

CSE 535 INFORMATION RETRIEVAL PROJECT 4

FALL 2022

THE CODE LINGUISTS

PRANEETH NEKKALAPUDI (PNEKKALA@BUFFALO.EDU)

SUMEDH KHODKE (SUMEDHK@BUFFALO.EDU)

JAY ASHOK LAL (JAYASHOK@BUFFALO.EDU)

EVA PRADHAN (EVAPRADH@BUFFALO.EDU)

ABSTRACT

The aim of this project is to build an effective retrieval based end to end chatbot able to converse in 5 defined topics (Politics, Environment, Education, Healthcare, Technology). Additionally the chatbot should also be able to carry out general conversations. Response depends on retrieval as opposed to generative chatbot models.

INDEX

- 1. Introduction 4**
- 2. Iterations and alternatives explored 5**
- 3. Methodology (Design and process) 6**
 - Concepts
 - Resources
 - Flow
- 4. Setup and use 13**
- 5. Observations 20**
- 6. Contributions 24**
- 7. Conclusion 25**

INTRODUCTION

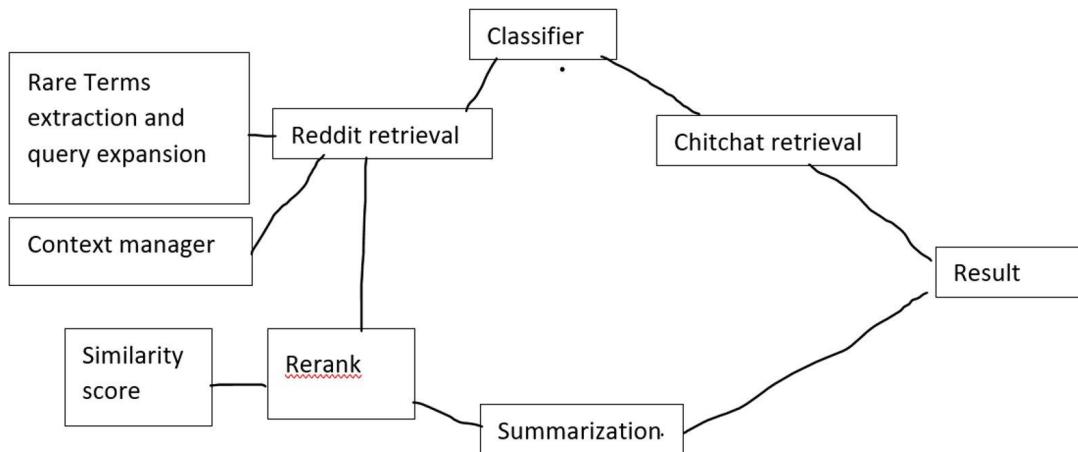
In this report we present our Multi-Topic Information Retrieval chatbot that caters to requests about information available on various subreddits on Reddit. The chatbot is designed to respond to general user queries on the whole knowledge base and focused queries that are directed to a restricted scope in the knowledge base.

Restricting the scope of search serves two purposes;

1. First, it can lead to more relevant information retrieved for the user since the user's intent is inherently captured when a particular topic is chosen.
2. Second, it serves as an evaluation measure of the chatbot by allowing it to capture the global statistics on the efficiency of the model. It creates the foundation for analytics.

Healthcare, Environment, Education, Politics and Technology are the topics that are made available for the user to choose from while interacting with the chatbot. Inferring context and tracking is a major challenge and is a necessary component for building a robust chatbot. Our chatbot includes reasonable heuristics to detect context, track and flush context through the chain of conversation. The chatbot is made available as a web application hosted on the cloud.

METHODOLOGY



WORKFLOW:

1. DATA SCRAPING AND PREPARATION:

The knowledge base is built on the data scraped from Reddit. Data was scraped using the Pushshift API. We've combined all our data to create the corpus. Data is cleaned, i.e filling any missing values by extracting them again and removing useless records. Additionally, we've also used a chitchat dataset (<https://github.com/BYU-PCCL/chitchat-dataset>) to handle general user queries. Our corpus consists of around 100k records from Reddit and around 10k records from ChitChat dataset. The data was indexed into 2 SOLR standalone cores, one consisting of the Reddit data and the other consisting of the Chitchat dataset. While they could be indexed together, enforcing separability diminishes chances of seeing irrelevant results upon searching in their individual indexes as opposed to searching them in a shared index, which would be more prone to junk irrelevant results. Having separated the data, we needed to classify incoming queries to search in their respective indexes.

2. QUERY INPUT AND INTENT CLASSIFIER:

For classifying a user's Intent we've used a simple logistic regression classifier, tuned for greater accuracy. The model is trained on the whole of Chitchat dataset (10k records) and 10k records of only Reddit submissions, excluding the comments from the Reddit dataset. This ensures that both the Chitchat dataset and the Reddit dataset are fairly represented to the classifier model. Furthermore, the 10k submissions from Reddit should be representative of the data collected from Reddit, since submissions set the course of discourse. The model takes document embeddings (centroid) computed from the pretrained glove word embeddings. Twitter-6B pretrained embeddings with a dimension of 100 were used for the task. Instead of plainly training the model on the document's centroid which assigns equal weight to every term, the document centroid is computed by multiplying the embedding vector with the corresponding IDF weight of the term. This ensures that the classifier picks up on the nature of Reddit dataset consisting of rarer terms compared to the Chitchat dataset. To achieve this, similar words must be identified and grouped together, so we've lemmatized the words using Spacy POS tagging. The IDF dictionary contains the root words. Therefore, training and classification also is performed on the root words. Testing it on a small left-out subset of corpus data yielded an accuracy of around 93%. The model's learned parameters were extracted and are being used for classifying user intent.

