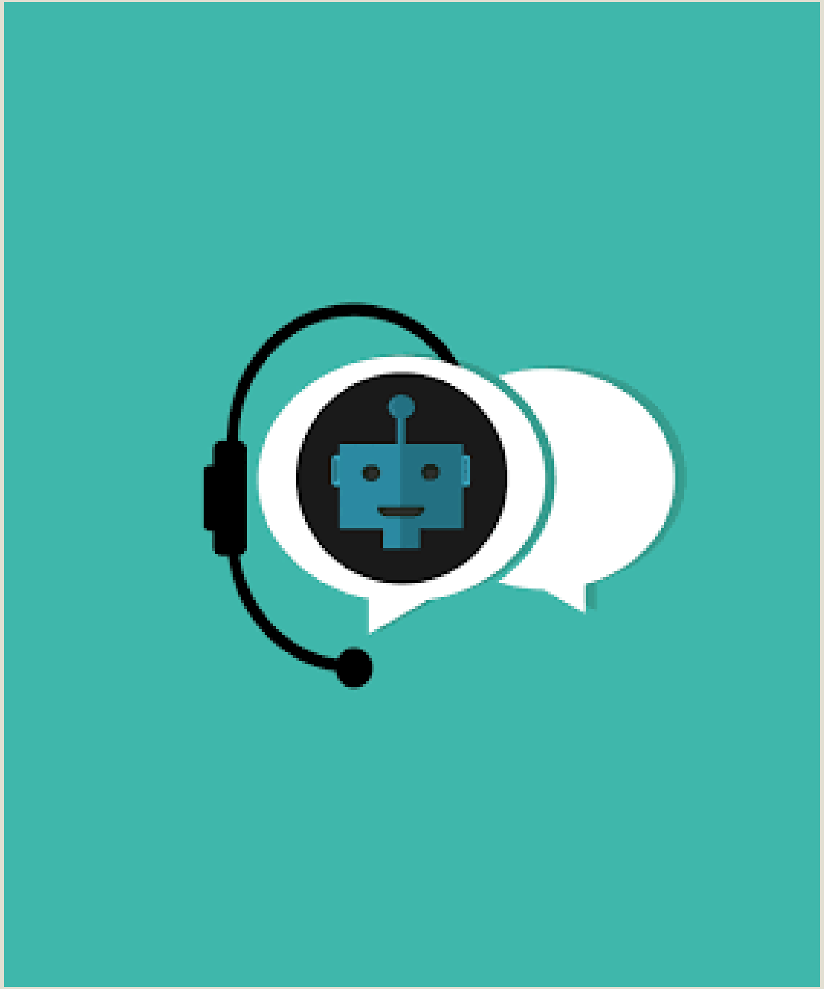
Retrieval-Augmented Generation (RAG) Chatbot



Mithun Senthil V
22PD22

Introduction

- In the ever-evolving digital landscape, Multi-Modal AI is revolutionizing the way we interact with information.
- By integrating various data types text, images, audio, and more. Multi-Modal AI systems enable more sophisticated and context-aware responses.
- This presentation delves into the innovative application of Multi-Modal AI in enhancing Retrieval-Augmented Generation (RAG) systems, ultimately driving more intelligent and dynamic user experiences.

PROJECT GOALS

1. Multimodal Chatbot purpose mainly as an assistant

2. Chatbot must fetch the relevant text and images from the knowledge base for the user query and present it, with minimal latency and hallucination

