INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Movie Subtitle Similarity - Search Engine

By,
Team ID– T211141
Shirisha Chintalapudi – IN1240659
Rakesh Kandi – IN1240894

Objective

The objective of this project is to develop an advanced search engine algorithm specifically tailored for retrieving video subtitles based on user queries. The primary focus is on enhancing search relevance and accuracy through the use of natural language processing (NLP) and machine learning (ML) techniques.

Introduction

The introduction outlines a project aimed at building an advanced search engine for movie and series subtitles using NLP and ML techniques to improve relevance and accuracy. The project emphasizes moving beyond basic keyword matching to understand the semantic context of user queries and subtitle content. By implementing this system in Python with libraries like pandas, scikit-learn, numpy, and Flask, the goal is to create a robust and efficient subtitle similarity search engine that enhances the accessibility and usability of video content.

Importing Libraries:

The process begins by importing necessary libraries such as zipfile, io, pandas, numpy, re, nltk, string, and scikit-learn.
These libraries are essential for data manipulation, text preprocessing, feature extraction, and similarity computation.

Preprocessing Text Data:

The clean_text() function is applied to the subtitle content to perform various text cleaning operations, including removing noise and standardizing the text format. This ensures that the text data is suitable for further processing and analysis.

Initializing TF-IDF Vectorizer:

A TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is initialized using the TfidfVectorizer class from scikit-learn. Parameters such as max_df, min_df, and max_features are specified to control the features extracted from the text and optimize performance.

Fit and Transform TF-IDF Vectors:

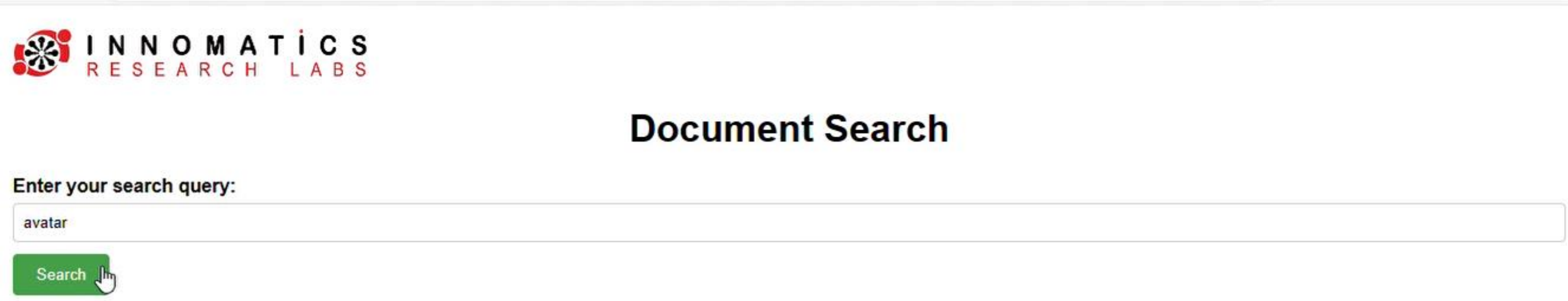
The TF-IDF vectorizer is fitted to the preprocessed subtitle content using the fit_transform() method. This converts the text data into numerical feature vectors, representing the importance of each word in the corpus.

Compute Cosine Similarity Matrix:

The cosine similarity matrix is computed between TF-IDF vectors using the cosine_similarity() function from scikit-learn. This calculates the cosine similarity between each pair of subtitle vectors, providing a measure of their similarity.

