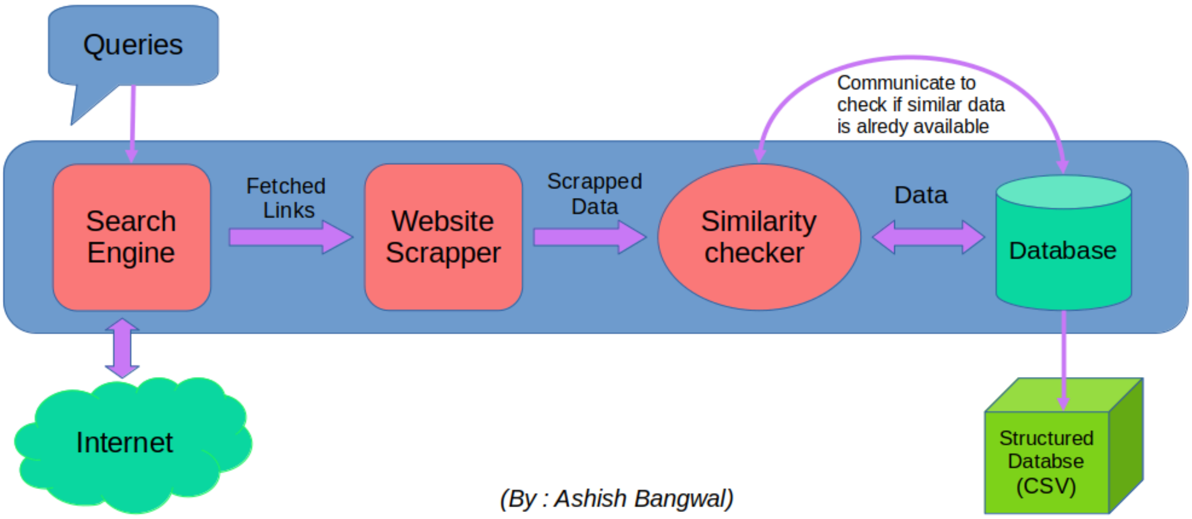
Retrieval System with Python

Internship Assignment 2024 Coding Challenge

Retrieval is the first step of the modern generation models called RAG which stands for retrieval augmented generation, these are a combination of a retrieval system and generation model such as Seq2Seq, LLM, etc. The retrieval part of the model aids and provides information that the generation model might not know, which leads to more accurate and updated responses along with fewer hallucinations and misinformation.

In the following document, I described how I built a system that can retrieve information from the internet, that can further be stored in a vector database ultimately helping RAG models.



(By : Ashish Bangwal)

(Approach / Complete pipeline implemented in Python)

Summary of the steps taken to complete the task

Step 1: Fetch links to the top result websites with the help of a search engine for this case I used DuckDuckGo, as it's open source and has a great Python API called `duckduckgo_search`.

Step 2: After collecting the top website's link I collected responses from those websites with the help of the Python `requests` library, then parsed the collected HTML response with `BeautifulSoup4` to extract the useful textual dataset.

Step 3: After receiving the scrapped data, I applied a similarity checker function which ensured that there are not too similar data for the same query to maintain diversity and new information. To implement this I use `spacy` 's `universal_sentence_encoder` NLP model to obtain a vector representation of data to compare and measure similarity.

Step 4: Once I have the scrapped and checked data I start storing them in a structured format with Python's `dict` object, once the data is collected and stored I convert `dict` into `pandas DataFrame` object which is an optimized tabular format for Python operations. Then finally, I exported it in CSV format with the pandas `.to_csv()` function.

Challenges faced and their solution

1) ***To choose from an ocean of available web scrapers*** : There are tons of web scrapers and all of them are unique in their way, after some research I narrowed it down to BeautifulSoup4, scrapy, and selenium, and after further brainstorming, I chose BeautifulSoup4 which has a legacy in web scrapping and a huge community.

2) ***Some sites have enabled anti-scrapping measures*** : Some sites have blocked bots accessing their website, which makes the scrapping a bit tricky a simple solution to this is to add `headers` with `user-agent` to bypass this.

