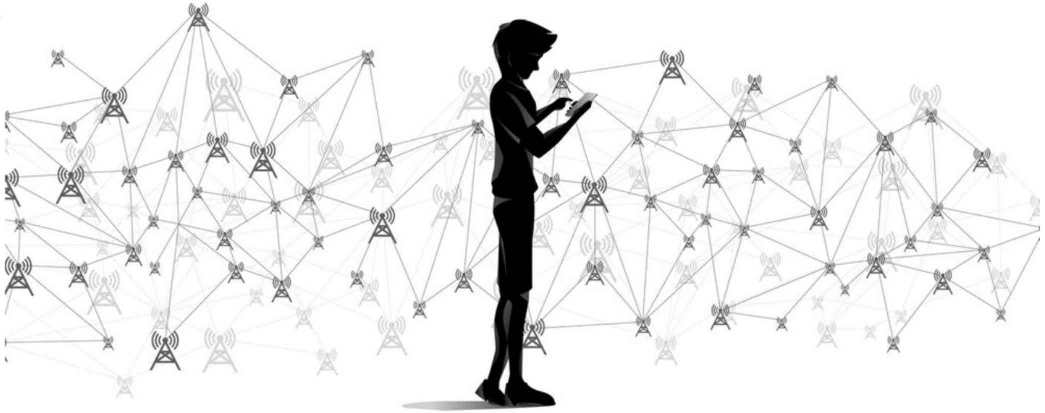
INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

SEMANTIC SEARCH ENGINE IMPLEMENTATION



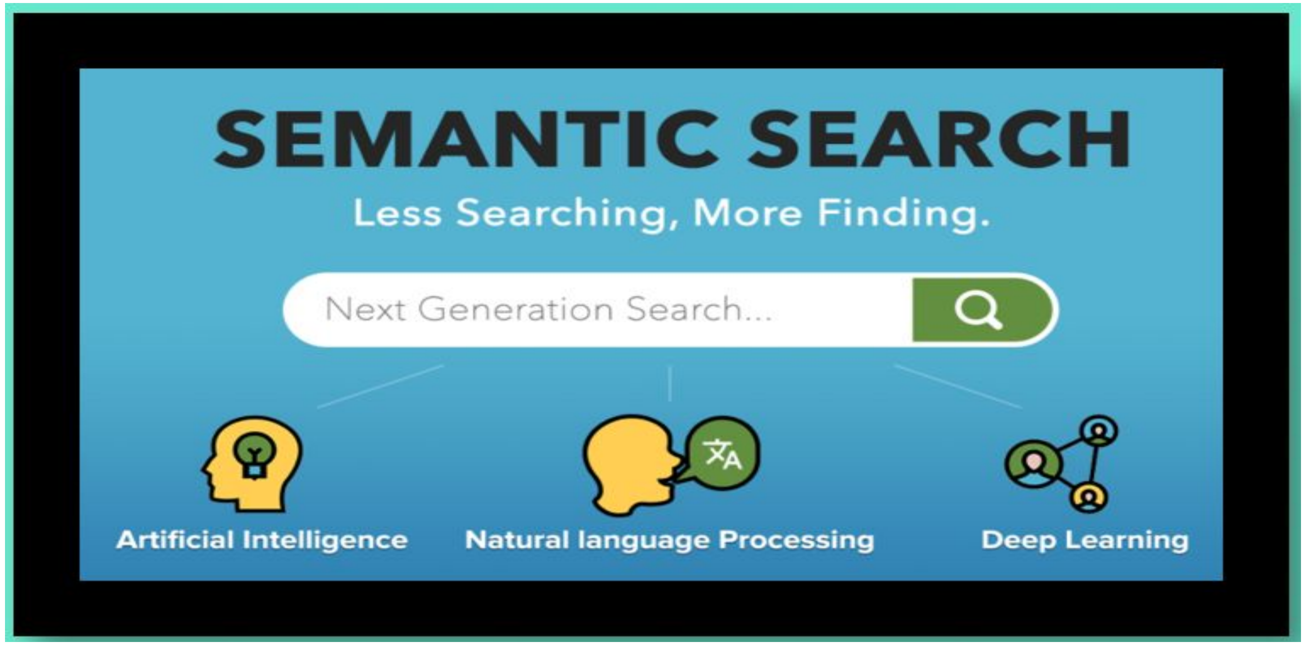
By

Ibteda Azeem

Degala Sravya

Problem Statement

- Create an innovative search engine algorithm designed to swiftly locate subtitles according to user queries, prioritizing the content within the subtitles themselves.
- Utilize advanced natural language processing and machine learning methodologies to optimize result accuracy and relevance.