TECH STACK

Language: Python

Technology: RAG

Embedding model: text-embedding-004

image model: gemini-2.0-flash

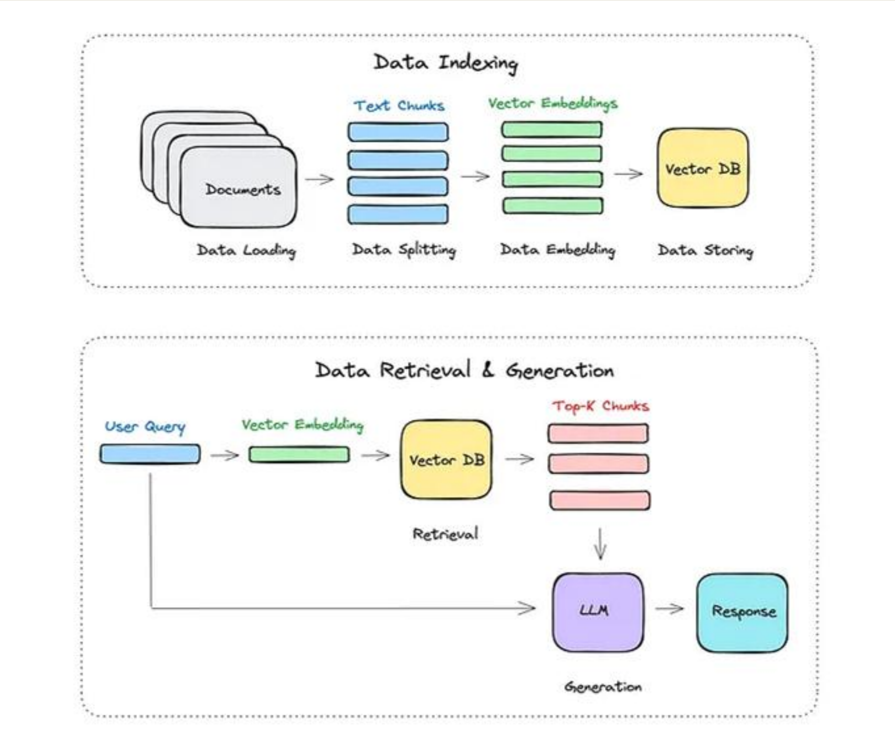
Chat model: Llama3-70b-8192

image processing: OpenCV

Vector DB: Pinecone Db

Note: No frameworks like langchian is used.

RAG PIPELINE



Vector Embedding : They are numerical representations of data points.

Vector Database : It is a specialized data management system designed to store, manage, and index high-dimensional vector embeddings.

Chunking : A process by which small individual pieces of information are bound together to create a meaningful whole in memory.

LLM : A large language model (LLM) is a type of artificial intelligence (AI) system capable of understanding, generating, and processing human language by training on vast amounts of text data.

Prompt : The input or instruction given to the model to elicit a specific response or output.

Project Overview

- This Retrieval-Augmented Generation (RAG) Chatbot uses Django framework to build a full-stack web application.
- Implemented a MultiModal RAG system, that queries from different document types like pdf, docs, csv, image and video, used different processing techniques for different document types.
- Knowledge Base:
  - Students Grading Dataset [CSV File]
  - Data Incident Management [PDF File]
  - Luxury Cars Image Dataset [Image Files]
  - Linear Regression in 3 Minutes [Video File]
  - Introduction to Machine Learning [DOC File]
- The user can also upload their own data into the system and can be queried from that also.

User Query Processing

- This is handled by rag_bot.py [rag\Utils\rag_bot.py]
- The **RAGbot** class has two methods **retrieve_context()** and **normal_response()**
- The **retrieve_context()** will take the user query as its input, the user query is embedding and searched in pinecone for all the related context and the function returns top 4 the matches.
- The **normal_response()** will take the user query and matches from the db and extracts all the context from the matches and feeds it to the **llama3-70b-8192** used via **groq API** to generate response for the user query. Used prompt engineering for optimal response.

Csv Query Processing

- If a csv content is present in the related context for the user query it will be processed differently.
- The summary of the csv along with the user query will be given to a LLM which will produce a pandas code for the user query, which will be executed and the pandas output is stored.
- The pandas code, pandas output along, csv summary along with the user query will be given to another LLM which will give a final response to the user query.

Insert file Processing

Each file type in handled differently to extract the context. Every file processing module is written as a class for modularity.

