AI-Enhanced Information Retrieval System with Big Data Analytics and NLP Transformers

YUNUS EMRE ISIKDEMIR

MOTIVATION

As the volume of data continues to grow, obtaining relevant information from it becomes more challenging, particularly since not all parts of the data are equally important. Therefore, it is necessary to handle, filter, and process the data before modeling it to obtain better inferences. In this regard, Big Data Tools are essential, as they can efficiently process vast amounts of data and retrieve relevant information in a matter of seconds. To this end, various AI-based technologies such as ChatGPT and BARD have emerged, which can extract abstractive information from this massive amount of data.

In response to the challenge of extracting relevant information from large volumes of data, I developed an AI-based information retrieval system in this study. The system enables rapid access to pertinent data and performs specific tasks based on user specifications.

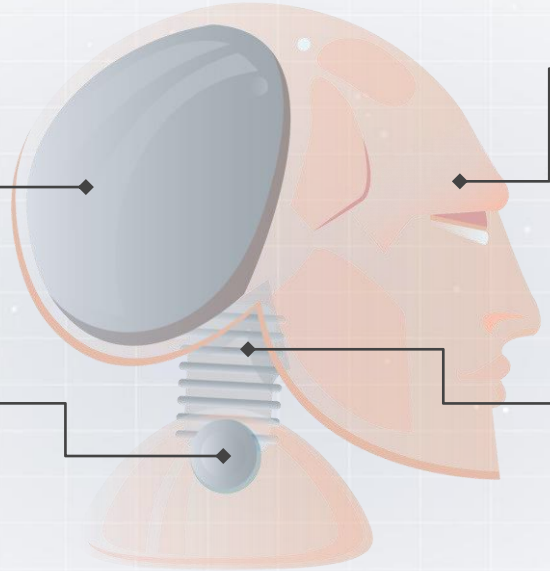
PROBLEM DESCRIPTION

NLU

Natural Language Understanding based on context

Q&A

Question answering from given passage



SIMILARITY

Identifying the most pertinent documents related to a query

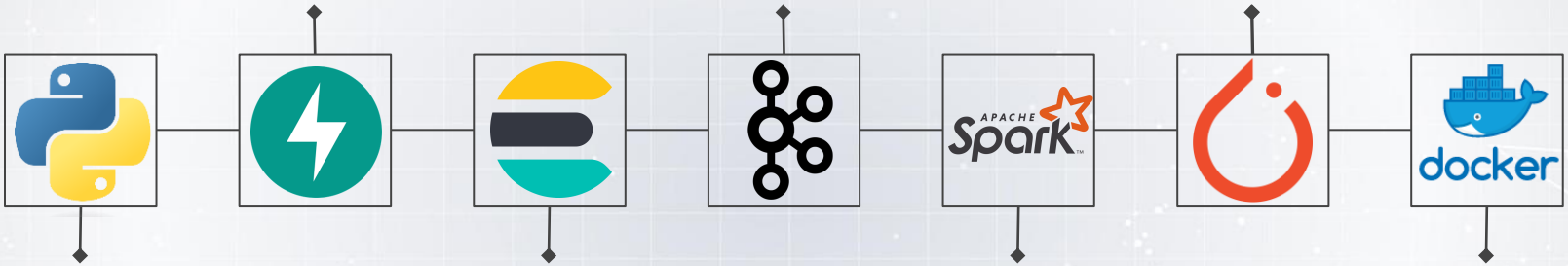
SUMMARIZE

Extraction of most informative sections from larger texts

01

TECHNOLOGIES

Languages, Tools and Frameworks



PYTHON
Language to logic implementaion

FASTAPI
HTTP requests for AI models

ELASTICSEARCH
Storing and indexing the documents

APACHE KAFKA
Data streaming and real-time data pipeline

APACHE SPARK
Real-time data processing and ETL

PYTORCH
Training and Finetuning the AI Models

DOCKER
Containerization and isolation of the applications

02

DESIGN SPECIFICATIONS

STAGES OF THE PIPELINE

01

ETL

The data is first preprocessed with Spark, then consumed from Kafka, before being utilized in Applications.

