Cross-lingual IR for Chanakya Neeti

CS657A PROJECT REPORT
Group 11
Deepak Raj-21111024-deepakr21@iitk.ac.in
Dinkar Tewari-21111025-dinkart21@iitk.ac.in
Divyansh Bisht-21111027-dbisht21@iitk.ac.in
Rohit Kushwah-21111053-krohit21@iitk.ac.in
Vikas-21111067-vikas21@iitk.ac.in

Indian Institute of Technology Kanpur (IIT Kanpur)

April 27, 2022

Abstract

Chanakya Neeti was written by Chanakya, a highly respected Indian economist, philosopher, teacher, and royal advisor of ancient times. In his treatise, Chanakya presented 455 aphorisms or 'sutras' for living an ideal life. Around 216 of these 415 sutras cover the rules for running a kingdom. This project implements cross-lingual information retrieval mechanism to search for relevant documents related to Chanakya Neeti. A cross lingual search engine is built which allows user to enter query in any user selected Indian language(English, Hindi, Gujarati) and also to retrieve document in any user selected Indian language. The search engine searches for relevant content in the multilingual database of shlokas and returns optimal documents to the user.

1 Problem Statement

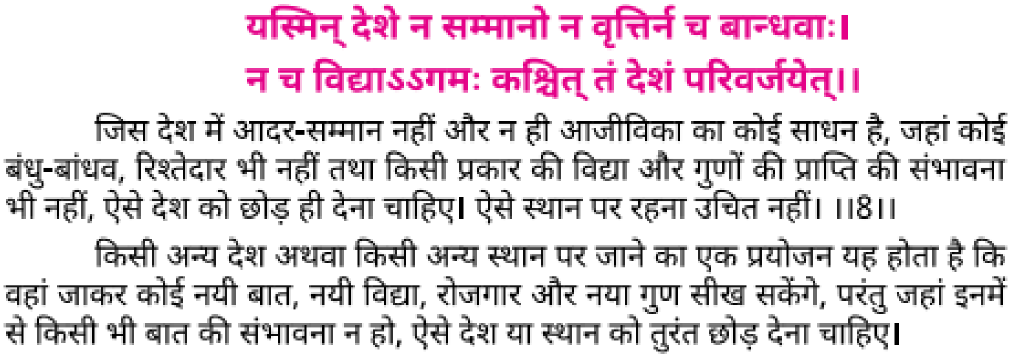
Chanakya Neeti is a very philosophical book that contains answers to almost all life problems. The shlokas and their meanings are present in different Indian languages. This leads to a slight difference in meanings of same shlokas. There are a lot of shlokas and it's tough for any user to get his required shloka and it's meaning. So, a search engine application is required to ease out this process. There must be a mechanism so that the user can search his query in his own mother tongue language and still get relevant shlokas and their meanings. Also, if the meanings that are retrieved are in a language that is not understandable by the user, then there must be an option so that the user can get output in his own mother tongue language. So, a basic cross lingual search engine is required to ease out the search process. This project works on this problem and attains required results.

2 Introduction

The basic aim of this project is to develop a cross lingual search engine for Chankya Neeti shlokas. The user must be able to search and receive information in the language he/she likes. Various IR algorithms are used in this project. The best one is selected at last and using it, we build a user-interface. Evaluation related to relevance is based on scoring as well as manual observation. The datasets used are of 3 languages mainly named English, Hindi and Gujarati. In the end the UI is designed using streamlit in a beautiful way that allows a lots of features to the user. Based on user's inputs, our model can adjust itself and generate the required outputs according to user's need.

3 Dataset

The datasets used in this project are three e-books of Chanakya Neeti in English, Hindi and Gujarati. The author of these books are B.K. Chaturvedi, Ashwani Parashar and Rajeshwar Mishra respectively. All these books are published under Diamond books publications. Our first challenge with this dataset was that the content of these e-books was in image format. So, we could not directly copy paste it's contents to form our required corpus. Hence, we used a software named Optical Character Recognition(OCR) that helped us to convert this image data to textual form. But, the OCR software was also not fully efficient. It failed to convert pages where the font in the image was not very clear. So, to handle those cases, we manually verified and cross checked all the data to ensure correctness of our dataset. After doing this process, we saved them into 3 files of text format. Then we manually removed some symbols, image data and other useless data from our dataset. Each document in our corpus has three main features: shlokas, their actual meaning and their interpretation according to the author.



