CS410 Project Progress Report, 23-Nov-2018

Holy Book Similarity Analysis

Graham D Chester: grahamc2@illinois.edu (coordinator)

John F Moran: jfmoran2@illinois.edu

Topic

To provide a tool (Jupyter notebook) that will analyze the text of some of the worlds key religious texts (along with benchmark texts from a similar period) and provide a 2D map of the comparative differences and similarities, as identified by topic analysis. Whilst the main tool will be the Jupyter notebook, the output will be made available as a publicly accessible web page with D3 functionality such as zoom and scroll.

The technology review is an NLTK tutorial on how to start from scratch with text information processing.

Progress to-date

1) A good first draft of the Technology Review is complete and in process of being reviewed

2) The following texts have been identified, downloaded and pre-processed for analysis

1. Christianity - Bible New and Old Testaments - 2.2 billion followers
2. Islam - The Quran – 1.6 billion followers
3. Buddhism – Tipitaka (subset) – 380 million followers
4. Hinduism – Bhagavad Gita - 1 billion followers
5. Sikhism - Guru Granth Sahib – 23 million followers
6. Judaism - Torah (also rest of Old Testament) – 14 million followers

3) A draft Jupyter notebook has been developed which performs the data cleaning on the above texts, word count vectorization, topic extraction with Latent Dirichlet Allocation (LDA), and 2-D topic mapping with t-distributed stochastic neighbor embedding (t-SNE). And lastly matplotlib has been used to generate a draft similarity map.
4) An early draft D3 dynamic similarity map has been generated and published on a website as a proof of technology.

Tasks Remaining

1) Complete review and edits of Technology Review
2) Research and Experiment with t-SNE and MDS to identify best solution
3) Consider if TFIDF transformation may be a better approach than raw word counts.
2) Add top 10 most and least similar books table to analysis, as the similarity map does not easily allow this to be visualized.
3) Identify additional books with a translation done in a similar era to provide more context (perhaps Jane Austen novels, Bhagavad Gita etc. to add some further variety to the map.)
4) Finalize similarity map parameter tuning and upload as D3 web page to publicly available website
4) Develop Project Final Report

Outstanding Issues

There are no significant outstanding issues and the project is slightly ahead of schedule, however t-SNE and MDS, and word count vs TFIDF need further exploration to understand which are the best solutions.