3. Solr schemas

All the relevant fields are defined and indexed appropriately. The core consisting of Reddit data is configured with BM25 global similarity with parameters for b set to 0.9 and k1 to 1.2. The core consisting of Chitchat dataset is configured with DFR similarity with the *basic model* parameter set to G, *after effect* set to L, *normalization* set to H2 and c hyperparameter set to 0.7. Due to the nature of the ChitChat dataset, short records and a comparatively small dataset, DFR's capability in handling short queries and small corpuses made more sense. However, no discernable difference was noticed when BM25 was configured the Chitchat core.

```
<schema name="default-config" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.BM25SimilarityFactory">
    <str name="k1">1.2</str>
    <str name="b">0.9</str>
  </similarity>
```

```
<schema name="default-config" version="1.6">
  <uniqueKey>id</uniqueKey>
  <similarity class="solr.DFRSimilarityFactory">
    <str name="basicModel">G</str>
    <str name="afterEffect">L</str>
    <str name="normalization">H2</str>
    <float name="c">7</float>
  </similarity>
```

```
<fieldType name="text_en_red" class="solr.TextField" positionIncrementGap="100">
  <analyzer type="index">
    <charFilter name="mapping" mapping="cmap.txt"/>
    <charFilter name="patternreplace" pattern="\\\" replacement="" />
    <charFilter name="patternreplace" pattern="US" replacement="USA" />
    <charFilter name="patternreplace" pattern="([A-Z]).(=[A-Z])" replacement="$1" />
    <tokenizer name="standard" />
    <filter name="phonetic" encoder="soundex" />
    <filter name="shingle" minShingleSize="2" maxShingleSize="4" />
    <filter name="stop" ignoreCase="true" words="lang/stopwords_en.txt" />
    <filter name="lowercase" />
    <filter name="englishPossessive" />
    <filter name="keywordrepeat" />
    <filter name="porterStem" />
    <filter name="removeduplicates" />
  </analyzer>
  <analyzer type="query">
    <charFilter name="mapping" mapping="cmap.txt" />
    <charFilter name="patternreplace" pattern="([A-Z]).(=[A-Z])" replacement="$1" />
    <tokenizer name="standard" />
    <filter name="phonetic" encoder="soundex" />
    <filter name="stop" ignoreCase="true" words="lang/stopwords_en.txt" />
    <filter name="lowercase" />
    <filter name="englishPossessive" />
    <filter name="keywordrepeat" />
    <filter name="porterStem" />
    <filter name="removeduplicates" />
  </analyzer>
</fieldType>
```

4. CONTEXT IDENTIFICATION AND TRACKING:

For managing context, heuristics were used to achieve the desired functionality. We are making a Markov assumption for tracking context, i.e, named entities, if any present, from each query are extracted using Spacy NER feature and the subsequent query is expanded with the entities from the previous query. But context should not be carried forward for each subsequent query. Context can be switched by the user. This is tracked using similarity search between contiguous queries from sentence transformers library, which effectively computes the cosine similarity between sentences based on their embeddings. If the similarity between a contiguous pair of queries meets a defined threshold, context continuity is assumed and the next query is expanded with the previous query's named terms, else the previous query's named entities are flushed. Additionally, when the user switches topics, context is flushed and if the user does not switch topics, context is carried forward. Finally, context is flushed at the end of each session.

5. QUERY FORMULATION:

Queries are formulated following the edismax syntax, as it offers a wider set of possibilities when it comes to boosting terms in the queries. Query is formulated based on whether the user chooses a corpus-wide general query or a topic-wise faceted search. Faceted search functionality is achieved using the filter query parameter. When formulating a query, both the question-and-answer fields are searched. Retrieval is based on searching the question, where rare terms and context terms are boosted and reranking is performed by searching the answer field for rare terms and context terms. Context terms refer to the named entities extracted from the previous query. Rare terms are the 2 rarest terms in any given query. These rare terms are further expanded by finding 3 top similar words to each of them from pretrained Glove embeddings.

6. Scoring

For retrieval, the documents are scored following the configured similarity rules, which is output by Solr. For reranking, the scores returned by solr and another additional similarity score computed between the query and response based on the pretrained embeddings is used.

$$\text{Score} = (1-\alpha) * \text{similarity score} + \alpha * \text{BM25 score.}$$

Alpha was set to 0.7 after tuning it based on subjective perception of the produced results.

No reranking is performed on results retrieved from Chitchat datasets. It performs well with basic retrieval given the nature of the data and user queries.

7. Summarization

Finally, the text from the best response is summarized so that useful information is captured and presented to the user. Pretrained BERT models for text summarization are used to achieve this.

8. Analytics

The analytics tab in the UI is designed for explainability. It also helps with debugging. It contains information on the index being searched, classifier output, rare terms being searched, context terms being searched. BM25 score, similarity score, total score, complete text of the summarized response presented to the user, and all the documents retrieved for the query.

9. User Feedback

The app also has provisions for user feedback. The user can provide feedback on whether the results were relevant to their information need. Currently we are using user feedback for the purpose of visualization.

10. Database

The app stores data on each turn in a MySQL server. Session id, User id, query, best response, all responses, scores, faceted topics, bot personality and user feedback are stored in the database and will be used for creating visualizations that explain the performance of our model.

SETUP AND USE

The web application has been hosted on <http://34.125.158.59:8000/TheCodeLinguists/>

This is what it looks like upon loading

The screenshot shows a web browser window titled "Chatbot by TheCodeLinguists". At the top, there are three tabs: "Chatbot" (selected), "Analytics", and "Visualizations". Below the tabs, there are two input fields: "Select a personality" (set to "Enthusiastic") and "Faceted search" (set to "None"). A large text area labeled "Chatbot" contains a message from the user: "Hi, how's it going?". Below this, there is a text input field labeled "Enter your message below" containing the text "Hi, how are you doing". To the right of this input field are two buttons: "Submit" and "Clear". Below the input field, there is a question "Was the last response of the chatbot relevant?" with two radio buttons: "Satisfactory" and "Not satisfactory". To the right of this section is a button labeled "Send feedback".

