INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

Enhancing Search Engine Relevance for Video Subtitles

Prepared By – Deepraj Vad and Vanita Deshmukh

INNOMATICS
RESEARCH LABS

About us

About Deepraj Vadhwane

After completing my BTech from MGM College of Engineering in Nanded, Maharashtra, I discovered my passion for Data Science during my exploration of various technological streams. Seeking to further my expertise, I joined Innomatics Research Lab in Hyderabad, where I engaged in intensive learning across Python, SQL, Power BI, Data Analysis, and Machine Learning. I tackled assignments, quizzes, and projects, including tasks like web scraping and data analysis.

Driven by the transformative impact of Data Science on modern society, I am fascinated by how AI expedites once time-consuming processes, enabling actionable insights and predictive modeling. Currently serving as a Data Scientist intern at Innomatics Research Lab, I continue to deepen my knowledge through real-time case studies and exploration of new concepts.

GitHub: https://github.com/Deepraj

LinkedIn: https://www.linkedin com in deeprajvadhwane

About Vanita Deshmukh

I'm Vanita Deshmukh, a graduate from GNDEC Bidar in 2023. My journey into data science began with a course at Innomatics Research Labs in Hyderabad. During my 4-month internship as a Data Scientist, I worked on AI and machine learning projects, achieving a sentiment analysis model and developing a chatbot for interviews using GenAI. I also contributed to creating AI apps using Python, Flask, and AWS EC2. With skills in Python, data handling, machine learning, and NLP, I'm passionate about using AI and ML to extract insights from data and develop innovative solutions. I'm eager to leverage my expertise in AI/ML to contribute to your projects.

GitHub: https://github.com/VanitaDeshmukh

LinkedIn: https://www.linkedin.com/in/vanitadeshmukh121/

Connect with me

INNOMATICS RESEARCH LABS

Table of Contents

INNOMATICS RESEARCH LABS

Project Overview

The digital landscape is rapidly evolving, and effective search engines are essential for connecting users with relevant content. This project aims to enhance the search relevance for video subtitles, thereby improving the accessibility of video content.

Objective

The primary objective of this project is to develop an advanced search engine algorithm that retrieves subtitles based on user queries. By leveraging natural language processing (NLP) and machine learning techniques, the project aims to enhance the relevance and accuracy of search results.

Keyword-based Search Engines

Keyword-based search engines rely on exact matches between the user query and the indexed documents. They are limited to matching keywords and may not capture the context or meaning behind the keywords.

Semantic Search Engines

Semantic search engines go beyond keyword matching to understand the meaning and context of user queries and documents. By analyzing and interpreting the content, semantic search engines deliver more relevant and meaningful search results.

Comparison Keyword-based vs Semantic Search Engines

While keyword-based search engines focus on exact matches, semantic-based search engines aim to understand the deeper meaning and context of user queries, providing more relevant and accurate search results.

Core Logic

To compare a user query against a video subtitle document, the core logic involves the following key steps.

1. Data Preprocessing

- Decode and clean subtitle data.
- Remove timestamps and other irrelevant information.

2. Text Vectorization

- Generate text vectors using Bag-of-Words (BOW)/TF-IDF for keyword-based search.
- Utilize BERT-based Sentence Transformers for semantic embeddings.

3. Document Chunking

- Divide large documents into smaller chunks for efficient embedding.
- Implement overlapping windows to maintain context across chunks.

4. Cosine Similarity Calculation

- Compute the cosine similarity between document and query embeddings.
- Determine the relevance of documents to the user's query based on similarity scores.

Step-by-Step Process

1. Read the Data

Objective:

The first step involves understanding the structure and content of the provided subtitle data.

Procedure:

- Open the database file using appropriate software or tools suitable for database files.
- Examine the dataset to identify the number of records, columns, and data types.
- Familiarize yourself with the dataset's layout and understand the kind of information it contains.

2. Decoding and Cleaning

Objective:

Before diving into the analysis, it's essential to preprocess the subtitle data to make it suitable for further steps.

Procedure:

- Refer to the README.txt file accompanying the dataset to identify any encoding methods used and decode the data accordingly.
- Implement cleaning procedures to remove any unnecessary characters, punctuation, and timestamps from the subtitle text. This step ensures that the text is consistent and ready for analysis.

3. Data Sampling

Objective:

When working with large datasets, especially with limited computational resources, it's beneficial to work with a subset of the data.

Procedure:

- Randomly select approximately 30% of the dataset to create a manageable sample.
- Ensure that the sampling process maintains the dataset's diversity and represents the entire range of subtitle content.

