Luis Ferrufino

CS 484-002

Final Project

5/8/20

Project Title:

Topic Analysis on Portions of the Bible

Team Member:

Luis Ferrufino (worked by myself)

For my project, I did topic analysis on the Bible. I created subdivisions of the Bible and applied LDA (Latent Dirichlet Allocation) to find the topics. I validated my parameters based on log likelihood and perplexity. I repeated this process for three granularities (for subdividing the Bible): verses, chapters, and books. The project title I used in my project declaration was “Clustering Verses from the Bible so as to Find Recurring Themes”--I changed it because 1) I am no longer doing clustering as LDA is a generative model, and 2) I am not just doing verses now.

My source code is in the folder sourceCode/ . There are four files: makeArray.py , clusterVerses.py , clusterChapters.py , clusterBooks.py . makeArray takes the dataset file t\_asv.txt, which contained the ASV (American Standard Version) and came from Kaggle ( <https://www.kaggle.com/oswinrh/bible> ; the uploader also included other versions of the Bible on that page), and turns it into three arrays, verses.npy , chapters.npy , and books.npy . As their names suggest, each entry in verses.npy was a verse from the Bible, each entry in chapters.npy was a chapter, and each entry in books.npy was a book. As the instructions specified not to include them in the zip file, I have not included them. ClusterVerses.py , clusterChapters.py , and clusterBooks.py all do the same thing, except with their own corresponding array (e.g. ClusterVerses.py uses verses.npy, thereby considering verses as the documents to be fed into the LDA model). In these python scripts, the appropriate array is loaded and then used to create a tf matrix; this tf matrix in turn is used to fit the LDA model at varying n\_components (which is the number of topics—this is a critical parameter); the values of n\_components with the best corresponding log likelihood is considered to be the best (the values compared go from 1 to 20, the best log likelihood is the one with the maximum value); the perplexities for each value of n\_components is also reported.

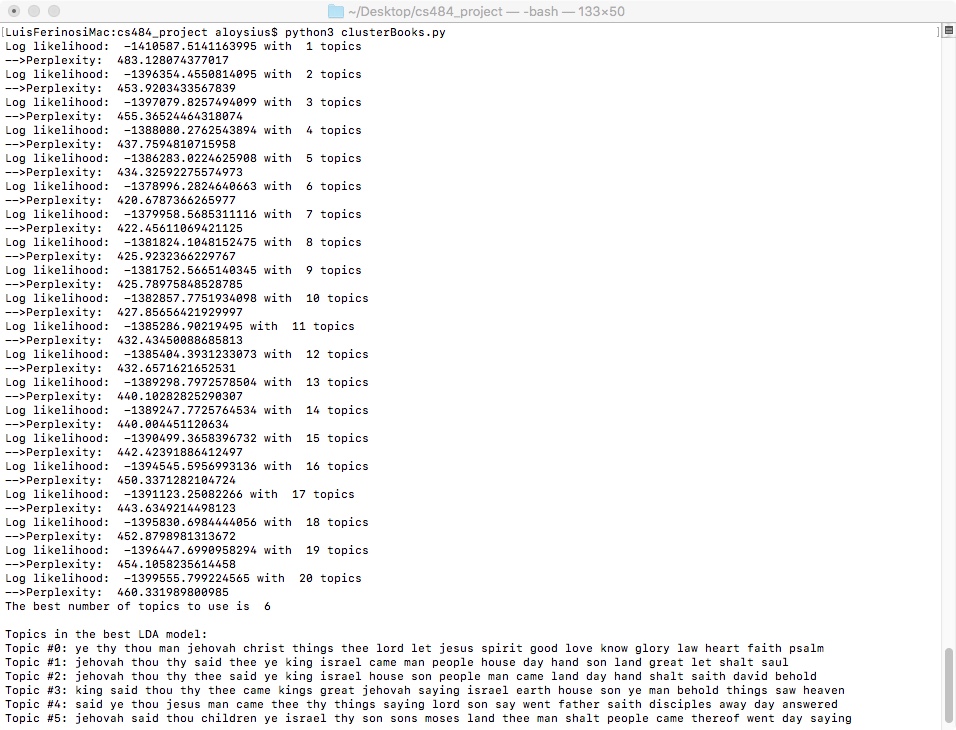
I chose the ASV because it is the most similar to the version I’m used to, the NRSV (New Revised Standard Version). I came to this conclusion after looking at the versions of the Bible included on the page linked above. However, I technically could have used a different one.

