['ft'] = self.df['From_Verse'].apply(self.
            generate_text)
```

7.5 Calculate similarities between embeddings

After all the text are ready for our next steps. We called our sentence transformer model to calculate the cosine similarities between each cross reference pairs. The code is as follows:

Listing 5: sentence transformer model import

```
from transformers import AutoTokenizer, AutoModel
import torch
import torch.nn.functional as F
from sklearn.metrics.pairwise import cosine_similarity

#Mean Pooling - Take attention mask into account for correct
    averaging
def mean_pooling(model_output, attention_mask):
    token_embeddings = model_output[0] #First element of model_output
        contains all token embeddings
    input_mask_expanded = attention_mask.unsqueeze(-1).expand(
        token_embeddings.size()).float()
    return torch.sum(token_embeddings * input_mask_expanded, 1) /
        torch.clamp(input_mask_expanded.sum(1), min=1e-9)

# Load model from HuggingFace Hub
tokenizer = AutoTokenizer.from_pretrained('sentence-transformers/all-
    MiniLM-L6-v2')
```

```
model = AutoModel.from_pretrained('sentence-transformers/all-MiniLM-
    L6-v2')


def get_embedding_similarity(sentences, tokenizer, model):
    # Tokenize sentences
    encoded_input = tokenizer(sentences, padding=True, truncation=
        True, return_tensors='pt')

    # Compute token embeddings
    with torch.no_grad():
         model_output = model(**encoded_input)

    # Perform pooling
    sentence_embeddings = mean_pooling(model_output, encoded_input['
        attention_mask'])

    # Normalize embeddings
    sentence_embeddings = F.normalize(sentence_embeddings, p=2, dim
        =1)
    # print(sentence_embeddings)
    # print(type(sentence_embeddings))
    # print(type(sentence_embeddings[0]))
    # compute the cosine similarity
    cos = torch.nn.CosineSimilarity(dim=0)
    similarity = cos(sentence_embeddings[0], sentence_embeddings[1])

    return similarity

def get_similarity_wrapper(sentence1, sentence2, tokenizer, model,
    bible):
    return get_embedding_similarity([sentence1, sentence2], tokenizer
        , model)
```

Since the deep learning model is quiet big, even though this one is relatively a small one, it takes around 9 microseconds for each calculation and we have around 300, 000 pairs so we added tqdm for our Pandas part.

Listing 6: tqdm pandas usage

