Cross-Lingual Information Retrieval System of Shrimad Bhagavad Gita

Authors:

Manik Arora (2020519), Amit Malik (2020493), Talhah Samar (2020343), Ujjwal Goel (2020545), Shivanshu Pandey (2020333), Praveen Singh Samota (2020104)

Problem Formulation:

Cross-lingual information retrieval (CLIR) refers to the process of finding relevant documents in a different language than the user's query. Although recent CLIR methods that use neural networks have achieved significant progress, there are still several obstacles that need to be overcome. These include the lack of resources for low-resource languages, developing effective translation models for languages with complex morphology and syntax, and establishing a standard evaluation framework for CLIR systems.

Methodology:

1. **Cross-lingual approaches**:CLIR requires matching information in the same representation space, even when the query and documents are in different languages. The main challenge in CLIR is to match terms that describe the same or similar meaning across different languages. Typically, machine translation is used as a means of mapping between different language representations.

   In CLIR this translation process can be done in many ways:
   - **Document Translation**: Map the document representation into the query representation space.
   - **Query translation**: Map the query representation into the document representation space
   - **Using common language** : Documents and query are both translated into some common language.

For this project, our primary approach will be to use the common language, which is a combination of the first two cases. This method is quite intricate and will require numerous translations to create an efficient model.

2. **Dataset preparation:**

The initial stage of our project involved preparing the dataset. During this phase, we manually sorted shlokas, their true meanings, and their interpretations according to the author. We followed specific spacing conventions, including leaving a single line gap between shlokas and their meanings, a two-line gap between shlokas and before the next shloka, and no line gap

between the shloka's actual meaning and its interpretation by the author. These conventions were designed to facilitate processing using newline symbols. With these conventions in place, our model will be able to retrieve any shloka and its meaning and determine the shloka's starting and ending points in the document. Overall, we will produce three final documents (in English, Hindi, and Gujarati) during this manual processing stage, completing our dataset preparation process.
In the baseline results to test the prototype, we have created two documents(in English and Hindi). We will create document in Gujarati language in the final stage of the project.

3. **Data preprocessing:**

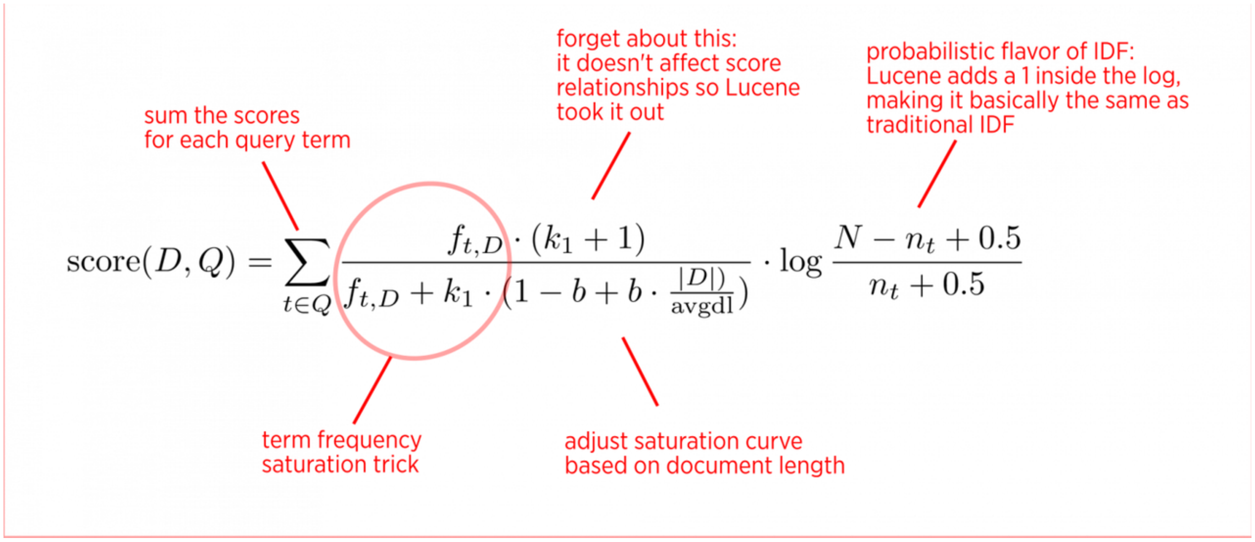
Our subsequent step involves data preprocessing techniques like tokenization, lemmatization, removing stop words.

