INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Enhancing Search Engine Relevance for Video Subtitles

Prepared by:
-> Aishwary Dakhare(IN1240614)

Team ID :- T211101

OBJECTIVE

Develop an advanced search engine algorithm that efficiently retrieves subtitles based on user queries, with a specific emphasis on subtitle content. The primary goal is to leverage natural language processing and machine learning techniques to enhance the relevance and accuracy of search results.

INTRODUCTION

Search Engine is a tool that helps to extract relevant contents from the huge amount of data. Google maintains a steadfast commitment to ensuring a seamless and precise search experience above all other considerations. Our project focuses on improving the search relevance for video subtitles, enhancing the accessibility of video content.

TYPES OF SEARCH ENGINE

Search engine can be categorized as

- **Keyword based Search Engines**
- **Semantic based Search Engines**

Keyword Based Search Engine: These search engines rely heavily on exact keyword matches between the user query and the indexed documents. It focus primarily on matching exact keywords in documents

Semantic Search Engines: Semantic search engines go beyond simple keyword matching to understand the meaning and context of user queries and documents. Aim to understand the deeper meaning and context of user queries to deliver more relevant and meaningful search results.

STEP BY STEP PROCESS

Part 1: Ingesting Documents

● Data Sampling

❖ Data preprocessing ● Document chunker

❖ Text vectorization

❖ Storing Embeddings

Part 2: Retrieving Documents

➢ Ingesting Documents

❖ *Data Sampling*

Dataset provided was in .db format. As it contained huge amount of data we took 30% of it and converted into .csv format which made us easy to preprocess on it.

❖ *Data Preprocessing*

Data cleaning is necessary before analysing it. It involved:

- Decoding the subtitles with “latin-1”
- Removing timestamps

    - Removing symbols and punctuation marks
    - Removing unwanted words

- *Document Chunker*

Subtitles are chunked into smaller chunks to ensure that no information is lost during embedding. Embeddings is done to convert words into numerical form that is machine friendly.

- *Text vectorization*

BERT is used for generating embeddings of the given subtitles.It is based on **SentenceTransformers** to generate embeddings which encode semantic information.

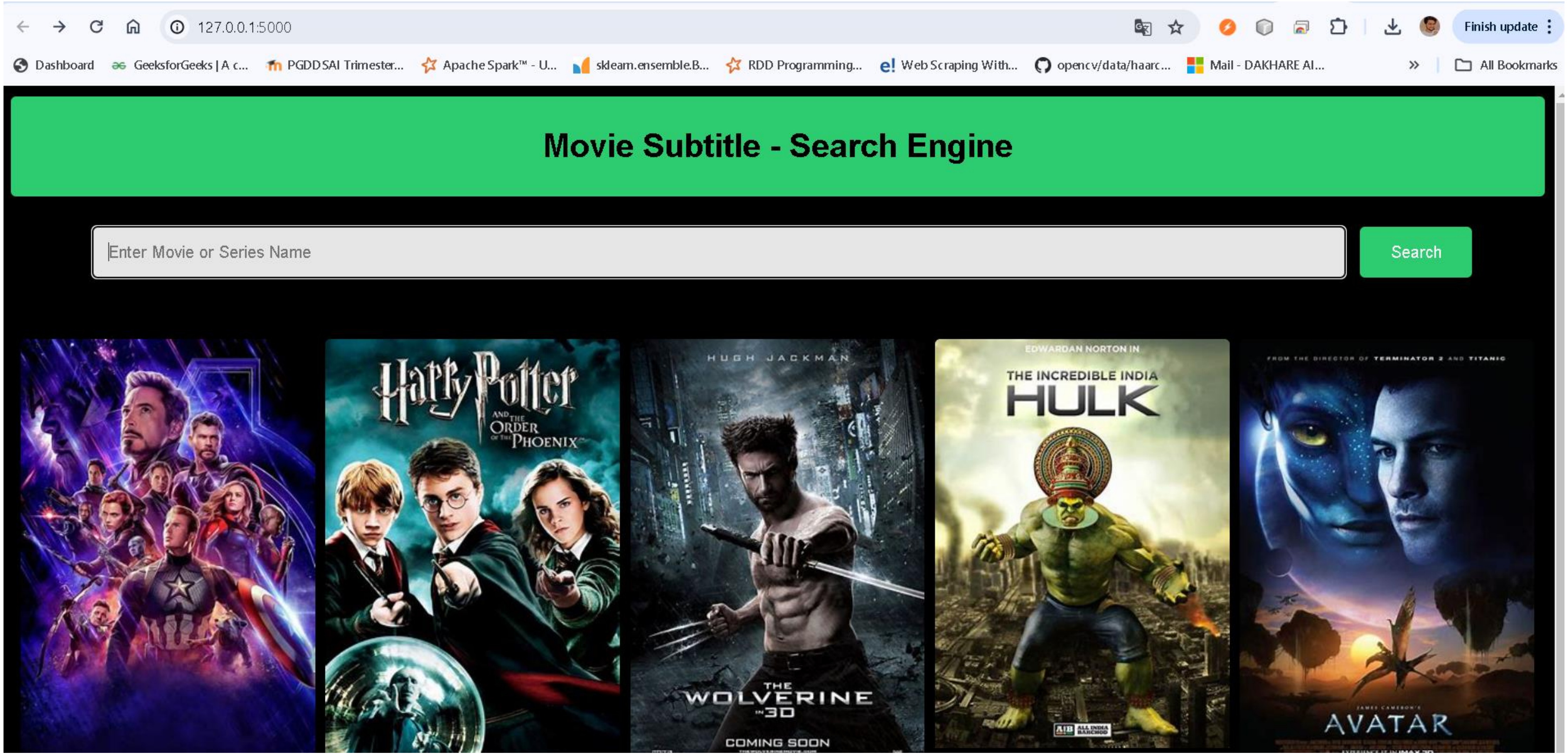
- *Storing Embeddings*

ChromaDB is used for storing the embeddings as it is suitable for storing the vector representations.

Retrieving Documents

- Take the user's search query.
- Preprocess the query (if required).
- Create query embedding.
- Used cosine distance to calculate the similarity score between embeddings of documents and user search query embedding.

RESULTS



Movie Subtitle - Search Engine

CONCLUSION

By following the step by step process as mentioned above our project “Search Engine web app “ on movie subtitle datasets was successfully built that enhanced searching within video subtitles.

THANK YOU

INNOMATICS RESEARCH LABS