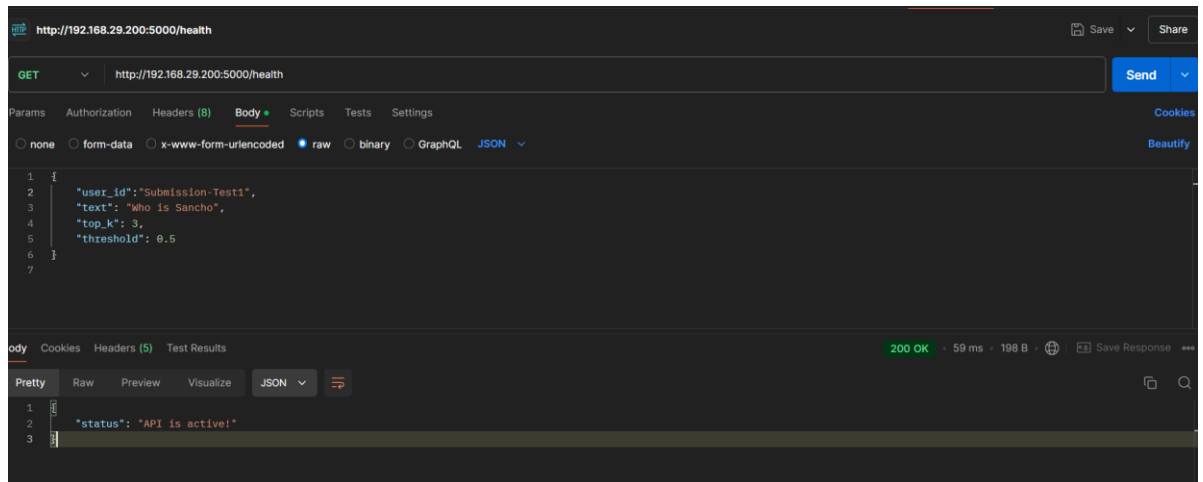


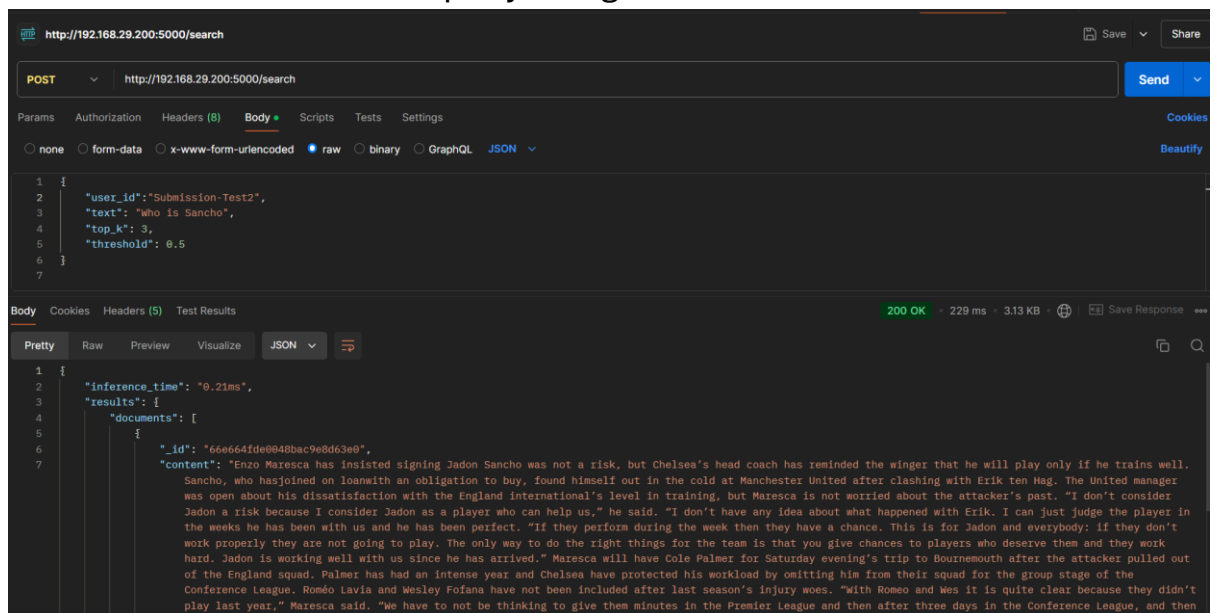
DOCUMENT - RETRIEVAL SYSTEM

1. API Design

- `health` endpoint:
Checks if the API is running or not



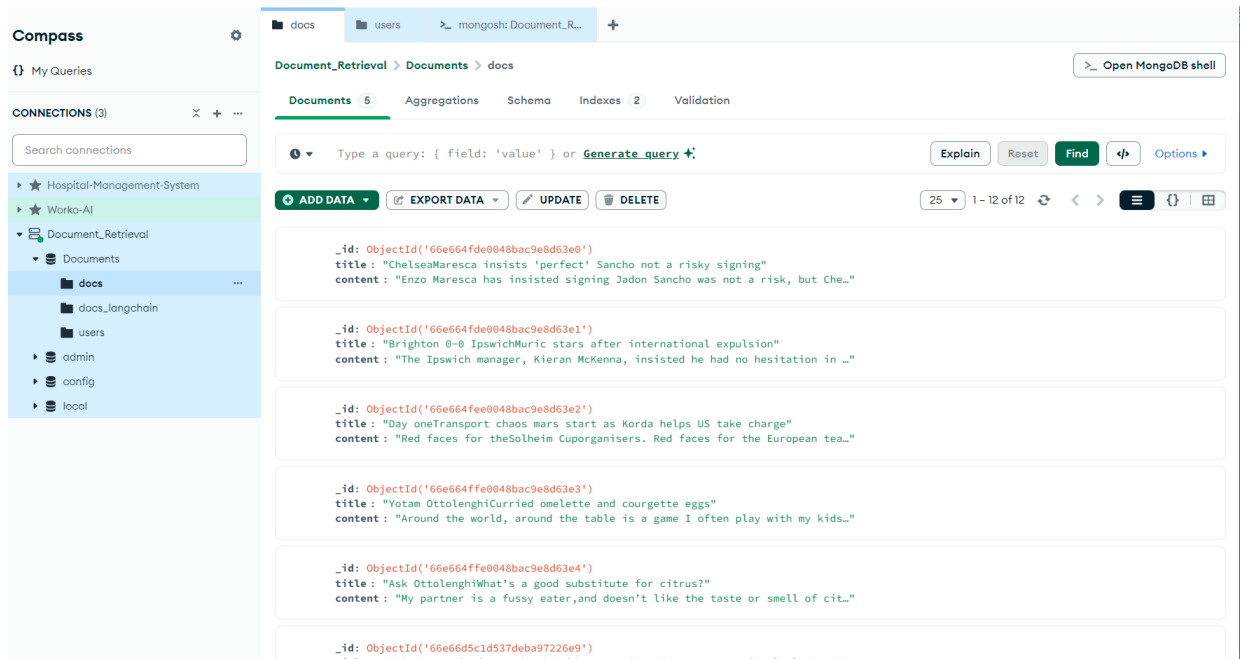
- `search` endpoint
Returns the results for the query along with inference time