- Removing Unnecessary characters: We remove unnecessary characters such as punctuation marks from the documents using string punctuation libraries.
- Tokenization: We then tokenize the text by splitting it by white space to obtain a set of tokens.
- Removal of stopwords: Since there is no direct library to remove stop words from Hindi language, we rely on a list of common stop words to eliminate them from our documents.
- Stemming/ Lemmatization: The most challenging part of the project was lemmatization, given that there is no direct library for Hindi word lemmatization. To overcome this challenge, we adopted various approaches to lemmatize the words.We have used help of shabdkosh dictionary for suffix and hindi stopwords list is large.

Model designing from scratch:

Our next step involves designing a model that scores documents and facilitates efficient retrieval. We will implement BM25 model for baseline results. For final stage of the project we will implement both TF-IDF and BM25 model and evaluate both model accuracy.
BM25 model:

BM25 is an enhanced version of the TF-IDF algorithm that takes into account not only term frequency and document frequency, but also factors in document length and term frequency saturation. While TF-IDF primarily rewards term frequency and penalizes document frequency, BM25 goes beyond this by incorporating additional variables to produce more accurate scoring of documents.



$$\mathrm{score}(D,Q) = \sum_{t \in Q} \frac{f_{t,D} \cdot (k_1 + 1)}{f_{t,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|)}{\mathrm{avgdl}})} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5}$$

sum the scores for each query term
forget about this: it doesn't affect score relationships so Lucene took it out
probabilistic flavor of IDF: Lucene adds a 1 inside the log, making it basically the same as traditional IDF
term frequency saturation trick
adjust saturation curve based on document length

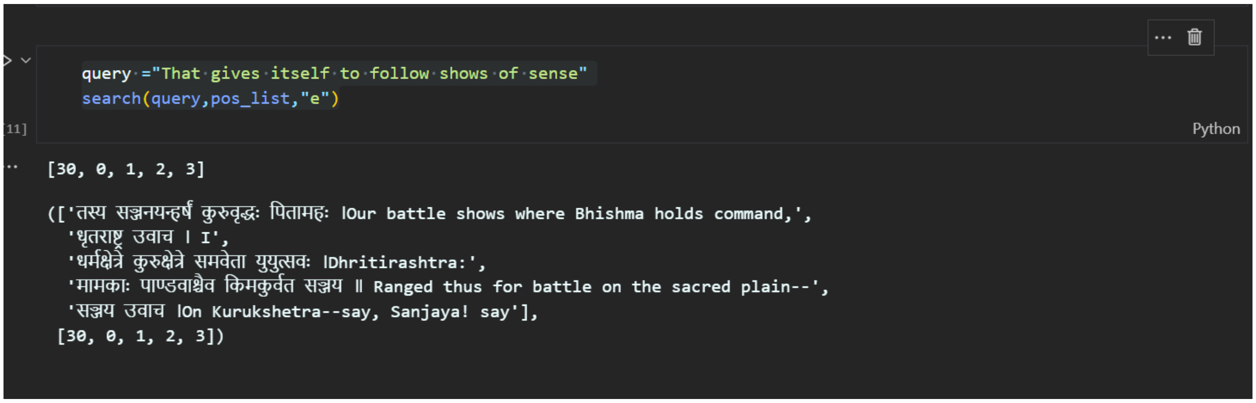
Source:https://kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm-25/

Literature Review:

Early approaches to CLIR used machine translation or bilingual dictionaries to translate queries into the language of the documents. However, these approaches suffered from low translation quality, especially for low-resource languages. Recent advances in CLIR have focused on neural network-based approaches that use deep learning techniques such as multilingual word embeddings or neural machine translation models to map queries and documents into a common space where similarity scores can be computed.

Evaluation And results:

We made **Two Model prototype** in **prototype 1** we have checked shlokas with only single user language (english) and applying bm25 model to rank documents and retrieve meaning of shlokas.



