



NATURAL LANGUAGE PROCESSING **PROJECT REPORT ON** **QUESTION ANSWERING IN HINDI**

Submitted to-

Dr. Pooja Jain

Assistant professor,CSE,IIITN

Submitted by-

Priya Gupta (BT19ECE011)

Apeksha Pandey(BT19ECE024)

INTRODUCTION :

This is a project on Question-Answering in Hindi language. This project uses Devanagari script for accepting queries and generating responses in Hindi language.

This type of Question-Answering is used in many rule specific chatbots that can answer queries related to train reservation, pizza delivery etc and it can also work as a personal medical therapist or personal assistant. Also many general purpose chat-bots can be seen that can answer queries on multiple fields, and there can be hybrid that can be engaged in both.

But, these are mostly available in English language and as in India, a huge population is comfortable in Hindi or their native languages, there is a need for the question answering to be available in Hindi and other languages. So, here we have tried to make a simple Question answering machine in Hindi language.

Here in this project, there are a specific set of rules. If the user query matches any rule, the answer to the query is generated, otherwise the user is notified that the answer to user query doesn't exist. One of the advantages of this is that they give accurate results most of the time. However, on the downside, they do not scale well. To add more responses, you have to define new rules.

In this, as the user enters a query, the query will be converted into vectorized form. All the sentences in the corpus will also be converted into their corresponding vectorized forms. Next, the sentence with the highest cosine similarity with the user input vector will be selected as a response to the user input.

If the proper match is not found, this will be handled by a hardcoded response and will again ask for the next input till the user inputs the query to stop the process.

FOR THE DATA SET:

The data set is needed to provide the response to the queries. The data set has been obtained by scrapping data from the webpage and processing.

Since we are using Devanagari script, for preprocessing, we have processed the scrapped data, which includes replacing “|” with “.” to identify the end of a sentence.

Then the sentences are separated and further broken down into words. Since the stop words interfere with the key-words, we have removed the stop words by matching the words with pre-defined stop words set.

FOR GENERATING THE RESPONSE:

- The tf-idf (term frequency-inverse document frequency) is used to weigh how important a word of a document in a document collection. NLTK does not support tf-idf. So, we have used scikit-learn. The scikit-learn has a built in tf-Idf implementation while we still utilize NLTK's tokenizer to preprocess the text. So, tf-idf weight for a term is the product of its tf weight and idf weight. It's the best known weighting scheme in information retrieval. Sometimes people denote it as tf.idf also.

We have pre-processed our text and removed punctuation already. So now, we initialize **TfidfVectorizer()**. We have passed the TfidfVectorizer our own function that performs custom tokenization and stemming.

$$tf-idf(t, d, D) = tf(t, d) \times idf(t, D)$$

where

$$tf(t, d) = f(t, d)$$

$$idf(t, d) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

```
word_vectorizer = TfidfVectorizer(tokenizer=get_processed_text,
stop_words=stop_words)
```

- Then, we call **fit_transform()** which does a few things: first, it creates a dictionary of 'known' words based on the input text given to it. Then it calculates the tf-idf for each term found in an article.

```
all_word_vectors = word_vectorizer.fit_transform(article_sentences)
```

- We use the **cosine_similarity** function to find the cosine similarity between the last item in the all_word_vectors list (which is actually the word vector for the user input since it was appended at the end) and the word vectors for all the sentences in the corpus.

We sort the list containing the cosine similarities of the vectors, the second last item in the list will actually have the highest cosine (after sorting) with the user input. The last item is the user input itself, therefore we did not select that.

```
similar_sentence_number = similar_vector_values.argsort()[0][-2]
```

- Finally, we flatten the retrieved cosine similarity and check if the similarity is equal to zero or not. If the cosine similarity of the matched vector is 0, that means our query did not have an answer.

In that case, we will simply print that we do not understand the user query.

so "मैं क्षमाप्रार्थी हूँ पर मैं आपके इस सवाल का उत्तर देने में असमर्थ हूँ ।" is returned.

Otherwise, if the cosine similarity is not equal to zero, that means we found a sentence similar to the input in our corpus.

In that case, we will just pass the index of the matched sentence to our "article_sentences" list that contains the collection of all sentences.

Prerequisites :

Python packages nltk and beautiful soup.

data.txt file was generated by data.py

(prerequisites - urllib, beautiful soup)

It will write the rendered text from the page to the data.txt file.

Python's regex library, re, will be used for some preprocessing tasks on the text.

