Information Retrieval System On Semantic Meaning Of Bhagavad Gita

Project Submitted By

Abhishek Sahu (21111002)

Alok Trivedi (21111008)

Ashutosh More (21111017)

Shivam Kharwar (21111058)

Tanikella Sai Kiran (21111061)

Motivation And Problem Statement

- Bhagavad Gita is considered the quintessence of the philosophy of life.
- This invaluable treasure of knowledge is limited to those who can understand its complex language and deep essence.
- Our Goal is to design a Multi-lingual Information Retrieval System which can understand the query of user i.e. semantics and provide with an appropriate solution from the Bhagavat Gita.

Data Collection

The data is scraped from various popular Gita web pages by web scrapping

Then the data is stored in a python dictionary

The Key to the dictionary is the “shloka” and value is its corresponding “commentary”

These Files still have some inconsistencies like missing commentaries, duplicate commentaries

URL for the web-scrapping

- Hindi: https://www.holy-bhagavad-gita.org/index/hi
- English: https://vedabase.io/en/library/bg/
- Odia: https://www.holy-bhagavad-gita.org/index/or
- Telugu: https://www.holy-bhagavad-gita.org/index/te

Basic Models

For this task we have used

- Boolean Retrieval System
- Tf-Idf System
- BM25 System

We have performed some Preprocessing task on them

- Cleaning
- Tokenization
- Stopwords removal
- Construction Of Posting List
- Construction Of Term Frequency List

Basic Models

Boolean Retrieval System

Given a query, we will do basic preprocessing and we will find the score from Boolean incidence matrix.

TF-IDF

Here we will get the score using cosine similarity between query vector and document vector

BM25

We have calculated the score using BM25 formula and selected high score values

Latent Semantic Analysis

01

Trained LSI model using 30 topics.

02

Preprocessing Of Text

03

Computing Document-Term Matrix

04

Training using SVD

05

Given a query Q, we embed it into vector

06

Using Cosine Similarity, we find the similarity score

Machine Learning Embedding Models

- Word2Vec – CBow
- Word2Vec – SkipGram
- Glove
- FastText

Sentence Embedding

To find sentence embedding, we have used the following strategy	Tokenize a sentence into words
	For each word get the embedding
	Find mean of all the embeddings of words present in the sentence

Passage Embedding

- MAX POOLING
- MEAN POOLING
- LINE BY LINE POOLING
- WHOLE POOLING

Common Preprocessing

- Download the pre-trained embeddings for different models
- Basic Preprocessing on the commentaries
- Computation Of Max Pooling Matrix
- Computation Of Mean Pooling Matrix
- Computation Of Line-by-Line 3d Matrix
- Computation Of Whole Pooling Matrix
- Pickle Them.
- **NOTE:** We have used trained Word2Vec (Skipgram and CBow) on Bhagavad Gita Dataset

Implementation

1. Given a query Q, we have to do basic preprocessing and find its embedding
2. Different Strategy have used for retrieval

- Max Pooling with Cosine Similarity and Euclidean Distance
- Mean Pooling with Cosine Similarity and Euclidean Distance
- Line-by-Line Pooling with Cosine Similarity and Euclidean Distance
- Whole Pooling with Cosine Similarity and Euclidean Distance
- KMeans Clustering with Max Matrix, Mean Matrix and Whole Pooling Matrix

Sentence BERT

- Used Sentence Bert transformer which is trained on 100 languages including Hindi and English
- Preprocessing

