INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

# PROJECT ON

# SEARCH ENGINE

# OBJECTIVE

The main goal is to develop a sophisticated search engine algorithm that optimizes the retrieval of video subtitles through user queries, focusing specifically on enhancing the accuracy and relevance of search results by leveraging advanced natural language processing (NLP) and machine learning (ML) techniques.

# Types of Search Engines

**Keyword-Based Search Engines:** These search engines function by identifying and returning results that contain exact matches with the words used in the search query. They index and search content based on specific keywords found within the text.

**Semantic Search Engines:** Semantic search engines utilize NLP and ML to go beyond mere keyword matching. They aim to understand the intent and contextual meaning behind user queries and the content within documents.

# DATA HANDLING

**Preprocessing:** To ensure the data is in the best format for analysis and processing, enhancing the effectiveness of subsequent steps.

Steps Involved:

**Removing Timestamps:** Subtitles often come with timestamps that are irrelevant for text analysis and search purposes. Removing these allows for cleaner data and more accurate vectorization.

**Normalization:** This includes converting all text to a uniform case (usually lowercase), removing special characters and punctuation, and handling whitespace effectively. Normalization ensures that the engine does not treat the same words differently due to superficial differences in formatting.

# Vectorization

To transform text data into a numerical format that machine learning algorithms can process, essentially turning raw text into feature vectors.

Techniques Used:

**BOW/TFIDF:** These methods create sparse vector representations of documents by counting word occurrences and weighing them by their importance across the document corpus. While effective for keyword-based search, they can overlook semantic relationships.

**BERT Embeddings:** Utilizes the BERT model to generate dense embeddings that capture deeper semantic meanings of phrases and sentences. This is particularly useful for understanding context and nuances in language, making it ideal for semantic search engines.

# Document Chunking:

**Chunk Size and Overlaps:** Documents are divided into segments (e.g., 500 tokens each), with overlapping windows to ensure that no contextual information is lost at the boundaries. This overlap helps maintain narrative flow and context across chunks, which is crucial for maintaining the integrity of semantic connections.

**Similarity Calculation:** Cosine similarity measures the cosine of the angle between two vectors. In the context of search engines, it is used to determine how similar a document's vector is to a query vector.

## Query Processing:

To ensure that user queries are accurately understood and matched against the subtitle documents in the database.

Steps Involved:

**Consistent Preprocessing:** Just like document data, user queries undergo similar preprocessing steps. This includes normalizing text (converting to lowercase, removing punctuation), which is crucial to ensure that the query matches the processed form of the documents.

