Andy Carson

Data 602: Final Project Part 2

# Overview

I am interested in learning more about what Islam teaches and what exactly is in the text of the Quran, especially because of its current political relevance both in the United States and abroad. Consequently, to begin to do this, my semester project will focus on text mining the text of the Quran. I intend to scrape the entire text from available sources online (e.g., http://www.clearquran.com/), grouping the text by chapter and verse.  And then for each chapter, I will find the top words, keywords, and sentiment. I will also group words by categories and attempt to find meaningful clusters of words. After analysis, I will visualize the data by looking at trends going across the entire Quran to get a sense of what the Quran/Islam is about.

# Interface

The code will be run from a Jupyter notebook. The text of the analysis, code, and visualizations will be organized into a readable document of analysis. A PDF or HTML file of the output will also be included in the final submission.

# Visualization Incorporation

The generated visualizations resulting from the code and analysis will be displayed in the notebook as plots, as well as saved to a folder for viewing later.

# Structure

The code will have the following structure:

* Get text data of Quran
  + For each chapter, verse:
    - Construct website link
    - Get website HTML
    - Extract text of verse
      * Clean text
    - Break into individual words
    - Store words by chapter and verse in dataframe
* Transform/Cleanup
  + Cast words to uppercase
  + Remove stopwords
  + Other cleanup transformations
* Analysis
  + Top words
    - Frequency within whole Quran
    - Trend over whole Quran
    - Normalize by length of chapter
  + Keywords
    - TFIDF over whole Quran
    - Trend over whole Quran
  + Groupings of words into larger categories
    - Within whole Quran
  + Sentiment by chapter
    - Trend over whole Quran
  + Clusters of words
    - Within Whole Quran
* Conclusion
  + Summary of analysis and insights

Please give me a full (at least one full page) specification of your proposed project.  This is not a software engineering course, so there is no defined format in which it needs to be written, but you may use one (UML, Z notation, etc) if you wish.  The document should be specific enough that I can understand completely the work you will be doing.  You should discuss the specifics you will use to create the program or programs.  Talk about how the interface will work, how the visualization will be incorporated, and what you will do with the data.  You will be, for the most part, bound to this description; this will be the program you write.  It will be used as a component to grade your final submission (how well you achieved what you propose). Due 4/24/2017