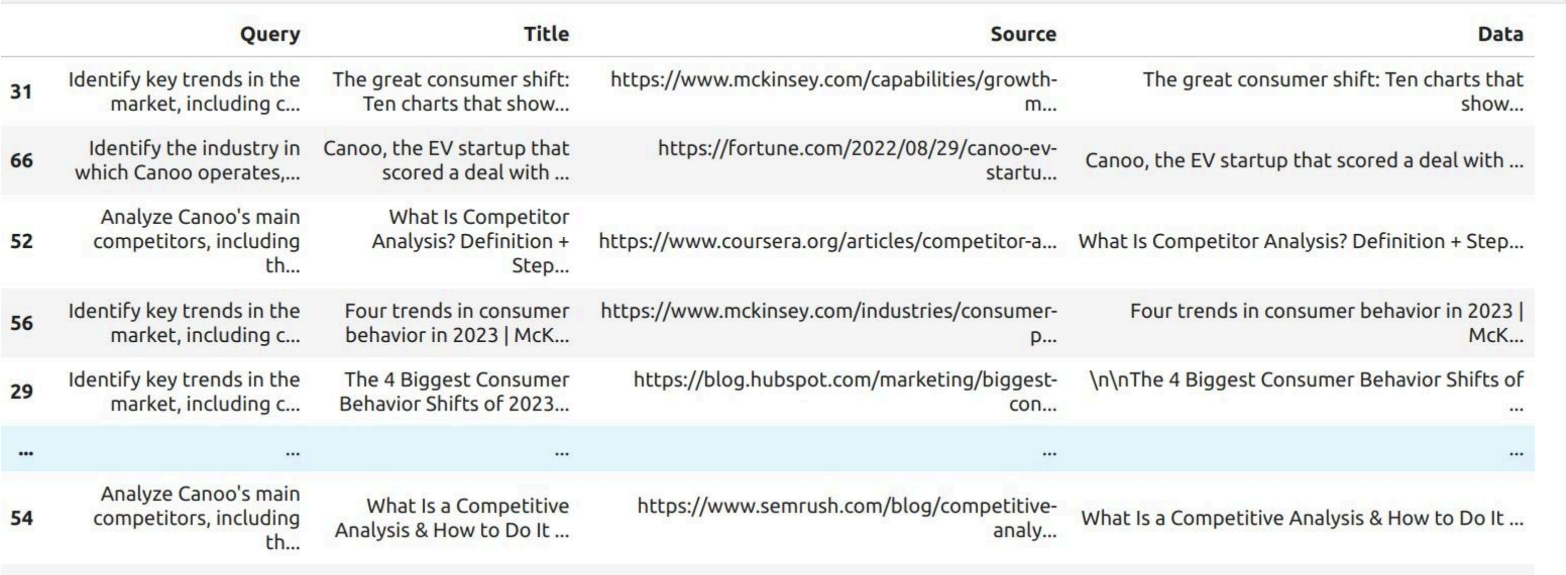
3) ***Extracting useful data from HTML response collected*** : After the collecting HTML response with the requests module, the next challenge is to extract useful information and remove all the HTML tags and other hyperlinks, for this, I use the BS4 `.getText()` method.

4) ***Multiple sources yielding similar data*** : Since I am scrapping the top 10 results from search engines for each query, there was a high chance of information repetition which will not add value to the database but consume memory resources. To solve this I ran similarity checks with SpaCy NLP models on scrapped data to not include repetitive content.

GitHub Link for deliverable :

https://github.com/ashish-2005/lizmotors-coding-assignment/tree/master

Sample Output :