Building the Flask Web App:

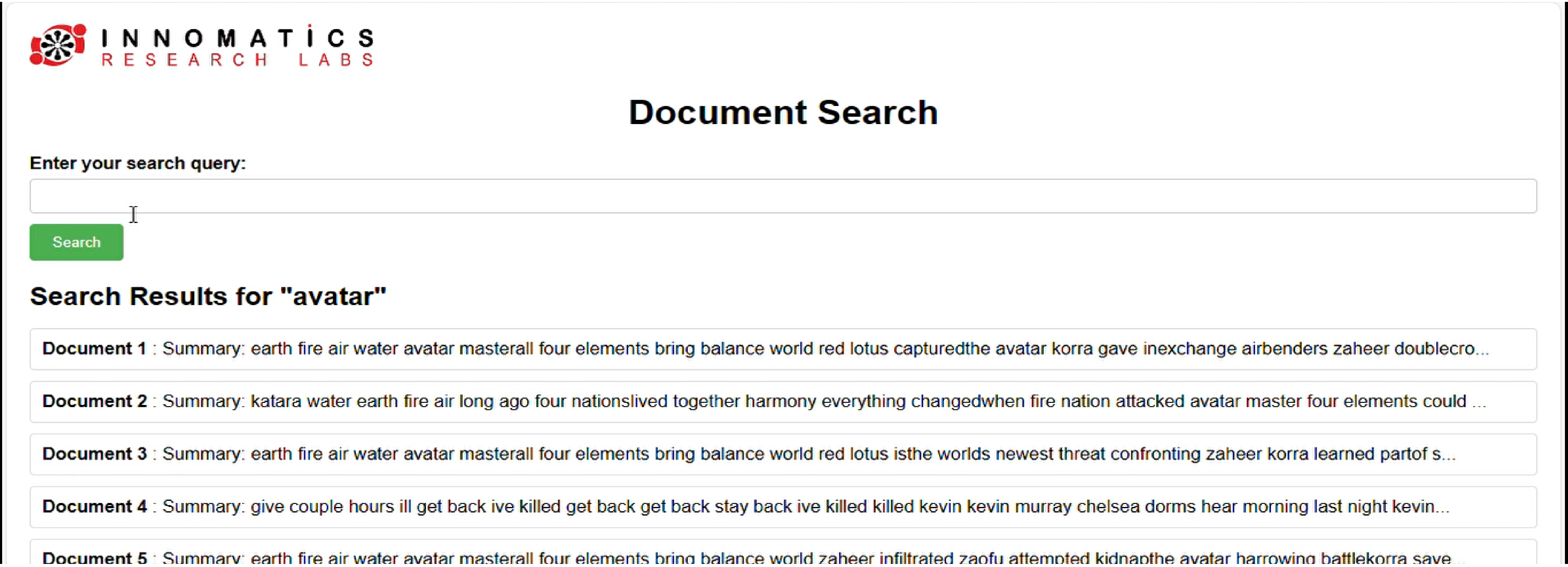
The main function of the application involves building a Flask web app for interactive movie and series subtitle similarity search. The app allows users to enter a query text, which is then compared against the subtitle corpus to find similar subtitles. The top matching subtitles, along with their similarity scores and snippets of content, are displayed to the user in the result page.



Document Search

Enter your search query:

avatar



Document Search

Enter your search query:

Search Results for "avatar"