```
from tqdm._tqdm_notebook import tqdm_notebook
```

```
import pandas as pd
tqdm_notebook.pandas()
similarity = cr.df.progress_apply(lambda df:get_similarity_wrapper(df
    ['ft'], df['tt'], tokenizer, model, cr.bible), axis=1)
```

8 Analysis

We had all the data really after around 2-hour calculation.

	From Verse	To Verse	Votes	Indicator	tt	tb	fb	ft	similarity
0	[Genesis.1.1]	[John.1.1, John.1.2, John.1.3]	217	False	In the beginning was the Word, and the Word wa...	John	Genesis	In the beginning God created the heaven and th...	0.517112
1	[Genesis.1.1]	[Psalms.104.30]	33	False	Thou sendest forth thy spirit, they are create...	Psalms	Genesis	In the beginning God created the heaven and th...	0.490014
2	[Genesis.1.1]	[Revelation.10.6]	37	False	And sware by him that liveth for ever and ever...	Revelation	Genesis	In the beginning God created the heaven and th...	0.576545
3	[Genesis.1.1]	[Psalms.121.2]	36	False	My help cometh from the LORD, which made heave...	Psalms	Genesis	In the beginning God created the heaven and th...	0.719329
4	[Genesis.1.1]	[1 Corinthians.8.6]	40	False	But to us there is but one God, the Father, of...	1 Corinthians	Genesis	In the beginning God created the heaven and th...	0.356431
...	...	...	...	...	...	...	...	...	...
344784	[Revelation.22.21]	[2 Corinthians.13.14]	5	False	The grace of the Lord Jesus Christ, and the lo...	2 Corinthians	Revelation	The grace of our Lord Jesus Christ be with you...	0.818452
344785	[Revelation.22.21]	[2 Thessalonians.3.18]	7	False	The grace of our Lord Jesus Christ be with you...	2 Thessalonians	Revelation	The grace of our Lord Jesus Christ be with you...	1.000000
344786	[Revelation.22.21]	[Romans.16.20]	5	False	And the God of peace shall bruise Satan under ...	Romans	Revelation	The grace of our Lord Jesus Christ be with you...	0.536710
344787	[Revelation.22.21]	[Revelation.1.4]	-1	False	John to the seven churches which are in Asia: ...	Revelation	Revelation	The grace of our Lord Jesus Christ be with you...	0.398616
344788	[Revelation.22.21]	[Romans.16.23]	-1	False	Gaius mine host, and of the whole church, salu...	Romans	Revelation	The grace of our Lord Jesus Christ be with you...	0.274833

344770 rows × 9 columns

Figure 6: Cleaned cross reference data

The statistics of this dataset is below:

	Votes	similarity
count	344770.000000	344770.000000
mean	4.759147	0.437165
std	11.897619	0.137123
min	-32.000000	-0.094726
25%	2.000000	0.346397
50%	3.000000	0.432726
75%	4.000000	0.522045
