



**INNOVATION. AUTOMATION. ANALYTICS**

## **PROJECT ON**

# **Enhancing Search Engine Relevance for Video Subtitles**

**By,**  
**Akula Nitish Kumar**  
**Project ID :T211126**  
**Intern ID: IN1241112**



# Objective

**Create a search algorithm leveraging NLP and ML to prioritize video subtitle analysis, enhancing relevance and accuracy of search results for user queries.**

1. **Data Collection:** Gather videos and subtitles.
2. **Preprocessing:** Clean and standardize subtitle text.
3. **Feature Extraction:** Capture word frequency, sentiment, and topics.
4. **Semantic Analysis:** Understand the meaning of subtitles using NLP.
5. **Query Processing:** Analyze user queries to understand intent.
6. **Matching and Ranking:** Compare query features with subtitle features to rank relevance.
7. **Feedback Loop:** Incorporate user feedback to refine results.
8. **Optimization:** Improve efficiency and scalability.
9. **Evaluation:** Measure performance and iterate for improvement.

By following these steps, you can develop a search engine algorithm that effectively prioritizes video subtitle analysis to deliver relevant results.

# Introduction

1. **Importance of Search Engines:** Highlight their crucial role in navigating the vast digital landscape.
2. **Google's Reputation:** Acknowledge Google's leadership in search technology and commitment to user satisfaction.
3. **Project Goal:** Aim to improve search relevance of videos using subtitle data.
4. **Accessibility Enhancement:** Emphasize the project's potential to make video information more accessible to diverse users.
5. **Broad Impact:** Consider the wider implications of enhancing search relevance for video content.

# Types of Search Engines

## **Keyword-Based Search Engine:**

- Matches search queries to web pages based solely on the presence of specific keywords.
- Doesn't consider the contextual or semantic meaning of the words in the query or on the web pages.
- Results may lack relevance or depth, as they rely heavily on keyword matches.

## **Semantic Search Engine:**

- Understands the intent behind user queries and the contextual meaning of web pages.
- Aims to provide more relevant results by considering the semantic relationships between words and the overall meaning of the content.
- Offers a deeper understanding of user intent, leading to more accurate and contextually relevant search results.

# Workflow

There are 2 different steps to implement search engine:

## 1. Data Ingestion and Preprocessing:

- **Data Sampling:** Select a representative subset of documents from the entire dataset for processing.
- **Data Preprocessing:** Clean and prepare the sampled documents for analysis, which may involve tasks like removing HTML tags, punctuation, and stop words.
- **Document Chuker:** Divide the preprocessed documents into smaller, manageable chunks to facilitate efficient processing.
- **Text Vectorization:** Convert the textual data into numerical representations (embeddings) using techniques such as TF-IDF or word embeddings.

## 2. Indexing and Retrieval:

- **Storing the Embeddings in Chroma DB:** Save the generated embeddings in a database optimized for similarity search, such as Chroma DB or other indexing systems.
- **Query Processing and Retrieval:** Process user queries, match them to the stored embeddings, and retrieve the most relevant documents based on similarity scores. This step involves ranking the results and presenting them to the user in a meaningful way.