PDF File handling:

- This is handled by process_pdf.py [rag\Utils\process_pdf.py]
- Main function **extract_content()** and **process_pdf()**
- The **extract_content()** read the pdf and extracts all the text, tables and images.
- The **process_pdf()** chunks the content in the file each image and tables is a separate chunk and the text is splited using **_split_text()** function with chunk_size=1000, chunk_overlap=300
- Finally all these chunks are embedded upserted along with some metadata into vector DB (pinecone)

Docx File handling:

- This is handled by process_doc.py [rag\Utils\process_doc.py]
- Main function **extract_content()** and **process_documents()**
- The **extract_content()** read the Doc or txt file and extracts all the text, tables.
- The **process_documents()** chunks the content in the file each image and tables is a separate chunk and the text is splited using **_split_text()** function with chunk_size=1000, chunk_overlap=200
- Finally all these chunks are embedded upserted along with some metadata into vector DB (pinecone)

Csv File handling:

- This is handled by process_csv.py [rag\Utils\process_csv.py]
- Main function **generate_csv_summary()** and **process_csv()**
- The **generate_csv_summary()** read the csv file and extracts all the important information.
- The **process_csv()** create embedding for the csv summary and it is upserted into the vector DB (pinecone)

Image File handling:

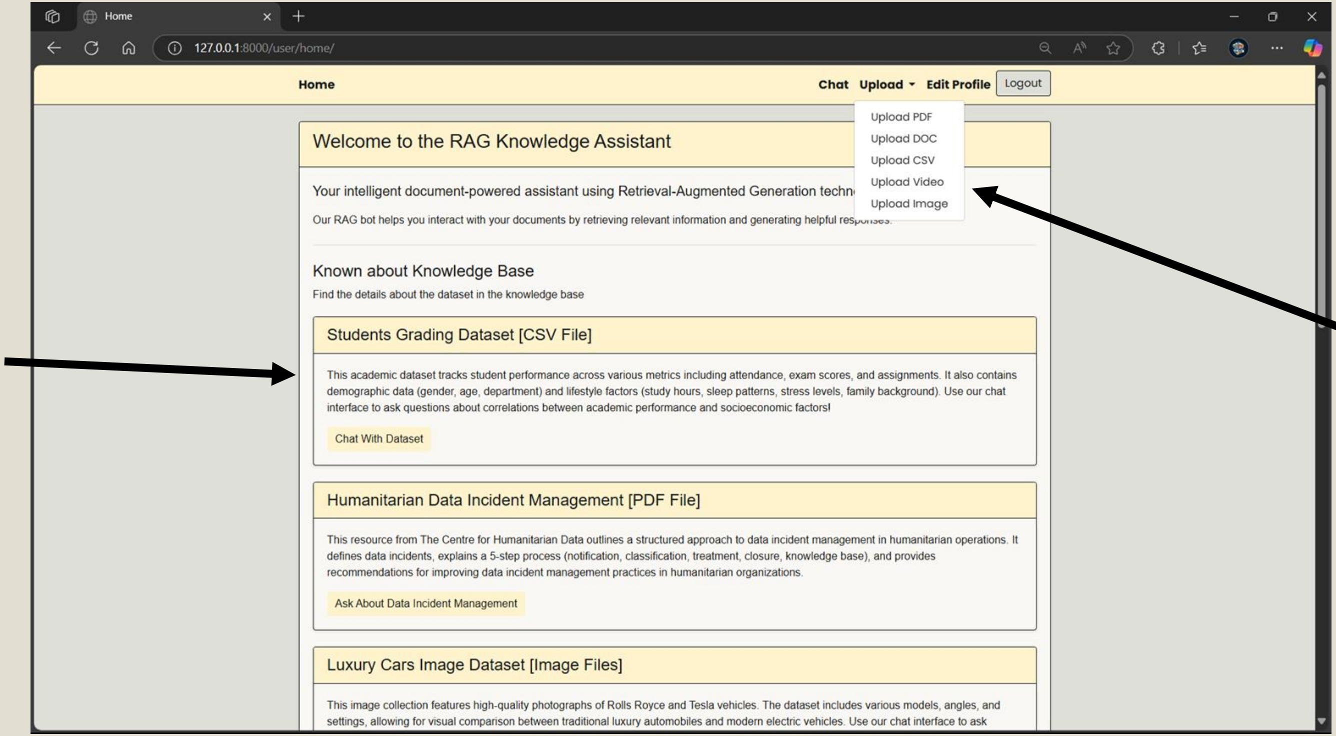
- This is handled by process_image.py [rag\Utils\process_ image.py]
- Main function **generate_image_summary()** and **process_image()**
- The **optimize_image()** is used to reduce the size of the image for storing, the **generate_image_summary()** is used to create a summary for the image using the **gemini-2.0-flash** which is the optimal model for image summarization.
- The image summary is embedded and upserted in the vector DB along with the metadata which contains the base64 encoding of the image file. This base64 encoding will be later used for displaying the image when required.