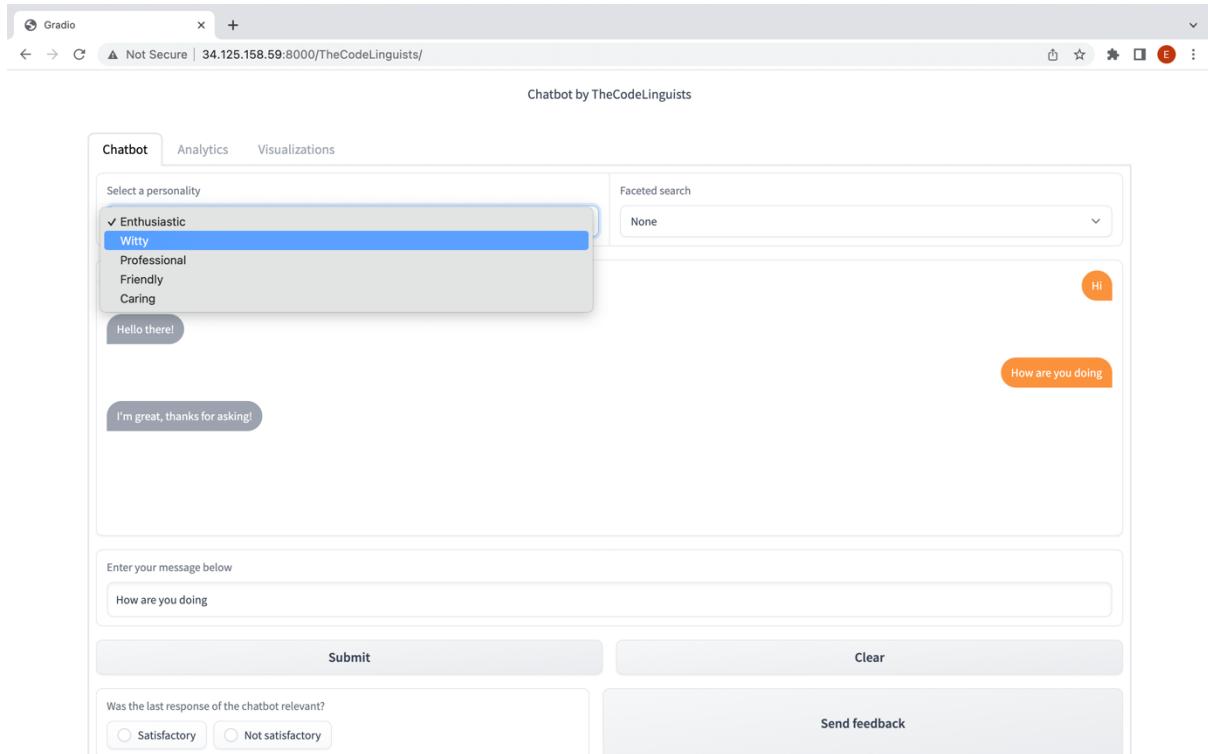
You will find 3 tabs on the top left: Chatbot, Analytics and Visualizations

On the Chatbot tab user can

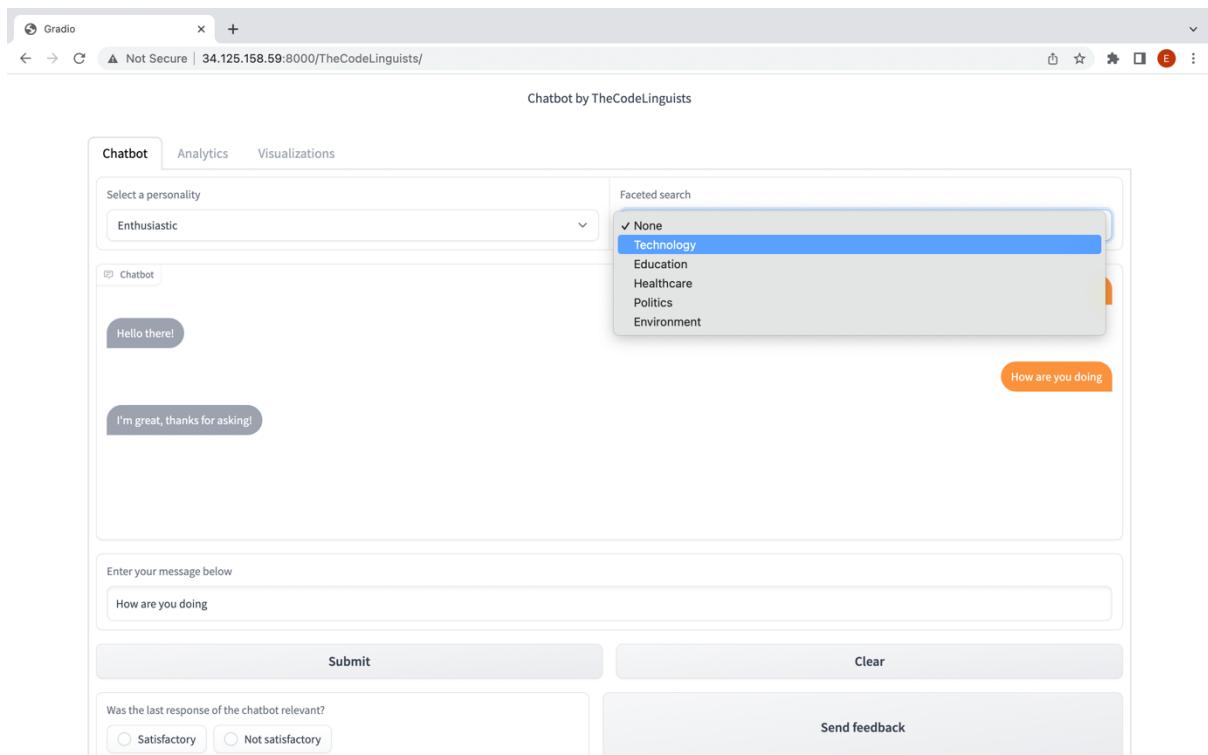
- input query under “Enter message below”

The screenshot shows the same web browser window as the previous one, but now with a visible conversation history. The user has typed "Hello there!" and the chatbot has responded with "Hi". The user has also typed "How are you doing" and the chatbot has responded with "I'm great, thanks for asking!". Below the conversation, the user's message "How are you doing" is highlighted with a blue border. The rest of the interface is identical to the first screenshot, including the tabs, personality selection, and feedback options.