यस्मिन् देशे न सम्मानो न वृत्तिर्न च बान्धवाः।
न च विद्याऽऽगमः कश्चित् तं देशं परिवर्जयेत्।।

जिस देश में आदर-सम्मान नहीं और न ही आजीविका का कोई साधन है, जहां कोई बंधु-बांधव, रिश्तेदार भी नहीं तथा किसी प्रकार की विद्या और गुणों की प्राप्ति की संभावना भी नहीं, ऐसे देश को छोड़ ही देना चाहिए। ऐसे स्थान पर रहना उचित नहीं। ।।8।।

किसी अन्य देश अथवा किसी अन्य स्थान पर जाने का एक प्रयोजन यह होता है कि वहां जाकर कोई नयी बात, नयी विद्या, रोजगार और नया गुण सीख सकेंगे, परंतु जहां इनमें से किसी भी बात की संभावना न हो, ऐसे देश या स्थान को तुरंत छोड़ देना चाहिए।

Figure 1: Data Structure

Whenever a query is fed to the search engine, various documents are returned according to our ranking algorithm. The content in each document are those three features as described above. After these processes, we moved ahead to designing our IR algorithms for smooth query processing.

4 Methodology

4.1 Cross-lingual approaches:

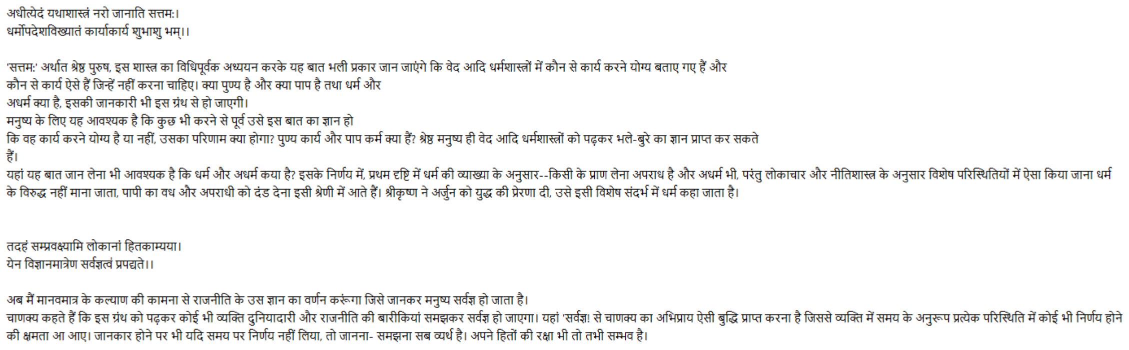
Cross Lingual Information Retrieval(CLIR) refers to the retrieval of documents that are in a language different from the one in which query is expressed. CLIR requires the ability to represent and match information in the same representation space even if the query and documents are in different languages. The fundamental problem in CLIR is to match terms in different languages that describe the same or similar meaning. The strategy of mapping between different language representations is usually machine translation. 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
- **Pivot language or Interlingua**: Documents and query are both translated into some common interlingua

We will mainly proceed with the Pivot language interlingua in this project. It's very complex as it's a mix of the first two cases. So, a lot of translations will be performed to build our effective model.

4.2 Dataset preparation:

Our first step towards our project was dataset preparation. In this phase, we manually seperated shlokas, their actual meaning and their interpretation according to the author. The spacing convention that we followed was that there must be a single line gap between the shlokas and their meanings. After shloka's meaning and before next shloka, there must be a gap of two lines. Between shloka's actual meaning and their interpretation according to the author, there must be no line gap. Just the interpretation will start from the immediate next line of shloka's meaning.



अधीत्येदं यथाशास्त्रं नरो जानाति सत्तमः।
धर्मोपदेशविख्यातं कार्याकार्य शुभाशु भम्।।