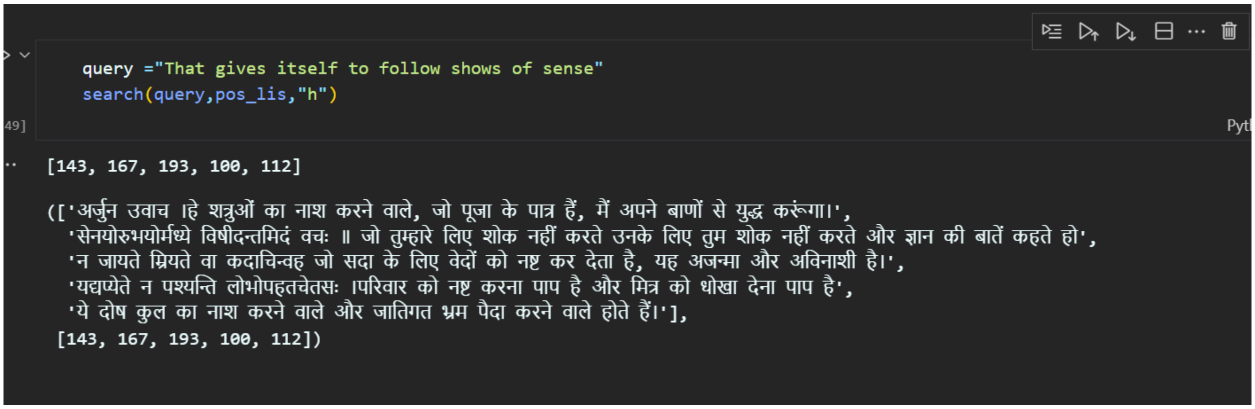
```
query ="That gives itself to follow shows of sense"
search(query,pos_list,"e")
```

```
[30, 0, 1, 2, 3]

(['तस्य सञ्जनयन्हर्षं कुरुवृद्धः पितामहः ।Our battle shows where Bhishma holds command,',
  'धृतराष्ट्र उवाच । I',
  'धर्मक्षेत्रे कुरुक्षेत्रे समवेता युयुत्सवः ।Dhritirashtra:',
  'मामकाः पाण्डवाश्चैव किमकुर्वत सञ्जय ॥ Ranged thus for battle on the sacred plain--',
  'सञ्जय उवाच ।On Kurukshetra--say, Sanjaya! say'],
 [30, 0, 1, 2, 3])
```

In prototype 2:

We have made for two user languages Hindi and english and retrieve documents according to user selected documents and use bm25 model on two languages and retrieve meaning of shlokas.



```
query ="That gives itself to follow shows of sense"
search(query,pos_lis,"h")
```

```
[143, 167, 193, 100, 112]

([ 'अर्जुन उवाच ।हे शत्रुओं का नाश करने वाले, जो पूजा के पात्र हैं, मैं अपने बाणों से युद्ध करूंगा।',
  'सेनयोरुभयोर्मध्ये विषीदन्तमिदं वचः ॥ जो तुम्हारे लिए शोक नहीं करते उनके लिए तुम शोक नहीं करते और ज्ञान की बातें कहते हो',
  'न जायते म्रियते वा कदाचिन्वह जो सदा के लिए वेदों को नष्ट कर देता है, यह अजन्मा और अविनाशी है।',
  'यद्यप्येते न पश्यन्ति लोभोपहतचेतसः ।परिवार को नष्ट करना पाप है और मित्र को धोखा देना पाप है',
  'ये दोष कुल का नाश करने वाले और जातिगत भ्रम पैदा करने वाले होते हैं।'],
 [143, 167, 193, 100, 112])
```

Results- we checked cross for more than 10 queries and we get Accuracy in Hindi language is more than in english language . In further days we will evaluate this basis upon precision, f-measure .

Conclusion:

Despite these advances, there are still challenges to overcome, including the lack of resources for low-resource languages, the need for effective translation models for languages with complex syntax and morphology, and the lack of a common evaluation framework. Future research should focus on developing effective CLIR systems for low-resource languages, improving the performance of CLIR for languages with complex syntax and morphology, and establishing a common evaluation framework for CLIR systems.