max	1221.000000	1.000000

Figure 7: statistics of the cross reference data

It is seen that the mean of votes is around 5, since it is a crowd-sourcing data, we put 4 as the threshold to see if this cross reference connection is credible. Also we can see that the average similarity is around 0.44.

Here are a histogram to show the distribution of the votes.

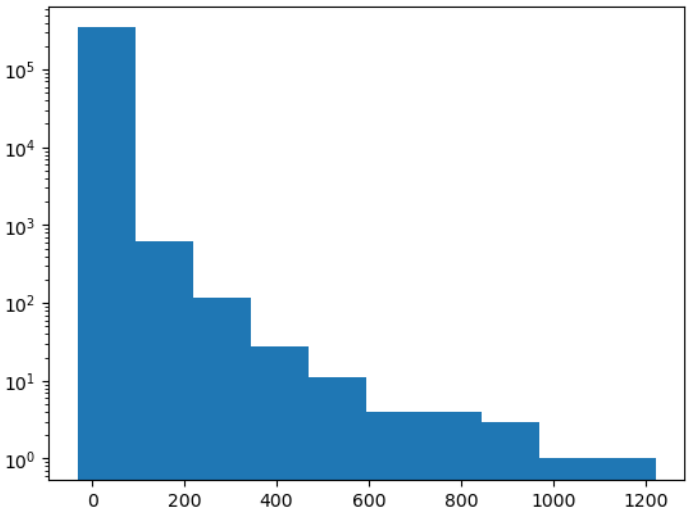


Figure 8: distribution of votes in cross reference data

8.1 Discovery 1: There are no statistically important relationship between votes and similarity of the text

Here is a plot between votes and similarity and it is obvious that there are no trend showing that there are any relationship between these two variables. And the correlation coefficient between Votes and similarity is 0.028, which shows these two variables are almost independent.

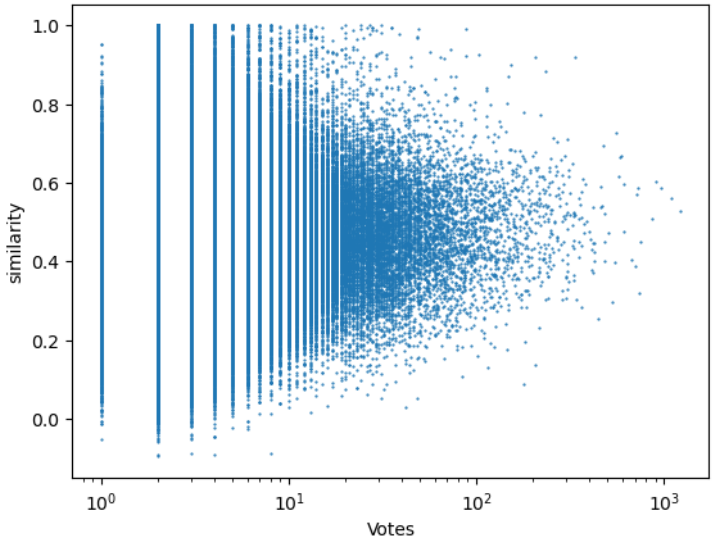


Figure 9: Similarity vs votes in cross reference data

One assumption to explain this can be that since this dataset is collected from viewers' feedback on these cross reference, it is natural there are more popular verses compared to others. To eliminate this effect, we normalized it to get a new column 'NVotes', which represents the weight of this "To verse" in all the "To verse", which is independent to whether a "From verse" is popular or not. So we draw the plot between NVotes and similarity and even though it is still the same but we can see it is not totally balanced plot but has more points on top-right part, which shows there are some relationship between normalized highly voted verses pairs and verses pairs with high similarity. And the correlation coefficient between NVotes and similarity increases to 0.196.

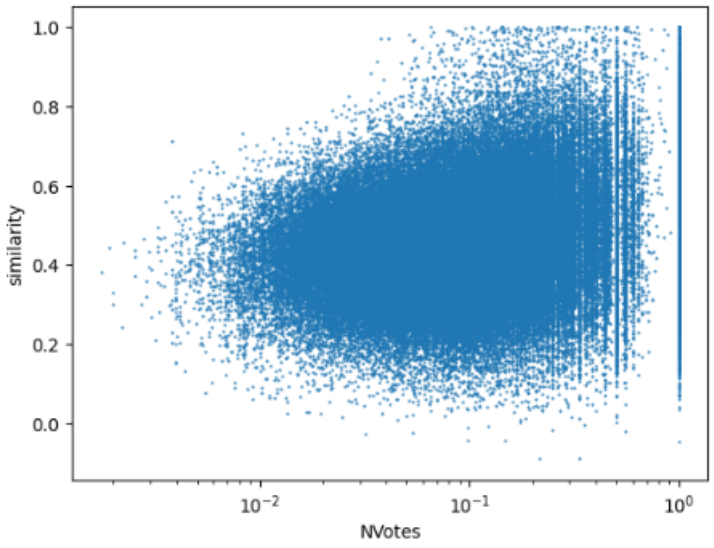


Figure 10: Similarity vs normalized votes in cross reference data

Also, due to the computing computing capacity, we didn't compute the similarities between all the verse pairs in the Bible and compare it to the Votes, or we may get a different perason coefficient.

8.2 Discovery 2: There are some verse pairs with relatively high similarity but negative votes

We did the query for negative votes pair with similarity large than 0.52, which is Q3 for similarity and found there are 107 cases. Even though in some cases, the bible uses the same word for a verse pair, they are not really cross references to each other like mentioned here "And the LORD spake unto Moses, saying" in Exodus.