02

QUERY

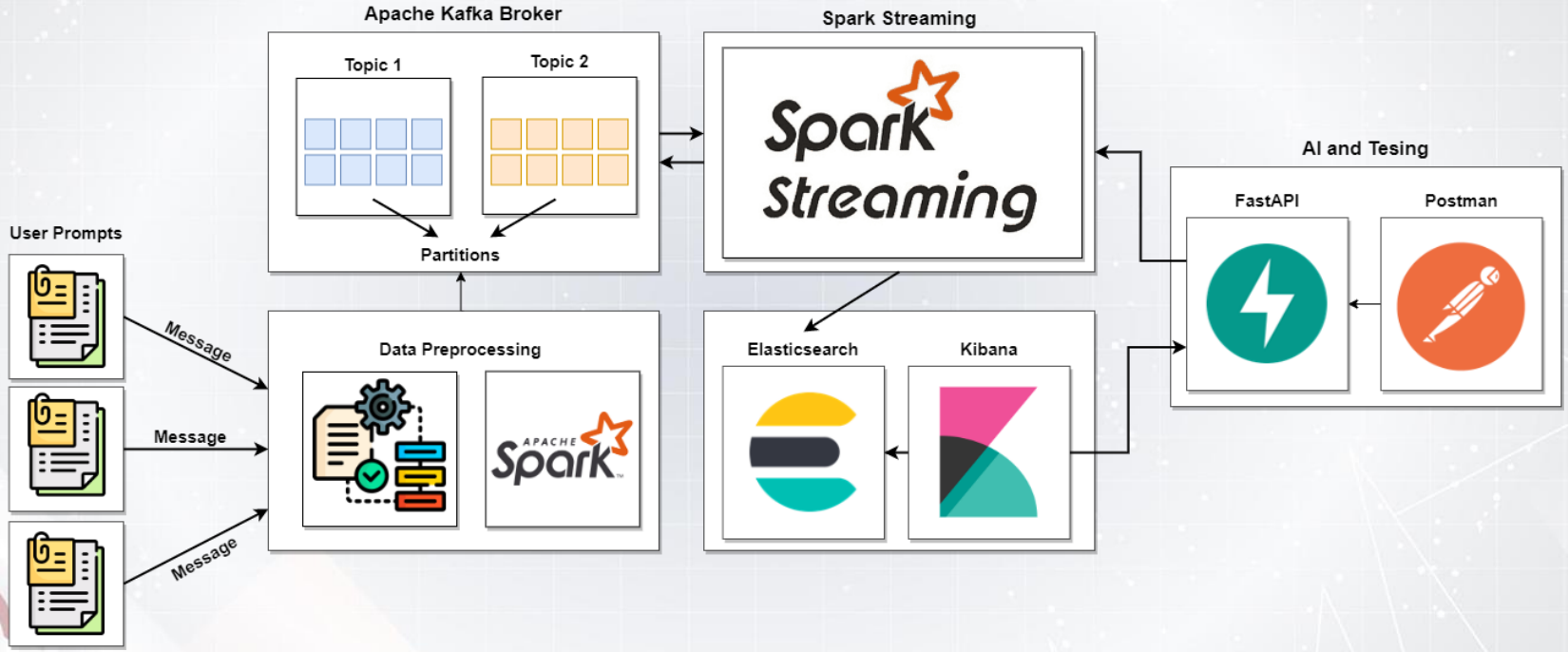
Prompt a query and send a HTTP request to trigger a particular operation

03

RETRIEVE

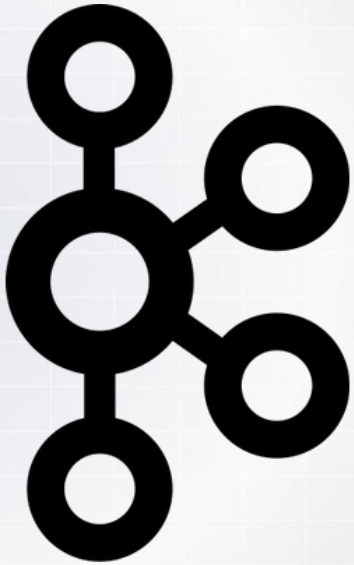
Retrieve relevant data from the database and perform specific actions based on the query.

DATA INGESTION PIPELINE



DATA STREAMING

Data is continuously streamed into Kafka in real-time, making it readily available for consumption by all applications.



