Joseph Lee
Software Design 2016 Spring
Mini Project 3: Text Mining Report/Reflection
02/24/2016

Project Overview:

The Douay Rheims Bible and the third edition of the Koran were collected from Gutenburg and compared based on word frequency analysis as well as through a simple implementation of TF-IDF analysis to work towards uncovering similarities and differences between the two religions.

Implementation:

The general implementation technique that I employed throughout the project was to start from my texts and the work my way up one line at a time, often running the code after writing each line or two to ensure that every line was behaving as expected. I prototyped and tested each function in the main body of the code first and then embedded it into a function when it was clear that it was working as intended. This development technique made for easy incremental development and allowed for extreme flexibility as far as what information or analysis techniques I wanted to employ. The project actually took shape as I went in a very exploratory manner. The two downsides of this approach are first, many of the functions are fairly redundant and the code would probably run more efficiently and would be more concise if several of the functions were generalized more and second, because the code was developed one line at a time instead of one module at a time, doctests are far less explicit.

The actual implementation was fairly straightforward: the Douay Rheims Bible and the Koran were each retrieved from their respective pickle files and returned as strings. A series of filters were then applied to these strings to make them more comparable and easier to work with. For instance, all of the punctuation was removed and all of the characters in the strings were made lowercase. A histogram was made from each text as well as a full list of words that appeared in that text. From there, TF-IDF analysis was performed for each and every word in the Douay Rheims Bible as well as each and every word in the Koran. For this TF-IDF analysis, the corpus was simply a compilation of all the words in both texts. This meant that words that appeared very frequently in both texts would return a low TF-IDF score. The histograms from each text were also sorted into a list of decreasing frequencies such that the most frequently used words of the two texts could be compared. One point where I had a huge number of options was when it came to actually implementing the TF-IDF analysis - for instance, should the corpus simply be a compilation of all the words in both, or should it be a compilation of words from a wider data set. Ultimately, I chose to implement the former due to its simplicity and relatively easy to interpret results. I do plan to pursue implementing the broader corpus and seeing how that data set sheds light on this topic. It seems likely that it would help to highlight differences between the two texts as well as differences between each text and writing outside of the religious sect instead of just the similarities.

Joseph Lee
Software Design 2016 Spring
Mini Project 3: Text Mining Report/Reflection
02/24/2016

Results:

Unfortunately, I ended up spending a huge amount of time in the implementation of actually getting the data and less time actually analyzing the data, however, there are still some pretty interesting results. The full tables of results can be found in the other files with the actual code. Some of the most interesting results have been detailed below:

The first text mining approach was to take a simple word frequency count of each of the texts. Excluding common words like ‘a’, ‘the’, ‘him’, ‘her’ etc., the 20 most frequently used words from each text are shown in the chart below:

Word Frequency Analysis

Most Frequently used Words in the Bible	Frequency of Appearance	Most Frequently used Words in the Koran	Frequency of Appearance
Lord	8373	God	3174
God	5754	Lord	916
man	3039	believe	453
Israel	2826	see	438
king	2690	people	423
son	2249	earth	405
children	2094	men	369
men	2086	sent	355
house	1968	truth	311
land	1790	fear	310
great	1407	signs	300
earth	1295	made	296
sons	1241	Muhammad	287
David	1164	gods	257
city	1118	man	240
name	1090	merciful	236
Jerusalem	1053	life	236
Jesus	1036	book	229
Christ	1025	koran	223
heart	977	apostle	223

Joseph Lee
Software Design 2016 Spring
Mini Project 3: Text Mining Report/Reflection
02/24/2016

The next approach taken to mine the texts for information was to perform TF-IDF analysis of all of the words in both texts with the corpus as a list of all the words contained in both texts. This TF-IDF analysis was performed twice, once for all the words in the Douay Rheims Bible and once for all the words in the Koran. One particularly interesting thing to look at with the TF-IDF analysis is to look specifically at all the words that appeared very frequently in either text. The analysis gives a quantitative measure of how different the importance of that word is to each text. Some of these findings are detailed in the table below:

TF-IDF Analysis Key Results

Word Analyzed with TF-IDF	TF-IDF Score for the Word in the Bible	TF-IDF Score for the Word in the Koran
Israel	0.0473069429048356	0.0473184044257788
king	0.0558387678160285	0.3058512651048367
son	0.0275489014975741	0.0275564769552983
children	0.1090750189737507	0.109096932071671
house	0.0943797828404444	0.0943898526562185
land	0.1535599942739774	0.1535693179692542
great	0.0849832758094456	0.0849949118121957
sons	0.0353678372540407	0.0353754127117649
David	0.1024158250772553	0.1024216891324954
city	0.0288504219371267	0.0288533431227464
name	0.0328210024747569	0.0328325137885954
Jerusalem	0.0713009670305494	0.0713180080134461
Jesus	0.0860309049686302	0.0860361533615973

To provide any sort of complete conclusions will require further analysis, and I think that the sample size for my TF-IDF function probably reduces the usability of the numbers it generates. However, it has been really interesting exploring this whole concept and it is also interesting to note that both the Koran and the Bible frequently spoke of God, which makes sense, but it is also interesting that the Bible seems to have a theme for words that speak of love/family, for instance, “children”, ”heart”, “house”, etc. The theme of the words frequently used in the Koran are less obvious and will require further analysis, but I found it really interesting that the word “fear” appears so frequently in the Koran. Sentiment analysis would be really useful in figuring out what the context around its common use is!

Joseph Lee
Software Design 2016 Spring
Mini Project 3: Text Mining Report/Reflection
02/24/2016

Reflection:

I think probably one the things that went best about this project is that I was super excited about it and interested in the results that I was working towards. As such, I ended up going far beyond what I had initially planned on doing (simply finding the most used words in a text) and in that process, I uncovered a lot of really interesting and useful things. Probably the best example of this is that in trying to figure out a good way to implement the TF-IDF analysis, I got to apply the information from the reading journal on class definitions to my project on something that would have been considerably more difficult/less efficient to implement without that knowledge. This project was also the first project that I really felt like I made something cool come together since there was no starter code for it and we had a lot of flexibility in choosing what we wanted to work on. The one area as far as my process was concerned that could have been better was the timeline of events. I should have gathered my texts from Gutenburg much sooner so that I would have had more time to implement my many ideas for improvement in the code itself.