Video File handling:

- This is handled by process_video.py [rag\Utils\process_video.py]
- The main functions **process_video()** and **process_frames_with_transcript()** handle video processing.
- The **extract_frames()** is used to extract key frames from the video, while **transcribe_audio()** is used to extract audio transcribe and these both are given to **process_frames_with_transcript()** which creates context with **gemini-2.0-flash** using both audio and frames.
- Finally all these are embedded upserted along with some metadata into vector DB (pinecone).

Outputs

Description of the data in the knowledge base is given



Welcome to the RAG Knowledge Assistant

Your intelligent document-powered assistant using Retrieval-Augmented Generation techn

Our RAG bot helps you interact with your documents by retrieving relevant information and generating helpful responses.

Known about Knowledge Base

Find the details about the dataset in the knowledge base

Students Grading Dataset [CSV File]

This academic dataset tracks student performance across various metrics including attendance, exam scores, and assignments. It also contains demographic data (gender, age, department) and lifestyle factors (study hours, sleep patterns, stress levels, family background). Use our chat interface to ask questions about correlations between academic performance and socioeconomic factors!

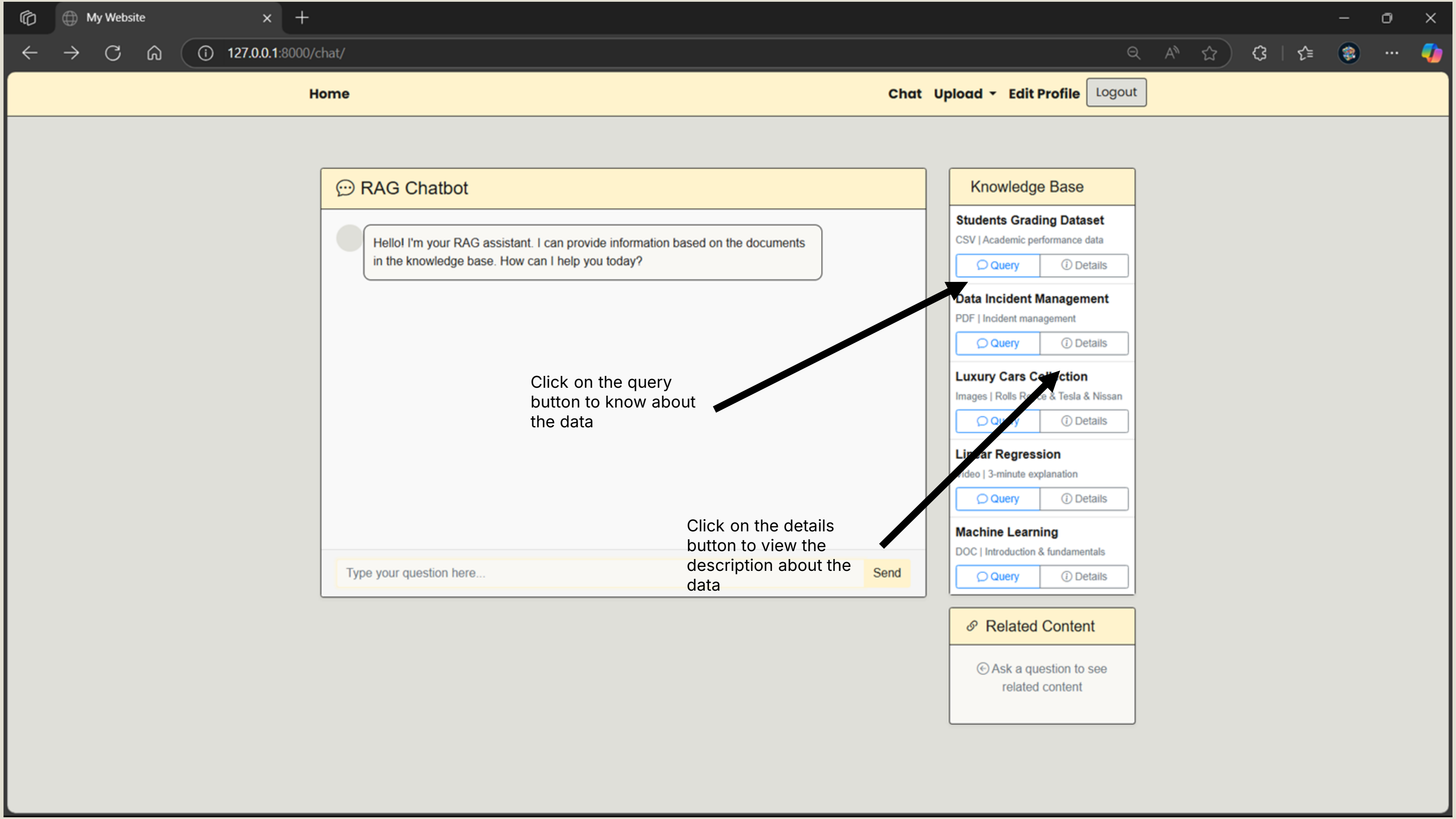
Humanitarian Data Incident Management [PDF File]

This resource from The Centre for Humanitarian Data outlines a structured approach to data incident management in humanitarian operations. It defines data incidents, explains a 5-step process (notification, classification, treatment, closure, knowledge base), and provides recommendations for improving data incident management practices in humanitarian organizations.

Luxury Cars Image Dataset [Image Files]

This image collection features high-quality photographs of Rolls Royce and Tesla vehicles. The dataset includes various models, angles, and settings, allowing for visual comparison between traditional luxury automobiles and modern electric vehicles. Use our chat interface to ask