How To Run:

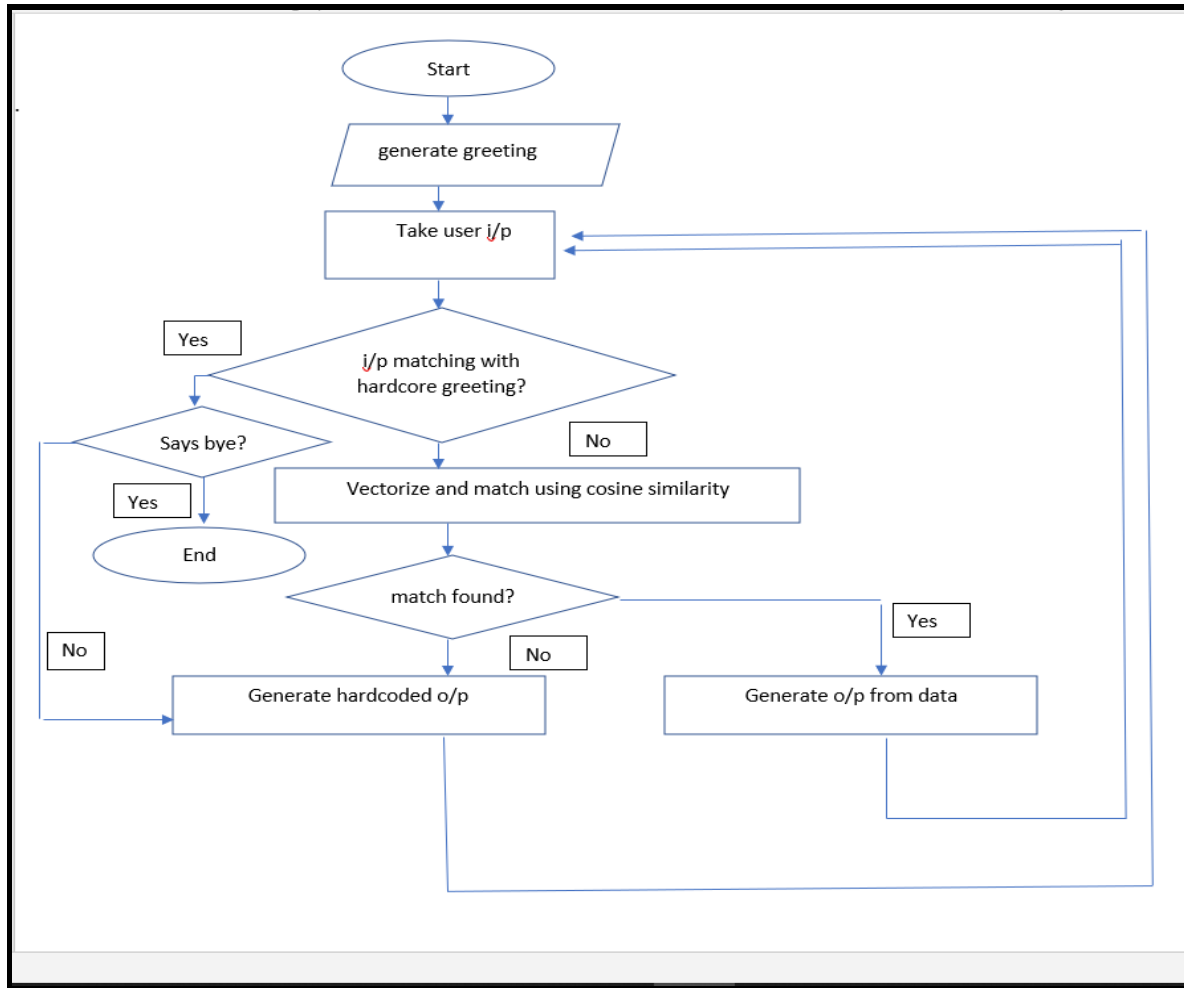
- Run data.py to extract data from the page and generate a data.txt file.
- Run main.py

You can copy and paste the queries from the queries.txt file to get the functionality of the Question answering.

You can also ask questions regarding the other topics ,it will give better output if the keyword is present in the data.txt.

FLOW OF THE CODE:

As we have generated our data.txt file,we now have our data ready.



