CSE 535: INFORMATION RETRIEVAL

MULTI TOPIC INFORMATION RETRIEVAL CHATBOT

Team: Caesars

Tejasri Karuturi - 50419936
Venkata Prathima Bhargavi Karri - 50419974
Eswara Venkata Pavan Kumar Parimi - 50441667

INTRODUCTION

Chatbots, also known as Conversational Agents or Dialog Systems, are a hot topic now. Microsoft is making big bets on chatbots, and so are companies like Facebook (M), Apple (Siri), Google, WeChat, and Slack.

Our Project is a Retrieval Based bot built on top of reddit and chitchat data set using Solr, Angular, Flask and NLTK. It uses a pre-built BM25 model in Solr to match the query from the indexed data with the keywords and retrieves the best found match which is relevant to the user query.

Various types of topics that the bot can converse with users are Politics, Education, Healthcare, Technology, Environment and basic chitchat.

METHODOLOGY

BackEnd

Dataset

- The data used in the application is brought from chitchat dataset and Reddit data.

Data Preprocessing

- Initially the raw data of the two datasets will be pre-processed to gather only the fields that are required for us in python.

Indexing

- Once data is processed, two cores will be created for each of the data in Solr. Chitchat data has fields <Question, Answer>. Reddit Data has fields <Parent_Body, Body, Topic>.
- These two files are indexed using the BM25 Model of Solr.

Query Preprocessing

- When the user sends the query, the query is preprocessed first like removing the special characters and all the words are brought to their root word using the Lemmatizer.

Query Classification

- These words are compared with a set of words which are extracted from chitchat data using the same preprocess done for user query. Based on the words obtained from the user query are compared to the preexisting words which can be used for classification of the query whether to send it to the reddit core or chitchat core. If the threshold value of the core is more than 0.85 query is sent to the chitchat core and if it is less than 0.85 query is sent to the reddit core.

Query Documents Retrieval

- Top 20 documents that are obtained from the corresponding core are used to re-rank the documents.

Documents Re-Ranking using Cosine Similarity

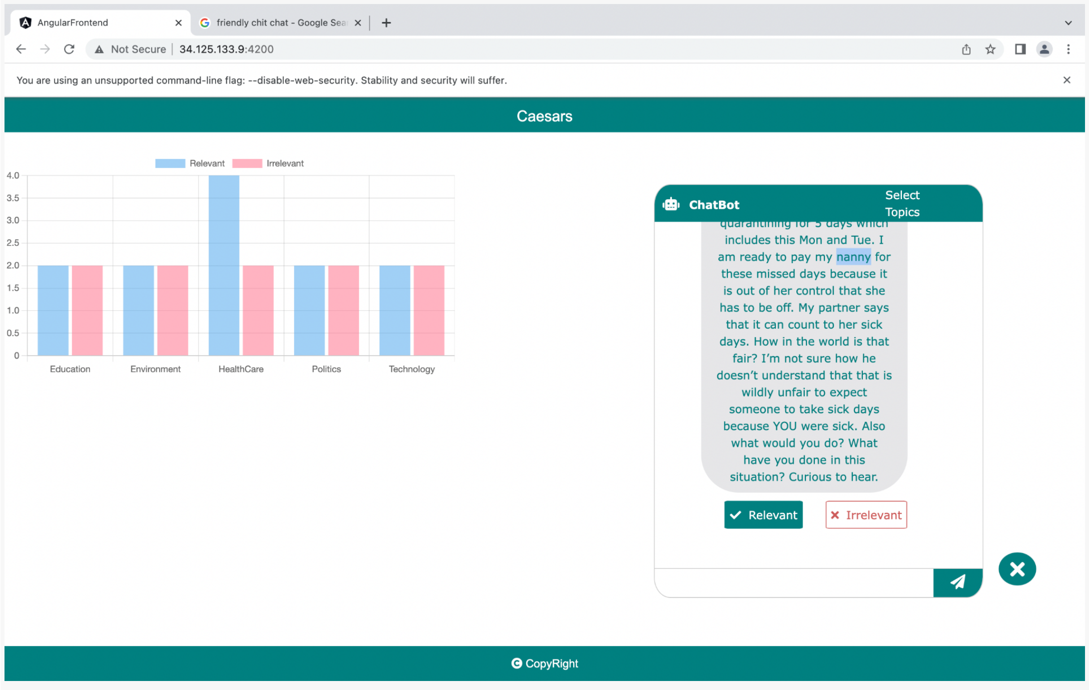
- Re-ranking of the documents are done using the cosine similarity by embedding those 20 documents as well as query.
- The corresponding body of the parent body of all the documents which are more similar to the user query is returned as a response to the user.