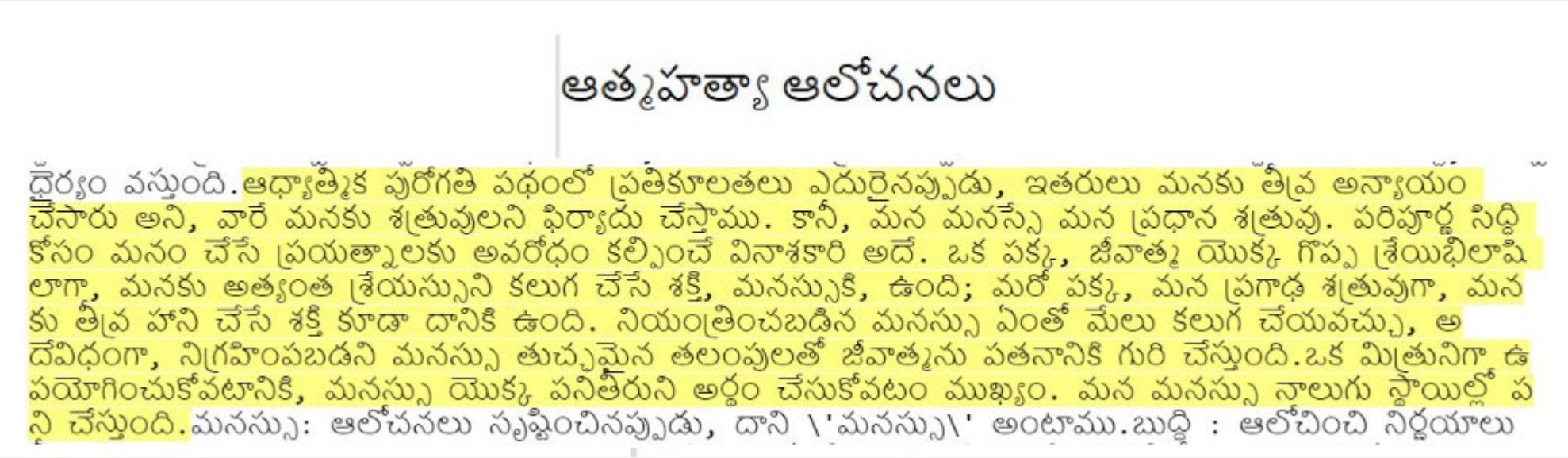
1. Cleaning Commentaries
2. Construction Of Max Pooling Matrix
3. Construction Of Mean Pooling Matrix
4. Construction Of Line-by-Line Pooling 3D Matrix
5. Construction Of Whole Pooling Matrix

Sentence BERT

1. Given a query Q either in Hindi or English, we have to do basic preprocessing and find its sentence embedding.
2. Different Strategy have used for retrieval

- Max Pooling with Cosine Similarity and Euclidean Distance
- Mean Pooling with Cosine Similarity and Euclidean Distance
- Line-by-Line Pooling with Cosine Similarity and Euclidean Distance
- Whole Pooling with Cosine Similarity and Euclidean Distance
- KMeans Clustering with Max Matrix, Mean Matrix and Whole Pooling Matrix

