Rule-based Chatbot

Tanish Lad and Jashn Arora

TOC

Problem Statement

Process

Definition

Demonstration

Future Work

Issues

References

Problem Statement

To Design a Rule-based Chatbot which interacts with the user about any topic in Hindi

What is a Chatbot ?

A computer program designed to simulate conversation with human users.

The Procedure Used

Briefly..

Scrap Raw Data

For any topic, scrap the whole page of the topic from Wikipedia

01

Filter Raw Data

Cleaning the data like removing empty spaces, lines and special characters, etc.

02

Parsing

Parsing each sentence of the filtered data

03

Lemmatize

Find root of each word of each sentence in the filtered data

04

User Query

Ask user for query regarding the topic

05

Cosine Similarity

Output Sentence(s) having maximum cosine similarity with the user query

06

Scrap Raw Data

01

We ask the user about the topic he wants to converse about.

We find the Wikipedia page regarding the topic.

Then we scrap the wikipedia page using beautiful soup Python library and store the data.

Filter Raw Data

02

The scraped data is in raw format.

A lot of cleaning is required to make the data look readable.

Using concepts of filtering and regex, we managed to clean a large part of the raw data

Parsing
03

We then send each sentence of the data one by one in a online parser for hindi sentences. Then we curl its output.

Lemmatization
04

We then find the root of each word of each sentence using the output of the chunked data.

We store all these data (sentence, lemmatized sentence, POS-tagged sentence) in text files which will serve as input to the chatbot program

User Query
05

The user is greeted and then asked for a query regarding the topic he chose.

The user inputs his query.

Alternatively, he can chose to bid adieu to the bot by typing out “नमस्ते”

Output Response
06

We chunk the input query of the user using the same technique.

We then lemmatize it and send the lemmatized sentence to find the cosine similarity with all sentences in our data.

The sentence(s) having the maximum cosine similarity with the input query is considered as the response.

Issues

1. Cosine Similarity is not a very good method of finding the best output for the user query

2. Chunking and then lemmatizing is a very slow method for finding root of a word and also requires internet.

Future Work

We plan on Running a Language Model to extract the best output.

References

1. M.Dahiya. ***A Tool of Conversation: Chatbot (2017)***

2. Ananthakrishnan Ramanathan, Durgesh D Rao. ***A Lightweight Stemmer for Hindi (2003)***

3. Lovely Sharma, Vijay Dhir and Kamaljeet Kaur. ***A New Model for Question-Answer based Dialogue System for Indian Railways in Hindi Language (2015)***

4. Nick Webb. ***Rule-Based Dialogue Management Systems (2001)***

Thank you.

Checkout our GitHub Repository on the work:
https://github.com/destinyson7/Chatbot-Hindi