Natural Language Processing Project

P10: Chat-Bot

GROUP NO.18

Mentored By: Udit Dharmin Desai

Supervised By: Prof. Sudeshna Sarkar

GROUP Members:

Name	Roll No.
Mamidi Baby Soumya	18CS30027
Manoranjan Behera	19CS91R04
Rajdeep Das	18CS30034
Venugopal Chakka	18CS30052

Table of Contents:

Introduction:

In this project we have implemented **Medibot!** which is a chatbot that can converse on health topics such as what are various diseases and how to cure those diseases.

Corpus:

The corpus is made with the informations parsed from this website:
https://www.niams.nih.gov/health-topics/all-diseases
Corpus link: data

Libraries Used:

We have used NLTK, SKLearn,Random,String etc to implement this Project.

Approach:

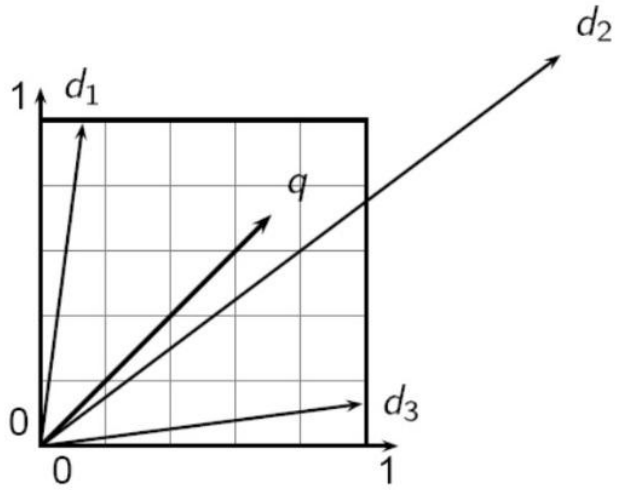
- The corpus contains details about various diseases and cures/treatments of those diseases. The main task of the chatbot is to get the query from the user and get the sentence from the corpus which best matches the query.

- For this we first splitted the corpus into sentences using sent_tokenize. Next step was to tokenize those sentences into words. Then we preprocessed those words by removing stopwords and applied stemming and lemmatization.

- Next step was to convert those sentences into vectors to calculate similarity. For vector representation we used “Bag of words”,which doesn’t take into consideration the order of occurrence of the words.

- Then we scaled the vectors by using TF-IDF values. Where TF denotes the term frequency of a word. And IDF is Inverse Document Frequency, DF represents in how much documents the word has occurred. This scaling was required mainly because there are Words which occurs many times have less significance than words which occurs less.

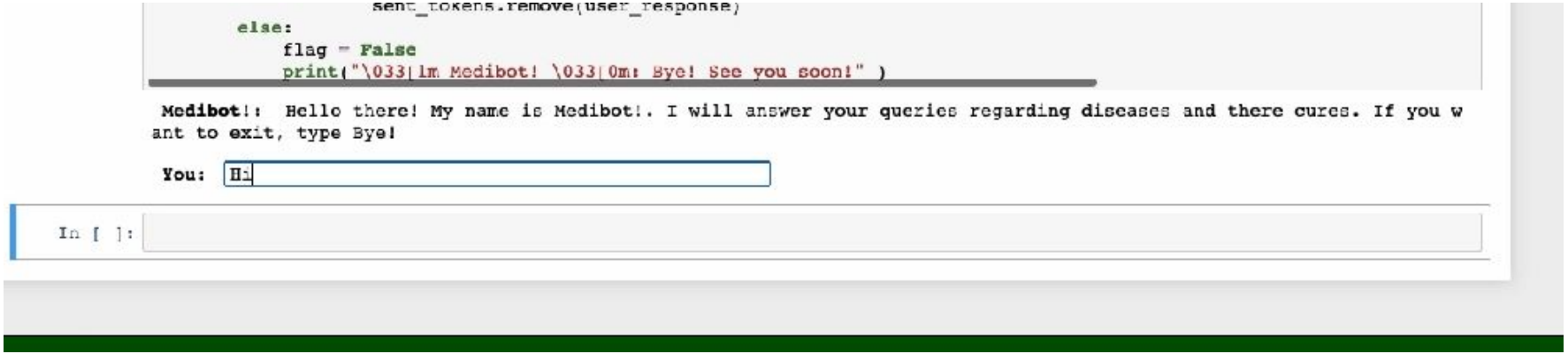
Continued...



- Now after generating vectors from all the Sentences and also from the query , our task was to get the sentence vector which is closest to query vector. For this we used cosine similarity.
- If we are Unable to find any similarity between the query and corpus sentences, then we simply printed out
  “I am sorry! I don't know about this disease”.
- In case if the user uses any greetings then we handled it differently.

How to Run The Code:

- First Download the Python Notebook file from this link: ipynb file
- Then Download the corpus from this link: data
- Import the code in Google Colab or Jupyter Notebook. Change the data file path accordingly. Run all the cells.
- Let's say X is a disease. Then you can ask questions like "What is X disease?" or "How to cure X?" or "How to do Treatment of X?" . If X is in Corpus, then the chatbot will reply with the Information about that Disease.



References

https://heartbeat.comet.ml/building-a-conversational-chatbot-with-nltk-and-tensorflow-part-1-f452ce1756e5

https://landbot.io/blog/natural-language-processing-chatbot

https://towardsdatascience.com/how-to-build-a-chatbot-a-lesson-in-nlp-d0df588afa4b