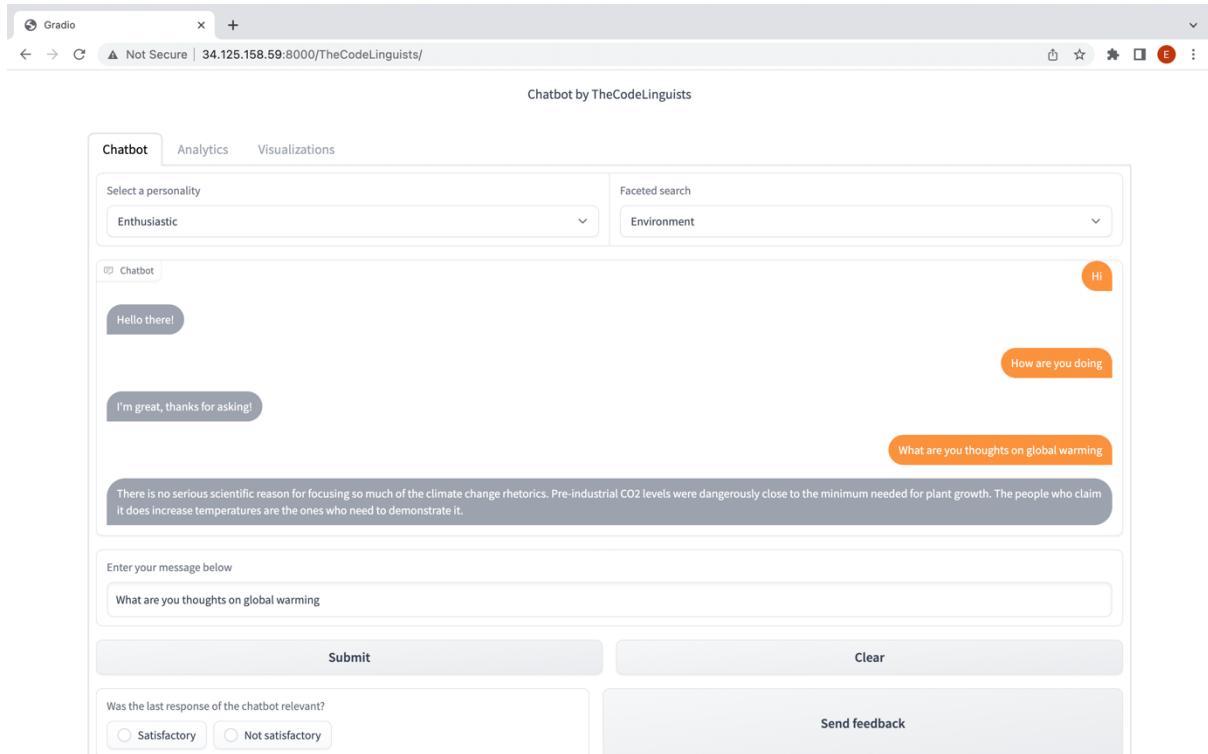
- Choose bot personalities under “Select a personality” (Witty, Enthusiastic, Friendly, Professional, Caring). Users can select a persona for the chatbot and the chatbot will align with the selected personality for general ChitChat queries.



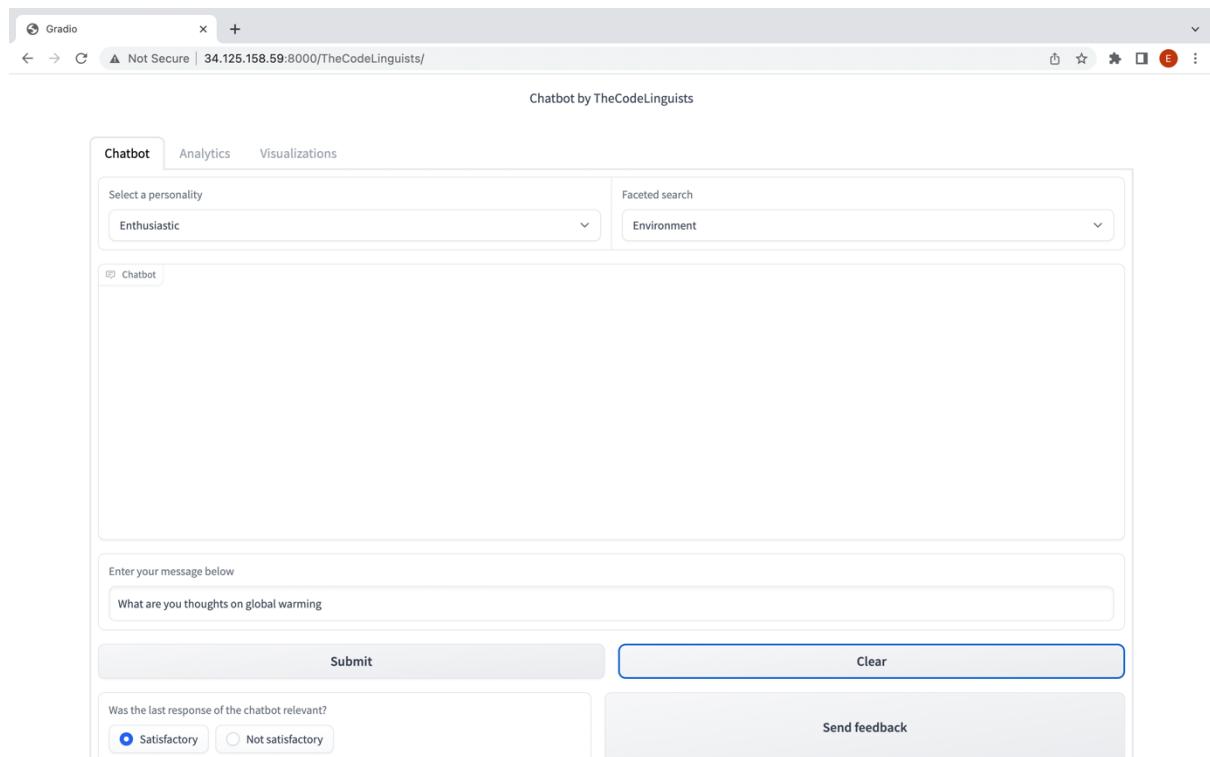
- Use a faceted search from the following topics: Healthcare, Education, Politics, Environment, Technology