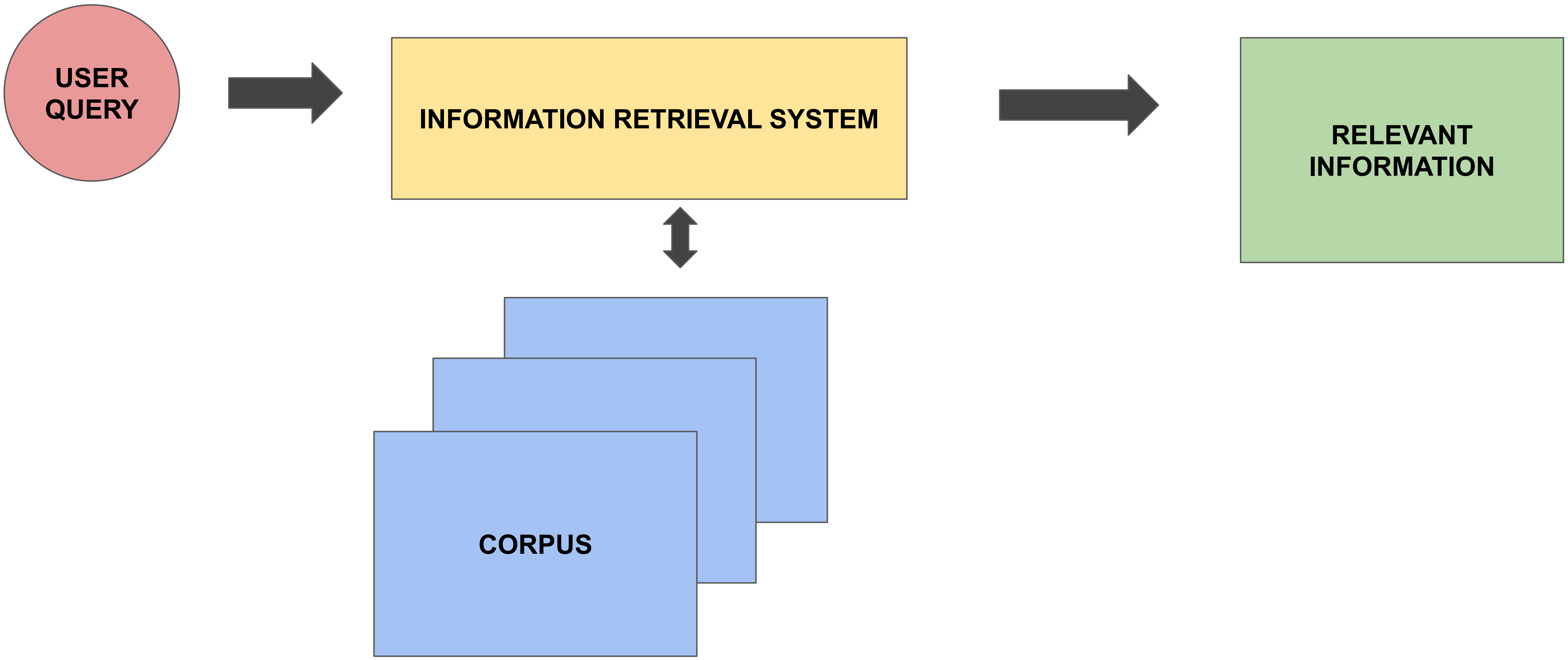

SEMANTIC SEARCH
Less Searching, More Finding.
Next Generation Search...
Artificial Intelligence
Natural language Processing
Deep Learning

Contents

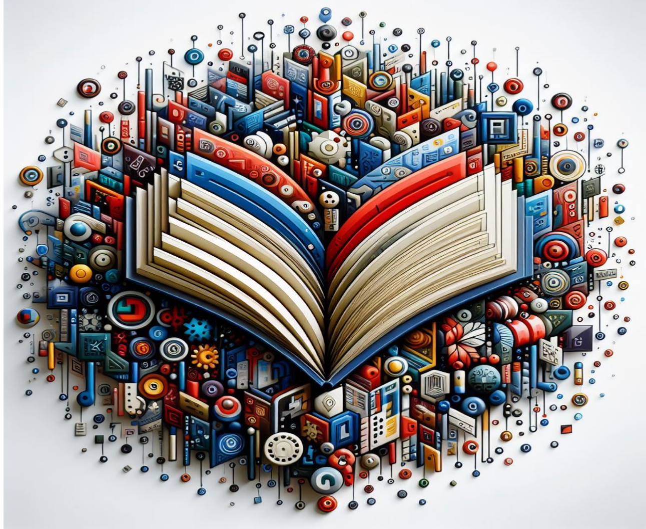
- Information Retrieval System
- Applications of Information Retrieval System
- Types Of Information Retrieval system
- Steps involved in Projects
- Data extraction
- Data Preprocessing
- Embeddings Generation
- Storing embeddings and metadata
- Retrieving the data

Information Retrieval System

- Task of finding relevant information (documents/book/product) that satisfies the user's Information need.