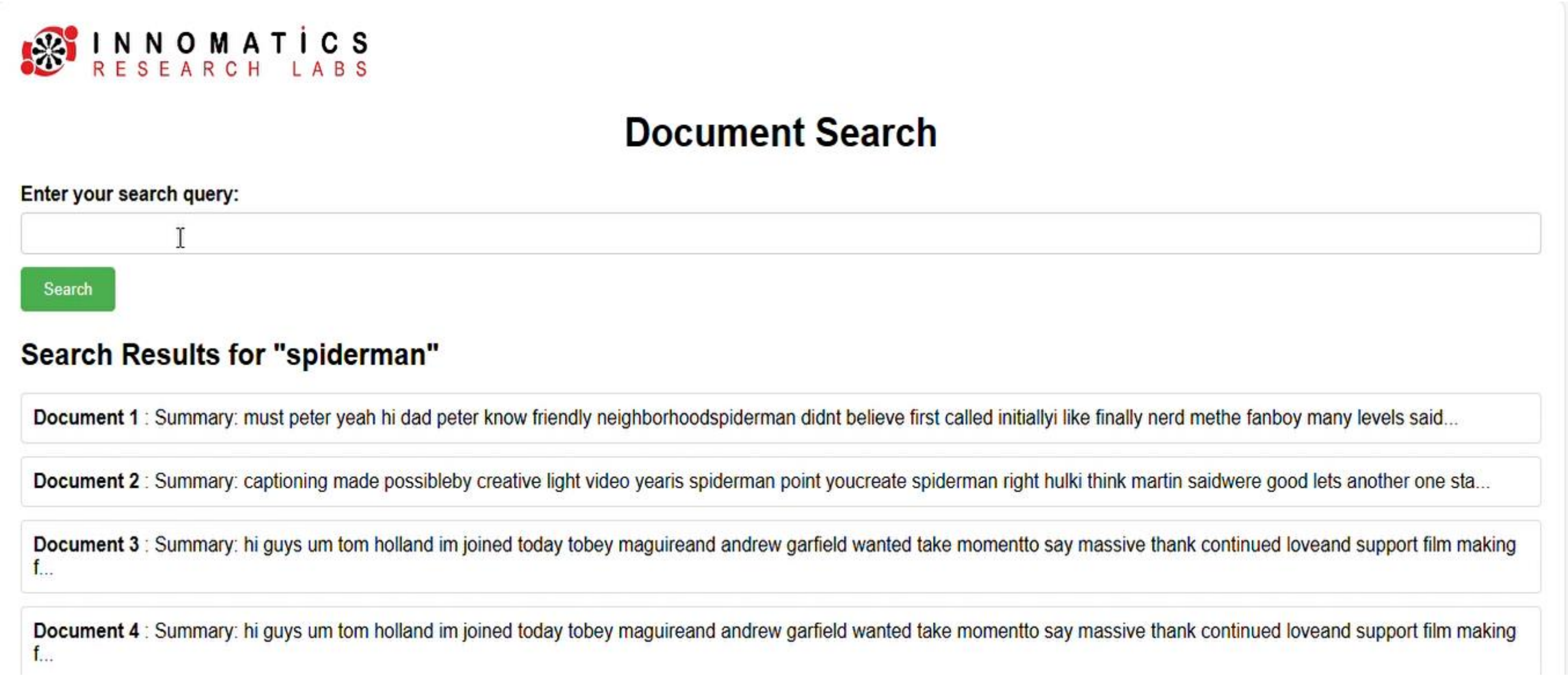
Document 1 : Summary: earth fire air water avatar masterall four elements bring balance world red lotus capturedthe avatar korra gave inexchange airbenders zaheer doublecro...

Document 2 : Summary: katara water earth fire air long ago four nationslived together harmony everything changedwhen fire nation attacked avatar master four elements could ...

Document 3 : Summary: earth fire air water avatar masterall four elements bring balance world red lotus isthe worlds newest threat confronting zaheer korra learned partof s...

Document 4 : Summary: give couple hours ill get back ive killed get back get back stay back ive killed killed kevin kevin murray chelsea dorms hear morning last night kevin...

Document 5 : Summary: earth fire air water avatar masterall four elements bring balance world zaheer infiltrated zaofu attempted kidnapthe avatar harrowing battlekorra save...



Document Search

Enter your search query:

Search Results for "spiderman"

Document 1 : Summary: must peter yeah hi dad peter know friendly neighborhoodspiderman didnt believe first called initiallyi like finally nerd methe fanboy many levels said...

Document 2 : Summary: captioning made possibleby creative light video yearis spiderman point youcreate spiderman right hulki think martin saidwere good lets another one sta...

Document 3 : Summary: hi guys um tom holland im joined today tobey maguireand andrew garfield wanted take momentto say massive thank continued loveand support film making f...

Document 4 : Summary: hi guys um tom holland im joined today tobey maguireand andrew garfield wanted take momentto say massive thank continued loveand support film making f...

THANK YOU