FrontEnd

Visualization

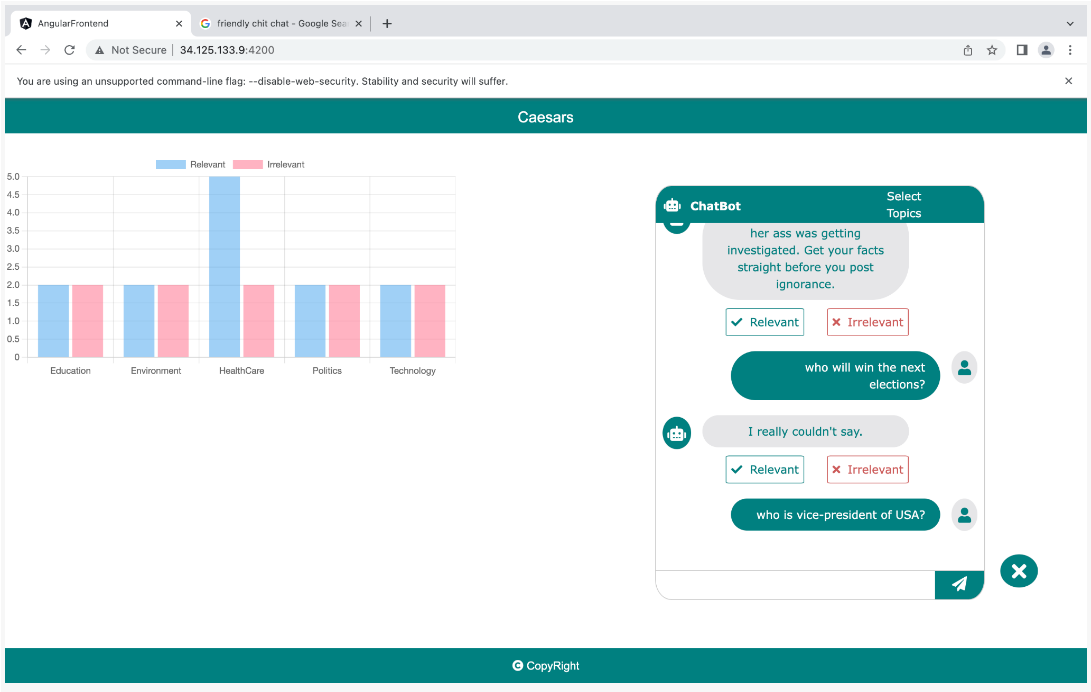
- We have used Charts.js library for visualization.
- We have taken the number of relevant and irrelevant answers that are given by the chatbot for every topic.

Screenshots



You are using an unsupported command-line flag: --disable-web-security. Stability and security will suffer.

quarantining for 5 days which includes this Mon and Tue. I am ready to pay my nanny for these missed days because it is out of her control that she has to be off. My partner says that it can count to her sick days. How in the world is that fair? I'm not sure how he doesn't understand that that is wildly unfair to expect someone to take sick days because YOU were sick. Also what would you do? What have you done in this situation? Curious to hear.



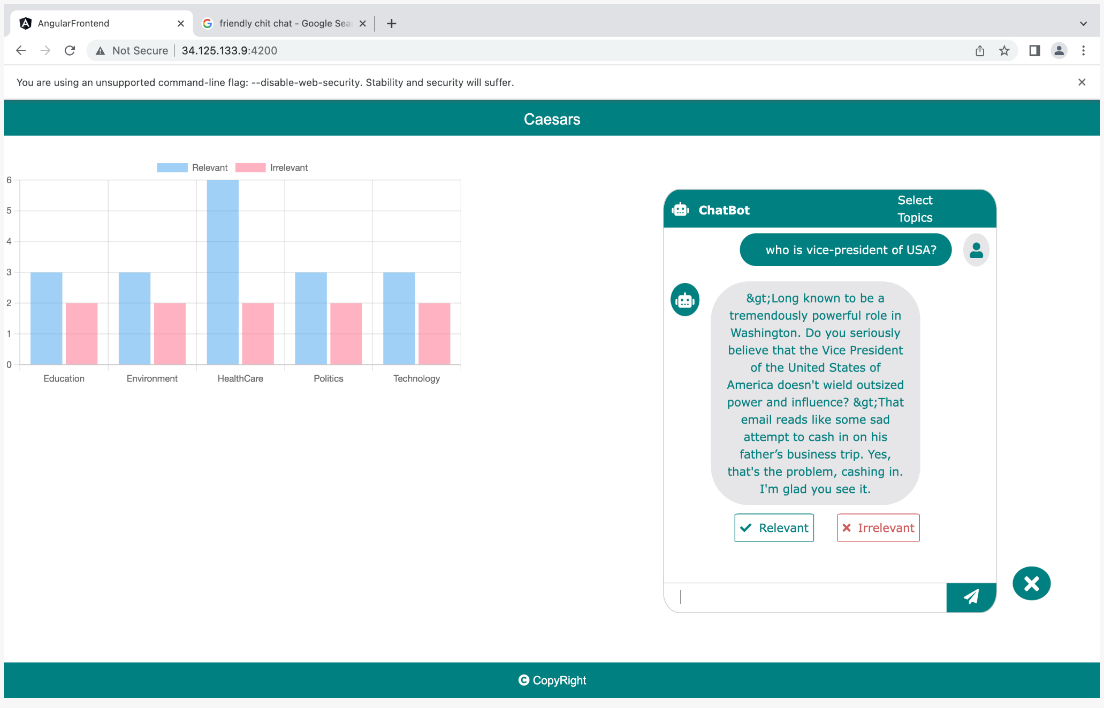
You are using an unsupported command-line flag: --disable-web-security. Stability and security will suffer.