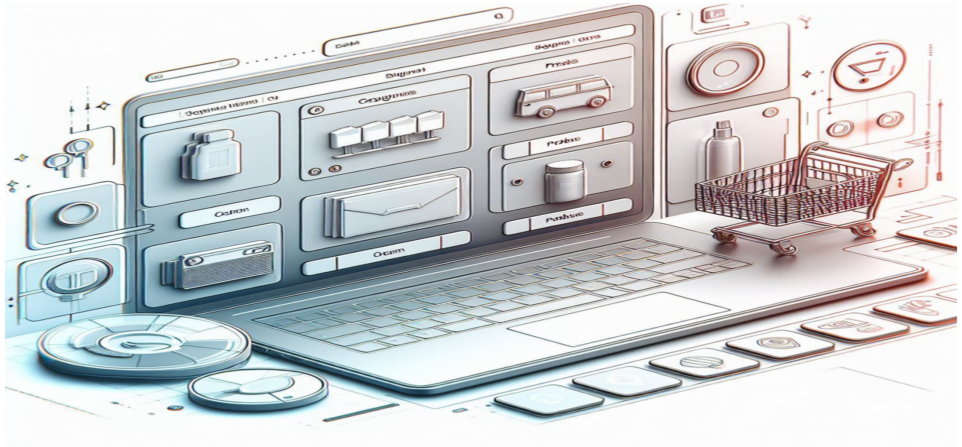


Applications Of Information Retrieval System



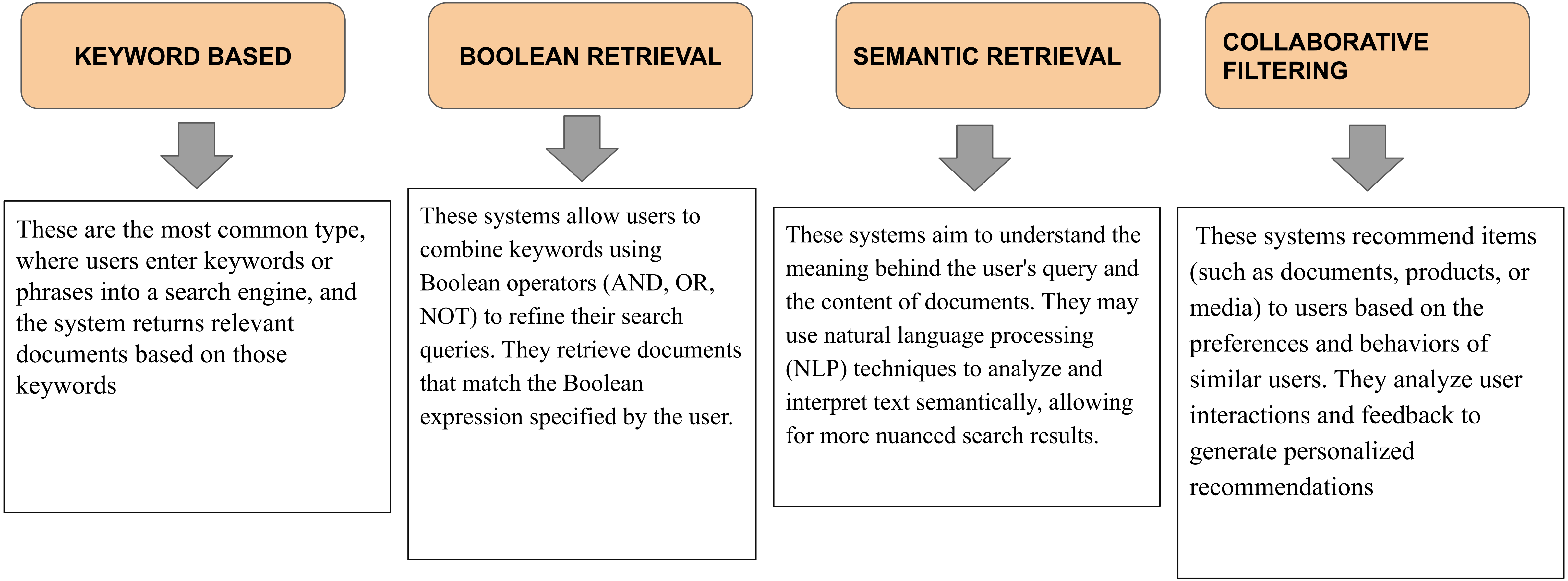
Online Book Store

Search Engine



E-Commerce Store

Types Of Information Retrieval System



KEYWORD BASED

These are the most common type, where users enter keywords or phrases into a search engine, and the system returns relevant documents based on those keywords

BOOLEAN RETRIEVAL

These systems allow users to combine keywords using Boolean operators (AND, OR, NOT) to refine their search queries. They retrieve documents that match the Boolean expression specified by the user.

