**Davis Busteed – LING 360 – HW #6**

Proposal Summary

For my final project, I want to analyze the differences between the most popular translations of the Bible. I will write a series of Python scripts that will create a corpus of the different Bible versions, analyze these different versions for textual variations, and create visualizations that will show exactly how and where these translations are different.

Project Goals

My goal for this project is to analyze the variations among the several translations of the Bible. I want to gain a greater understanding of this ancient religious document and how it is read today. Thanks to my current religion class, I have become more interested in the Bible and its many translations. Lately, I have often wondered whether the different translations have a unitary distribution of discrepancies, or whether some translations are radically different from others in certain books and/or chapters. I am also curious how they vary. Are entire verses or chapters missing? Or is it just a matter of different word choice? Since the Church of Jesus Christ of Latter-day Saints uses the King James Version of the Bible, I am also interested in understanding how the KJV compares to other translations, as well as finding out which popular translation is most similar to the KJV.

This project will not only be interesting in terms of religious scholarship, but I think that it will provide beneficial opportunities for developing career skills. This summer I will be working as a data engineer, where I will be tasked in “pipelining” data from a team of data scientists to a team of application operators. I believe this experience of working with six fairly large corpora will help me feel more comfortable working with large sets of data.

Data

I plan on conducting my analysis on the following translations of the Bible:

* English Standard Version
* King James Version
* New American Standard
* New International Version
* New King James Version
* New Living Translation

As explained below, I will write a web scraper that will compile a corpus for each of the above translations to be used in my analysis.

Project Plan

I plan on beginning this project with data collection and corpus creation. To do this, I will write a web scraper that will grab the text of the different Bible translations. Although this might change upon further inspection of the HTML structure, I am currently planning on scraping the Bible texts from [www.biblegateway.com](http://www.biblegateway.com). This site contains all the Bible translations I am interested in, and also appears to have a straightforward way of accessing different translations, books, and chapters through the URL’s query string. The scraper will clean the text, and save it to a file to be used later in the analysis. The exact format of these corpus files will be decided later.

After the corpora has been created, I will work on writing the Python script that will analyze the texts and make note of all differences. I don’t have this step perfectly planned out (if I did, this project wouldn’t be that interesting), but I imagine that I will develop some method of comparing the texts on a verse-by-verse basis, and record the results as each text is compared. After all the translations have been compared, the script will create a set of summary statistics and figures and save them to a file.

I will then use the data outputted by the script to visualize the results, so that a summary of the differences between the many translations can be quickly understood. During this step, I will also measure the overall accuracy of my analysis. When I decide on the actual algorithm that will be used to identify differences, I will take a random sample of verses from the different translations and compute these textual differences by hand, and then compare it to the script’s performance and measures.

Required Skills

Even though the Bible is quite interesting to me, I chose this project because I believe that it will challenge my programming skills and give me an opportunity to learn better programming practices. I plan on completing this project with the Python scripting language, and although my proficiency with Python will increase, I also expect to become a better programmer altogether. That said, the following list outlines the general programming concepts required for the project as well as the corresponding Python modules / features used to accomplish it:

* Web scraping
  + Python modules BeautifulSoup, Requests
* Corpus creation
  + Python module OS
  + File I/O
* Algorithms and data structures
  + Lists, two-dimensional lists
  + Dictionaries within dictionaries
  + For loops and functions
* Batch execution
  + Python module Progressbar
  + Error handling with try/except
* Data visualization
  + Python modules Seaborn, MatPlotLib