CS410 Project Proposal, 03-Oct-2018

Holy Book Similarity Analysis

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

John F Moran: jfmoran2@illinois.edu

Topic

In an age of divisiveness how different are we fundamentally? In particular how different are the world's religions to which many belong, and how different are their key texts? For this project we will analyze some of the worlds key religious texts for similarity in terms of topic content and vocabulary.

Whilst there are a number of tools available on the internet that perform text analysis on the Bible and Quran in particular (many with an agenda to prove one point or another), there are none we could find that look at the overall similarities or differences between a broad range of Holy Books.

Functionality

We will take at least one Holy Book from each of the world's major religions and analyze their differences in both subject matter and vocabulary. Since clearly the translation of these works will have an impact on their similarity and vocabularies, we will also utilize some of the major works of Shakespeare, and some classic novels from the 19th century which were written approximately around the time of the translations, in an attempt to provide some context for the Holy Book relative similarities.

Each Holy Book will be decomposed into its key chapters, and text pre-processing such as stop word removal, stemming etc. performed. We will then project the resulting high-dimensional vectors onto 2 dimensions for plotting using either Multidimensional Scaling (MDS) and/or t-Distributed Stochastic Neighbor Embedding (t-SNE). Additionally we will chart the size of the vocabularies. The MDS/t-SNE projection will allow us to visualize how the chapters of each Holy Book cluster (or conversely don't cluster), and will show how the difference between each Holy Book compares to the differences between each chapter.

Audience

The audience is intended to be the general public with delivery via either charts on a static website, or a dynamic website as per the Scope/Extensions section below.

Datasets/Toolsets

We will obtain copies of most or all of the following (depending on availability and licensing) from sources on the internet such as https://www.holybooks.com. Given the variety of sources and formats, it is expected that the text preprocessing will be substantial.

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

We will use Python, and Jupyter notebooks with the Natural Language Tool Kit (NLTK), and scikit-learn libraries for text processing and modelling respectively.

Scope/Extensions

The core tool will be provided as static charts from a Jupyter Notebook, however if time is available we will look to implement this as a Flask-based website, potentially with the ability for a user to upload a book for comparison to the books described above.

Approximate Timeline

Week 6-10:	Source and cleanse data
Week 11-13:	Develop core functionality / Progress Report
Week 14-15:	Perform Analysis on Books and adjust functionality
Week 16:	Finalize Project and Submit