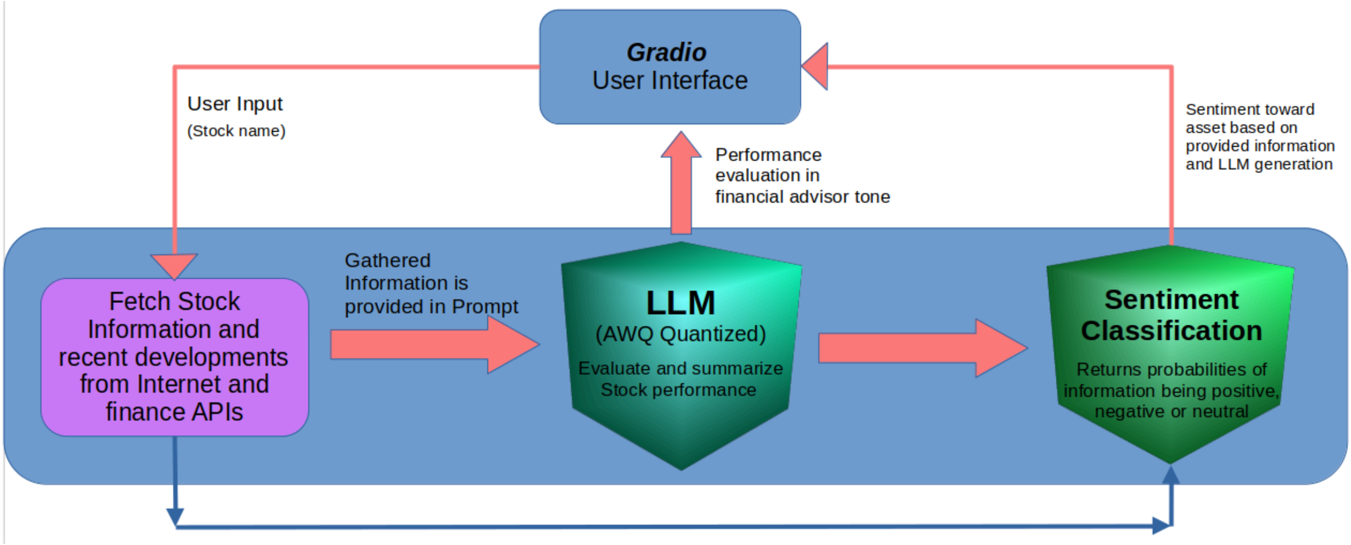


	Query	Title	Source	Data
31	Identify key trends in the market, including c...	The great consumer shift: Ten charts that show...	https://www.mckinsey.com/capabilities/growth-m...	The great consumer shift: Ten charts that show...
66	Identify the industry in which Canoo operates,...	Canoo, the EV startup that scored a deal with ...	https://fortune.com/2022/08/29/canoo-ev-startu...	Canoo, the EV startup that scored a deal with ...
52	Analyze Canoo's main competitors, including th...	What Is Competitor Analysis? Definition + Step...	https://www.coursera.org/articles/competitor-a...	What Is Competitor Analysis? Definition + Step...
56	Identify key trends in the market, including c...	Four trends in consumer behavior in 2023 | McK...	https://www.mckinsey.com/industries/consumer-p...	Four trends in consumer behavior in 2023 | McK...
29	Identify key trends in the market, including c...	The 4 Biggest Consumer Behavior Shifts of 2023...	https://blog.hubspot.com/marketing/biggest-con...	\n\nThe 4 Biggest Consumer Behavior Shifts of ...
...	...	...	...	...
54	Analyze Canoo's main competitors, including th...	What Is a Competitive Analysis & How to Do It ...	https://www.semrush.com/blog/competitive-analy...	What Is a Competitive Analysis & How to Do It ...

Submitted By:
Ashish Bangwal
bangwalashish41@gmail.com
+91 9871639868

(PTO)

Similar RAG projects implemented by me :