DATA WRANGLING

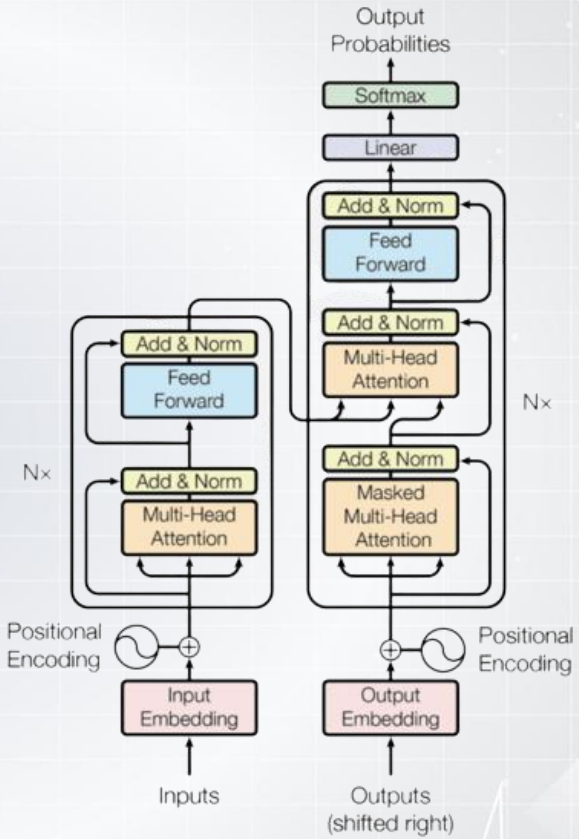
Apache Spark is employed to preprocess text data prior to publishing it, as well as for performing read and write operations with Elasticsearch. By processing queries similarly to indexed documents, the data can be efficiently prepared for AI model execution.

DOCUMENT INDEXING

Elasticsearch is used to store and index documents, allowing for efficient retrieval of information relevant to a given query.

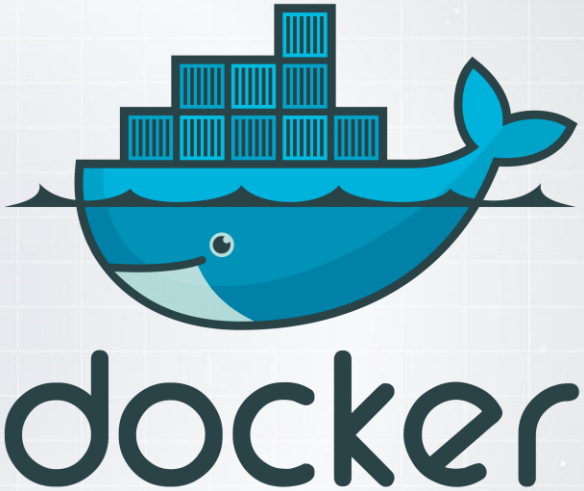
NLP AND NLU

Cutting-edge NLP architectures are utilized to comprehend the provided prompt and carry out a particular task.



CONTAINERIZATION

The services are containerized and isolated using Docker, a platform that allows for efficient and consistent deployment of applications across different environments.