Outputs



ఆత్మహత్యా ఆలోచనలు

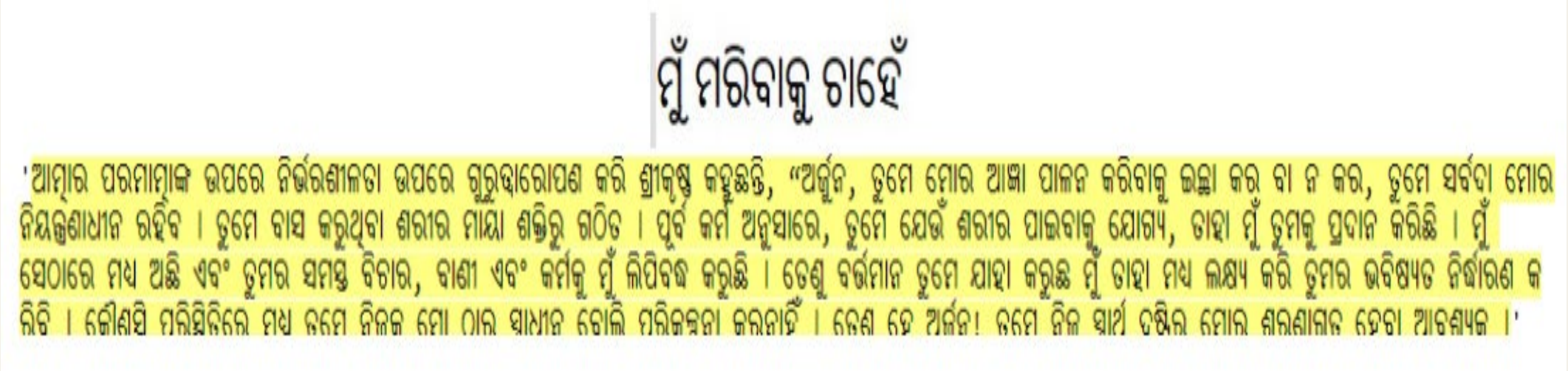
ధైర్యం వస్తుంది.ఆధ్యాత్మిక పురోగతి పథంలో ప్రతికూలతలు ఎదురైనప్పుడు, ఇతరులు మనకు తీవ్ర అన్యాయం చేసారు అని, వారే మనకు శత్రువులని ఫిర్యాదు చేస్తాము. కానీ, మన మనస్సే మన ప్రధాన శత్రువు. పరిపూర్ణ సిద్ధి కోసం మనం చేసే ప్రయత్నాలకు అవరోధం కల్పించే వినాశకారి అదే. ఒక పక్క, జీవాత్మ యొక్క గొప్ప శ్రేయిభిలాషి లాగా, మనకు అత్యంత శ్రేయస్సుని కలుగ చేసే శక్తి, మనస్సుకి, ఉంది; మరో పక్క, మన ప్రగాఢ శత్రువుగా, మన కు తీవ్ర హాని చేసే శక్తి కూడా దానికి ఉంది. నియంత్రించబడిన మనస్సు ఏంతో మేలు కలుగ చేయవచ్చు, అ దేవిధంగా, నిగ్రహింపబడని మనస్సు తుచ్ఛమైన తలంపులతో జీవాత్మను పతనానికి గురి చేస్తుంది.ఒక మిత్రునిగా ఉ పయోగించుకోవటానికి, మనస్సు యొక్క పనితీరుని అర్థం చేసుకోవటం ముఖ్యం. మన మనస్సు నాలుగు స్థాయిల్లో ప ని చేస్తుంది.మనస్సు: ఆలోచనలు సృష్టించినప్పుడు, దాని \'మనస్సు\' అంటాము.బుద్ధి : ఆలోచించి నిర్ణయాలు

Outputs

मैं मरना चाहता हूँ

'आत्मा की भगवान पर निर्भरता पर बल देते हुए श्रीकृष्ण कहते हैं-"अर्जुन! भले ही तुम मेरी आज्ञा का पालन करो या न करो, तुम्हारी स्थिति सदैव मेरे प्रभुत्व में रहेगी। जिस शरीर में तुम रहते हो वह यंत्र मेरी माया शक्ति से निर्मित है। तुम्हारे पूर्व जन्मों के अनुसार मैंने तुम्हारी पात्रता के अनुसार तुम्हें शरीर प्रदान किया है। मैं इसमें स्थित रहता हूँ और तुम्हारे विचारों, शब्दों, और कर्मों का लेखा-जोखा रखता हूँ। इस प्रकार से वर्तमान में जो कर्म तुम करते हो उसका आंकलन करते हुए मैं तुम्हारे भविष्य का निर्णय करता हूँ। यह मत सोचो कि तुम मुझसे किसी भी प्रकार से स्वतंत्र हो। इसलिए अर्जुन तुम्हारे हित में यही है कि तुम मेरी शरण ग्रहण करो।'