SEMANTIC RETRIEVAL

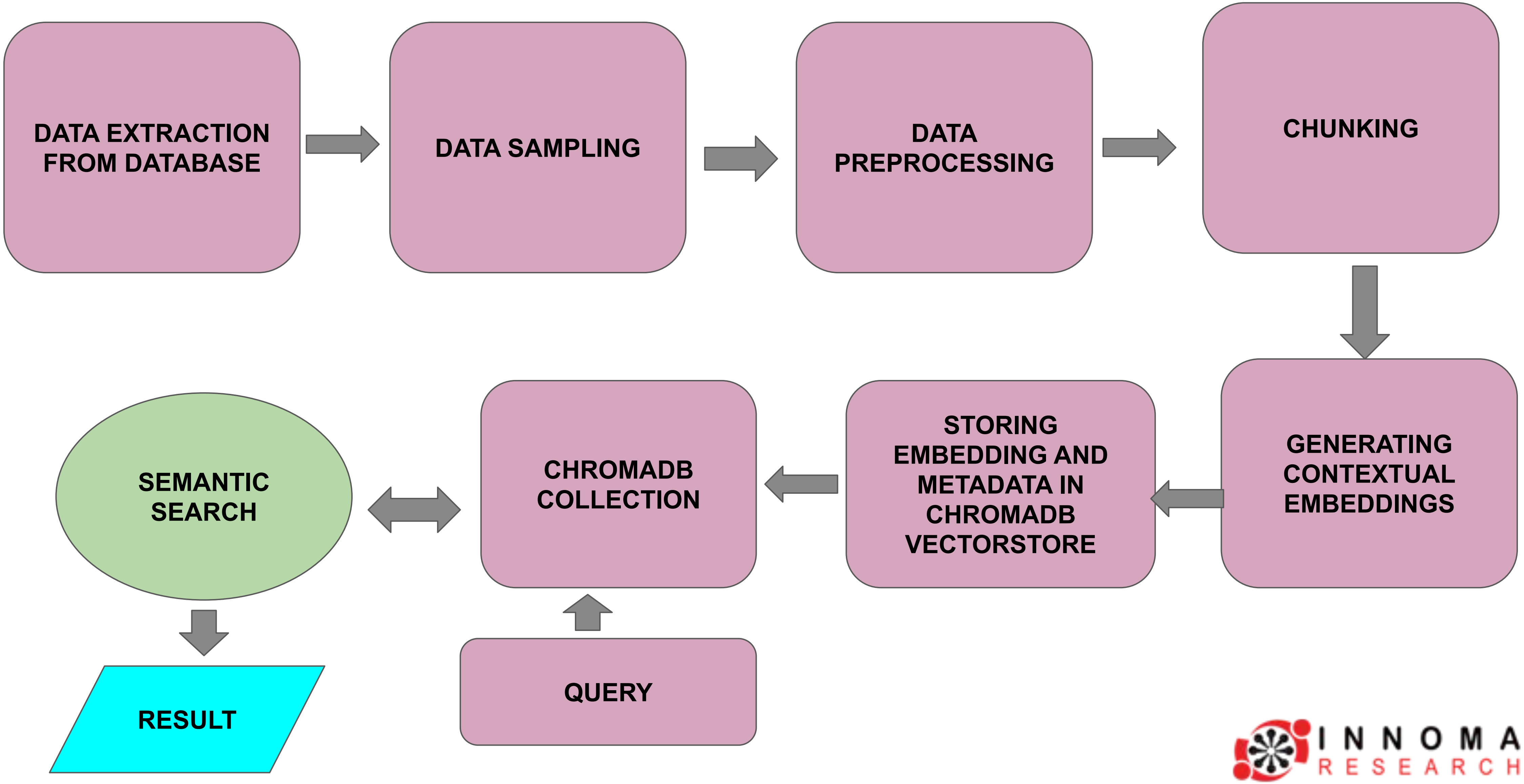
These systems aim to understand the meaning behind the user's query and the content of documents. They may use natural language processing (NLP) techniques to analyze and interpret text semantically, allowing for more nuanced search results.

COLLABORATIVE FILTERING

These systems recommend items (such as documents, products, or media) to users based on the preferences and behaviors of similar users. They analyze user interactions and feedback to generate personalized recommendations