'सत्तमः' अर्थात श्रेष्ठ पुरुष, इस शास्त्र का विधिपूर्वक अध्ययन करके यह बात भली प्रकार जान जाएंगे कि वेद आदि धर्मशास्त्रों में कौन से कार्य करने योग्य बताए गए हैं और
कौन से कार्य ऐसे हैं जिन्हें नहीं करना चाहिए। क्या पुण्य है और क्या पाप है तथा धर्म और
अधर्म क्या है, इसकी जानकारी भी इस ग्रंथ से हो जाएगी।
मनुष्य के लिए यह आवश्यक है कि कुछ भी करने से पूर्व उसे इस बात का ज्ञान हो
कि वह कार्य करने योग्य है या नहीं, उसका परिणाम क्या होगा? पुण्य कार्य और पाप कर्म क्या हैं? श्रेष्ठ मनुष्य ही वेद आदि धर्मशास्त्रों को पढ़कर भले-बुरे का ज्ञान प्राप्त कर सकते
हैं।
यहां यह बात जान लेना भी आवश्यक है कि धर्म और अधर्म कया है? इसके निर्णय में, प्रथम दृष्टि में धर्म की व्याख्या के अनुसार--किसी के प्राण लेना अपराध है और अधर्म भी, परंतु लोकाचार और नीतिशास्त्र के अनुसार विशेष परिस्थितियों में ऐसा किया जाना धर्म
के विरुद्ध नहीं माना जाता, पापी का वध और अपराधी को दंड देना इसी श्रेणी में आते हैं। श्रीकृष्ण ने अर्जुन को युद्ध की प्रेरणा दी, उसे इसी विशेष संदर्भ में धर्म कहा जाता है।

तदहं सम्प्रवक्ष्यामि लोकानां हितकाम्यया।
येन विज्ञानमात्रेण सर्वज्ञत्वं प्रपद्यते।।

अब मैं मानवमात्र के कल्याण की कामना से राजनीति के उस ज्ञान का वर्णन करूंगा जिसे जानकर मनुष्य सर्वज्ञ हो जाता है।
चाणक्य कहते हैं कि इस ग्रंथ को पढ़कर कोई भी व्यक्ति दुनियादारी और राजनीति की बारीकियां समझकर सर्वज्ञ हो जाएगा। यहां 'सर्वज्ञ! से चाणक्य का अभिप्राय ऐसी बुद्धि प्राप्त करना है जिससे व्यक्ति में समय के अनुरूप प्रत्येक परिस्थिति में कोई भी निर्णय होने
की क्षमता आ आए। जानकार होने पर भी यदि समय पर निर्णय नहीं लिया, तो जानना- समझना सब व्यर्थ है। अपने हितों की रक्षा भी तो तभी सम्भव है।

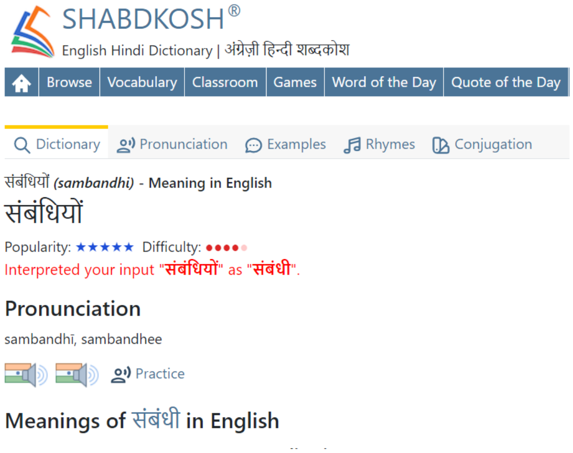
Figure 2: Spacing Structure

Such conventions were made in order to process it later using our code using '/n' newline symbols. So, whenever our model will retrieve any shloka and it's meaning, then it will know the shloka's starting and ending point in the document using these spacing conventions. So, overall we will have 3 final documents(English, Hindi and Gujarati) manually processed in this step. This completes our dataset preparation process.

4.3 Data preprocessing:

Our next step involves data preprocessing that includes methods such as tokenization, lemmatization, stop words removal and non-ASCII characters removal.

- **Removing Unnecessary characters**: All irrelevant characters like punctuation marks are removed from documents using libraries like regex, string.punctuation.
- **Tokenization**: We have extrated set of token by splitting the text by white space.
- **Removal of stopwords**: Since there is no direct library for hindi language to remove stopwords, we have used a list of common stopwords to remove them from documents.
- **Stemming/ Lemmatization**: This was the most crucial and difficult part of the project. Since there is no direct library to do lemmatization on hindi words, we have used different kind of approach to lemmatize them. We have scraped lemmatized word from website(https://www.shabdkosh.com/) using libraries like request, beautifulsoup, selenium, etc.



संबंधियों *(sambandhi)* - Meaning in English

संबंधियों

Popularity: ★★★★★ Difficulty: ●●●●●

Interpreted your input "**संबंधियों**" as "**संबंधी**".