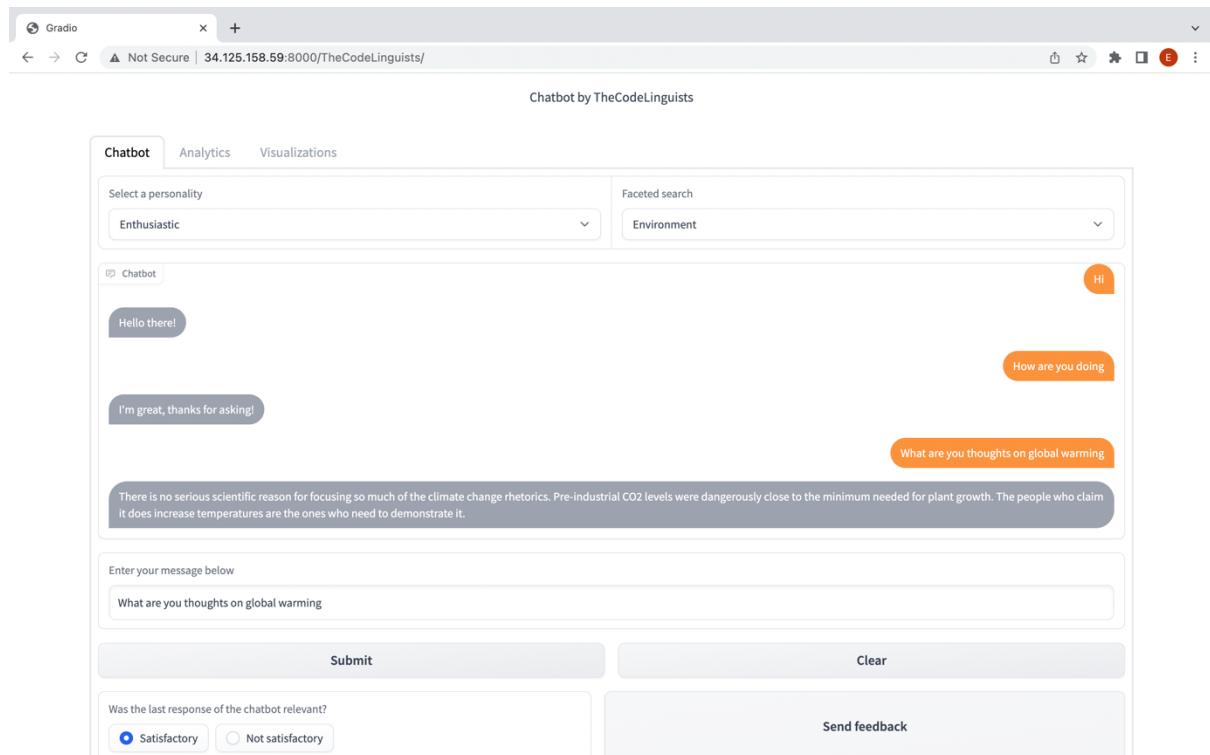
- Submit a query for response. The chatbot comes up with the most relevant response which is shown in the same user interface in a conversational style.



- Clear the chats for a new user/session



- Provide user feedback per query response as Satisfactory or Not satisfactory



On the Analytics tab the user can

- View a detailed analytics data related to a query

- Important statistics and information regarding the last retrieved response are displayed here. Some of the significant statistics that are displayed here for each response are searched index, classifier probability for the searched index, boosted terms (rare terms and context terms) along with displaying the top 50 retrieved documents.

Chatbot by TheCodeLinguists

Analytics

Analytics of the retrieved response

Detailed analytics for the chatbot's last response

Query : How do you enhance technology
 Response : NASA is a bloated government agency. Free market capitalism is the force you seek. Smaller start up companies are able to disrupt if allowed to operate and not get strong armed by crony capitalism. SpaceX, Tesla, Uber etc all have great talent and they are at the perfect size to innovate fast and efficiently.
 Index searched : Reddit
 Classifier probability for searched index : 1.0
 Rare terms boosted : ['tech', 'innovation', 'systems', 'improve', 'maximize', 'enhances']
 Entities boosted : []
 Context terms boosted : []
 Top retrieved docs : [body : Who Else just scrolled for the comments ? id : hdhmmhf parent_body : New US Crypto Regulation Far More Invasive Than We Thought
 US Congress intends to regulate crypto on a level far deeper than currently understood—They will:
 * Designate Bitcoin, Ether, and their hard-forks as commodities and regulate their transactions accordingly;
 * Create legal uncertainty for all other crypto projects and ICOs by allowing them to be labeled as securities;
 * Ban the use of (unauthorized) stablecoins;
 * Introduce penalties for the use of mixers and privacy coins;