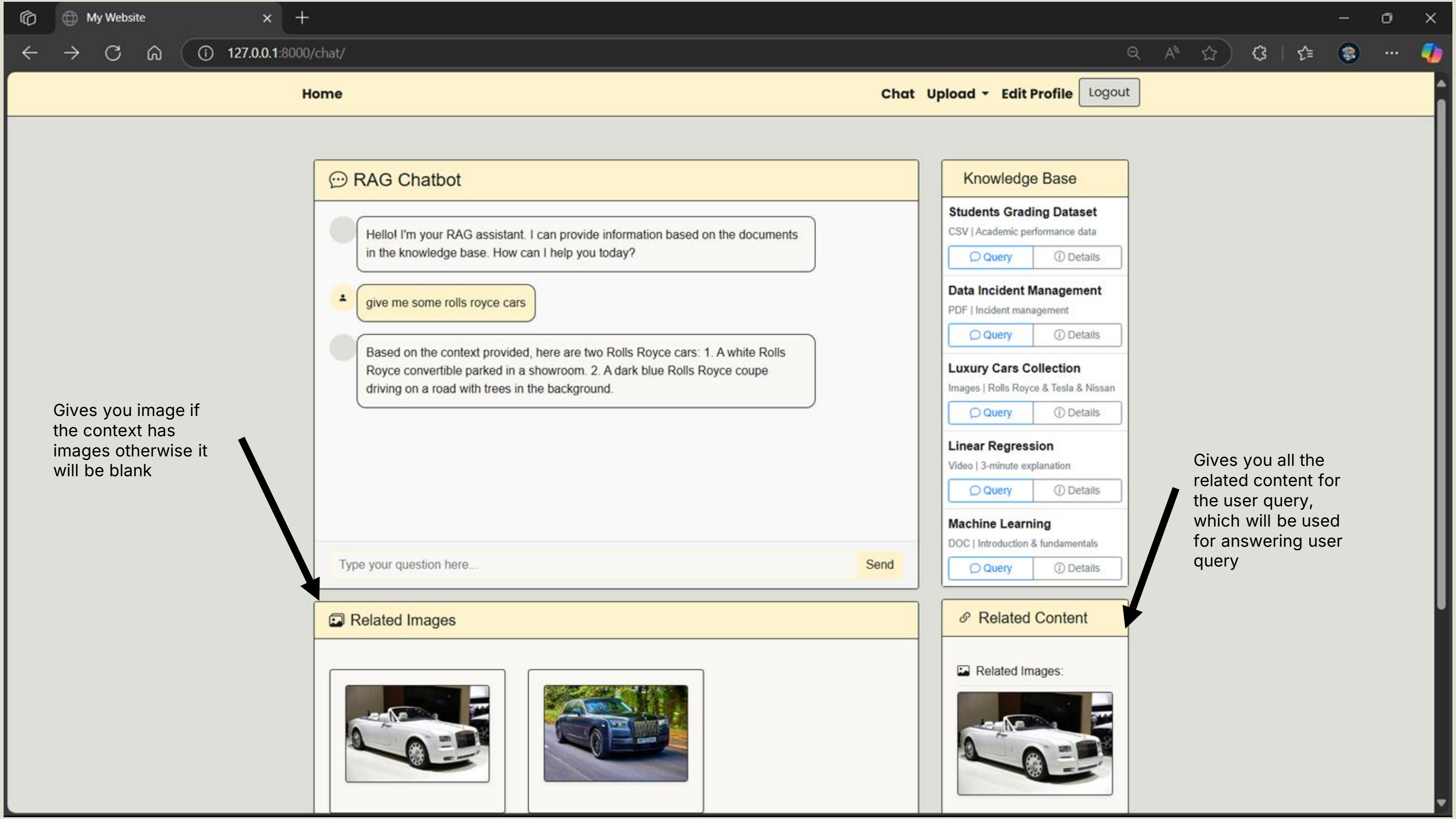
Different types of documents can be upload here, which will be processed and stored in pinecone (Vector DB)



Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

Click on the query button to know about the data

Click on the details button to view the description about the data



Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

give me some rolls royce cars

Based on the context provided, here are two Rolls Royce cars: 1. A white Rolls Royce convertible parked in a showroom. 2. A dark blue Rolls Royce coupe driving on a road with trees in the background.

Type your question here...

