Document Similarity

Project: Comparing the Quran, bible, and Tanakh using Natural Language Processing, python, data mining, and machine learning
Project Manager: Farhana Akter, PhD, Certified IBM Data Science Professional
Research Team: AIsoftsolution

The Aim Of The Project:
1. Compare documents (Quran, King James Bible, and Tanakh) similarity using Python | NLP,
2. Compare the Quran (Muslim scripture), Tanakh (Jewish scripture), & King James Bible (Christian scripture) using machine learning text comparison (similarity),
3. find text similarity (Quran, Tanakh & King James Bible) using NLP and machine learning,
4. find text matching (Quran, bible, and Tanakh) with Deep Learning.

- **Step For Document Similarity:**

- **Data Reading :**
  We have used pandas and PyPDF2 for data reading. And save it into a list instance accordingly.

- **Data Cleaning :**
  Then we clean our data through the NLTK package and regular expression. We clean everything other than text from our data. In this process, we removed punctuation, special characters, and numeric data from our text.

- **Model Selection :**
  Then we studied different models to convert text into vectorization so we can find the distance between documents.

  Final Model:
  We finalized the "Fast Text model" which is an open-source, free, lightweight library that allows users to learn text representations and text classifiers. It works on standard, generic hardware. Models can later be reduced in size to even fit on mobile devices.

  In this model, we first generate the dictionary from our textual data that contains the weight according to every word. Then, we find the similarity of the whole document accordingly from where we got a similarity matrix against every book in our case. Then we find the cosine similarity between documents.

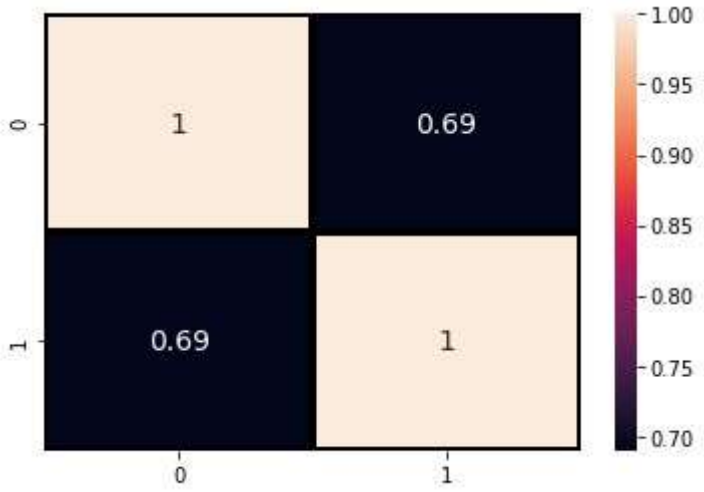
- **Result and Visualization:**

Then we plot cosine distance on heat map for distance visualization. Human error or technical error may affect the accuracy and adequacy of the results of this research.

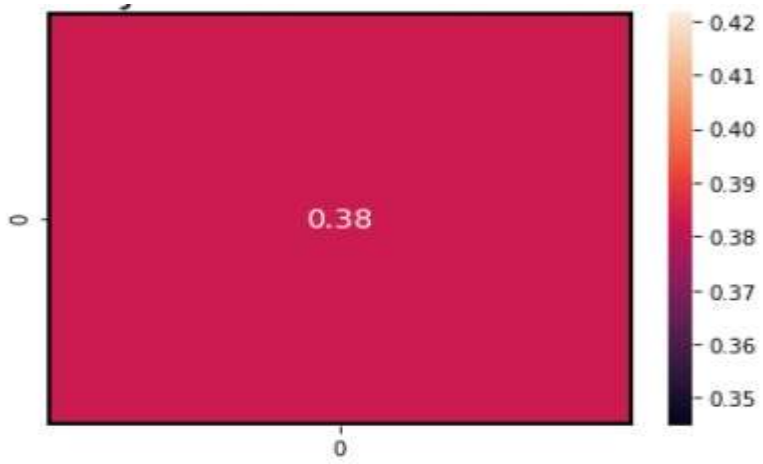
Supported tools: Jupyter notebook
Working language: Python
Machine learning model: Fast Text model (Deep-learning architect)
Distance finding: Cosine similarity

Similarity Result Between Quran, Tanakh, And Bible:

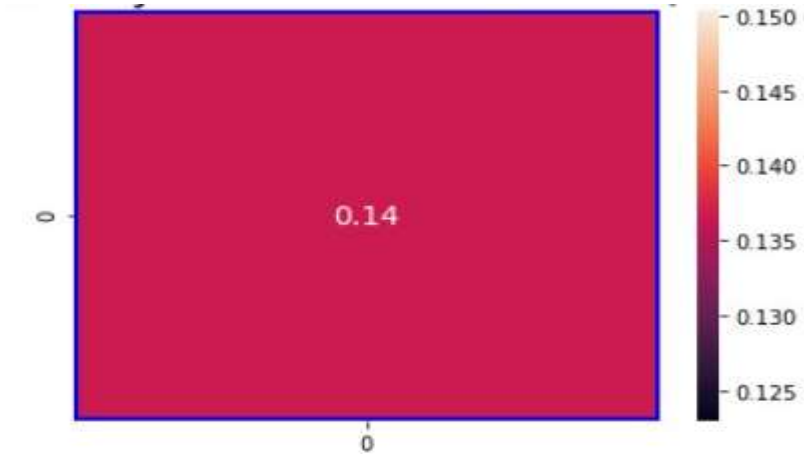
This heat map shows us that the Quran and King James bible have a 69% similarity between them.



The following heat map shows us that Tanakh and King James bible have a 38% similarity between them.



The following heat map shows us that Tanakh and Quran have a 14% similarity between them.



Conclusion:

The findings of this research demonstrate that the Quran and King James Bible (English version) are 69% similar that is the highest similarity among the three books.