Chatbot by TheCodeLinguists

Analytics

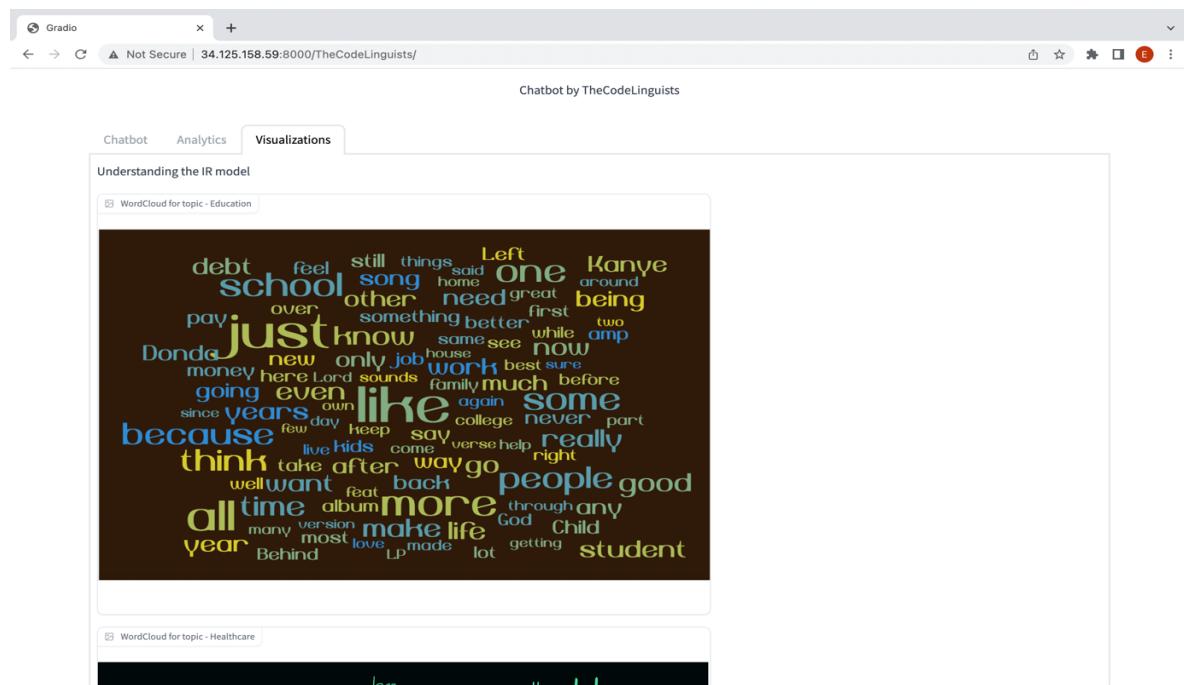
Analytics of the retrieved response

Detailed analytics for the chatbot's last response

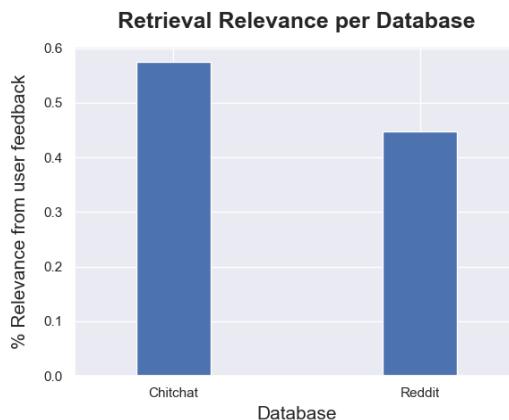
Query : What are your thoughts on global warming
 Response : There is no serious scientific reason for focusing so much of the climate change rhetorics. Pre-industrial CO2 levels were dangerously close to the minimum needed for plant growth. The people who claim it does increase temperatures are the ones who need to demonstrate it.
 Index searched : Reddit
 Classifier probability for searched index : 1.0
 Rare terms boosted : ['social', 'asia', 'economic', 'climate', 'heating', 'climatechange']
 Entities boosted : []
 Context terms boosted : []
 Top retrieved docs : [body : A group of deep sea divers discover the ruins of a large structure in what was once called " Puratchi Thalaivar Dr. M.G. Ramachandran Central Railway Station" . id : hdadm7a parent_body : Articles about global warming are a little unsettling knowing that what will happen if we don't tackle climate change. So what do you think will happen to Chennai if we don't take global warming seriously? score : 0.0010901636285128853 bm25 : 0.003140505682401554 desm : -0.0014355247840285301], [body : I guess there won't be any Chennai. id : hdac9k7 parent_body : Articles about global warming are a little unsettling knowing that what will happen if we don't tackle climate change. So what do you think will happen to Chennai if we don't take global warming seriously? score : 0.0024846424670639298 bm25 : 0.003140505682401554 desm : 0.006084454245865345], [body : they are not elite. Never were. Not smarter, prettier or a better human.

Born into luck.

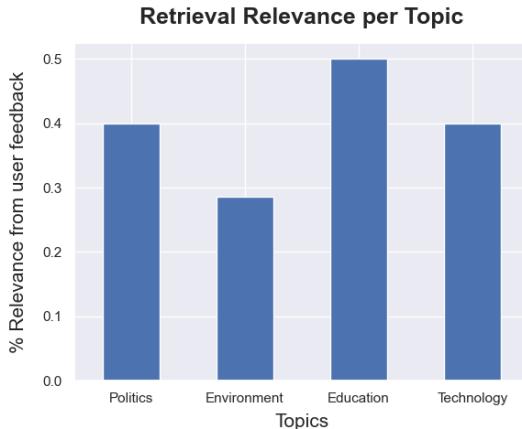
On the visualizations tab the user can



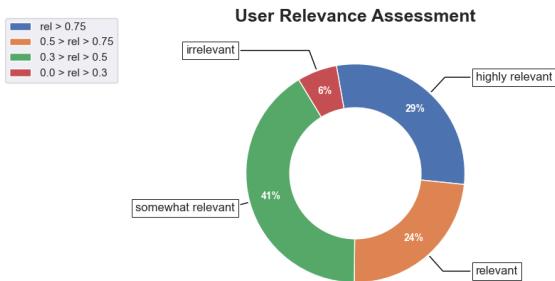
- View the word clouds for each topic: Healthcare, Education, Politics, Environment, Technology By allowing the user to provide feedback for each response, we generate several statistics that indicate the performance of our chatbot across different topics, databases and from the perspectives of different users.



- Aggregating the feedback received for each database queried, we calculate the relevance w.r.t each. We see that, when querying the chitchat dataset, more relevant responses have been found as compared to ‘Reddit’ dataset. This implies that the Reddit dataset is more complex and must cater to more diverse information needs making relevant retrieval difficult.



- Another statistic calculated is the relevance for each topic when the user has turned on faceted search. The above chart shows that the chatbot performed the best when the topic was ‘Education’, followed by ‘Technology’ and ‘Politics’. ‘Environment’ related queries were the hardest for the chatbot.