WORKING:

Now as we know, it uses cosine similarity to check the similarities between two statements and considers the best similar statements from the data with the human query as the result. We have also included some hard-coded greetings and exit statements for better interface experience.

```
Run: main
C:\Users\hp\AppData\Local\Programs\Python\Python39\python.exe "C:/Users/hp/PycharmProjects/NLP - Copy/main.py"
नमस्कर ।
मे आपकी सहायता कैसे कर सकता हूँ ?
Warning: Your stop_words may be inconsistent with your preprocessor.
warnings.warn(
भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर (संक्षिप्त रूप में IIITN) भारतीय सूचना प्रौद्योगिकी संस्थानों (IIIT) में से एक है और नागपुर, महाराष्ट्र में स्थित एक राष्ट्रीय महत्व का संस्थान है। संस्थान ने जुलाई 2016 से काम करना शुरू कर दिया था, संस्थान का उद्देश्य है भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर (संक्षिप्त रूप में IIITN) भारतीय सूचना प्रौद्योगिकी संस्थानों (IIIT) में से एक है और नागपुर, महाराष्ट्र में स्थित एक राष्ट्रीय महत्व का संस्थान है। संस्थान ने जुलाई 2016 से काम करना शुरू कर दिया था, संस्थान का उद्देश्य है भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर (संक्षिप्त रूप में IIITN) भारतीय सूचना प्रौद्योगिकी संस्थानों (IIIT) में से एक है और नागपुर, महाराष्ट्र में स्थित एक राष्ट्रीय महत्व का संस्थान है। संस्थान ने जुलाई 2016 से काम करना शुरू कर दिया था, संस्थान का उद्देश्य है
```

So as we see here, "भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर" is getting matched irrespective of the query being asked.

If the keyword in the query is not present in the data then it won't give the desired output and would return the standard hardcoded output. Including more data could solve this issue.

```
Warning:
वैद्य: मैं क्षमाप्रार्थी हूँ पर मैं आपके इस सवाल का उत्तर देने में असमर्थ हूँ ।
वैद्य: अलविदा
Process finished with exit code 0
```

If the keyword in the query is not in hindi language, then it won't give the desired output and would return the standard hardcoded output.

```
मे आपकी सहायता कैसे कर सकता हूँ ?
नमस्कर ।
वैद्य: नमस्कर। मे आपकी सहायता कैसे कर सकता हूँ ?
वैद्य: नमस्कर ।
वैद्य: मैं क्षमाप्रार्थी हूँ पर मैं आपके इस सवाल का उत्तर देने में असमर्थ हूँ ।
```

Some of the queries and their responses are shown below:



```
नमस्कार ।
मेें आपकी सहायता कैसे कर सकता हूँ ?
नमस्कार
वैष्म्य: नमस्कार। मेें आपकी सहायता कैसे कर सकता हूँ ?
'कोविड रोग' के बारे में कुछ बताएं
वैष्म्य: 'कोविड रोग' का अर्थ है पदोवरण में बढ़ता तापमान ,सूर्य की गर्मी से धरती लगातार गर्म हो रही है और इसका मुख्य कारण है पदोवरण में कार्बन डाईऑक्साईड की मात्रा में वृद्धि ।इस कार्बन डाईऑक्साईड के स्तर के बढ़ने का मुख्य कारण है, धरत
'स्वच्छ भारत अभियान' का अर्थ है पदोवरण में बढ़ता तापमान ,सूर्य की गर्मी से धरती लगातार गर्म हो रही है और इसका मुख्य कारण है पदोवरण में कार्बन डाईऑक्साईड की मात्रा में वृद्धि ।इस कार्बन डाईऑक्साईड के स्तर के बढ़ने का मुख्य कारण है, धरत
वैष्म्य: 'स्वच्छ भारत अभियान' या 'स्वच्छ भारत मिशन' की शुरुआत 2 अक्टूबर 2014 को भारत के वर्तमान प्रधानमंत्री श्री नरेन्द्र मोदी ने की थी : देश के राष्ट्रपिता का 'क्लीन इंडिया' का सपना स्वच्छ भारत अभियान के माध्यम से पूरा किया जाएगा।
'स्वच्छ भारत अभियान' के बारे में बताएं
वैष्म्य: 'स्वच्छ भारत अभियान' या 'स्वच्छ भारत मिशन' की शुरुआत 2 अक्टूबर 2014 को भारत के वर्तमान प्रधानमंत्री श्री नरेन्द्र मोदी ने की थी : देश के राष्ट्रपिता का 'क्लीन इंडिया' का सपना स्वच्छ भारत अभियान के माध्यम से पूरा किया जाएगा।
'दीपावली' का त्योहार प्रति वर्ष कार्तिक मास की अमावस्या को देश के एक कोने से दूसरे कोने तक बड़ी पूजा-पाज से मनाया जाता है। इस त्योहार की पूजा-पाज कार्तिक कृष्ण त्रयोदशी से कार्तिक शुक्ल द्वितीय अर्थात् पाँच दिनों तक रहती है। दीपावली से
अधिक जानकारी लें
वैष्म्य: सचिन तेंदुलकर क्रिकेट के इतिहास में विश्व के सर्वश्रेष्ठ बल्लेबाजों में गिने जाते हैं। भारत के सर्वोच्च नागरिक सम्मान भारत रत्न से सम्मानित होने वाले यह सर्वप्रथम खिलाड़ी और सबसे कम उम्र के व्यक्ति हैं। राजीव गांधी खेल रत्न पुरस्कार से सम्मानित
'स्वच्छ भारत अभियान' के बारे में बताएं
वैष्म्य: 23 दिसम्बर 2012 को सचिन ने वन-डे क्रिकेट से सन्यास लिया और वहीं 16 नवम्बर 2013 को मुम्बई के अपने अंतिम टेस्ट मैच में उन्होंने 74 रनों की पारी खेलकर टेस्ट क्रिकेट से सन्यास लिया।
'कोविड रोग' के बारे में कुछ बताएं
वैष्म्य: रोग से बचाव करने में ही समझदारी है, क्योंकि यह एक संक्रामक रोग है जो बहुत ही तेजी से एक दूसरे में फैलता है।इन्फु एच ओ ने कुछ सावधानियों की सूची निकाली है और यह भी बताया है की कोरोना से बचाव के ये मूल मंत्र हैं।सदैव बहार से अ
अपने को कोविड रोग से बचाने के लिए सावधान रहें
वैष्म्य: अलविदा
Process finished with exit code 0
```