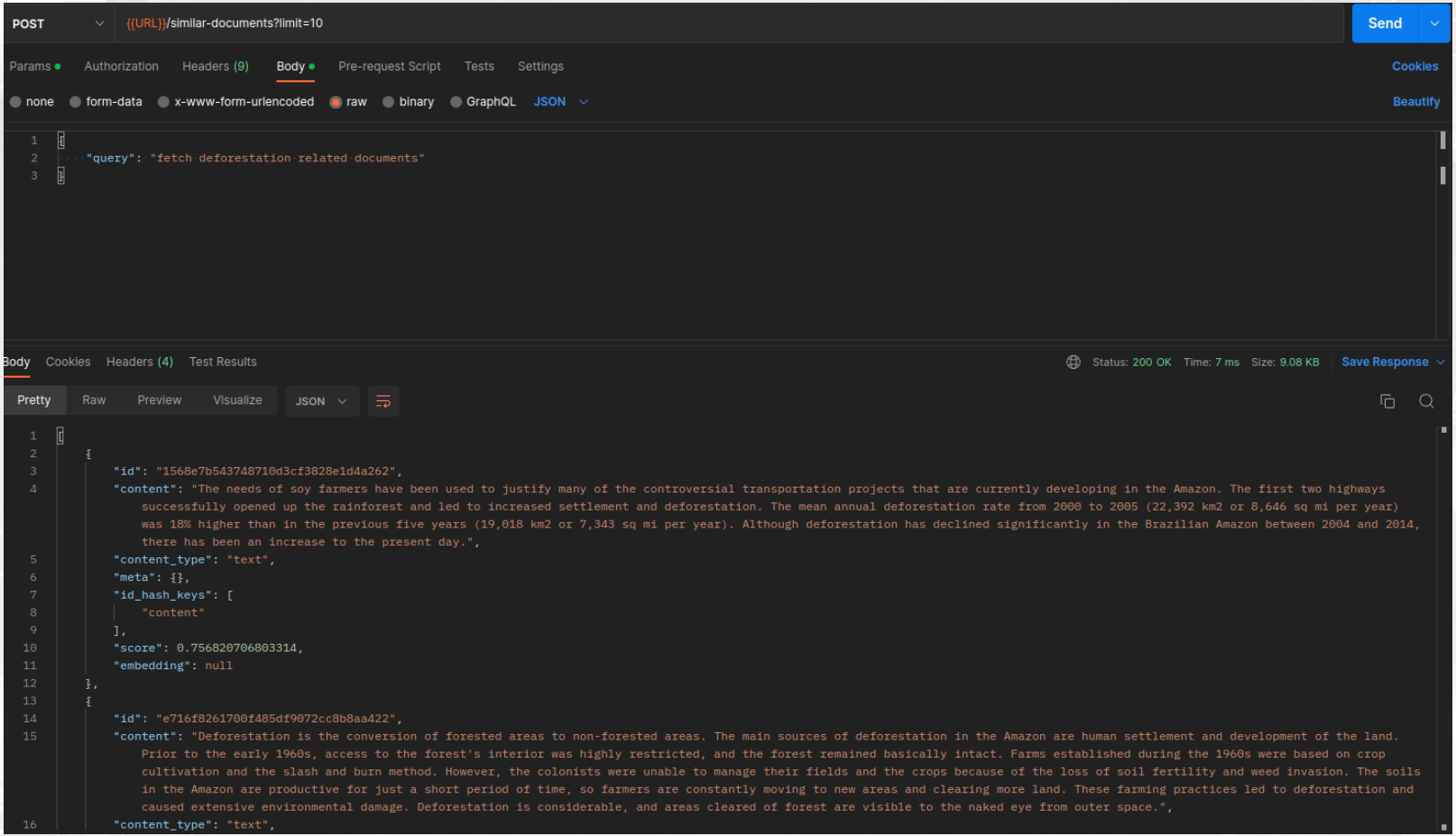


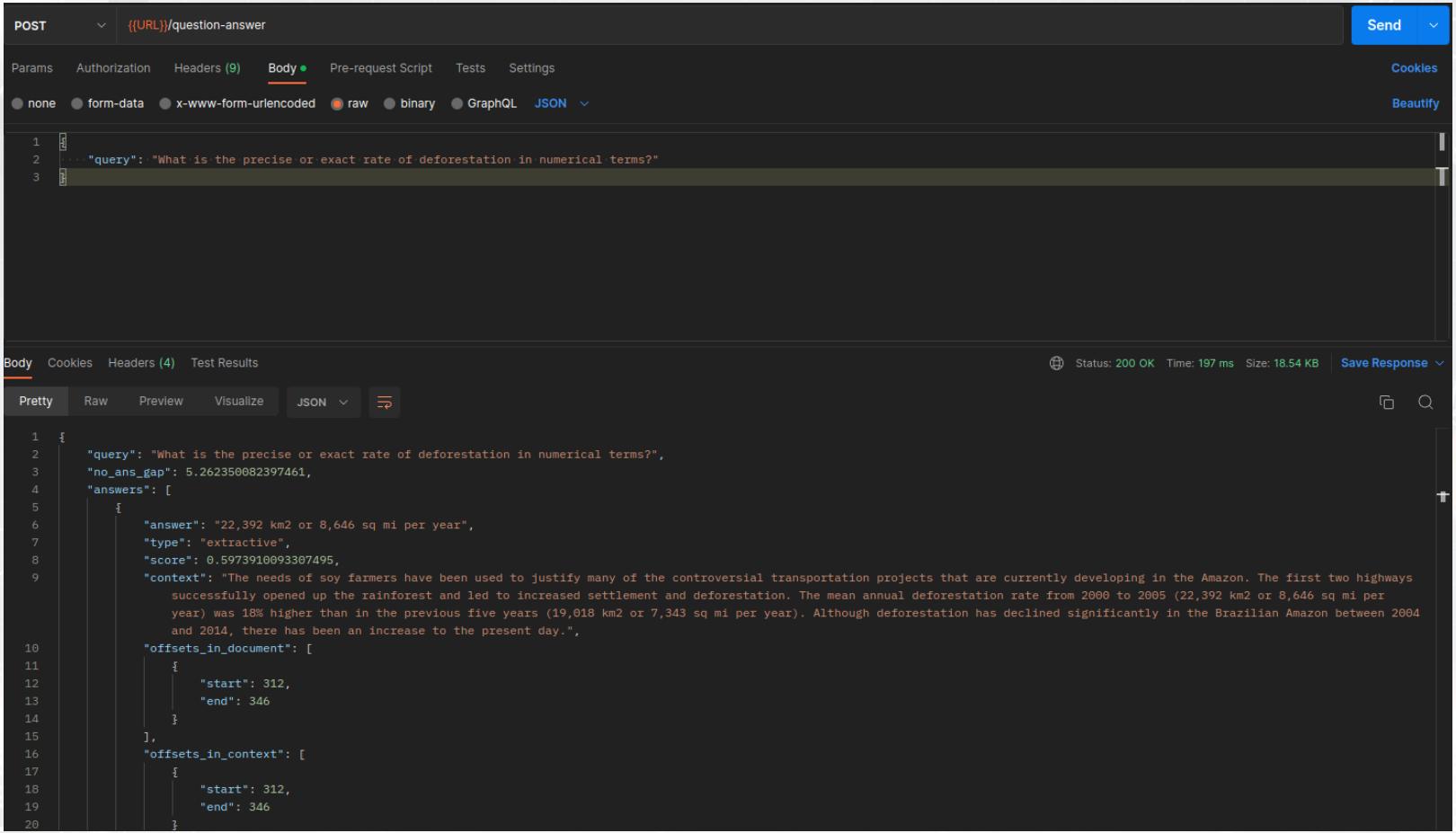
API DEVELOPMENT

The provision of communication between the services is facilitated through HTTP requests via FastAPI.

03

EXPERIMENTAL RESULTS


POST
{{URL}}/similar-documents?limit=10
Send
Params
Authorization
Headers (9)
Body
Pre-request Script
Tests
Settings
Cookies
none
form-data
x-www-form-urlencoded
raw
binary
GraphQL
JSON
Beautify
{
"query": "fetch deforestation related documents"
}
Body
Cookies
Headers (4)
Test Results
Status: 200 OK
Time: 7 ms
Size: 9.08 KB
Save Response
Pretty
Raw
Preview
Visualize
JSON
[
{
"id": "1568e7b543748710d3cf3828e1d4a262",
"content": "The needs of soy farmers have been used to justify many of the controversial transportation projects that are currently developing in the Amazon. The first two highways successfully opened up the rainforest and led to increased settlement and deforestation. The mean annual deforestation rate from 2000 to 2005 (22,392 km2 or 8,646 sq mi per year) was 18% higher than in the previous five years (19,018 km2 or 7,343 sq mi per year). Although deforestation has declined significantly in the Brazilian Amazon between 2004 and 2014, there has been an increase to the present day.",
"content_type": "text",
"meta": {},
"id_hash_keys": [
"content"
],
"score": 0.756820706803314,
"embedding": null
},
{
"id": "e716f8261700f485df9072cc8b8aa422",
"content": "Deforestation is the conversion of forested areas to non-forested areas. The main sources of deforestation in the Amazon are human settlement and development of the land. Prior to the early 1960s, access to the forest's interior was highly restricted, and the forest remained basically intact. Farms established during the 1960s were based on crop cultivation and the slash and burn method. However, the