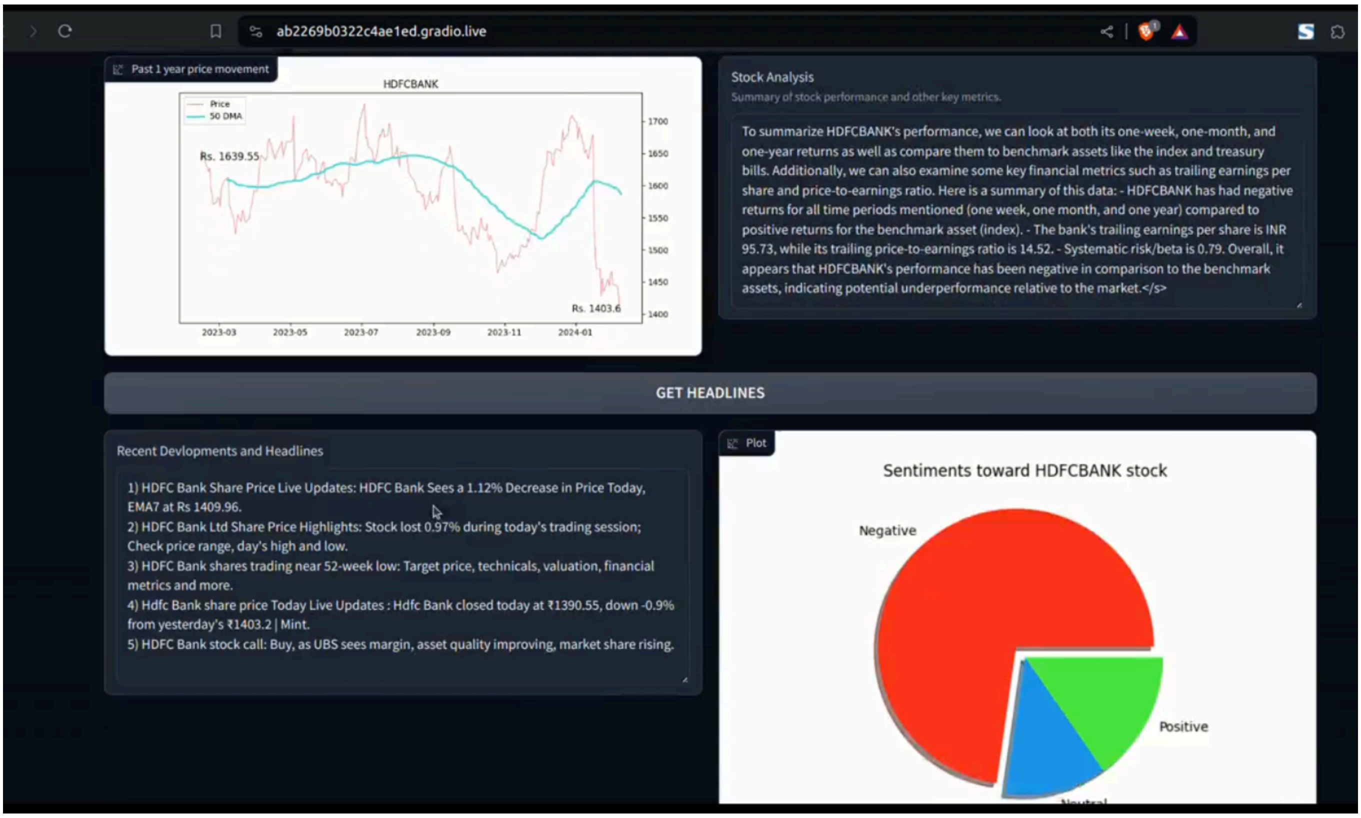

Gradio
User Interface
User Input
(Stock name)
Performance evaluation in financial advisor tone
Sentiment toward asset based on provided information and LLM generation
Fetch Stock Information and recent developments from Internet and finance APIs
Gathered Information is provided in Prompt
LLM
(AWQ Quantized)
Evaluate and summarize Stock performance
Sentiment Classification
Returns probabilities of information being positive, negative or neutral

I recently built a similar project called Pet-Analyst, which is an LLM-powered stock analyst for publicly traded companies in NSE.

It fetches live data about the queried stock like PE DE ratio, EPS, Dividend yield, returns in different time-frames, etc from my custom-built API which is backed by Yahoo-finance and Google.

After retrieving data It passes data to LLM in the form of a prompt and lets it explain stock performance like a financial advisor.

I used [finance-chat-AWQ]() an LLM with 1.3B parameters which is fine-tuned on financial data and quantized with AWQ strategy for farter inference.



Stock Analysis

Summary of stock performance and other key metrics.

To summarize HDFCBANK's performance, we can look at both its one-week, one-month, and one-year returns as well as compare them to benchmark assets like the index and treasury bills. Additionally, we can also examine some key financial metrics such as trailing earnings per share and price-to-earnings ratio. Here is a summary of this data: - HDFCBANK has had negative returns for all time periods mentioned (one week, one month, and one year) compared to positive returns for the benchmark asset (index). - The bank's trailing earnings per share is INR 95.73, while its trailing price-to-earnings ratio is 14.52. - Systematic risk/beta is 0.79. Overall, it appears that HDFCBANK's performance has been negative in comparison to the benchmark assets, indicating potential underperformance relative to the market.</s>

Recent Devlopments and Headlines

1) HDFC Bank Share Price Live Updates: HDFC Bank Sees a 1.12% Decrease in Price Today, EMA7 at Rs 1409.96.
2) HDFC Bank Ltd Share Price Highlights: Stock lost 0.97% during today's trading session; Check price range, day's high and low.
3) HDFC Bank shares trading near 52-week low: Target price, technicals, valuation, financial metrics and more.
4) Hdfc Bank share price Today Live Updates : Hdfc Bank closed today at ₹1390.55, down -0.9% from yesterday's ₹1403.2 | Mint.
5) HDFC Bank stock call: Buy, as UBS sees margin, asset quality improving, market share rising.