INNOMATICS
RESEARCH LABS

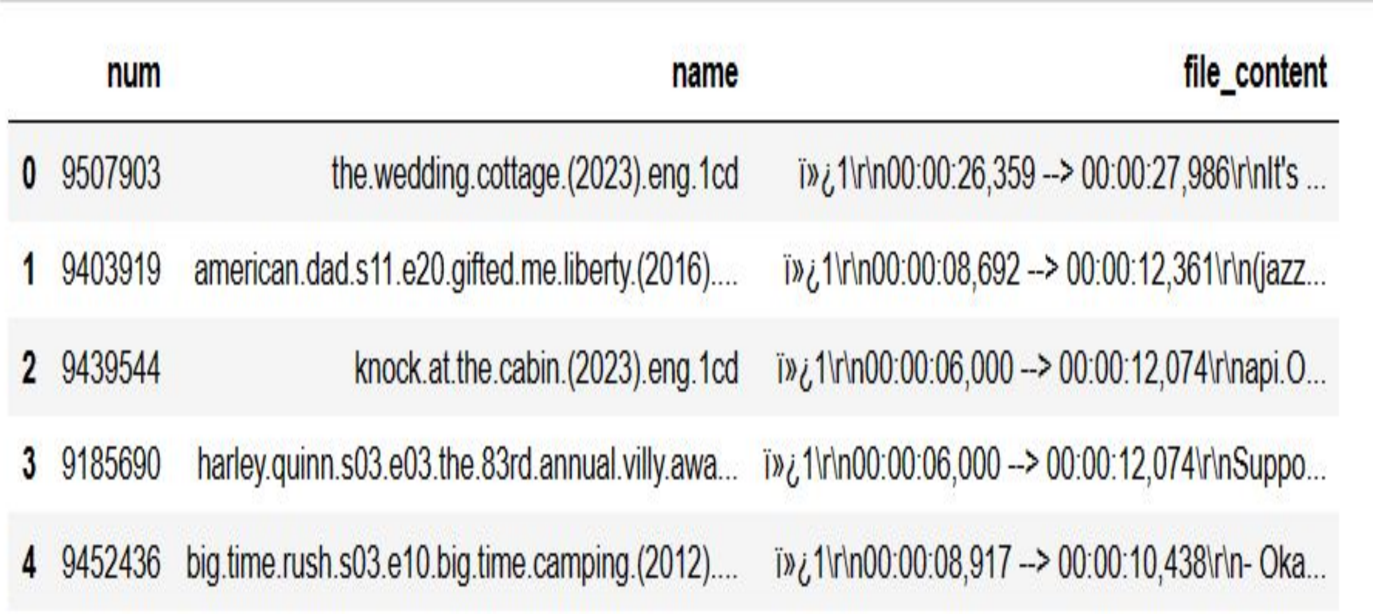
Steps Involved In The Project


DATA EXTRACTION FROM DATABASE
DATA SAMPLING
DATA PREPROCESSING
CHUNKING
GENERATING CONTEXTUAL EMBEDDINGS
STORING EMBEDDING AND METADATA IN CHROMADB VECTORSTORE
CHROMADB COLLECTION
SEMANTIC SEARCH
QUERY
RESULT

INNOMATICS RESEARCH LABS

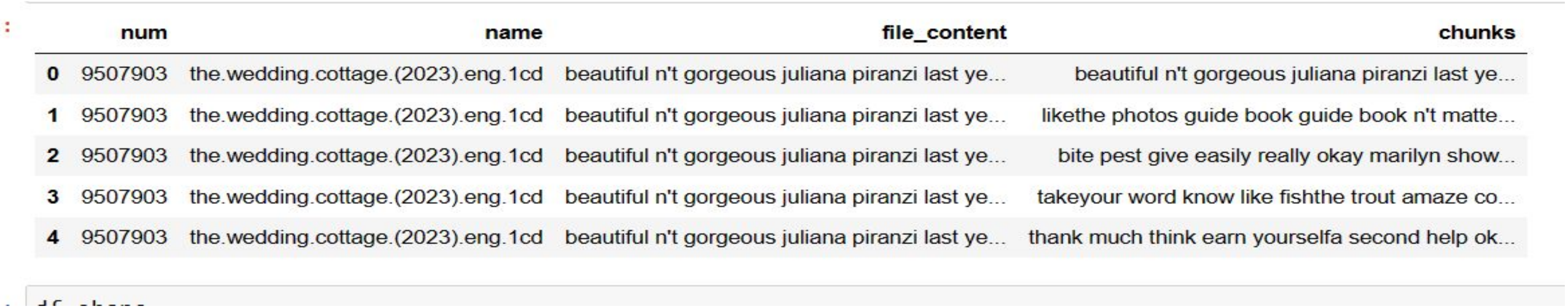
Data Extraction and Sampling

- Extracting the data from the database.
- Originally data consist of 82498 srt files
- It consist of the number,name of the file and file content.
- Taking 30 percent of data using random sampling because of computational constraint.
- Now our data has 24749 files.
- Decompressing the data



	num	name	file_content
0	9507903	the.wedding.cottage.(2023).eng.1cd	ï»¿1\r\n00:00:26,359 --> 00:00:27,986\r\nIt's ...
1	9403919	american.dad.s11.e20.gifted.me.liberty.(2016)....	ï»¿1\r\n00:00:08,692 --> 00:00:12,361\r\n(jazz...
2	9439544	knock.at.the.cabin.(2023).eng.1cd	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\napi.O...
3	9185690	harley.quinn.s03.e03.the.83rd.annual.villy.awa...	ï»¿1\r\n00:00:06,000 --> 00:00:12,074\r\nSuppo...
4	9452436	big.time.rush.s03.e10.big.time.camping.(2012)....	ï»¿1\r\n00:00:08,917 --> 00:00:10,438\r\n- Oka...

