

Mid-Term Project Review 1

Sports History and Records Archives

(Group 45)



AUTHORS:

VANSHAJ SHARMA(MT23103)

PULKIT RIHANI (MT23066)

BHARAT NAGDEV (MT23029)

RITESH RAJPUT (MT23075)

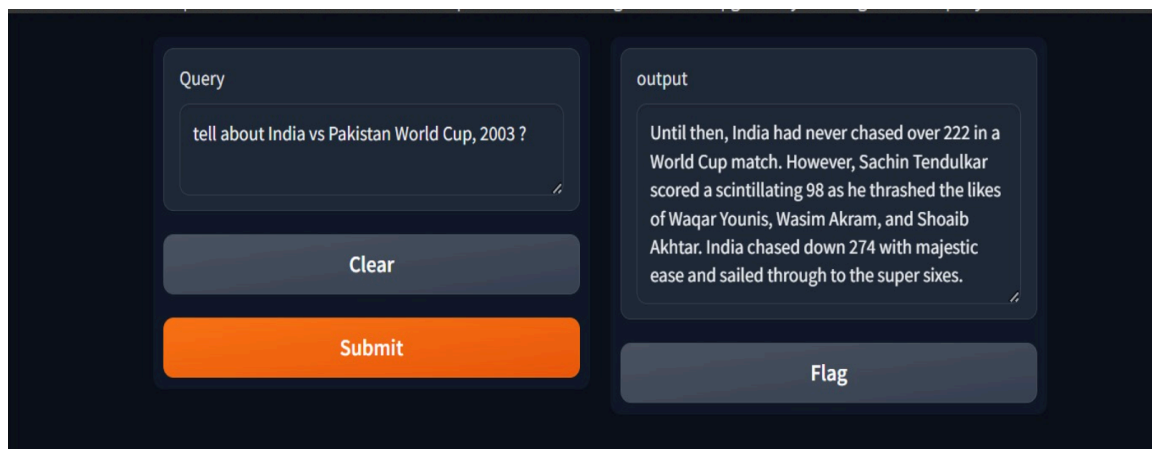
LOKESH SAINI (MT23120)

SIDDHANT JHA (MT23097)

Problem

The identified problem revolves around the lack of a dedicated chatbot focused on retrieving sports history and records, despite the increasing interest and importance of memorable sports events. Our project highlights the need for an efficient and optimized chatbot solution capable of providing users with relevant sports history information through simple interactions. Our proposed solution is utilizing the power of LLMs and RAG model to provide rank based retrieval and relevant results by understanding user's need by the feedback mechanism. So far, no rank-based retrieval model along a chatbot has been developed for the domain, and based on our project, our system will be leveraging the rank-based retrieval of url for query-specific tasks along with the url it will extract a short summary for the same along with a special event date and its headline.

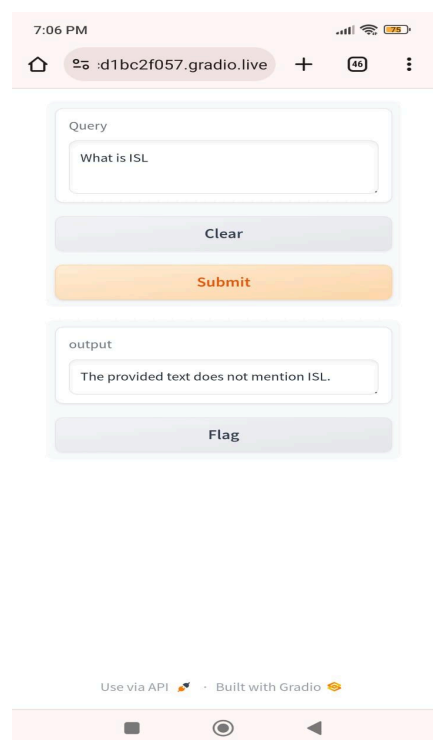
Results Till baseline



The screenshot shows a chatbot interface with a dark background. On the left, there is a 'Query' input field containing the text 'tell about India vs Pakistan World Cup, 2003 ?'. Below the input field are two buttons: 'Clear' and 'Submit'. On the right, there is an 'output' field containing the text: 'Until then, India had never chased over 222 in a World Cup match. However, Sachin Tendulkar scored a scintillating 98 as he thrashed the likes of Waqar Younis, Wasim Akram, and Shoaib Akhtar. India chased down 274 with majestic ease and sailed through to the super sixes.' Below the output field is a 'Flag' button.