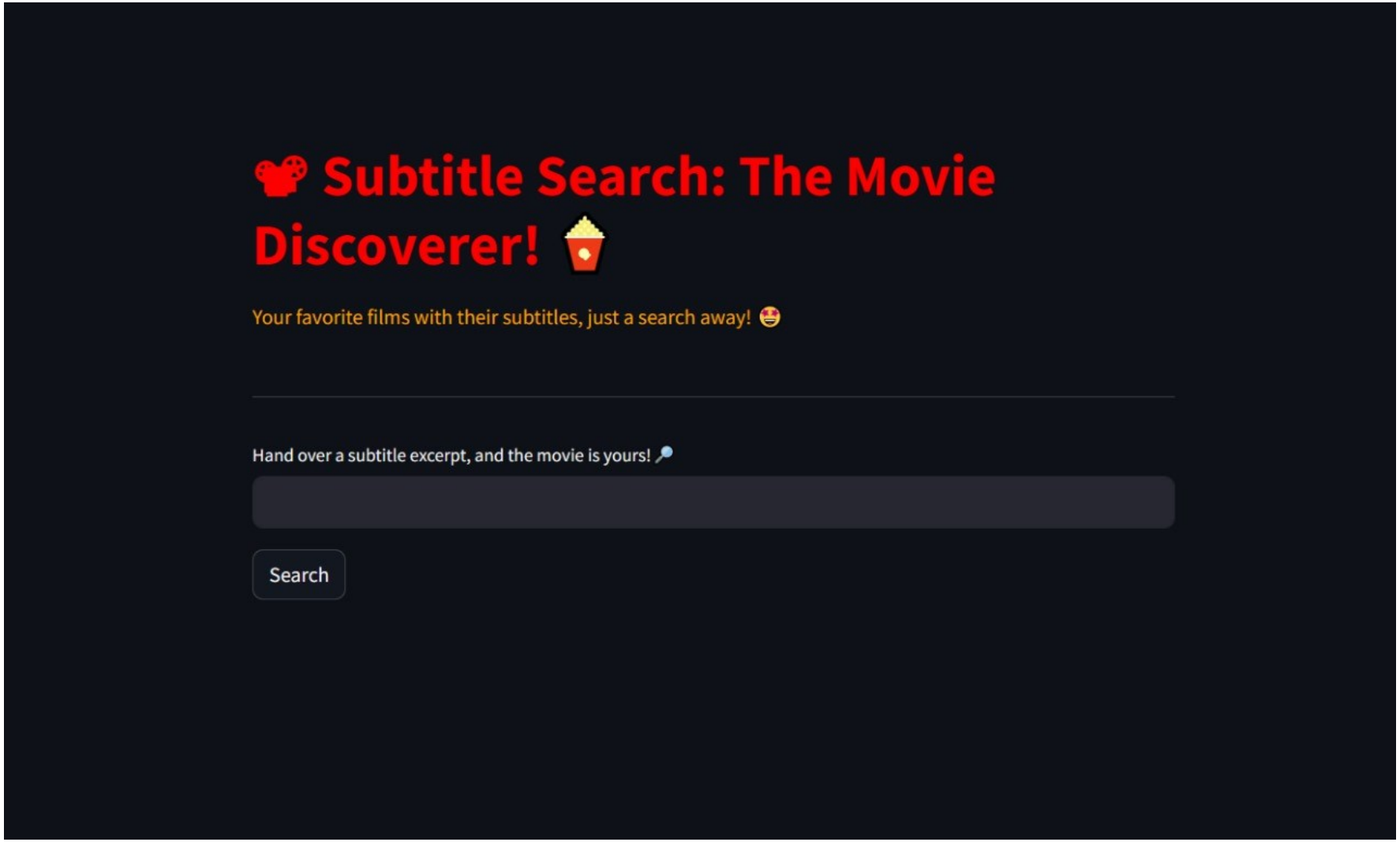
**Embedding the Query:** The preprocessed query is then converted into an embedding using the same method (BOW/TFIDF or BERT) as the documents. This ensures that both the query and the documents are represented in the same vector space, enabling effective similarity comparisons.

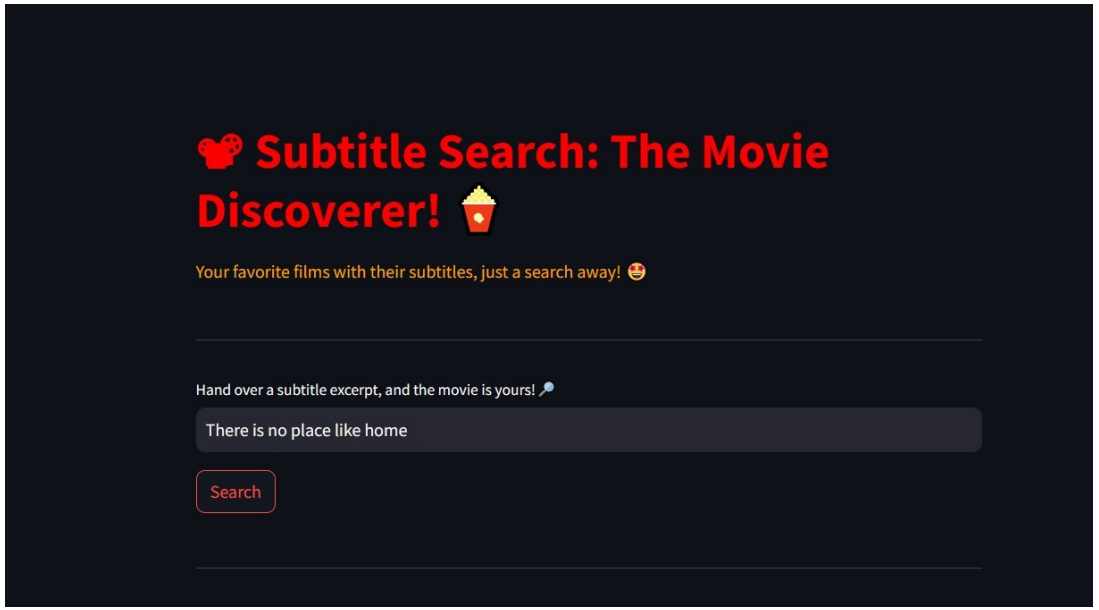
## Cosine Distance Calculation:

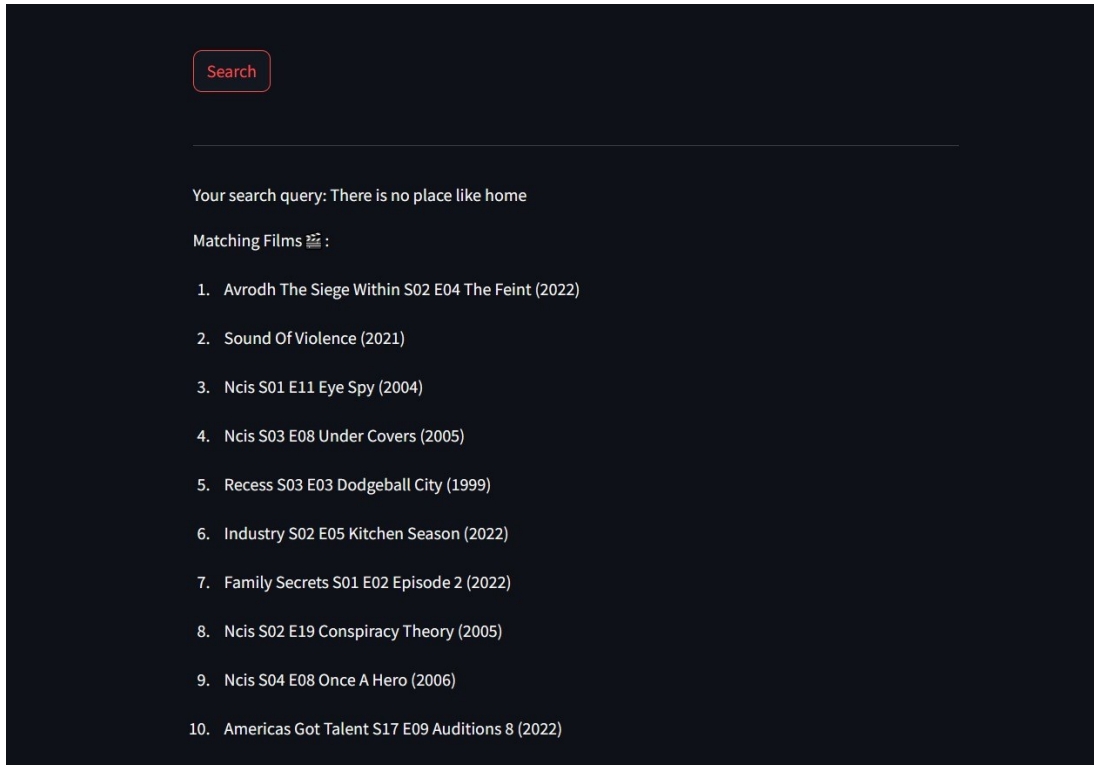
Utilizing cosine similarity to assess the closeness between the query embedding and each of the document embeddings

The cosine distance (1 - cosine similarity) between the query vector and each document vector is computed. A lower cosine distance indicates a higher similarity.

# RESULTS


Subtitle Search: The Movie Discoverer!
Your favorite films with their subtitles, just a search away!
Hand over a subtitle excerpt, and the movie is yours!
Search


Subtitle Search: The Movie Discoverer!
Your favorite films with their subtitles, just a search away!
Hand over a subtitle excerpt, and the movie is yours!
There is no place like home
Search


Search
Your search query: There is no place like home
Matching Films :
1. Avrodh The Siege Within S02 E04 The Feint (2022)
2. Sound Of Violence (2021)
3. Ncis S01 E11 Eye Spy (2004)
4. Ncis S03 E08 Under Covers (2005)
5. Recess S03 E03 Dodgeball City (1999)
6. Industry S02 E05 Kitchen Season (2022)
7. Family Secrets S01 E02 Episode 2 (2022)
8. Ncis S02 E19 Conspiracy Theory (2005)
9. Ncis S04 E08 Once A Hero (2006)
10. Americas Got Talent S17 E09 Auditions 8 (2022)

# CONCLUSION

The above are  effectively communicates the detailed process of how the search engine handles and retrieves documents based on user queries, emphasizing the sophistication of the underlying technology and its impact on user experience.

# THANK YOU!

INNOMATICS
RESEARCH LABS