Capstone Final Report

Bible Themes

September 2017

Problem

The Bible has been the the best selling book of all time, having sold over [1]6 billion copies and counting. At least a portion of it has been translated in [2]3,223 languages all over the word.

As the journalist Daniel Radosh points out, [3]"The familiar observation that the Bible is the best-selling book of all time obscures a more startling fact: the Bible is the best selling book of the year, every year."

Even with this, Daniel Silliman's observes, [4]"There are persistent concerns that despite being so widely available, people struggle to understand the Bible."

The problem really is that it is hard to understand the Bible. And because of this, people rely on the expertise of other individuals to communicate its message. It is also being used as a guide on how to live life.

This paper will seek to determine the top themes in the Bible to drive conversations and communicate its message.

Client

It is mainly written for a Christian group who conducts regular study groups on the Bible on a weekly or bi-weekly basis. The results of the study will help to facilitate their discussions about it and its many lessons. It would help navigate these discussions and provide insights on what message they need to emphasize.

[1] http://brandongaille.com/27-good-bible-sales-statistics/

[2] http://www.wycliffe.net/statistics

[3] http://www.newyorker.com/magazine/2006/12/18/the-good-book-business

[4] https://www.washingtonpost.com/news/acts-of-faith/wp/2015/08/28/the-most-popular-bible-of-the-year-is-probably-not-what-you-think-it-is/?utm_term=.9d1a59ff9fd1

Data Description

With the many translations of the Bible, it can be classified into two main categories; Word-for-Word and Thought-for-Thought. For the purpose of this study, a Word-for-Word translation will be used. Of all the Word-for-Word translations, the King James Version (KJV) will be used. This is after taking into account Copyright laws and how it fares with other translations.

KJV is available to the public domain and that [5]" … the KJV remains the translation powerhouse." This is true even after looking at a current search in [6]google trends on the common bible translations available.

The KJV Bible will come from GitHub (https//github.com/scrollmapper/bibble_databases).

The data is using a verse ID system, using a unique key in the combination of Book, Chapter and Verse and it is available in CSV format. For example:

Genesis 1:1 (Genesis chapter 1, verse 1) = 01001001 (01 001 001)

After these fields, the corresponding verse follows. For example:

01 001 001 In the beginning God created the heaven and the earth

"01001001","1","1","1","In the beginning God created the heaven and the earth."

Also, it has to be noted that the data's language is in the Early Modern English. And as such, the words in the language may need to be updated to get a better understanding of the message.

[5] http://www.christianitytoday.com/news/2014/march/most-popular-and-fastest-growing-bible-translation-niv-kjv.html

[6] https://trends.google.com/trends

Data Wrangling

Here are the data wrangling steps done on the project data set to solve the problem:

A. All available data was read and compiled into one data frame by merging the data together on a common column such as “Title” and “GenreCode”. This includes the following csv files:
   - kjv.csv
   - key_english.csv
   - key_genre_english.csv
   *all data are available under the data folder

B. Main preparations were done by renaming the columns and adding a separate column strictly devoted to words.

C. A quick data analysis can is done by looking at the summary statistics of the data.

D. Cleaning the data for NLP was done by correcting the format required by spaCy and gensim to process the data.
   1. This includes turning the text into an entire list and changing it to unicode format for use in spaCy.
   2. For gensim use, words were lemmatize and punctuations and stops words were removed. Stopwords are words that do not add meaning to the topics. Also, a dictionary and corpus was created to feed on to the topic modeling algorithms.

Data Statistics

The following is a statistics summary of the data.

KJV Bible Statistics Summary

Books	66
Verses	31103
Words	789635
Unique Words	13719
Unique Words to Total Words Used in percentage	1.74%

The data is further divided into two parts. The Old and New Testaments.

Old Testament Statistics Summary

Books	39
Books as a percentage the Bible	59.09%
Verses	23145
Verses as a percentage the Bible	74.41%
Words	609253
Words as a percentage the Bible	77.16%
Unique Words	1157
Unique Words to Total Words Used in the Old Testament in percentage	1.89%

New Testament Statistics Summary

Books	27
Books as a percentage the Bible	40.91%
Verses	7958
Verses as a percentage the Bible	25.59%
Words	180382
Words as a percentage the Bible	22.84%
Unique Words	6555
Unique Words to Total Words Used in the Old Testament in percentage	3.63%

[7]Top 10 Words

	Words	Counts
1	lord	8004
2	say	5413
3	god	4715
4	man	4402
5	come	4147
6	son	3486
7	king	2873
8	day	2611
9	israel	2575
10	house	2160

[7] Words in this list have been processed for natural language processing and does not include stop words such as the, and, I, etc. For an extensive list of the top words, please view Capstone_Final_Code or Capstone Notebook-Statistics for the list of the top words including the stop words.

Interesting Insights and Findings

This section includes some interesting findings on the data.

There were more instances of the word “I” than “lord” in the Bible. If we take “lord” and “god” as one and the same and combine “I”, “me” and “my” together, the group of personal pronouns still have more instances as compared to the group of “lord” and “god”. These words do not appear on the “Top 20 Words” in this report because these words were considered as stop words. From this observation, we may say that the Bible is one’s personal story in relation to God rather than just God Himself.