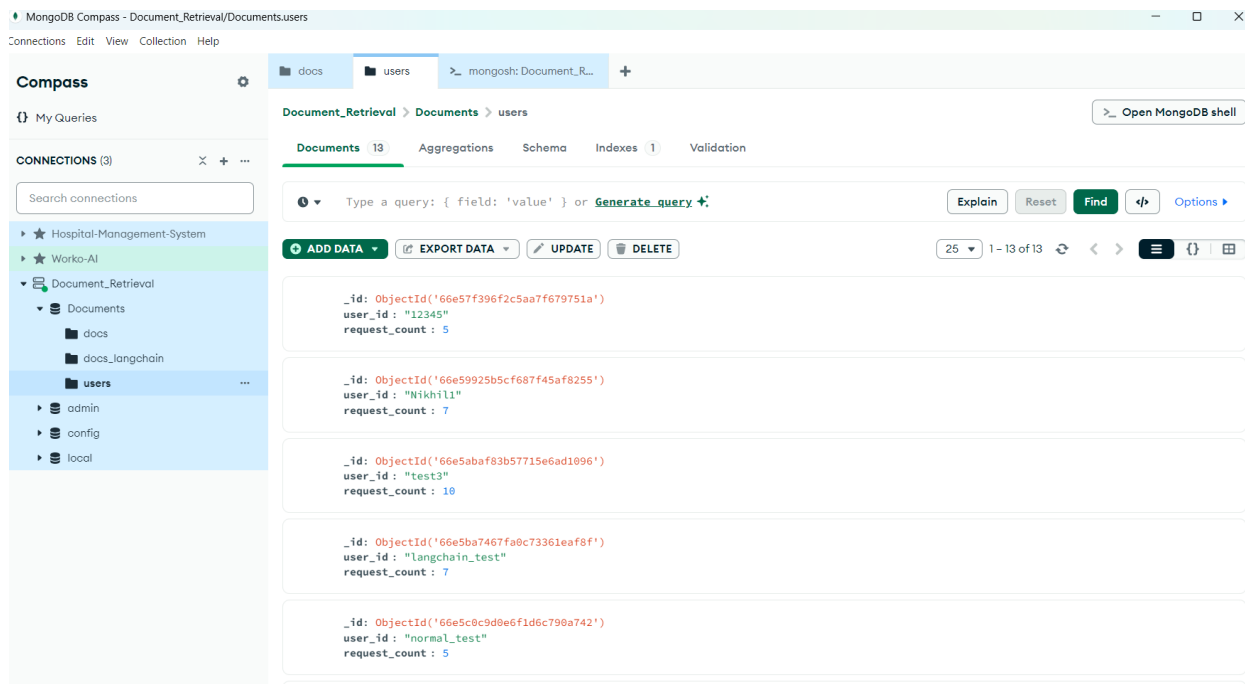
2. Database Interaction

- The database used is **MongoDB** document database
- Two collections namely docs and docs and users are created. Docs containing the news articles web scraped from **The Guardian News**, the thread for which will be running in the background when server runs.
- Users collection contains the user_id of the Users and the no. of requests the user gave for the search endpoint of the API.

- Docs Collection



- Users Collection



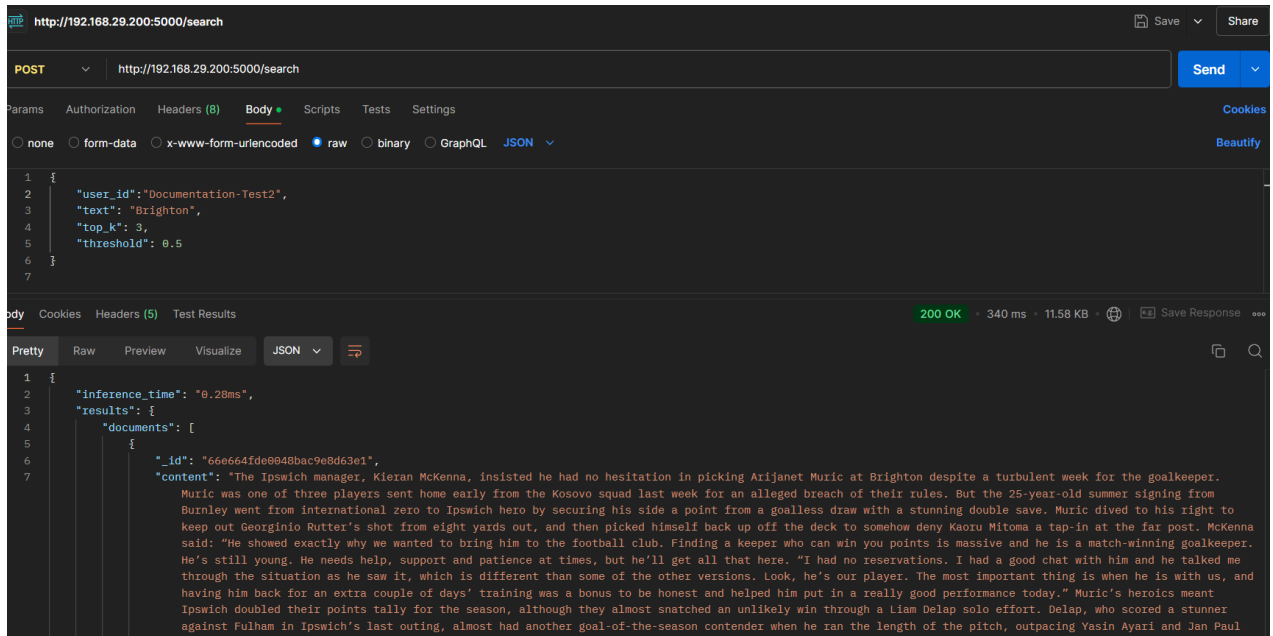
3. Concurrency

Threads has been used to scrape through the News Website and add the scraped content to the database.

4. Caching and Optimization

This project uses Redis as the caching mechanism to improve the performance of a document retrieval system. Caching is essential in reducing the load on the database and improving the overall speed of repeated queries. This section explains why Redis was chosen and the caching method implemented.