```
[165]: crs[(crs['Votes'] < 0) & (crs['similarity'] >0.52)].sort_values(['similarity'])
```

[165]:

	From Verse	To Verse	Votes	Indicator	tt	tb	fb	ft	similarity
244690	['Mark.3.1']	['1 Kings.13.4']	-2	False	And it came to pass, when king Jeroboam heard ...	1 Kings	Mark	And he entered again into the synagogue; and t...	0.520243
269055	['John.10.34']	['John.8.17']	-5	False	It is also written in your law, that the testi...	John	John	Jesus answered them, Is it not written in your...	0.520700
269061	['John.10.34']	['1 Corinthians.14.21']	-11	False	In the law it is written, With men of other to...	1 Corinthians	John	Jesus answered them, Is it not written in your...	0.520810
165168	['Isaiah.40.4']	['Proverbs.2.15']	-1	False	Whose ways are crooked, and they froward in th...	Proverbs	Isaiah	Every valley shall be exalted, and every mount...	0.521997
264272	['John.3.17']	['Luke.9.56']	-2	False	For the Son of man is not come to destroy men'...	Luke	John	For God sent not his Son into the world to con...	0.522026
...	...	...	...	...	...	...	...	...	...
145	['Genesis.1.8']	['Genesis.1.19']	-1	False	And the evening and the morning were the fourt...	Genesis	Genesis	And God called the firmament Heaven. And the e...	0.724262
170506	['Isaiah.54.5']	['Isaiah.48.2']	-1	False	For they call themselves of the holy city, and...	Isaiah	Isaiah	For thy Maker is thine husband; the LORD of ho...	0.724979
147	['Genesis.1.8']	['Genesis.1.13']	-2	False	And the evening and the morning were the third...	Genesis	Genesis	And God called the firmament Heaven. And the e...	0.750118
16354	['Exodus.14.1']	['Exodus.12.1']	-1	False	And the LORD spake unto Moses and Aaron in the...	Exodus	Exodus	And the LORD spake unto Moses, saying,	0.874669
16353	['Exodus.14.1']	['Exodus.13.1']	-1	False	And the LORD spake unto Moses, saying,	Exodus	Exodus	And the LORD spake unto Moses, saying,	1.000000

107 rows × 9 columns

Figure 11: Some verse pairs with relatively high similarity but negative votes

8.3 Discovery 3: The Bible is consistent among different books even though they are written by different people in different periods

After filtering with verse pairs with more than 3 votes, we got 116, 843 pairs left. Among them there are 86,884 pairs are verse pair from different books. However, the average similarity for pairs drops from 0.46 to 0.44, which means the Bible is consistent among different books given they are written by different people in different period.

Also, if we use the mask that verse pairs come from different books, and we sort it by the similarity, we can find a lot of direct quote between books as shown below:

```
[29]: fcrs[fcrs.tb != fcrs.fb].sort_values('similarity').tail(20)
```

[29]:

	From Verse	To Verse	Votes	Indicator	tt	tb	fb	ft	similarity	NVotes
18223	Exodus.20.7	['Deuteronomy.5.11']	26	False	Thou shalt not take the name of the LORD thy G...	Deuteronomy	Exodus	Thou shalt not take the name of the LORD thy G...	0.995554	0.209677
39366	Deuteronomy.5.11	['Exodus.20.7']	28	False	Thou shalt not take the name of the LORD thy G...	Exodus	Deuteronomy	Thou shalt not take the name of the LORD thy G...	0.995554	0.736842
231071	Matthew.3.10	['Luke.3.9']	14	False	And now also the axe is laid unto the root of ...	Luke	Matthew	And now also the axe is laid unto the root of ...	0.996810	0.104478
252116	Luke.3.9	['Matthew.3.10']	5	False	And now also the axe is laid unto the root of ...	Matthew	Luke	And now also the axe is laid unto the root of ...	0.996810	0.151515
192403	Jeremiah.52.1	['2 Kings.24.18']	4	False	Zedekiah was twenty and one years old when he ...	2 Kings	Jeremiah	Zedekiah was one and twenty years old when he ...	0.997585	0.500000
244245	Mark.1.24	['Luke.4.34']	5	False	Saying, Let us alone; what have we to do with ...	Luke	Mark	Saying, Let us alone; what have we to do with ...	0.998869	0.185185
252702	Luke.4.34	['Mark.1.24']	7	False	Saying, Let us alone; what have we to do with ...	Mark	Luke	Saying, Let us alone; what have we to do with ...	0.998869	0.259259
96401	Nehemiah.7.65	['Ezra.2.63']	4	False	And the Tirshatha said unto them, that they sh...	Ezra	Nehemiah	And the Tirshatha said unto them, that they sh...	0.998995	0.181818
241604	Matthew.24.46	['Luke.12.43']	4	False	Blessed is that servant, whom his lord when he...	Luke	Matthew	Blessed is that servant, whom his lord when he...	1.000000	0.266667
261204	Luke.21.33	['Mark.13.31']	24	False	Heaven and earth shall pass away: but my words...	Mark	Luke	Heaven and earth shall pass away: but my words...	1.000000	0.169014
344766	Revelation.22.21	['2 Thessalonians.3.18']	7	False	The grace of our Lord Jesus Christ be with you...	2 Thessalonians	Revelation	The grace of our Lord Jesus Christ be with you...	1.000000	0.411765
257280	Luke.12.34	['Matthew.6.21']	11	False	For where your treasure is, there will your he...	Matthew	Luke	For where your treasure is, there will your he...	1.000000	0.366667
83442	1 Chronicles.16.22	['Psalms.105.15']	7	False	Saying, Touch not mine anointed, and do my pro...	Psalms	1 Chronicles	Saying, Touch not mine anointed, and do my pro...	1.000000	0.291667
249131	Mark.13.31	['Luke.21.33']	12	False	Heaven and earth shall pass away: but my words...	Luke	Mark	Heaven and earth shall pass away: but my words...	1.000000	0.122449
232977	Matthew.6.21	['Luke.12.34']	27	False	For where your treasure is, there will your he...	Luke	Matthew	For where your treasure is, there will your he...	1.000000	0.141361
18298	Exodus.20.13	['Deuteronomy.5.17']	20	False	Thou shalt not kill.	Deuteronomy	Exodus	Thou shalt not kill.	1.000000	0.180180
39407	Deuteronomy.5.17	['Exodus.20.13']	31	False	Thou shalt not kill.	Exodus	Deuteronomy	Thou shalt not kill.	1.000000	0.632653
164696	Isaiah.37.32	['2 Kings.19.31']	4	False	For out of Jerusalem shall go forth a remnant,...	2 Kings	Isaiah	For out of Jerusalem shall go forth a remnant,...	1.000000	0.500000
192608	Jeremiah.52.24	['2 Kings.25.18']	7	False	And the captain of the guard took Seraiah the ...	2 Kings	Jeremiah	And the captain of the guard took Seraiah the ...	1.000000	0.259259
249477	Mark.14.26	['Matthew.26.30']	5	False	And when they had sung an hymn, they went out ...	Matthew	Mark	And when they had sung an hymn, they went out ...	1.000000	0.555556

Figure 12: Some verse pairs that reflect direct quotations in Bible

So this embedding did a good job on finding direct quotation in the Bible.

8.4 Discovery 4: The allusions between different books are caught by similarities between the embeddings

To test how the embedding can help finding the allusions in bible. We picked up some allusion examples to test. Here is one example from John 3:14,

> 'And as Moses lifted up the serpent in the wilderness, even so must the Son of man be lifted up:'

which uses the allusion in Numbers 21:7-9,

> 'Therefore the people came to Moses, and said, We have sinned, for we have spoken against the LORD, and against thee; pray unto the LORD, that he take away the serpents from us. And Moses prayed for the people.And the LORD said unto Moses, Make thee a fiery serpent, and set it upon a pole: and it shall come to pass, that every one that is bitten, when he looketh upon it, shall live.And Moses made a serpent of brass, and put it upon a

pole, and it came to pass, that if a serpent had bitten any man, when he beheld the serpent of brass, he lived.'

Our embeddings did a good job in finding allusions in the Bible. By attention mechanism used in our sentence transformer, as shown in the search result, this related result ranked first among all the result.

'He removed the high places, and brake the images, and cut down the groves, and brake in pieces the brazen serpent that Moses had made: for unto those days the children of Israel did burn incense to it: and he called it Nehushtan.'

Also, we can see that the verses from 2 Kings.18.4 are also retrived and ranked high at the 4th place, which shows the attention mechanism succesfully catches the key word in the query verse.

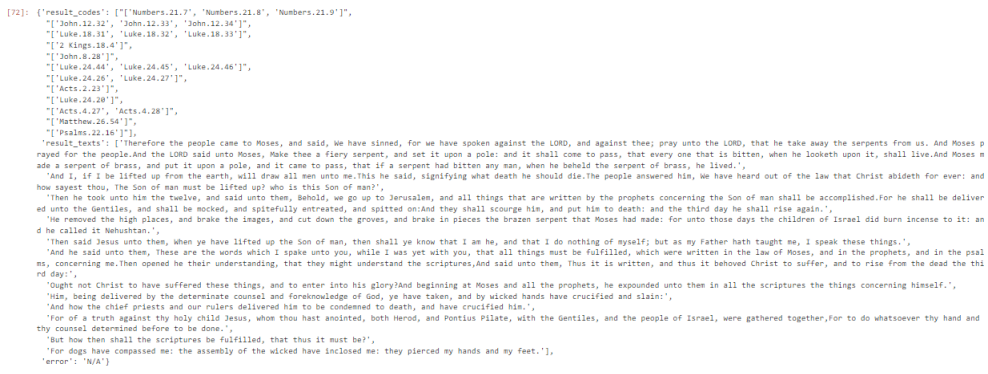


Figure 13: Results return by our search API through the similarity

To improve our search API, the next step can be combining other information retrival technique like BM25 to build a better search engine.

9 Conclusion

In this project, we focus on how to use the sentence transformer model to grab the semantic meaning of the bible and use this information on the analysis of cross reference. Here are a few findings: 1. There are no statistically important relationship between votes and similarity of the text even though we normalized the votes. 2. There are some verse pairs with relatively high similarity but negative votes. 3. The Bible is consistent among different books even though they are written by different people in different periods. 4. The allusions between different books are caught by similarities between the embeddings. The future work can be done is to calculate the all the 9.6 million verse pairs and use the correlation coefficent as an indicator of how well our sentence transformer understands the semantic of the Bible. Another thing is to combine other information retrival technique like BM25 to build a better search engine.