# Data Ingestion and Preprocessing

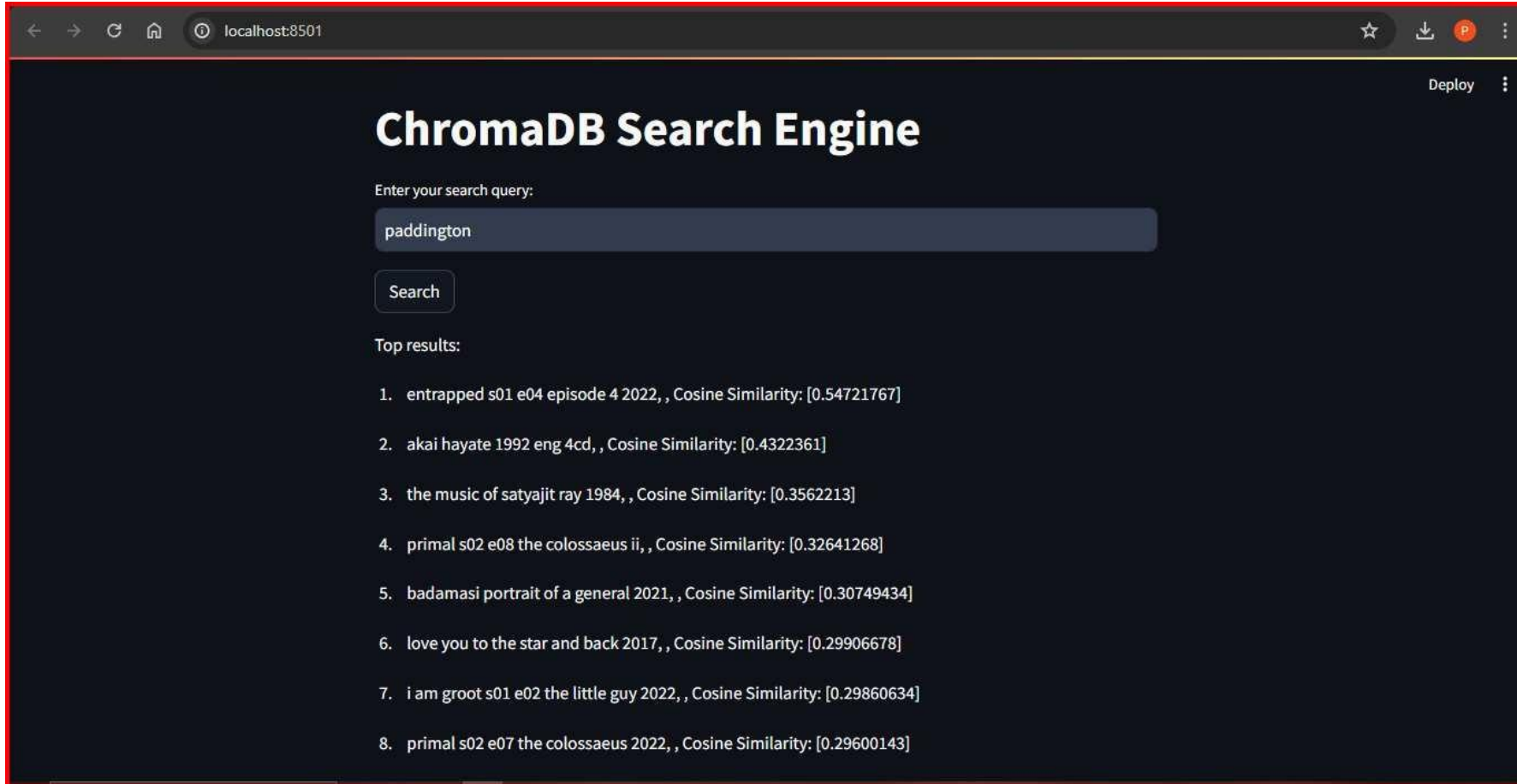
1. **Data Acquisition:** Gather documents from various sources such as websites, databases, or file systems.
2. **Data Sampling:** Select a representative subset of documents for processing to manage computational resources effectively.
3. **Data Preprocessing:** Clean and prepare the documents by removing noise, formatting inconsistencies, and irrelevant content.
4. **Document Chunking:** Divide large documents into smaller chunks or segments to facilitate efficient processing and analysis.
5. **Text Extraction:** Extract relevant text content from different document formats, such as HTML, PDF, or Word documents.
6. **Metadata Extraction:** Extract metadata such as title, author, publication date, and keywords to enrich document representations and facilitate search filtering.

# Indexing and Retrieval

1. **Indexing:** Create a searchable index of the ingested documents by organizing them into a structured format optimized for efficient retrieval.
2. **Text Analysis:** Analyze the textual content of the documents to extract key terms, phrases, and other relevant information for indexing.
3. **Indexing Algorithms:** Utilize algorithms such as inverted indexing to map terms to the documents they appear in, enabling fast lookup during retrieval.
4. **Query Parsing:** Parse user queries to identify keywords, phrases, and other search parameters.
5. **Document Retrieval:** Retrieve relevant documents from the index based on the user's query, using techniques such as term frequency-inverse document frequency (TF-IDF) or cosine similarity.
6. **Ranking:** Rank the retrieved documents based on their relevance to the user's query, considering factors like keyword frequency, document popularity, and metadata attributes.

# Web Interface

## Search Engine Page



The screenshot shows a web browser window with the address bar set to localhost:8501. The page title is "ChromaDB Search Engine". Below the title, there is a search input field containing the text "paddington" and a "Search" button. Underneath the search field, the text "Top results:" is displayed. A list of eight search results follows, each showing a title and its Cosine Similarity score in brackets. The results are as follows:

Rank	Search Result	Cosine Similarity
1.	entrapped s01 e04 episode 4 2022, ,	[0.54721767]
2.	akai hayate 1992 eng 4cd, ,	[0.4322361]
3.	the music of satyajit ray 1984, ,	[0.3562213]
4.	primal s02 e08 the colossaeus ii, ,	[0.32641268]
5.	badamasi portrait of a general 2021, ,	[0.30749434]
6.	love you to the star and back 2017, ,	[0.29906678]
7.	i am groot s01 e02 the little guy 2022, ,	[0.29860634]
8.	primal s02 e07 the colossaeus 2022, ,	[0.29600143]



# Wrap-Up

1. **Content-Centric Approach:** Prioritizes analyzing the textual content of video subtitles to enhance search accuracy and relevance.
2. **User-Friendly Experience:** Aims to improve the overall search experience by delivering more precise results aligned with user expectations.
3. **Increased Accessibility:** Enhances accessibility to video subtitle information for a broader audience, including those who rely on subtitles for comprehension.
4. **Substantial Improvement:** Introduces a significant advancement in the effectiveness of video search algorithms by leveraging subtitle content.
5. **Alignment with User Intent:** Focuses on understanding and fulfilling user intent by providing highly relevant search results tailored to the content of subtitles.
6. **Enhanced Engagement:** Encourages increased user engagement with video content through more accurate and relevant search outcomes.

THANK YOU