I used the numpy and scikit-learn modules to complete my project. Numpy was used to extract the verses from the csv file (numpy.genfromtxt), and store them in arrays. Scikit-learn provided LDA and tf\_vectorizer.

The running time for clusterVerses.py was around 25 minutes. The running time for clusterChapters.py was around 10 minutes. The running time for clusterBooks was around 3 minutes. So the overall running time was about 38 minutes.

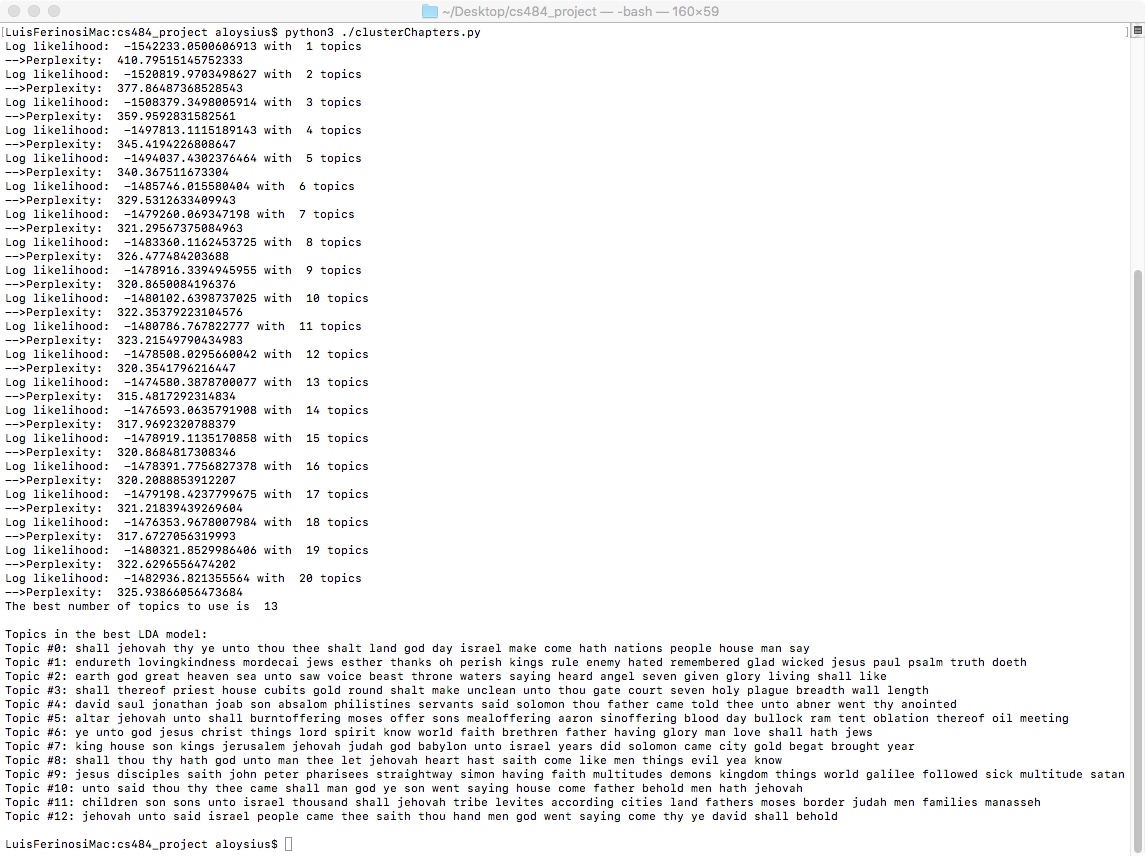
The output of the three python scripts have been included in the Report/ folder as versesOutput.txt , chaptersOutput.txt , booksOutput.txt .