4. Text Vectorization

Objective:

To analyze and compare subtitle documents effectively, we need to convert the text data into a numerical format.

Procedure:

- Utilize text vectorization techniques such as Bag-of-Words (BOW) or Term Frequency-Inverse Document Frequency (TF-IDF) to transform the cleaned subtitle text into numerical vectors.
- Alternatively, employ advanced embedding techniques like BERT-based SentenceTransformers to capture semantic meanings and relationships within the subtitle content.

5. Document Chunking

Objective:

Handling long and extensive subtitle documents efficiently poses a challenge due to computational limitations.

Procedure:

- Divide each subtitle document into smaller, more manageable chunks based on a predefined token or character limit.
- Implement overlapping windows between adjacent chunks to ensure continuity and maintain context throughout the document.

6. Store Embeddings

Objective:

After generating the text embeddings, it's crucial to store them systematically for easy retrieval and future analysis.

Procedure:

- Utilize a suitable database system, such as ChromaDB, to store the generated BOW, TF-IDF, and BERT embeddings along with their corresponding document IDs.
- Organize the database to facilitate efficient querying and retrieval of embeddings based on user search queries or other criteria.

Part 2: Retrieving Documents - Enhancing User Experience

Understanding the User Query

The Essence: Every user query is a unique string of words, representing a quest for specific information or context.

Our Approach: To understand this quest deeply, we clean, normalize, and break down the query into its fundamental parts. This ensures that the query is ready for the next stage, where it will be matched against our subtitle database.

Translating Words to Meaningful Vectors

The Essence: Words have meaning, and we aim to capture this essence in a way a computer can understand.

Our Approach: We employ advanced techniques, specifically SentenceTransformers, to transform the user's query into a series of numbers. These numbers, or vectors, represent the semantic meaning of the query, allowing us to compare it with our subtitle database more effectively.

Finding the Best Matches

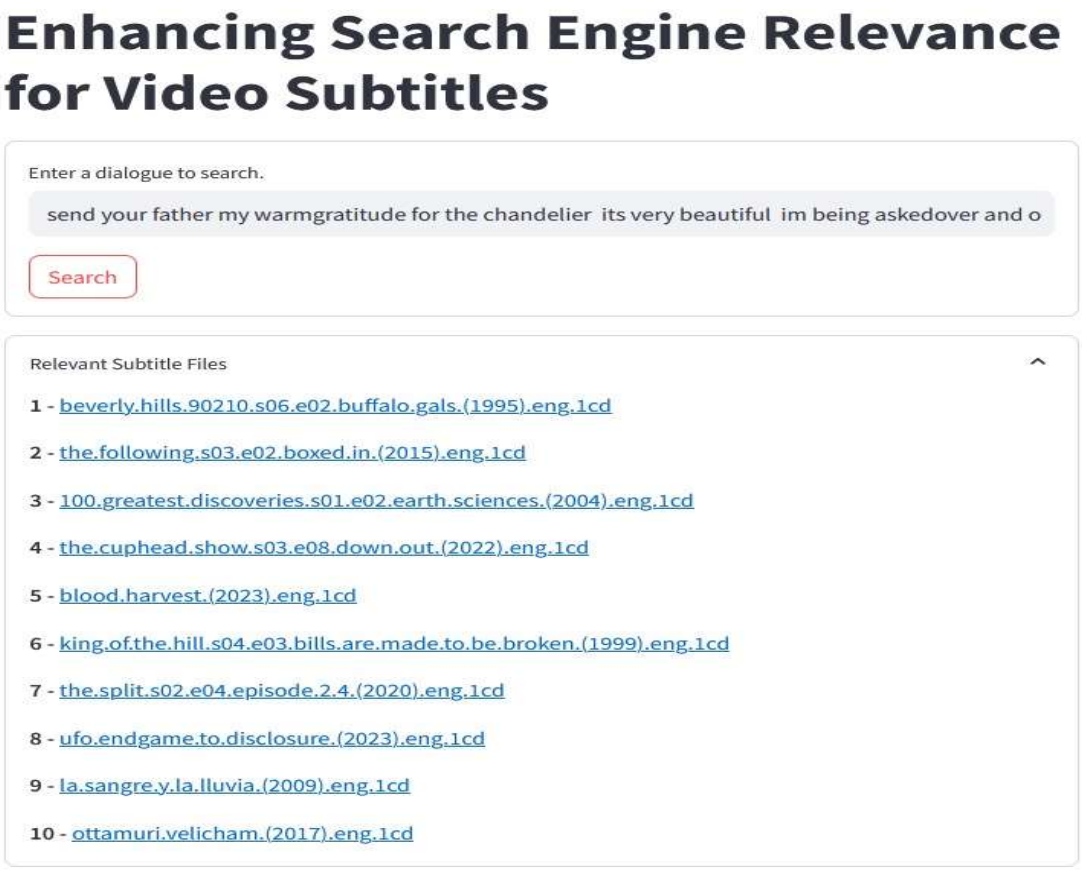
The Essence: In the vast sea of subtitles, how do we find the ones that resonate most with the user's query?