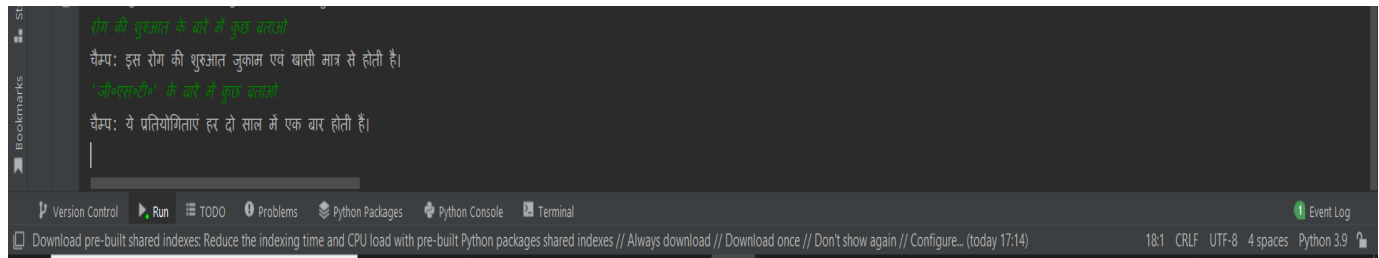
Some improvement work that could be done:

The Question-answering project doesn’t understand the words semantically.

Certain errors regarding absolute text matching can occur if more than one occurrence is there in different contexts.

A Lemmatizer in place of stemmer would better the results for example

Here in this, बारे में Was searched as बार leading to wrong result.



```
कोविड रोग के बारे में कुछ बताएं
वैष्म्य: इस रोग की शुरुआत जुकाम एवं खासी मात्र से होती है।
'कोविड रोग' के बारे में कुछ बताएं
वैष्म्य: ये प्रतियोगिताएं हर दो साल में एक बार होती हैं।
```

CONCLUSION and Results:

Hence,our project is successfully working for the data that has been processed and saved as a file.This project could further be improved by enhancing the dataset,making the library rich will enhance the chances of matching the keywords.Further we can

improve it as a chatbot with interface that can be used as specific rule based or general purpose.

REFERENCES:

<https://www.analyticsvidhya.com/blog/2020/01/3-important-nlp-libraries-indian-languages-python/>

https://research.variancia.com/hindi_stemmer/

<https://github.com/goru001/inltk>

<https://studymachinelearning.com/cosine-similarity-text-similarity-metric/>

<https://www.geeksforgeeks.org/python-measure-similarity-between-two-sentences-using-cosine-similarity/>

<https://www.analyticsvidhya.com/blog/2021/10/hands-on-hindi-text-analysis-using-natural-language-processing-nlp/>

<https://stackabuse.com/python-for-nlp-creating-a-rule-based-chatbot/>

<https://data.mendeley.com/datasets/bsr3frvvc/1>