The next three pages include images of the output for each script and (albeit subjective) interpretations as to what the topics were about.

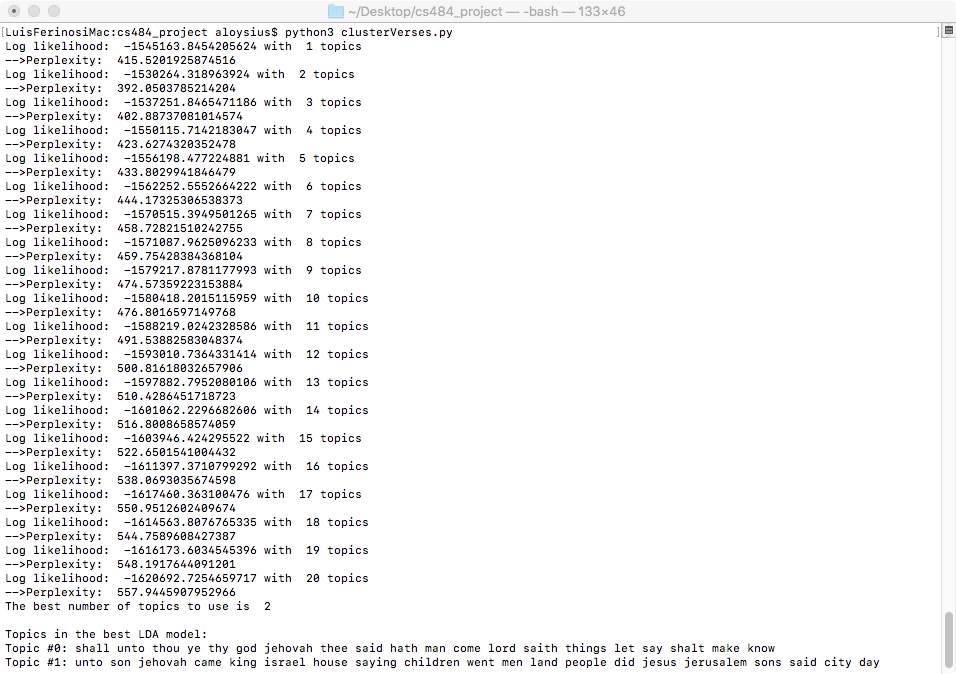


Above are the results from clusterBooks.py . The best model had a n\_components value of 6, a log likelihood of -1378996.2824640663 and a perplexity of 420.6787366265977. Here are some interpretations I made using the top words that were printed out for each topic:

|  |  |
| --- | --- |
| Topic # | Interpretation |
| 0 | Faith, Believing |
| 1 | Kingdom of Israël |
| 2 | House of David |
| 3 | Kingdom of Heaven |
| 4 | Apostles, Jesus’ ministry |
| 5 | Progeny of Moses, Children of Israël, Heritage |

 Above are the results from clusterChapters.py . The best model had a n\_components value of 13, a log likelihood of -1474580.3878700077 and a perplexity of 315.4817292314834 Here are some interpretations I made using the top words that were printed out for each topic:

|  |  |
| --- | --- |
| Topic # | Interpretation |
| 0 | Glory of Israël above other nations |
| 1 | Righteousness |
| 2 | Endtimes |
| 3 | Temple, Priesthood |
| 4 | Ancestry of the Jews |
| 5 | Sacrifice, Burnt offerings |
| 6 | Jesus’ Ministry |
| 7 | Solomon, Pagansim (?) |
| 8 | Righteousness |
| 9 | Jesus’ Passion |
| 10 | Relationship between the Father and The Son |
| 11 | The Twelve Tribes, the People of Israël |
| 12 | The Covenant |



Above are the results from clusterVerses.py . The best model had a n\_components value of 2, a log likelihood of -1530264.318963924 and a perplexity of 392.0503785214204. Here are some interpretations I made using the top words that were printed out for each topic:

|  |  |
| --- | --- |
| Topic # | Interpretation |
| 0 | God making declarations, Prophecy |
| 1 | (New) Jersualem, (New) Israël |