Our Approach: We use a mathematical measure called cosine similarity. Imagine each subtitle as a point in space and the user query as another point. Cosine similarity calculates how close these points are, helping us identify subtitles that are conceptually close to what the user is looking for

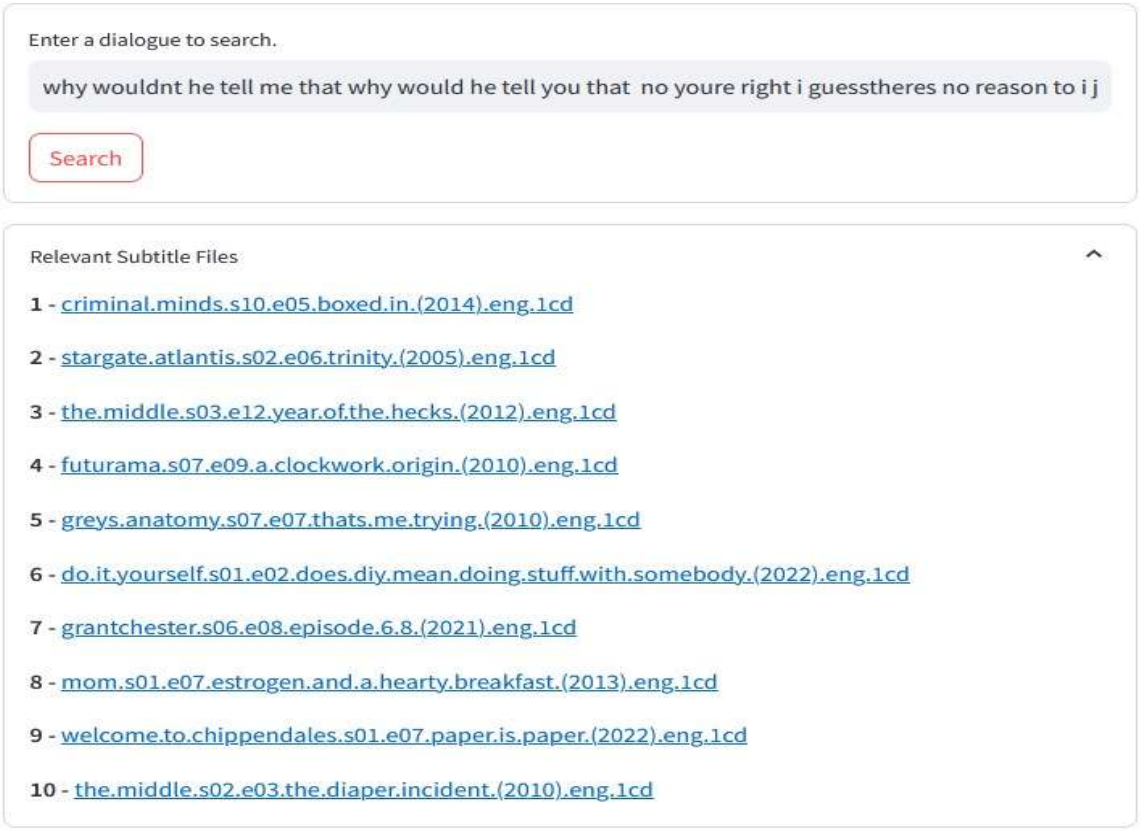
Delivering the Most Relevant Results

The Essence: Our ultimate goal is to present the user with subtitles that are not just relevant but resonate with the context of their search.

Our Approach: Based on the similarity scores, we rank the subtitle documents. The higher the score, the more relevant the subtitle. We then present these top-ranking subtitles to the user, providing them with a concise yet comprehensive overview to aid in their quest for information.



Enhancing Search Engine Relevance for Video Subtitles



Conclusion

This project demonstrates the potential of leveraging NLP and machine learning techniques to enhance search relevance for video subtitles. By focusing on semantic understanding and context, the developed search engine offers a more intuitive and accurate search experience for users, making video content more accessible and user-friendly.