INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Enhancing Search Engine Relevance for Video Subtitles

Prepared by:
-> Ajay Chaudhary(IN1240153)
-> Sapna Limbu(IN1240692)
Team ID :- T211009

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. In the current landscape, 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

Based on how searching action is done, 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

Part 1: 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
  - Standardizing text to lowercase
  - 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.

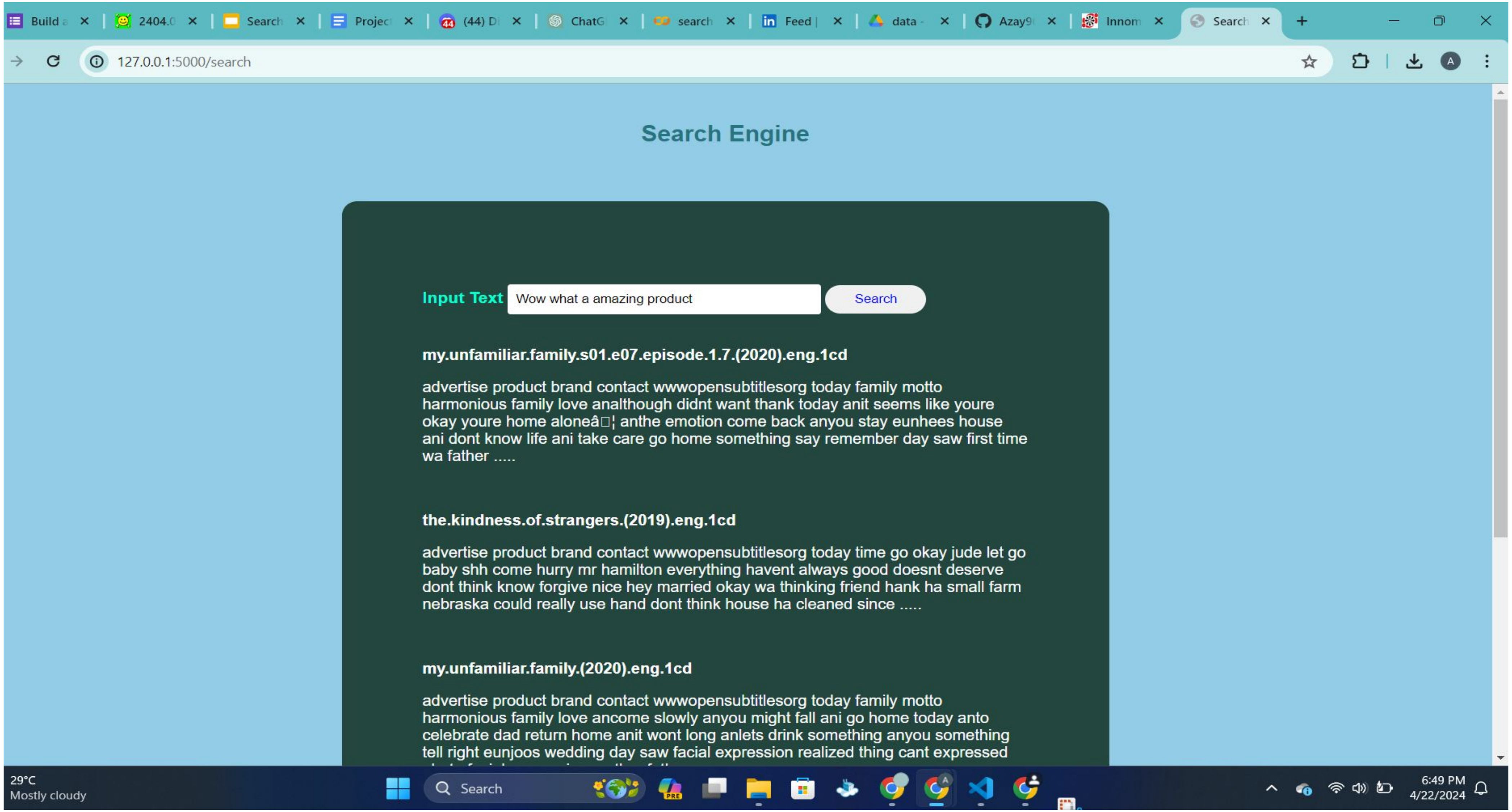
- *Storing Embeddings*

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

Part 2: 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



Search Engine

Input Text: Wow what a amazing product

Search

my.unfamiliar.family.s01.e07.episode.1.7.(2020).eng.1cd

advertise product brand contact wwwopensubtitlesorg today family motto harmonious family love analthough didnt want thank today anit seems like youre okay youre home aloneâ□¦ anthe emotion come back anyou stay eunhees house ani dont know life ani take care go home something say remember day saw first time wa father .....

the.kindness.of.strangers.(2019).eng.1cd

advertise product brand contact wwwopensubtitlesorg today time go okay jude let go baby shh come hurry mr hamilton everything havent always good doesnt deserve dont think know forgive nice hey married okay wa thinking friend hank ha small farm nebraska could really use hand dont think house ha cleaned since .....

my.unfamiliar.family.(2020).eng.1cd

advertise product brand contact wwwopensubtitlesorg today family motto harmonious family love ancome slowly anyou might fall ani go home today anto celebrate dad return home anit wont long anlets drink something anyou something tell right eunjoos wedding day saw facial expression realized thing cant expressed

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