- Lastly, to measure the overall satisfaction of users and their information needs, we aggregate data for 20+ different chat sessions. We notice that the majority of the users found the responses somewhat relevant. Almost 1/3 of the users found the chatbot’s responses highly relevant (more than 2/3 of the responses in a conversation marked relevant). Further only 6% of total users did not have their information needs met and found most of the responses in their conversation irrelevant.

- To understand our dataset and the most retrieved documents across different topics, we generate word clouds.

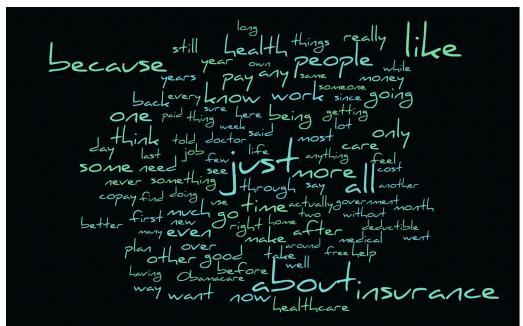
Education



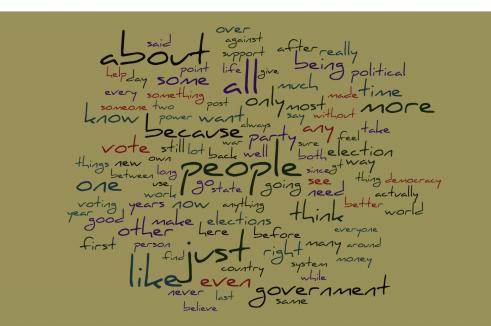
Environment



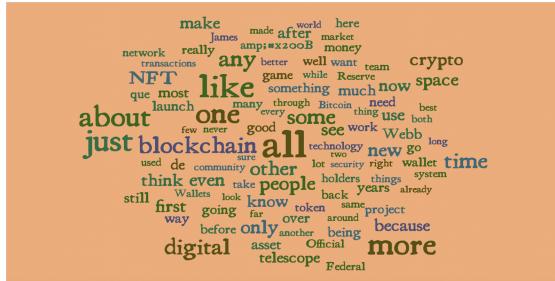
Healthcare



Politics



Technology



OBSERVATIONS:

Given below is a string of 17 query responses from the chatbot with a few of them invoking faceted search. The chatbot does fairly well when the queries correspond to the domain of the data scraped. Reranking is not flawless but works in most cases. There is a potential risk of query drift when expanding the query based on similar words to the rare words in the query. This could lead to irrelevant results. Facet search also works fairly well; sometimes it improves the relevancy of the responses, at other times, it leads to fetching irrelevant information. A trivial workaround would be to simply fetch documents boosted on the rare terms from the query. While this could work, it does not generalize well when catering to the user's information needs. It may or may not improve precision depending on the query, but it will diminish the recall of the system. Lastly, the chatbot has a latency of between 5-10 seconds when fetching results from the Reddit index. This is due to the complexity of the model which has been described above. We wanted to build a complex model which would present a greater opportunity to learn.

Chatbot by TheCodeLinguists

The screenshot shows the 'Chatbot' interface by TheCodeLinguists. At the top, there are three tabs: 'Chatbot' (selected), 'Analytics', and 'Visualizations'. Below the tabs, there are two dropdown menus: 'Select a personality' (set to 'Witty') and 'Faceted search' (set to 'None'). A large orange button labeled 'Hi' is positioned on the right side of the interface. The main area shows a conversation history:

- User: Hello there!
- Chatbot: Hi
- User: How nice!
- Chatbot: What's the weather like today?
- User: To be honest, I am on the fence
- Chatbot: Do you believe in god
- User: When is the US election

At the bottom, there is a text input field with placeholder text 'Enter your message below' and a message 'what's good about James Webb Telescope'. Below the input field are two buttons: 'Submit' and 'Clear'. A feedback section asks 'Was the last response of the chatbot relevant?' with radio buttons for 'Satisfactory' (selected) and 'Not satisfactory'. A 'Send feedback' button is located to the right of this section.

Chatbot by TheCodeLinguists

Chatbot Analytics Visualizations

Select a personality Faceted search

Witty None

Chatbot *Just, I am on the fence*

"Truth is in the middle" is marker of three categories of people: paid trolls, people trying to justify their deeds and shallow-minded victims. The criminals who steal the elections are not boys who throw-in ballots, but people from the commissions who steal it and fake records in books.

When is the US election

Compare alternatives to FPTP on Wikipedia and check out [ElectoWiki] to better understand the idea of election methods. Consider finding a good place for your contribution in the [EndFPTP subreddit wiki]. I am a bot, and this action was performed automatically.

Who will win the election

Donald Trump is not fit for office and doesn't have the competence to carry out the job, according to John Bolton, Trump's appointed National Security Advisor. Donald Trump is the first president in my lifetime who does not try to unite the American people, but instead, he tries to divide them.

Who is Donald Trump

Enter your message below

what's good about James Webb Telescope

Submit Clear

Was the last response of the chatbot relevant?

Satisfactory Not satisfactory

Send feedback

Chatbot by TheCodeLinguists

Chatbot Analytics Visualizations

Select a personality Faceted search

Witty None

Chatbot *people, but instead, he tries to divide them.*

Mobilecoin foundation declares there will be 250,000,000 mobilecoins and the bitcoin inventor declares 21,000 bitcoins. I'm not a bitcoin maximalist. I think deflationary fixed-supply coins don't make sense in the modern world and blockchains are inefficient linked lists. Mining has the side-effect of costing money.

Is bitcoin secure

This is the best I could find. The sun is going to supernova and destroy all living things on earth in about 5 billion years. The only thing that matters is having enough food, shelter, people we love, and people who love us back. The people who suffered in Nazi KZs may have had shelter, a little bit of food, some people they loved and some people who loved them back, but it wasn't worth living.

to the moon

There is a UN article that says global deserts are expanding. If someone questions whether the Amazon rainforest is threatened based on review of satellite maps, I would give them a link to the additional evidence that is available to support the claim, not ask them a snarky question.

What do you think about deforestation in the Amazon forest

Enter your message below

what's good about James Webb Telescope

Submit Clear

Was the last response of the chatbot relevant?