Students Grading Dataset
CSV | Academic performance data

Data Incident Management
PDF | Incident management

Luxury Cars Collection
Images | Rolls Royce & Tesla & Nissan

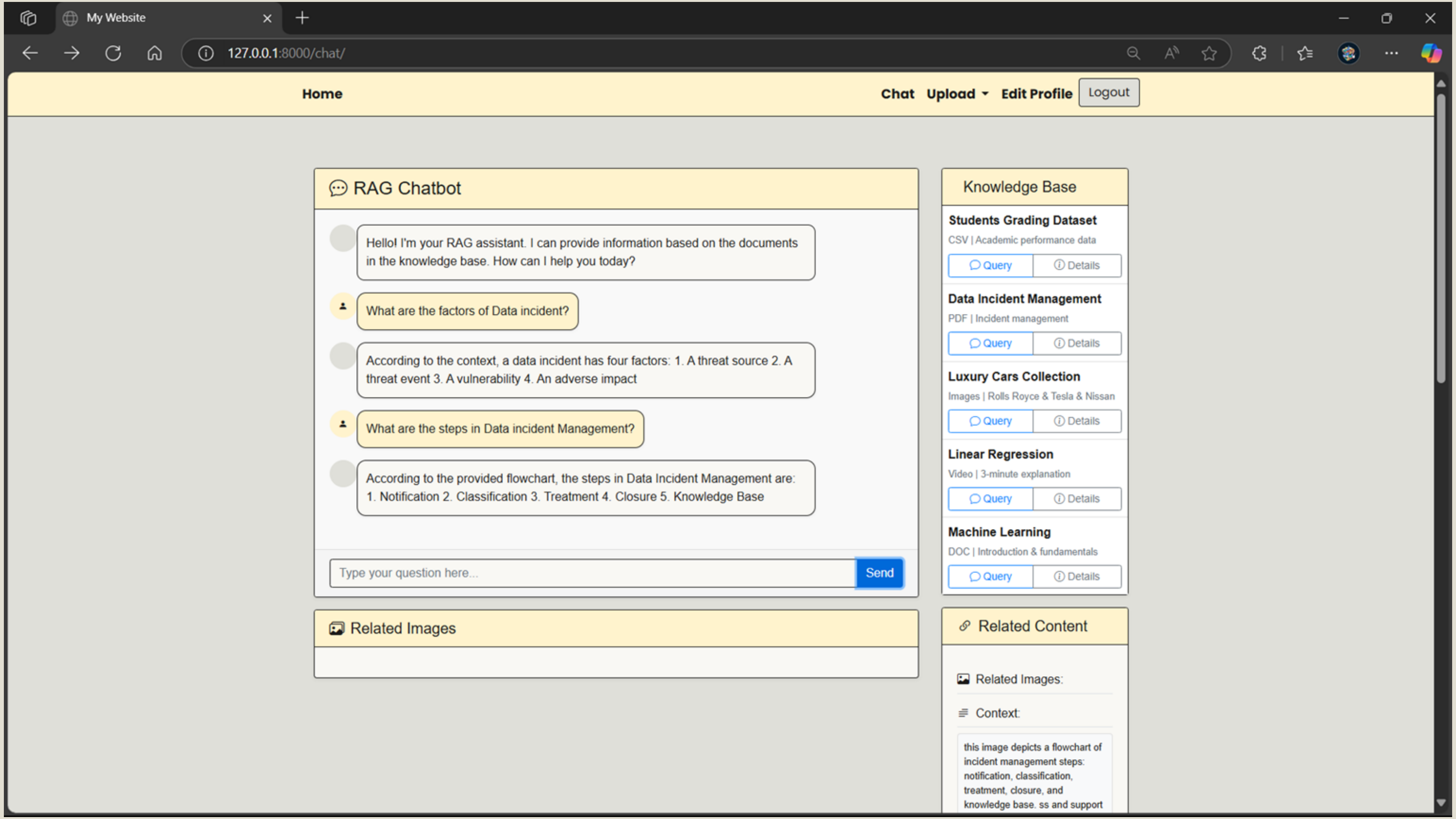
Linear Regression
Video | 3-minute explanation

Machine Learning
DOC | Introduction & fundamentals

Gives you image if the context has images otherwise it will be blank

Gives you all the related content for the user query, which will be used for answering user query

Query from Pdf Data



RAG Chatbot

Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

What are the factors of Data incident?

According to the context, a data incident has four factors: 1. A threat source 2. A threat event 3. A vulnerability 4. An adverse impact

What are the steps in Data incident Management?

According to the provided flowchart, the steps in Data Incident Management are: 1. Notification 2. Classification 3. Treatment 4. Closure 5. Knowledge Base

Type your question here...