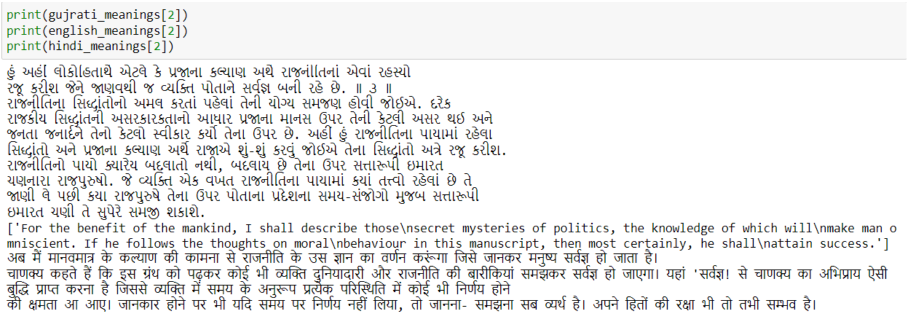
Pronunciation

sambandhī, sambandhee

Meanings of संबंधी in English

Figure 3: Lemmatized word on Shabdkosh.com

- **Mapping of Corpora**: The mapping of all the datasets i.e. english, hindi and gujrati is done manually. The mapping is done such that all the shlokes are parallel to each other that is all shlokas in all language are present at the same indices. Ex- if a shloka and its meaning in hindi is present at index 3, then its corresponding english and gujarati meanings are also present at the same index, 3.



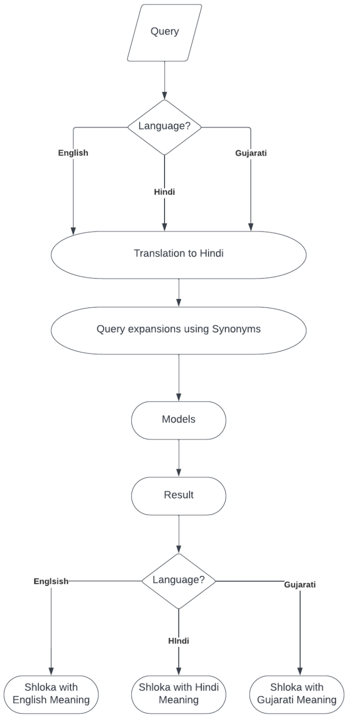
```
print(gujrati_meanings[2])
print(english_meanings[2])
print(hindi_meanings[2])
```

```
હું અહીં લોકોહિતાર્થે એટલે કે પ્રજાના કલ્યાણ અર્થે રાજનીતિનાં એવાં રહસ્યો
રજૂ કરીશ જેને જાણવથી જ વ્યક્તિ પોતાને સર્વજ્ઞ બની રહે છે. ॥ ૩ ॥
રાજનીતિના સિદ્ધાંતોનો અમલ કરતાં પહેલાં તેની યોગ્ય સમજણ હોવી જોઈએ. દરેક
રાજકીય સિદ્ધાંતની અસરકારકતાનો આધાર પ્રજાના માનસ ઉપર તેની કેટલી અસર થઇ અને
જનતા જનાર્દને તેનો કેટલો સ્વીકાર કર્યો તેના ઉપર છે. અહીં હું રાજનીતિના પાયામાં રહેલા
સિદ્ધાંતો અને પ્રજાના કલ્યાણ અર્થે રાજાએ શું-શું કરવું જોઈએ તેના સિદ્ધાંતો અત્રે રજૂ કરીશ.
રાજનીતિનો પાયો ક્યારેય બદલાતો નથી, બદલાય છે તેના ઉપર સત્તારૂપી ઇમારત
ચણનારા રાજપુરુષો. જે વ્યક્તિ એક વખત રાજનીતિના પાયામાં કયાં તત્ત્વો રહેલાં છે તે
જાણી લે પછી કયા રાજપુરુષે તેના ઉપર પોતાના પ્રદેશના સમય-સંજોગો મુજબ સત્તારૂપી
ઇમારત ચણી તે સુપેરે સમજી શકાશે.
['For the benefit of the mankind, I shall describe those\nsecret mysteries of politics, the knowledge of which will\nmake man o
mniscient. If he follows the thoughts on moral\nbehaviour in this manuscript, then most certainly, he shall\nattain success.']
अब मैं मानवमात्र के कल्याण की कामना से राजनीति के उस ज्ञान का वर्णन करूंगा जिसे जानकर मनुष्य सर्वज्ञ हो जाता है।
चाणक्य कहते हैं कि इस ग्रंथ को पढ़कर कोई भी व्यक्ति दुनियादारी और राजनीति की बारीकियां समझकर सर्वज्ञ हो जाएगा। यहां 'सर्वज्ञ! से चाणक्य का अभिप्राय ऐसी
बुद्धि प्राप्त करना है जिससे व्यक्ति में समय के अनुरूप प्रत्येक परिस्थिति में कोई भी निर्णय होने
की क्षमता आ आए। जानकार होने पर भी यदि समय पर निर्णय नहीं लिया, तो जानना- समझना सब व्यर्थ है। अपने हितों की रक्षा भी तो तभी सम्भव है।
```