colonists were unable to manage their fields and the crops because of the loss of soil fertility and weed invasion. The soils in the Amazon are productive for just a short period of time, so farmers are constantly moving to new areas and clearing more land. These farming practices led to deforestation and caused extensive environmental damage. Deforestation is considerable, and areas cleared of forest are visible to the naked eye from outer space.",
"content_type": "text",


POST
{{URL}}/question-answer
Send
Params
Authorization
Headers (9)
Body
Pre-request Script
Tests
Settings
Cookies
none
form-data
x-www-form-urlencoded
raw
binary
GraphQL
JSON
Beautify
{
"query": "What is the precise or exact rate of deforestation in numerical terms?"
}
Body
Cookies
Headers (4)
Test Results
Status: 200 OK
Time: 197 ms
Size: 18.54 KB
Save Response
Pretty
Raw
Preview
Visualize
JSON
{
"query": "What is the precise or exact rate of deforestation in numerical terms?",
"no_ans_gap": 5.262350082397461,
"answers": [
{
"answer": "22,392 km2 or 8,646 sq mi per year",
"type": "extractive",
"score": 0.5973910093307495,
"context": "The needs of soy farmers have been used to justify many of the controversial transportation projects that are currently developing in the Amazon. The first two highways successfully opened up the rainforest and led to increased settlement and deforestation. The mean annual deforestation rate from 2000 to 2005 (22,392 km2 or 8,646 sq mi per year) was 18% higher than in the previous five years (19,018 km2 or 7,343 sq mi per year). Although deforestation has declined significantly in the Brazilian Amazon between 2004 and 2014, there has been an increase to the present day.",
"offsets_in_document": [
{
"start": 312,
"end": 346
}
],
"offsets_in_context": [
{
"start": 312,
"end": 346
}