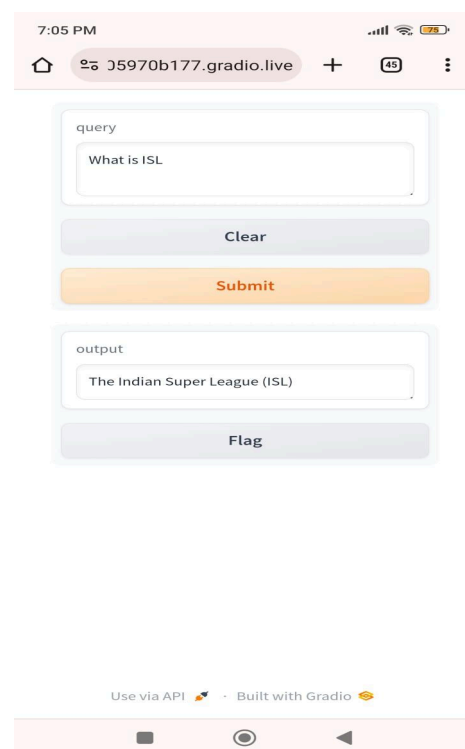
Results at mid term project review:

Old Model on csv wiki cricket dataset



The screenshot shows the 'Old Model' chatbot interface. The 'Query' input field contains 'What is ISL'. The 'output' field contains the text: 'The provided text does not mention ISL.' The interface includes 'Clear', 'Submit', and 'Flag' buttons. The status bar at the top shows the time as 7:06 PM and the battery level at 98%.

New Model on csv wiki cricket dataset



The screenshot shows the 'New Model' chatbot interface. The 'query' input field contains 'What is ISL'. The 'output' field contains the text: 'The Indian Super League (ISL)'. The interface includes 'Clear', 'Submit', and 'Flag' buttons. The status bar at the top shows the time as 7:05 PM and the battery level at 97%.

Old Model on csv dataset of football

Query

Roman Abramovich



Clear

Submit

output


The provided context does not mention Roman Abramovich.