Figure 4: Mapping of Corpora

After completing the data preprocessing step, we move on to implementing ur scoring models from scratch. We'll be implementing TF-IDF and BM25 information retrieval models from scratch in this project.

4.4 Steps



Query
Language?
English
Hindi
Gujarati
Translation to Hindi
Query expansions using Synonyms
Models
Result
Language?
Englsish
HIndi
Gujarati
Shloka with English Meaning
Shloka with Hindi Meaning
Shloka with Gujarati Meaning

Figure 5: Steps

Steps are as follows:

1. First query is translated to hindi language.
2. next Query is optimized using synonyms to fetch better results.

3. Query is then passed to models to get relevant result.
4. At last, the result is shown in language desired by user.

4.5 Model designing from scratch:

Next, we move to designing our model that helps us in documents scoring and efficient retrieval. We have implement two models: TF-IDF and BM25.

1. TF-IDF:

   TF-IDF (term frequency-inverse document frequency) can be thought of as a numerical metric that reflects how important a word is in a collection of corpus. Words that are frequent in a document but not across documents tend to have high TF-IDF score.
   Mathematically:

$$TF_{ij} = \frac{f_{ij}}{n_j}$$

$$IDF_i = 1 + log(\frac{N}{c_i})$$

$$w_{ij} = TF_{ij} * IDF_i$$

   where f_{ij} is the frequency of term i in document j. n_j is the total no of words in document j. N is the total no of documents in corpus. c_i is the no of documents that contain word i. w_{ij} is the TFIDF score of term i in document j.

2. BM25:

   This is an improvement of TFIDF algorithm. TF-IDF rewards term frequency and penalizes document frequency. BM25 goes beyond this to account for document length and term frequency saturation.
   Mathematically,

$$score(q,d) = \sum_{i=1}^{|q|} idf(q_i).\frac{tf(q_i,d)(k+1)}{tf(q_i,) + k(1 - b + b.\frac{|d|}{avgdl})}$$

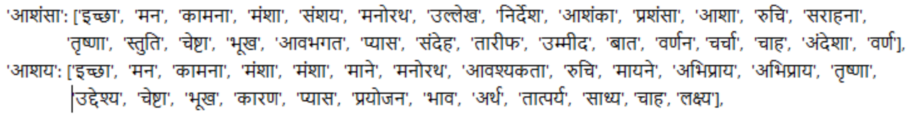
$$idf(q_i) = log\frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

4.6 Query Expansion using Synonyms

We will now try to optimize our query using query expansion techniques. It is necessary to do so because if a user feeds in a query, then it doesn't mean that he need only those documents that actually contain those words of query only. The documents desired may contain some words that are synonymous to the words in the query. So, we try to implement the query expansion technique in order to obtain better and efficient results. We try to find the synonyms of the words in query and expand our query with it. On evaluation, if the documents contain these synonym words, then that document too would be retrieved. The detailed description about the process is as follows:

1. First step is to translate the given query into hindi language if not already.
2. We then found all the unique words in our hindi corpus.
3. Next, we found all the synonyms of each unique word using package - pyiwn(https://github.com/cfiltnlp/pyiwn).
4. We have created a dictionary whose keys will be the synonym words and values will be the list of words of hindi corpus with synonym as key.

