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)

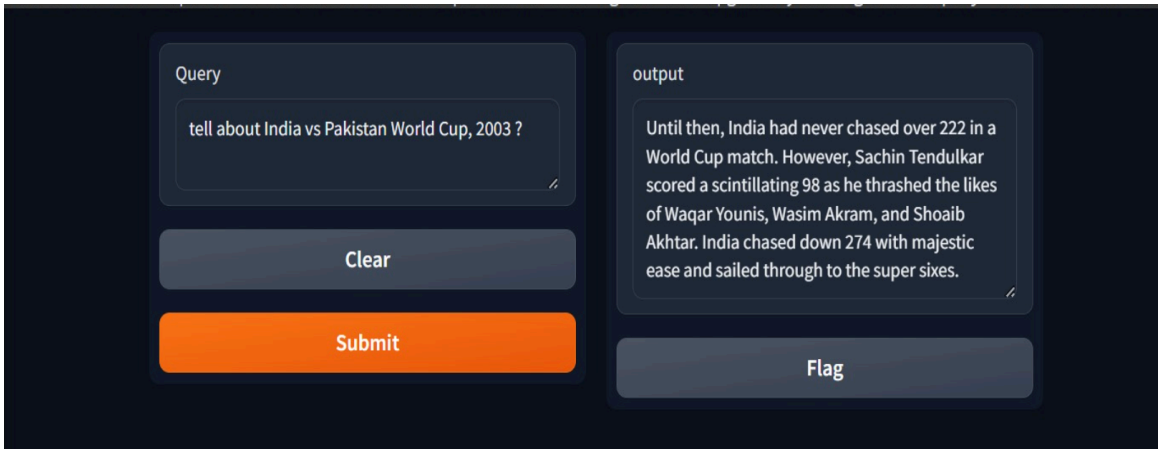
LOKESH SAINI (MT23120)

SIDDHANT JHA (MT23097)

Problem

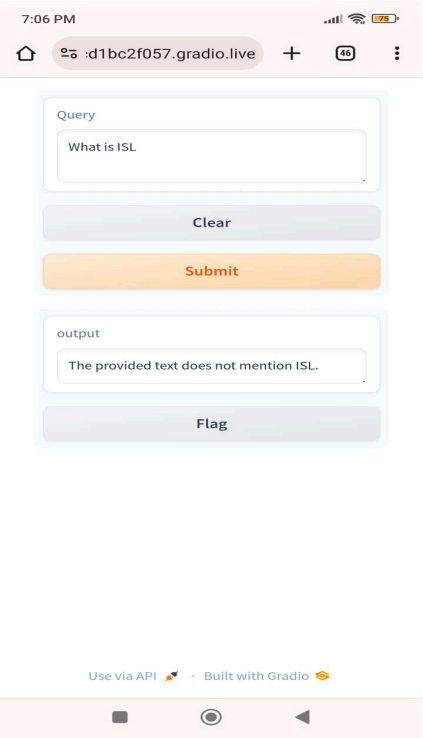
The identified problem revolves around the lack of a dedicated chatbot focused on retrieving sports history and records, despite the increasing interest and importance of memorable sports events. Our project highlights the need for an efficient and optimized chatbot solution capable of providing users 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

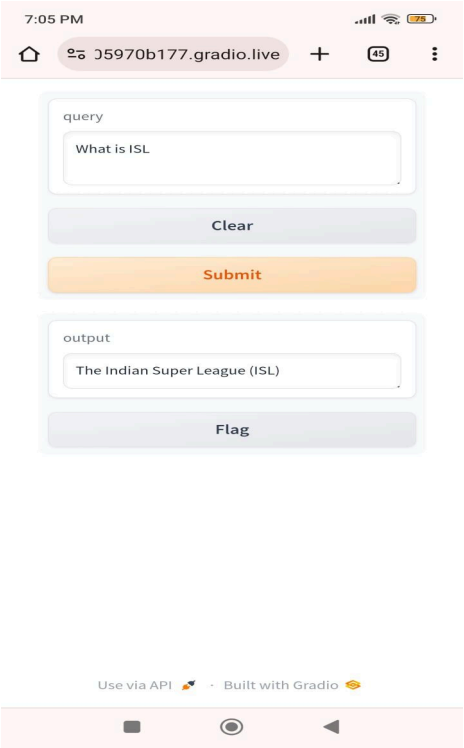


Results at mid term project review:

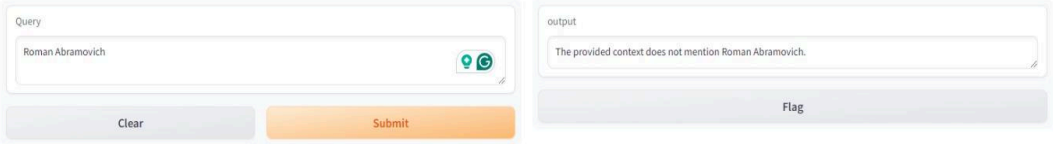
Old Model on csv wiki cricket dataset



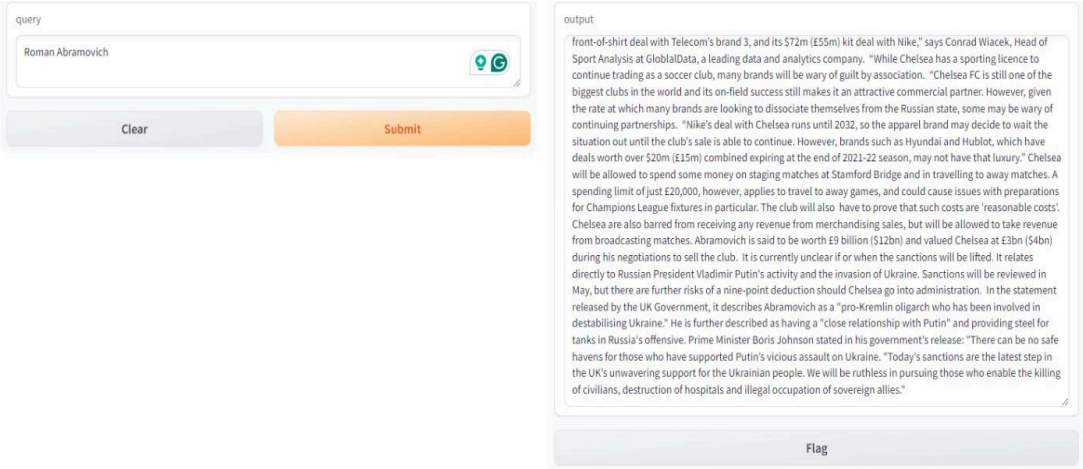
New Model on csv wiki cricket dataset



Old Model on csv dataset of football



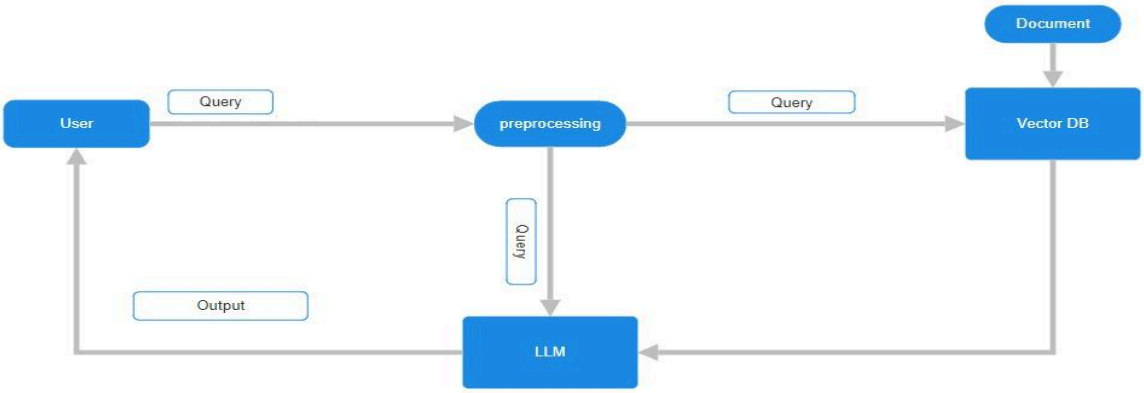
New Model on csv dataset of football



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

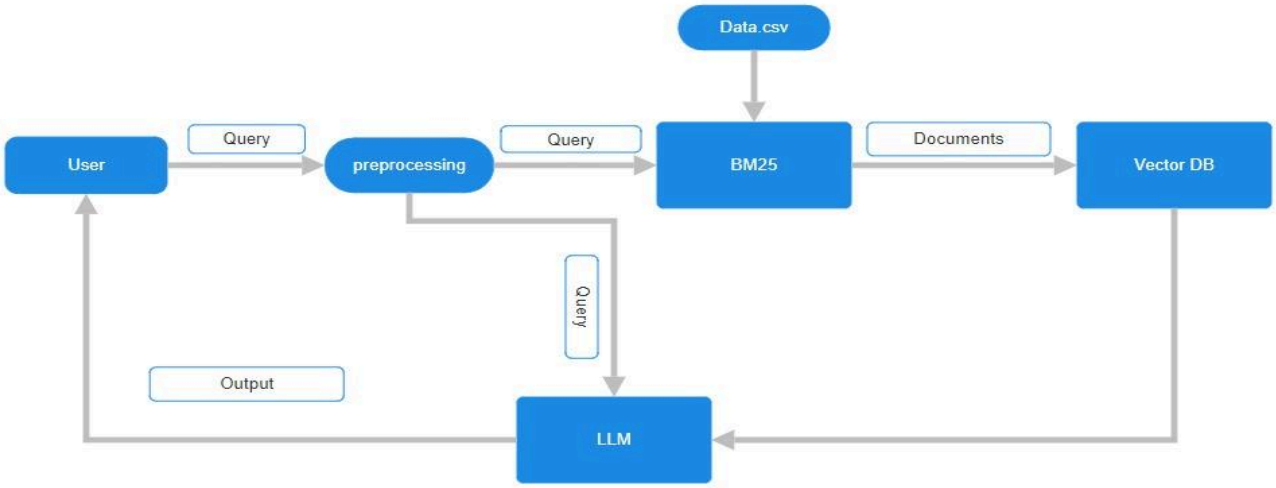
Methodology:

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
```
[28] Python

```
acc_old1
```
[29] Python

```
39.891270144666
```

```
acc_new1
```
[30] Python

```
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
```
Python

```
acc_old2
```
Python

```
20.40603998204463
```

```
acc_new2
```
Python

```
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/1UCedZzTzsOOooYzWlYiR738Q4FW5-h-pOwgEvrdNG9o/edit#gid=0

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