In-Memory Storage for High Performance: Redis stores data in memory, which allows for incredibly fast data access. This makes it an ideal choice for caching, as results can be retrieved much faster than querying a database. Since search results in our document retrieval system may be requested multiple times, the speed advantage of Redis significantly reduces the response time for repeated queries.

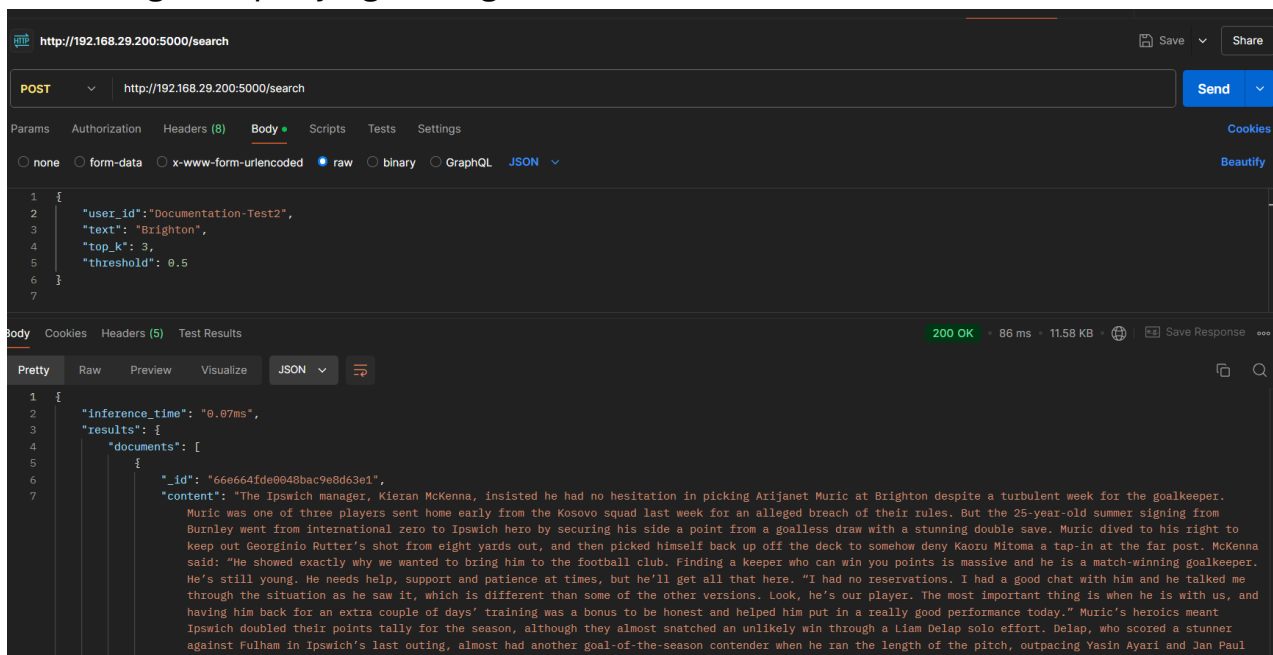


```
POST http://192.168.29.200:5000/search

{
  "user_id": "Documentation-Test2",
  "text": "Brighton",
  "top_k": 3,
  "threshold": 0.5
}
```

```
{
  "inference_time": "0.20ms",
  "results": {
    "documents": [
      {
        "_id": "66e664fde0048bac9e8d63e1",
        "content": "The Ipswich manager, Kieran McKenna, insisted he had no hesitation in picking Arjanet Muric at Brighton despite a turbulent week for the goalkeeper. Muric was one of three players sent home early from the Kosovo squad last week for an alleged breach of their rules. But the 25-year-old summer signing from Burnley went from international zero to Ipswich hero by securing his side a point from a goalless draw with a stunning double save. Muric dived to his right to keep out Georginio Rutter's shot from eight yards out, and then picked himself back up off the deck to somehow deny Kaoru Mitoma a tap-in at the far post. McKenna said: 'He showed exactly why we wanted to bring him to the football club. Finding a keeper who can win you points is massive and he is a match-winning goalkeeper. He's still young. He needs help, support and patience at times, but he'll get all that here. 'I had no reservations. I had a good chat with him and he talked me through the situation as he saw it, which is different than some of the other versions. Look, he's our player. The most important thing is when he is with us, and having him back for an extra couple of days' training was a bonus to be honest and helped him put in a really good performance today.' Muric's heroics meant Ipswich doubled their points tally for the season, although they almost snatched an unlikely win through a Liam Delap solo effort. Delap, who scored a stunner against Fulham in Ipswich's last outing, almost had another goal-of-the-season contender when he ran the length of the pitch, outpacing Yasin Ayari and Jan Paul"
      }
    ]
  }
}
```