her ass was getting investigated. Get your facts straight before you post ignorance.

who will win the next elections?

I really couldn't say.

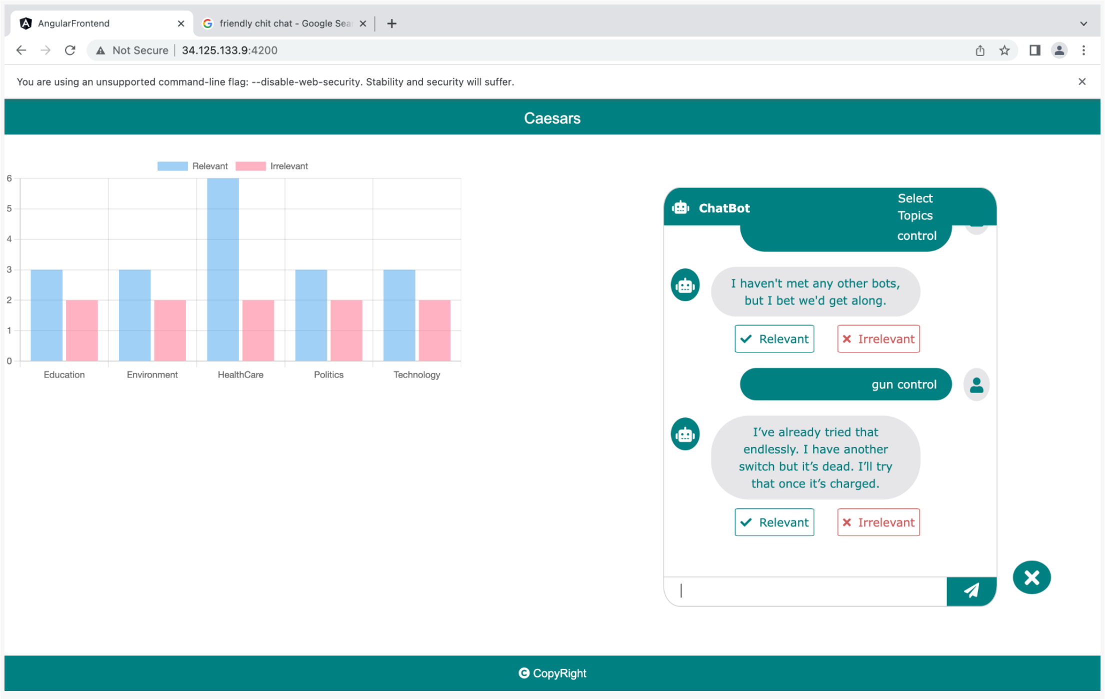
who is vice-president of USA?



You are using an unsupported command-line flag: --disable-web-security. Stability and security will suffer.

who is vice-president of USA?

>Long known to be a tremendously powerful role in Washington. Do you seriously believe that the Vice President of the United States of America doesn't wield outsized power and influence? &gt;That email reads like some sad attempt to cash in on his father's business trip. Yes, that's the problem, cashing in. I'm glad you see it.



You are using an unsupported command-line flag: --disable-web-security. Stability and security will suffer.

control

I haven't met any other bots, but I bet we'd get along.

gun control

I've already tried that endlessly. I have another switch but it's dead. I'll try that once it's charged.

Contribution

Name	Contribution
Tejasri Karuturi	**Frontend Using Angular**(User Interface, Visualization using ng2-charts)
Prathima Bhargavi	**Data Collection**(450,000), **Indexing** Chit Chat and Reddit
Pavan Kumar	**Backend Using Flask** (Query Classification, Query Documents Retrieval, Documents Re-Ranking using Cosine Similarity.)

Conclusion

- The Multi Topic Information Retrieval chatbot that is currently designed can chit-chat and talk to the user on a few topics: Politics, Education, HealthCare, Environment and Technology.
- Chatbot is performing quite well in classification between chit-chat and reddit data as well as the responses that are obtained for chit-chat are very good and maintain a bit of context as well. But the responses that are obtained from the reddit data are just matching to the relevant query terms and some of the responses are not relevant to the question.
- The reasons behind this is because of insufficient data and in a way of understanding the question and matching to the specific response, which can be improved using the few NLP models.