Satisfactory Not satisfactory

Send feedback

Chatbot by TheCodeLinguists

Chatbot Analytics Visualizations

Select a personality: Witty | Faceted search: None

Chatbot: *task them a snarky question.*

I only answer to the call of destiny.

who's your boss

Sure. Take me to city hall. See what happens.

will you marry me

I'm a high school teacher in the US. Florida just gave starting teachers a \$7,000 raise, but they are having trouble filling the positions. I'm one of the few people at my site who went to college to specifically be a teacher.

What are the best local schools

Will student loans be forgiven

Enter your message below: what's good about James Webb Telescope

Submit | Clear | Send feedback

Was the last response of the chatbot relevant?

Satisfactory | Not satisfactory

Chatbot by TheCodeLinguists

Chatbot Analytics Visualizations

Select a personality: Witty | Faceted search: None

Chatbot: *I'm a high school teacher in the US. Florida just gave starting teachers a \$7,000 raise, but they are having trouble filling the positions. I'm one of the few people at my site who went to college to specifically be a teacher.*

Will student loans be forgiven

They better not come out with student forgiveness, maybe student protection but not forgiveness. It's unfair to those that payed off student debt or have never taken on student debt. - John. He doesn't want to see student forgiveness or protection.

what's good about James Webb Telescope

This is the best I could find. Bryn Mawr is a good area to look at. It's still listed as pending, so I don't think it sold for 225k. I bought a place for 275k near Powderhorn that brings in 2725 gross (1375 and 1350). If I were you I'd focus on location and rental yields for your first property.

Did obamacare make healthcare more accessible

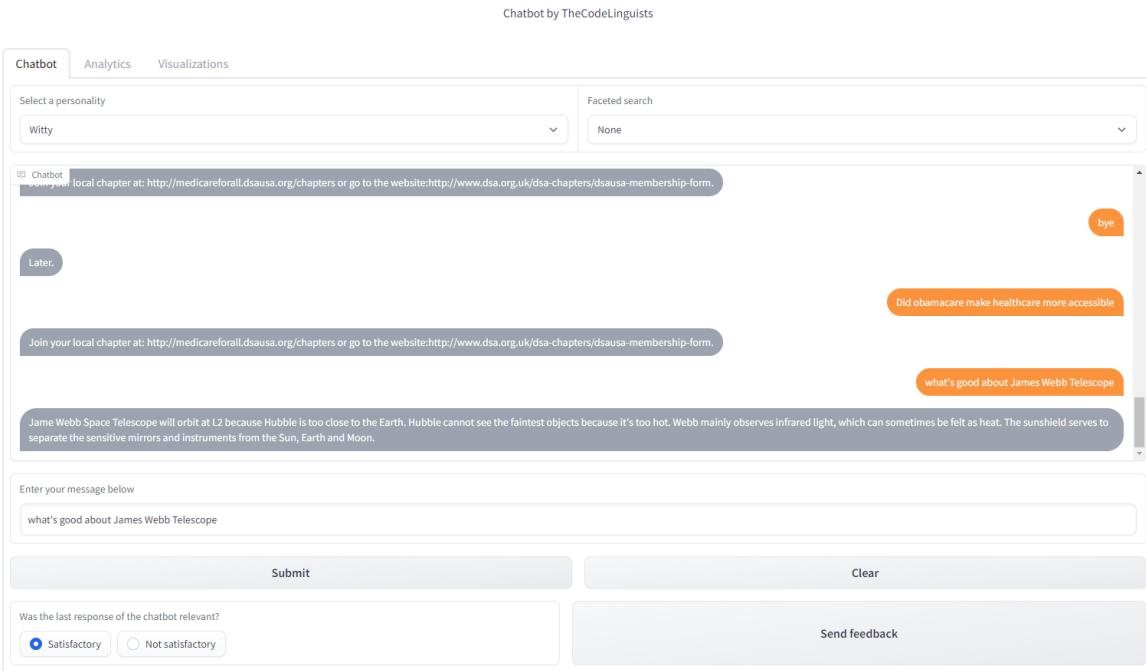
Join your local chapter at: <http://medicareforall.dsusa.org/chapters> or go to the website: <http://www.dsa.org.uk/dsa-chapters/dsusa-membership-form>.

Enter your message below: what's good about James Webb Telescope

Submit | Clear | Send feedback

Was the last response of the chatbot relevant?

Satisfactory | Not satisfactory



Alternatives explored

- We implemented Dual Embedding Space Model, but we've hit a snag with accessing output layer embeddings on pretrained Glove models. You can use the out-layer embeddings by training on the data available, but the performance wasn't good enough compared to the pre-trained embeddings.
- We explored Solr's Learning to Rank feature, however, since we don't have the relevance judgement data to achieve that. We contemplated constructing random queries from the corpus and using the similarity scores as a proxy for relevancy, but it seemed like an inefficient approach which would take a long time to build and train the models with random queries.
- For context management, we've attempted to base the context continuity on the similarity between the query and previous response, but this has proven to be ineffective compared to taking the similarity between contiguous query pairs. Due to the size of the responses, and presence of many other terms, similarity is not being robustly captured.

CONTRIBUTION

Praneeth Nekkalapudi:

Data scraping/cleaning, Solr schema configuration and indexing, Implemented the classifier, retrieval, reranking and context tracking logic.

Eva Pradhan:

Integrated the Database, Visualisations, alternatives, report

Sumedh Khodke:

Frontend and Backend for the chatbot.

Jay Lal

Visualizations, Deployment, Assisted in designing classifier and context logic.

CONCLUSION

We were able to implement a decent chatbot, that performs several actions which we've learned in the course (Indexing, Retrieving, Ranking, Reranking, Summarization, Embeddings, Lemmatization, POS tagging, Named Entity detection, boosting, indexing and query filters). The chatbot caters to the information needs of the user moderately well. Plenty of scope to improve the model in every area, especially context tracking and reranking which will be explored in the future.

RESOURCES:

- Twitter glove embeddings (100 D) for word expansions
- Sentence transformer for phrase summarization and reformulation:
<https://huggingface.co/docs/transformers/index>