The inference time is 0.20ms, for the first search. Now the search results would be stored in the cache memory and on subsequent searches, the server would fetch the results from the cache memory, reducing the time and optimizing the searching and querying strategies.



```
POST http://192.168.29.200:5000/search

{
  "user_id": "Documentation-Test2",
  "text": "Brighton",
  "top_k": 3,
  "threshold": 0.5
}
```

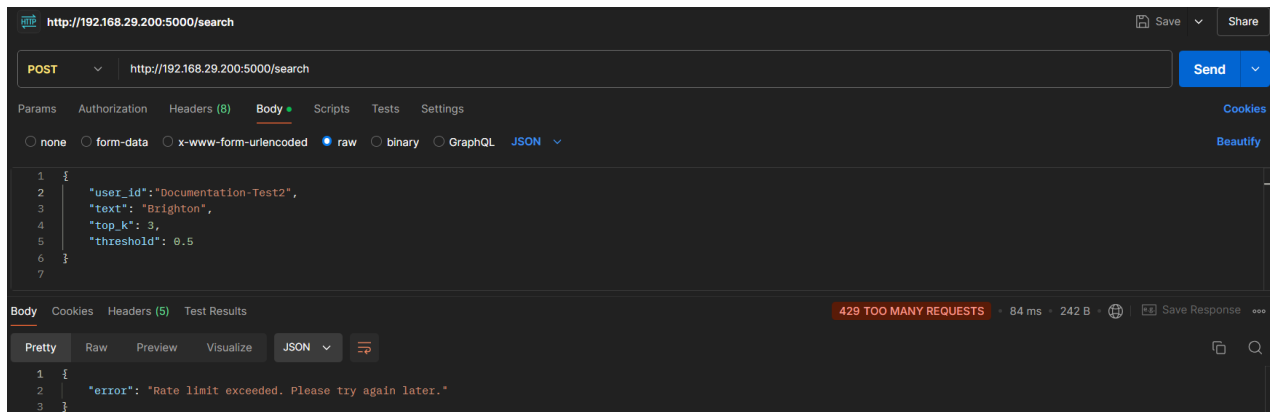
```
{
  "inference_time": "0.07ms",
  "results": {
    "documents": [
      {
        "_id": "66e664fde0048bac9e8d63e1",
        "content": "The Ipswich manager, Kieran McKenna, insisted he had no hesitation in picking Arjanet Muric at Brighton despite a turbulent week for the goalkeeper. Muric was one of three players sent home early from the Kosovo squad last week for an alleged breach of their rules. But the 25-year-old summer signing from Burnley went from international zero to Ipswich hero by securing his side a point from a goalless draw with a stunning double save. Muric dived to his right to keep out Georginio Rutter's shot from eight yards out, and then picked himself back up off the deck to somehow deny Kaoru Mitoma a tap-in at the far post. McKenna said: 'He showed exactly why we wanted to bring him to the football club. Finding a keeper who can win you points is massive and he is a match-winning goalkeeper. He's still young. He needs help, support and patience at times, but he'll get all that here. 'I had no reservations. I had a good chat with him and he talked me through the situation as he saw it, which is different than some of the other versions. Look, he's our player. The most important thing is when he is with us, and having him back for an extra couple of days' training was a bonus to be honest and helped him put in a really good performance today.' Muric's heroics meant Ipswich doubled their points tally for the season, although they almost snatched an unlikely win through a Liam Delap solo effort. Delap, who scored a stunner against Fulham in Ipswich's last outing, almost had another goal-of-the-season contender when he ran the length of the pitch, outpacing Yasin Ayari and Jan Paul"
      }
    ]
  }
}
```

On subsequent search, the inference time is reduced drastically, inferring that the results have been fetched from the cache memory.

5. Error Handling and logging

- Rate limits:

If a user searches the same query for 5 times in a row in a time window of 60 seconds, rate limit will be exceeded and the server returns a 429 error.



- Logging:

The server logs the API requests made by the user in a separate database named users. It stores the user id and the number of requests made by the user.

- Error Handling:

Proper error-handling mechanisms have been implemented, and all the edge cases have been taken account of.