After our model is ready, we test some random inputs on our model. We manually look if the documents being retrieved are correct or not. We find our model to work fairly well for all the general queries we fed into our model. The detailed description about the model's evaluation process is given in the next subsection.



```
'आशंसा': ['इच्छा',  'मन',  'कामना',  'मंशा',  'संशय',  'मनोरथ',  'उल्लेख',  'निर्देश',  'आशंका',  'प्रशंसा',  'आशा',  'रुचि',  'सराहना',
         'तृष्णा',  'स्तुति',  'चेष्टा',  'भूख',  'आवभगत',  'प्यास',  'संदेह',  'तारीफ',  'उम्मीद',  'बात',  'वर्णन', 'चर्चा',  'चाह',  'अंदेशा',  'वर्ण'],
'आशय': ['इच्छा',  'मन',  'कामना',  'मंशा',  'मंशा',  'माने',  'मनोरथ',  'आवश्यकता',  'रुचि',  'मायने',  'अभिप्राय',  'अभिप्राय',  'तृष्णा',
        'उद्देश्य',  'चेष्टा',  'भूख',  'कारण',  'प्यास',  'प्रयोजन',  'भाव',  'अर्थ',  'तात्पर्य',  'साध्य', 'चाह', 'लक्ष्य'],
```

Figure 6: Synonyms Dictionary

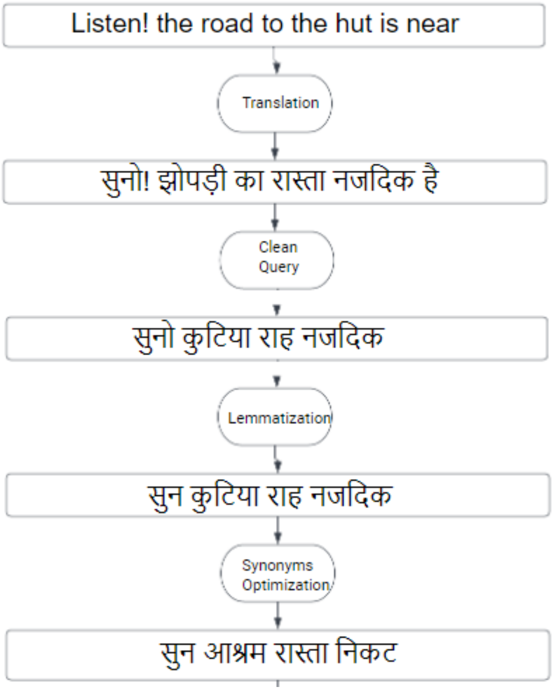


Figure 7: Example of Query Expansion

5 Model Evaluation

We created a test case of total 30 queries, where 10 queries are in english language, 10 in hindi language and 10 in gujrati language. Our model returns top 5 relevant documents based on the query. If these top 5 documents containing the ground truth document, we say that model has predicted correctly, if not present, we say model has predicted wrongly.

6 Results

We have evaluated each language query individually. We are testing our model on total of 10 queries in each language. For english queries, we are getting an accuracy of 40%, For Hindi queries, we are getting 95% accuracy, For Gujrati Queries, we are getting a total of 50% accuracy. So the average accuracy of our improved BM25 model is 61.66%.

7 Conclusion

1. The given project generates an effective cross-lingual search engine for Chanakya neeti in English, Hindi and Gujarati languages.
2. The TF-IDF and BM25 models perform very well in the retrieval task.
3. A frontend web app is successfully created as an end result of this project that is very user friendly.

4. The user can search the query in his desired language and also retrieve the documents in his desired language, even though the retrieved documents are originally of some other language.
5. The query expansion technique using synonyms and semantically similar words proves to be very effective in efficient searching of the documents.

8 Discussion and Future Work

The same system can be implemented on other long datasets too. Also, the query optimization approach can further be improved if proper dataset is provided.

9 Individual Contributions

Deepak Raj (21111024)

- Web scraping of Chanakya neeti database
- Data integration and file management
- Jupyter Files
  - Finding synonyms
- Streamlit work
- Project Report

Dinkar Tewari (21111025)

- Web scraping of Chanakya neeti database
- Jupyter Files
  - implementation of bm25 and tfidf
- Streamlit work
- Project Report
- Presentation

Divyansh Bisht (21111027)

- Data preprocessing
- Query optimization implementation
- Streamlit work
- Project Report

Rohit Kushwah (21111053)

- Query optimization implementation
- Model evaluation
- Streamlit work
- Presentation

Vikas (21111067)

- Model evaluation
- Results calculations
- Streamlit work
- Presentation
- Readme and other files

Every member of our group has contributed equally to this project.

10 Web app link

We created a web app using streamlit to generate better and visual insights to the user. The link of our web app is given below:

https://share.streamlit.io/divyansh009/ir/main/main2.py

- https://github.com/cfiltnlp/pyiwn
- https://www.shabdkosh.com/