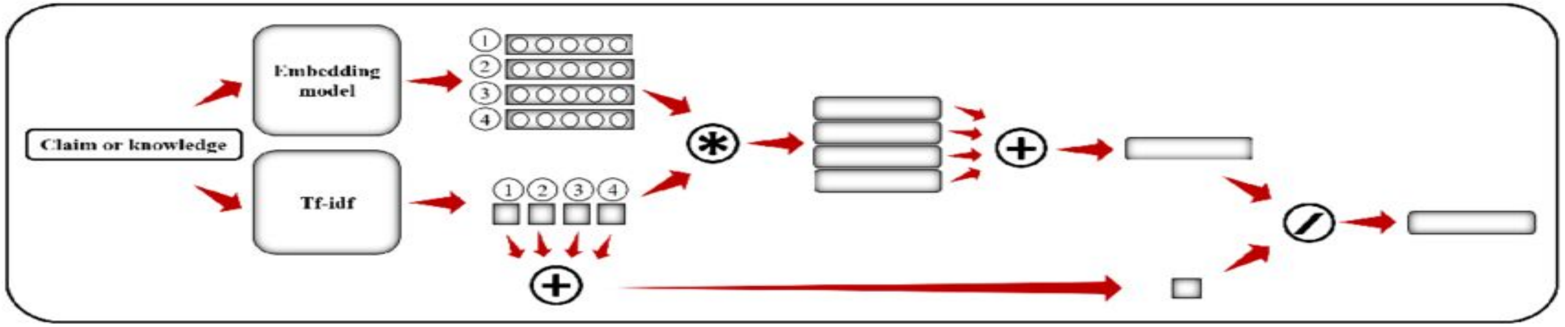
Data Preprocessing and Chunking



	num	name	file_content	chunks
0	9507903	the.wedding.cottage.(2023).eng.1cd	beautiful n't gorgeous juliana piranzi last ye...	beautiful n't gorgeous juliana piranzi last ye...
1	9507903	the.wedding.cottage.(2023).eng.1cd	beautiful n't gorgeous juliana piranzi last ye...	likethe photos guide book guide book n't matte...
2	9507903	the.wedding.cottage.(2023).eng.1cd	beautiful n't gorgeous juliana piranzi last ye...	bite pest give easily really okay marilyn show...
3	9507903	the.wedding.cottage.(2023).eng.1cd	beautiful n't gorgeous juliana piranzi last ye...	takeyour word know like fishthe trout amaze co...
4	9507903	the.wedding.cottage.(2023).eng.1cd	beautiful n't gorgeous juliana piranzi last ye...	thank much think earn yourselfa second help ok...

- we require file content column for data preprocessing.
- Removed special characters and digits .
- Removed stopwords
- Lemmatization
- Created chunks with chunk size of 500 tokens and overlap of 20 tokens.
- Finally we got 136536 rows after chunking.

Embeddings for Document Representation



- **Bag-of-Words (BOW) / TF-IDF**
  - Represents documents as vectors based on word counts and importance.
  - Suitable for keyword-based search engines.
- **Document Chunker**
  - Divides large documents into manageable chunks (e.g., 500-token windows) with overlaps.
  - Ensures efficient embedding without losing context.
- **BERT-based Sentence Transformers**
  - Utilizes pre-trained BERT models to encode semantic information.
  - Enables development of a semantic search engine capturing contextual meaning.

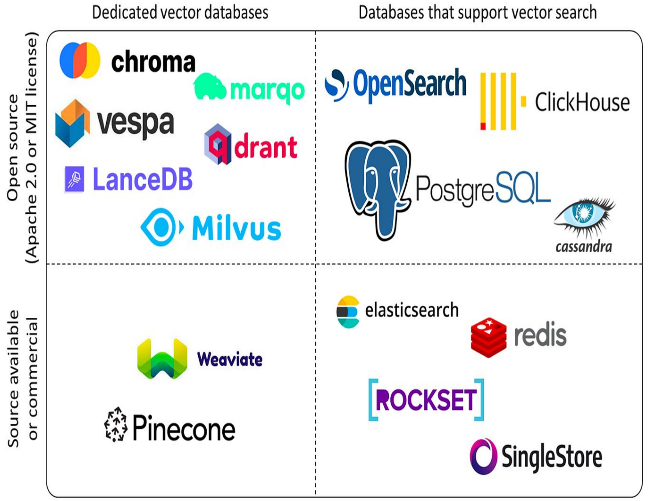
TF-IDF is having issue called higher dimensionality and no semantic meaning. So, preferred BERT based search engine due to less dimensions and semantic context.

Introduction to Vector Databases

Vector databases are specialized databases designed to efficiently store and query high-dimensional vector data, often used in applications involving machine learning, natural language processing, and image analysis.

Types of Vector Databases:

1. **Open Source Vector Databases:** are freely available and typically developed collaboratively by the community.Examples: Chroma DB, Milvus, Annoy, Marqo, etc.
2. **Commercial Vector Databases:**are proprietary solutions offered by companies for enterprise use.Examples: Pinecone DB, ArangoDB, etc.


Dedicated vector databases
Databases that support vector search
Open source (Apache 2.0 or MIT license)
Source available or commercial
chroma
marqo
vespa
qdrant
LanceDB
Milvus
OpenSearch
ClickHouse
PostgreSQL
cassandra
Weaviate
Pinecone
elasticsearch
redis
ROCKSET
SingleStore

Use Cases of Vector Databases:

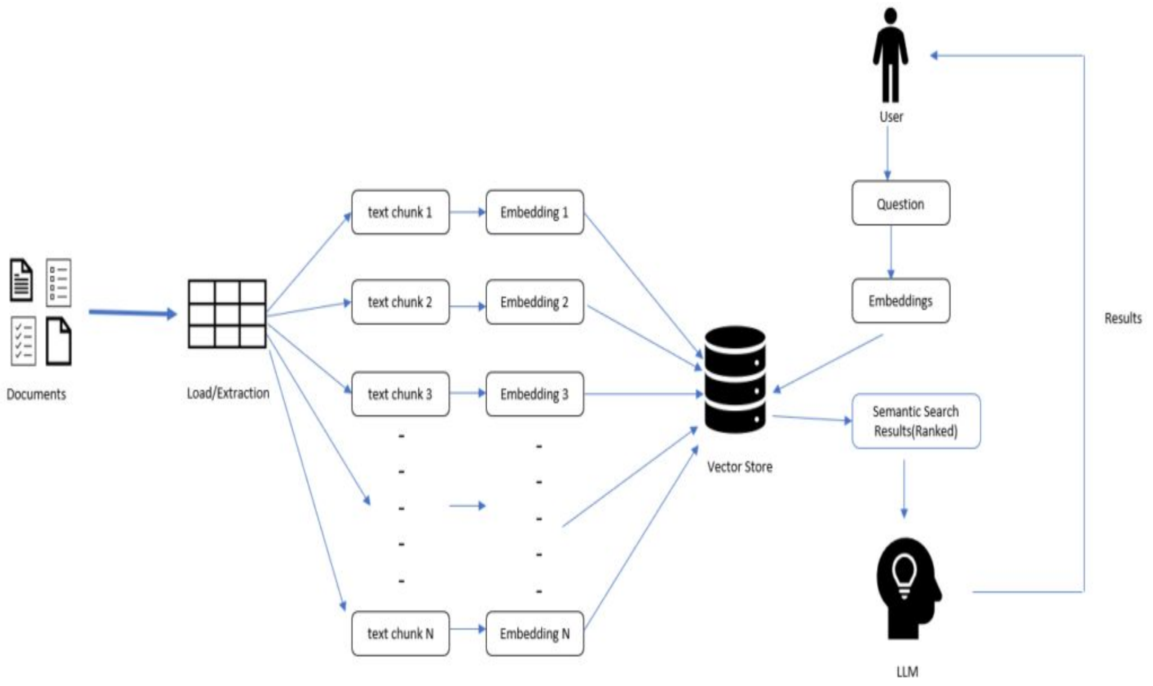
- Search Engines:
- Recommendation Systems:
- Image and Video Analysis:

In our project we used open source Vector database: Chroma

Storing Embeddings into Chroma DB

- **Description:**
  - Designed for storing and querying high-dimensional vector data efficiently.
  - Optimized for similarity search operations.
- **Key Features:**
  - Supports nearest neighbor search based on **cosine similarity**.
  - Efficient indexing and retrieval of embeddings.
- **Use Case:**
  - Ideal for search engines using advanced embeddings (e.g., BERT-based) to retrieve similar documents.

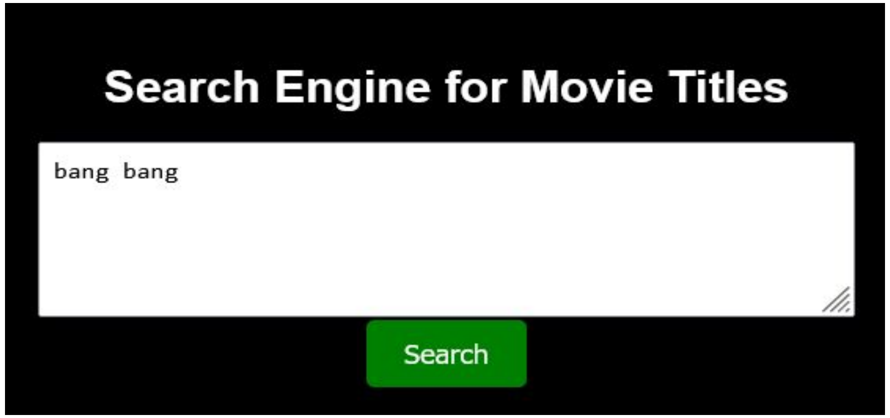
Retrieving Documents



Process Overview:

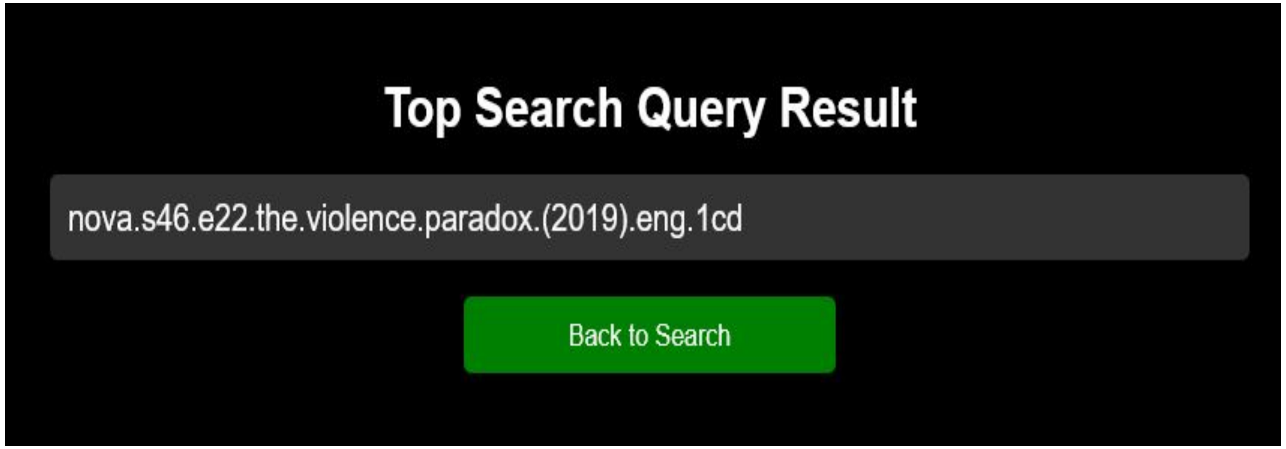
1. **User Query Processing:**
   - User's search query.
   - **Preprocessing (if Required):**
     - Tokenization: Splitting the query into individual words or tokens.
     - Cleaning: Removing stop words, punctuation, and other noise.
     - Normalization: Converting words to lowercase, handling synonyms, etc.
2. **Query Embedding:**
   - **Embedding Technique:** Utilize BERT-based SentenceTransformers or similar methods.
   - **Purpose:** Convert the processed query into a numerical vector representation capturing semantic meaning.
3. **Cosine Similarity Calculation:**
   - **Definition:** Cosine similarity measures the angle between two vectors.
   - **Use:** Calculate cosine similarity between the query embedding vector and document embedding vectors.
4. **Ranking Documents:**
   - **Scoring Mechanism:** Higher cosine similarity scores indicate greater relevance.
   - **Result:** Retrieve and rank documents based on similarity scores.

Flask User Interface



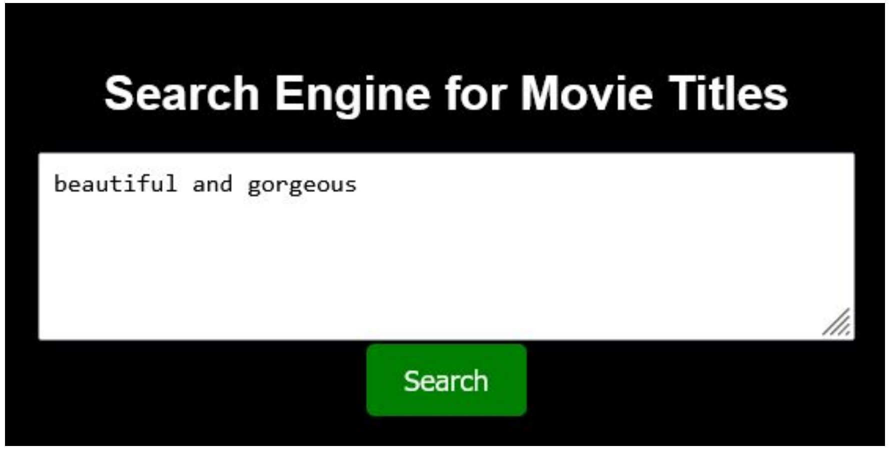
Search Engine for Movie Titles

bang bang



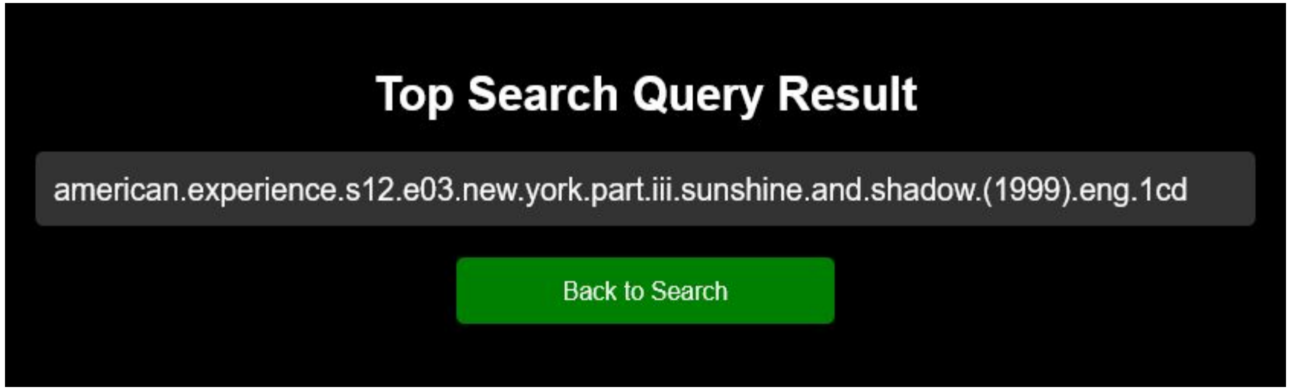
Top Search Query Result

nova.s46.e22.the.violence.paradox.(2019).eng.1cd



Search Engine for Movie Titles

beautiful and gorgeous



Top Search Query Result

american.experience.s12.e03.new.york.part.iii.sunshine.and.shadow.(1999).eng.1cd

INNOMATICS RESEARCH LABS

Q & A

Thank you!