Also, it was noticeable that there was a decrease in the usage of the word “lord” from the Old to the New Testament. An increase in the usage of the word “god” may explain this. It is possible though that other terms were used instead of God.

There were only 13,719 unique words that were used in the entire Bible. This made up about 1.74% of the total words used. Majority of the words, around 77%, are in the Old Testament and the remaining 23% can be found on the New Testament. This may mean that we can get a lot of insights can be found on that part of the Bible.

The fewer word totals in the New Testament as compared to the Old Testament can be explained by having majority of the books in the Old Testament. As a matter of fact, 39 out of the total 66 books are in the Old Testament.

Proposed Solution

In order to provide the top themes in the Bible, topic modeling needs to be applied on the text.

[8]Topic modeling is a simple way to analyze large volumes of unlabelled text. A "topic" consists of a cluster of words that frequently occur together. It is by using this contextual clues that we can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Basically, we will be using an unsupervised machine learning algorithm to form word clusters using the frequencies of words appearing next to each other. From here, we can rank a cluster's size using the total frequency of the top words in that particular cluster.

Python has several tools that can be used for Natural Language Processing(NLP).

[9]spaCy is one of them. [10]spaCy, according to its author is written to help get things done while nltk was created to support education. It is said to be much faster and more accurate compared to what is commonly used.

spaCy for this paper's purpose would be used to parse the text and to lemmatize the words. Lemmatization is a process of returning the base or dictionary form of a word. It also does a good job of lemmatizing words. For example, if you look at the word "dove", it can be lemmatize to the word "dive" or retain as "dove" depending on the contextual use of the word. Other nlp tools may simply return the word "dive".

[8] http://mallet.cs.umass.edu/topics.php

[9] https://spacy.io

[10] https://www.quora.com/What-are-the-advantages-of-Spacy-vs-NLTK

Another useful tool for the paper's purpose is the [11]gensim library which is considered to a great tool for topic modeling. It uses different topic modeling algorithms to determine clusters. In order to help determine which is the best algorithm to use for our data we can use the library's Coherence Model to determine coherence. The [12]Coherence Model takes a look at different factors such as segmentation, probability estimation, confirmation measure and aggregation in determining a coherence value. The higher the coherence value the better the algorithm's fit is for our data.

Essentially, the following is the exact steps we can use to arrive at a solution:

1. Clean the data and prepare it for spaCy to parse
2. Lemmatize the document
3. Create a dictionary of the unique words to be used by genesis
4. Create a corpus converted in the bag-of-words format
5. Use gensims' topic modeling algorithms and check for highest coherence model score
6. Fine-tune and customize the model to arrive at consistent results
7. Use the modifications and run the model
8. Determine the top topics using word frequencies of the top 20 words in the topic cluster
9. Label the topics
10. Generate word clouds for easier understanding of the data

Limitations

There are inherent limitations to this data:

1. The copyright laws limit us from exploring more recent translations which may or may not be accurate but can provide us with other data sets.
2. The recent translations may also work better with current natural language processing models though that hasn't been looked at yet. The data's English may have archaic text which might pose some problems with analysis.
3. As the data is a translation, some part of its meaning or message may be lost.

[11] https://radimrehurek.com/gensim/

[12] https://rare-technologies.com/what-is-topic-coherence/

4. The data is only limited to words and meanings coming from sentence structures and combination of certain words would not be extracted or used for this study.

Results

The following are the top themes based on the words' frequencies:

1. The Lord God
2. God's Grace despite Man's Choice to Sin
3. Satan and his Destructive Behavior
4. Continuous Denial and Self-Sacrifice
5. Appointed Messengers and their Mission
6. Worship of False Gods
7. Humble Service
8. Faith in God
9. Man's Rebellion Against God
10. Sin and Overcoming It

These themes were arrived at after searching for the word's meanings in [13]google and [14]bible study tools. There are also instances where name meanings or etymologies were used to get a better idea about a topic. When available, archaic meanings of the words were prioritized in the study to determine an appropriate label.

[13] www.google.com

[14] http://www.biblestudytools.com

Recommendations

While it is not surprising that the top theme in the Bible is about God, it is interesting that based on the measures implemented in these study false gods and sin are among the top themes.

Here are some recommendations for the use of these results:

- Use these themes as a reminder of which messages to drive on in their discussions.
- Tweak discussions about modern day topics in a way that is coherent with these themes. For example, if one commits a fault which burdens their consciousness remind them that God's grace will more than compensate to heal them.
- Use themes to prioritize which ones to focus on and give more time in laying out a plan of discussion.
- The results can also be in presented to other groups as a guide.

Suggestions for Further Research

In order to improve the results of the study some changes and additional steps may be used.

1. Use ESV or any other bible which uses Modern English, though concerns about the copyrights for the materials need to be taken consideration of before moving forward with such a study.
2. Phrase Modeling can also be applied on the data before processing it using the topic modeling algorithms.
3. Run the model with different parameters to get a better coherence score and label the topics from those results.
4. Use the labeled topics to analyze texts and describe its contents