Flag

New Model on csv dataset of football

query

Roman Abramovich



Clear

Submit

output

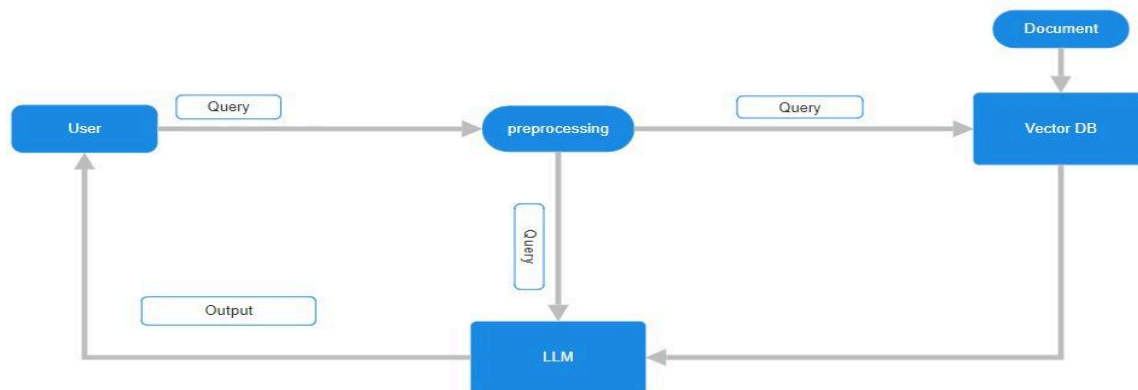
front-of-shirt deal with Telecom's brand 3, and its \$72m (£55m) kit deal with Nike," says Conrad Wiacek, Head of Sport Analysis at GlobalData, a leading data and analytics company. "While Chelsea has a sporting licence to continue trading as a soccer club, many brands will be wary of guilt by association. "Chelsea FC is still one of the biggest clubs in the world and its on-field success still makes it an attractive commercial partner. However, given the rate at which many brands are looking to dissociate themselves from the Russian state, some may be wary of continuing partnerships. "Nike's deal with Chelsea runs until 2032, so the apparel brand may decide to wait the situation out until the club's sale is able to continue. However, brands such as Hyundai and Hublot, which have deals worth over \$20m (£15m) combined expiring at the end of 2021-22 season, may not have that luxury." Chelsea will be allowed to spend some money on staging matches at Stamford Bridge and in travelling to away matches. A spending limit of just £20,000, however, applies to travel to away games, and could cause issues with preparations for Champions League fixtures in particular. The club will also have to prove that such costs are 'reasonable costs'. Chelsea are also barred from receiving any revenue from merchandising sales, but will be allowed to take revenue from broadcasting matches. Abramovich is said to be worth £9 billion (\$12bn) and valued Chelsea at £3bn (\$4bn) during his negotiations to sell the club. It is currently unclear if or when the sanctions will be lifted. It relates directly to Russian President Vladimir Putin's activity and the invasion of Ukraine. Sanctions will be reviewed in May, but there are further risks of a nine-point deduction should Chelsea go into administration. In the statement released by the UK Government, it describes Abramovich as a "pro-Kremlin oligarch who has been involved in destabilising Ukraine." He is further described as having a "close relationship with Putin" and providing steel for tanks in Russia's offensive. Prime Minister Boris Johnson stated in his government's release: "There can be no safe havens for those who have supported Putin's vicious assault on Ukraine. "Today's sanctions are the latest step in the UK's unwavering support for the Ukrainian people. We will be ruthless in pursuing those who enable the killing of civilians, destruction of hospitals and illegal occupation of sovereign allies."

Flag

Difference Between the approach of model at baseline and mid term 1

Methodology:

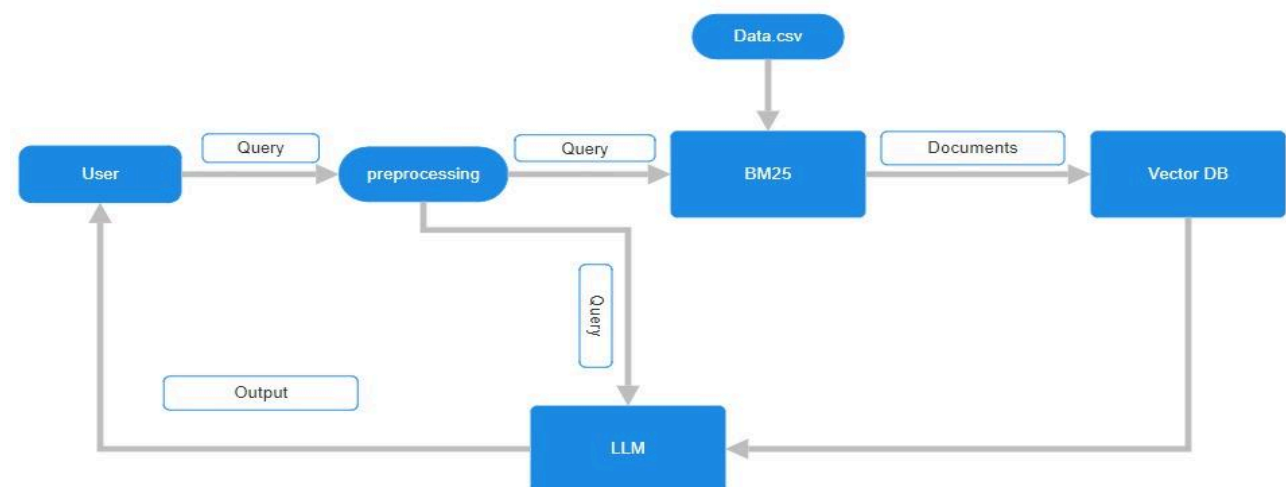
Pipeline diagram of RAG without BM25:



Pipeline Description:-

Over the pipeline on receiving the user query, preprocessing is done on the received query, based on the user query and the context from the document leading to vector DB and passed to the LLM for the result fetching over the user query.

Pipeline diagram of RAG + BM25 :-



Pipeline Description:-

Over the pipeline on receiving the user query, preprocessing is done on the received query and based on the user query and content in the Data.csv file containing the dataset having 2 columns: title and content, based on the user query and BM25 ranking top 2 most relevant result are fetched from the dataset and are combined and leading to vector DB and finally passing onto LLM for efficient query result retrieval.

Evaluation/Accuracy Results:

Accuracy metric used :- Rough Score

For the evaluation Rouge Result is used which is the result of average of all 3 extracted rouge scores i.e. - (Rouge-1,Rouge-2,Rouge-3) and leading to the following results.

Accuracy results for comparing old model and new model:

Using of wiki cricket dataset:

```
For Dataset-1

sum_new1 = data1["New Rouge Score"].sum()
count1= data1.shape[0]
acc_new1 = (sum_new1/count1)*100
sum_old1 = data1["Old Rouge Score"].sum()
acc_old1 = (sum_old1/count1)*100

acc_old1

39.891270144666

acc_new1

76.47619005882858
```

Using of Web crawled Football Dataset:

```
For Dataset-2

sum_new2 = data2["New Rouge Score"].sum()
count2 = data2.shape[0]
acc_new2 = (sum_new2/count2)*100
sum_old2 = data2["Old Rouge Score"].sum()
acc_old2 = (sum_old2/count2)*100

acc_old2

20.40603998204463

acc_new2

84.21052589473685
```

Proposed Method

Topic modeling using LDA:

For data analysis we are going to use LDA i.e Latent Dirichlet Allocation for Topic modeling. Currently we are using a dataset with two columns that are URL and Content. We are using BM25 to get the top 2 most relevant results based on relevancy score. But we get only 84 percent accuracy measured using Rouge as accuracy metrics on a small dataset only. So for improving the results we are planning to do topic modeling and create another column called Headings based on the values retrieved from the LDA model.

Llama 2 as LLM:

Also we are currently using Gemini pro as LLM but the thing is we cannot fine tune it because its parameter is known google not to us. So dealing with this situation we are planning to implement Llama2 as our new LLM in place of gemini pro. As Llama2 provides us the parameters using which we can fine tune it and make it useful for specific purposes in our case it is for a sports archive model.

Mongo DB as vector database:

Also Currently we are using Chroma db as the vector database but for more features and better vector indexing, we are planning to shift towards Mongo db because it provides more features and latest updates about vector indexing. Also Mongo db currently releases more features about multimodal data, so that we can give multimodal outputs as well very easily.

Also we are currently using BM25 for filtering the documents over content and giving documents to the Vector database but as BM25 is not context aware so we are getting like 84 percent accuracy even on 20 questions testing the dataset. So we are planning to use some kind of Context filtering technique so that we could increase our model accuracy.

Test-dataset over football dataset:

<https://docs.google.com/spreadsheets/d/1UCedZzTzs00ooYzWlYiR738Q4FW5-h-pOwgEvrdNG9o/edit#gid=0>

Github: <https://github.com/siddhantJH/Information-Retrieval-Project>