Outputs



ମୁଁ ମରିବାକୁ ଚାହେଁ

'ଆତ୍ମାର ପରମାତ୍ମାଙ୍କ ଉପରେ ନିର୍ଭରଶୀଳତା ଉପରେ ଗୁରୁତ୍ୱାରୋପଣ କରି ଶ୍ରୀକୃଷ୍ଣ କହୁଛନ୍ତି, "ଅର୍ଜୁନ, ତୁମେ ମୋର ଆଜ୍ଞା ପାଳନ କରିବାକୁ ଇଚ୍ଛା କର ବା ନ କର, ତୁମେ ସର୍ବଦା ମୋର ନିୟନ୍ତ୍ରଣାଧୀନ ରହିବ । ତୁମେ ବାସ କରୁଥିବା ଶରୀର ମାୟା ଶକ୍ତିରୁ ଗଠିତ । ପୂର୍ବ କର୍ମ ଅନୁସାରେ, ତୁମେ ଯେଉଁ ଶରୀର ପାଇବାକୁ ଯୋଗ୍ୟ, ତାହା ମୁଁ ତୁମକୁ ପ୍ରଦାନ କରିଛି । ମୁଁ ସେଠାରେ ମଧ୍ୟ ଅଛି ଏବଂ ତୁମର ସମସ୍ତ ବିଚାର, ବାଣୀ ଏବଂ କର୍ମକୁ ମୁଁ ଲିପିବଦ୍ଧ କରୁଛି । ତେଣୁ ବର୍ତ୍ତମାନ ତୁମେ ଯାହା କରୁଛ ମୁଁ ତାହା ମଧ୍ୟ ଲକ୍ଷ୍ୟ କରି ତୁମର ଭବିଷ୍ୟତ ନିର୍ଦ୍ଧାରଣ କ ରିବି । କୌଣସି ପରିସ୍ଥିତିରେ ମଧ୍ୟ ତୁମେ ନିଜକ ମୋ ଠାରୁ ସ୍ୱାଧୀନ ବୋଲି ପରିକଳ୍ପନା କରନାହିଁ । ତେଣୁ ହେ ଅର୍ଜୁନ! ତୁମେ ନିଜ ସ୍ୱାର୍ଥ ଦୃଷ୍ଟିରୁ ମୋର ଶରଣାଗତ ହେବା ଆବଶ୍ୟକ ।'

Outputs

Query=”Importance of Bhagavad Gita”

'The Lord was acting as the spiritual master of Arjuna. Therefore it was His duty to inquire from Arjuna whether he understood the whole Bhagavad-gītā in its proper perspective. If not, the Lord w as ready to re-explain any point, or the whole Bhagavad-gītā if so required. Actually, anyone who hears Bhagavad-gītā from a bona fide spiritual master like Kṛṣṇa or His representative will find t hat all his ignorance is dispelled. Bhagavad-gītā is not an ordinary book written by a poet or fic tion writer; it is spoken by the Supreme Personality of Godhead. Any person fortunate enough to he ar these teachings from Kṛṣṇa or from His bona fide spiritual representative is sure to become a l iberated person and get out of the darkness of ignorance.'

Evaluation

Created Ground-Truth values for a set of 30 Queries

Each Query contains related chapter number and verse number pairs

Around 10 outputs for each query

We have found the Ground-Truth values by verifying outputs of different model

Evaluation Metrics: Mean Average Precision (mAP)

Findings

Boolean Retrieval System and TF-IDF are good models for exact keyword searching for all four languages. However, they are not efficient for semantic searching.

BM25 and LSI model work well for all four languages; also, it is very lightweight and fast compared to other models.

BERT models are performing very well for semantic and cross-lingual searching tasks for Hindi and English. It works well with Mean Pooling and Cosine Similarity metrics.

For all other models, Clustering with Max pooling performs well. Also, Mean Pooling with cosine similarity shows high values in the evaluations.

Future Work

Including better commentaries for different languages

Improving word embeddings which are highly specific for Bhagavad Gita

Construction of Knowledge graph, which helps to find a better result

Text Summarisation features which will help to find a summary of a relevant commentary

Extending this work for more languages

Use of State-Of-Art model like BERT, which can capture more context

Conclusion

- The Multi-Lingual Information Retrieval System used for the Semantic meaning of the Bhagavad Gita provided a diverse range of results based on the Language and Model.
- Sentence BERT shows good results when it comes to the semantic
- LSA and BM25 are light-weight, fast and shows some good results
- Mean pooling with cosine similarity outperform other similarity measures
- KMeans clustering with Max Pooling shows good results

Contributions

Abhishek

- LSA (H,O) ,Glove (H,O) ,Word2Vec (E,O,H) , SentBert (H, E) , FastText (O) ,Basic (H, O)

Alok

- LSA (E,H) , Glove (E) , Word2Vec(H) , FastText (E) ,Basic (H)

Ashitosh

- LSA (E,H) , Glove (H) , Word2Vec (E, H) ,FastText (H,E), Basic (H, E)

Shivam

- LSA (E,H) , Glove (E,H) , Word2Vec (E, H) ,FastText (H,E), Basic (H, E)

Sai Kiran

LSA (E,T) , Glove (T) , Word2Vec (T) ,FastText (T), SentBert (H, E), Basic (T, E)


THANK YOU