Chatbot for college Q&A

Satyam,Shahid,Debarghya, Nikhil

Data collection for Question-Answering

For the QA data set we have collected data from various sources including reddit, college dunia website and FAQ website of IIT Bhilai.

We used python's Reddit API wrapper Praw to obtain an instance of a read-only reddit instance. Next we obtained a instance of our subreddit by passing the subreddit's name to our reddit instance.

Further we did web scraping using beautifulsoup to get more FAQ data from college dunia website and IIT Bhilai official FAQ website.

Corpus data collection

Using wget we retrieved all the HTML contents from the IIT Bhilai official websites

```
wget \
     --recursive \
     --no-clobber \
     --page-requisites \
     --html-extension \
     --no-check-certificate \
     --convert-links \
     --restrict-file-names=unix \
     --domains iitbhilai.ac.in\
     --no-parent \
         https://www.iitbhilai.ac.in
```

Data preprocessing

For preprocessing the corpus data we used Pup a command line tool for preprocessing HTML data. Further we removed any CSS content from the data using sed and extracted all the text into iit-corpus.txt file.

```
for i in ${ls}
do
cat $i | pup 'div#content' text{} | sed '/^[[:space:]]*$/d' >> iit-corpus.txt
done
```

Baseline model of chatbot

- The baseline model is trained with question and answers taken from FAQs in IIT Bhilai website.
- The training data is preprocessed using Wordnetlemmatizer,RegexpTokenizer and stopwords
- To get the word embeddings from training data we used models like Fasttext and Word2Vec imported from gensim package

Input & output of model

- The input to the embedding model is list of questions and the model returns an object using which we can get the embeddings of words and can also check the similarity between words and so on..
- To answer the given question we checked the word Mover's distance between the asked question and the available question in the training data.
- we recorded the minimum distance from all the available questions and the answer for the particular question with minimum distance.

Contd.

- If the minimum distance is below a threshold then we answered the question with the recorded answer.

- If the minimum distance is above the threshold then we answered as “sorry I don’t have answer.Can you please rephrase the query”.

- Have a look at the sample output of the baseline model.

Snap of output of baseline model

```
Enter query
where is IIT Bhilai
0.4746478945044717
No, it has an in-transit campus at GEC, Raipur.



Enter query
What is the best branch in IIT Bhilai
0.5693715981810334
It depends on the performance of the student in the first year. Less than 10% of students get to change their branch.



Enter query
Can i change my branch in IIT bhilai
0.288298674415518
It depends on the performance of the student in the first year. Less than 10% of students get to change their branch.



Enter query
Which is better NIT or IIT
0.5662757347565934
IIT Bhilai offers quality placements. An average package of 5-7LPA is offered and a maximum o 24 LPA (as of 2021).



Enter query
```

BERT MODEL

- **Domain specific pretraining:** Pretrained the BERT ‘bert-base-uncased’ model on IIT Bhilai corpus on Masked Language Modeling(MLM)

- **Next sentence prediction (NSP) training on IIT Bhilai FAQ:** Trained the pretrained model on the question and answer pairs such that the model can predict the correct answer given the question.

BERT Inference

- During inference given the query we matched the query with all the available answers in the training data
- From all the matchings recorded the most probable answer.

Frontend Part

- Worked on Frontend Part the page design in Html & CSS. This page is static because we only render the template website of bot and client.
- We also used Javascript v8 engine to perform the action of the buttons and the hold the dynamic properties which will be beneficial.
- Beside of this if the model get stuck on some part the frontend part automatically render the part of some templated value that is stored in hard coded of the code

Backend Part

- The backend used in the project is Django Framework which enables the backend part and client server connection.
- The Django framework under the python library which ensure the server rendering the templates & statics pages.
- All the API written under the Django framework and context switch to HTML template.
- Under this there is very feasibility of the program that could run so smooth because of the programme also written in python

User Interface