Related Images

Knowledge Base

Students Grading Dataset
CSV | Academic performance data

Data Incident Management
PDF | Incident management

Luxury Cars Collection
Images | Rolls Royce & Tesla & Nissan

Linear Regression
Video | 3-minute explanation

Machine Learning
DOC | Introduction & fundamentals

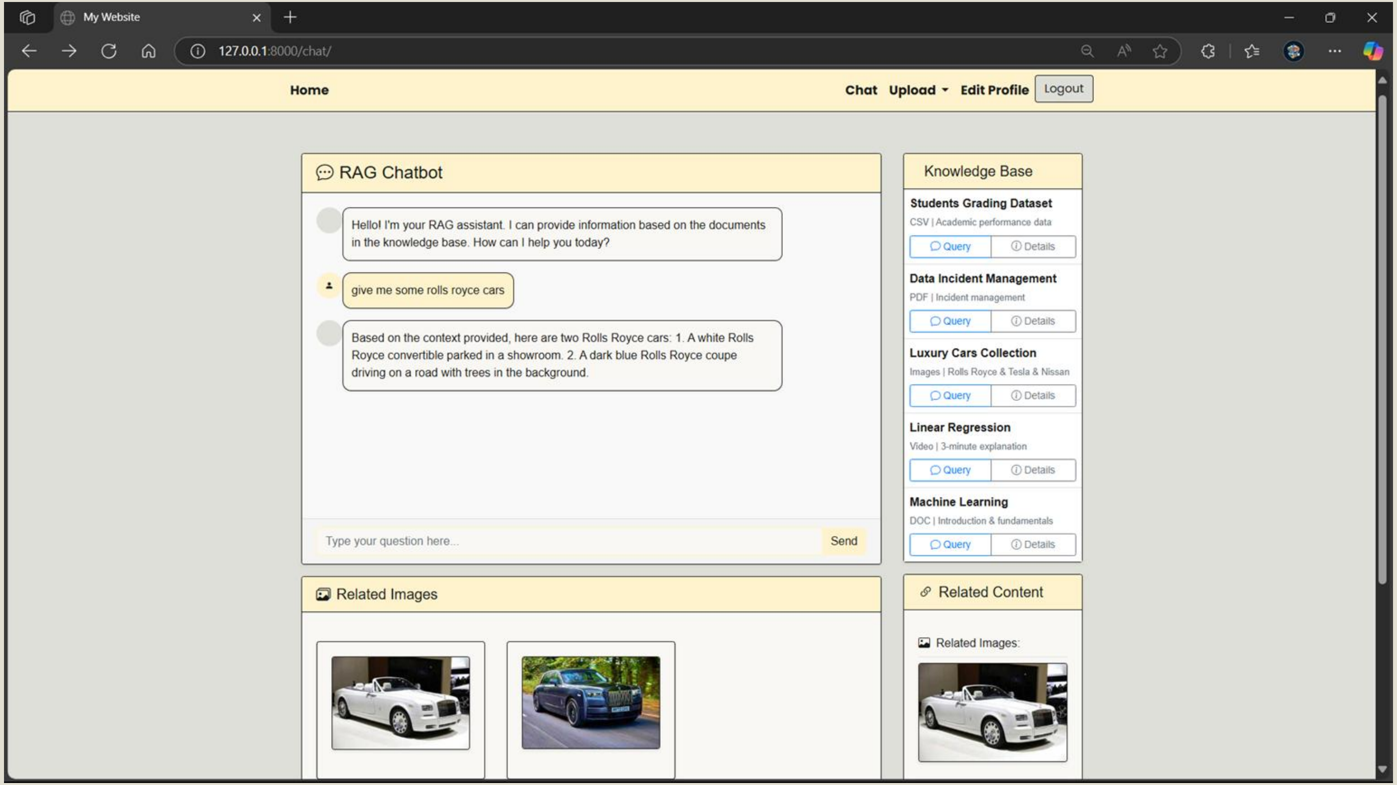
Related Content

Related Images:

Context:

this image depicts a flowchart of incident management steps: notification, classification, treatment, closure, and knowledge base. ss and support

Query from Image Data

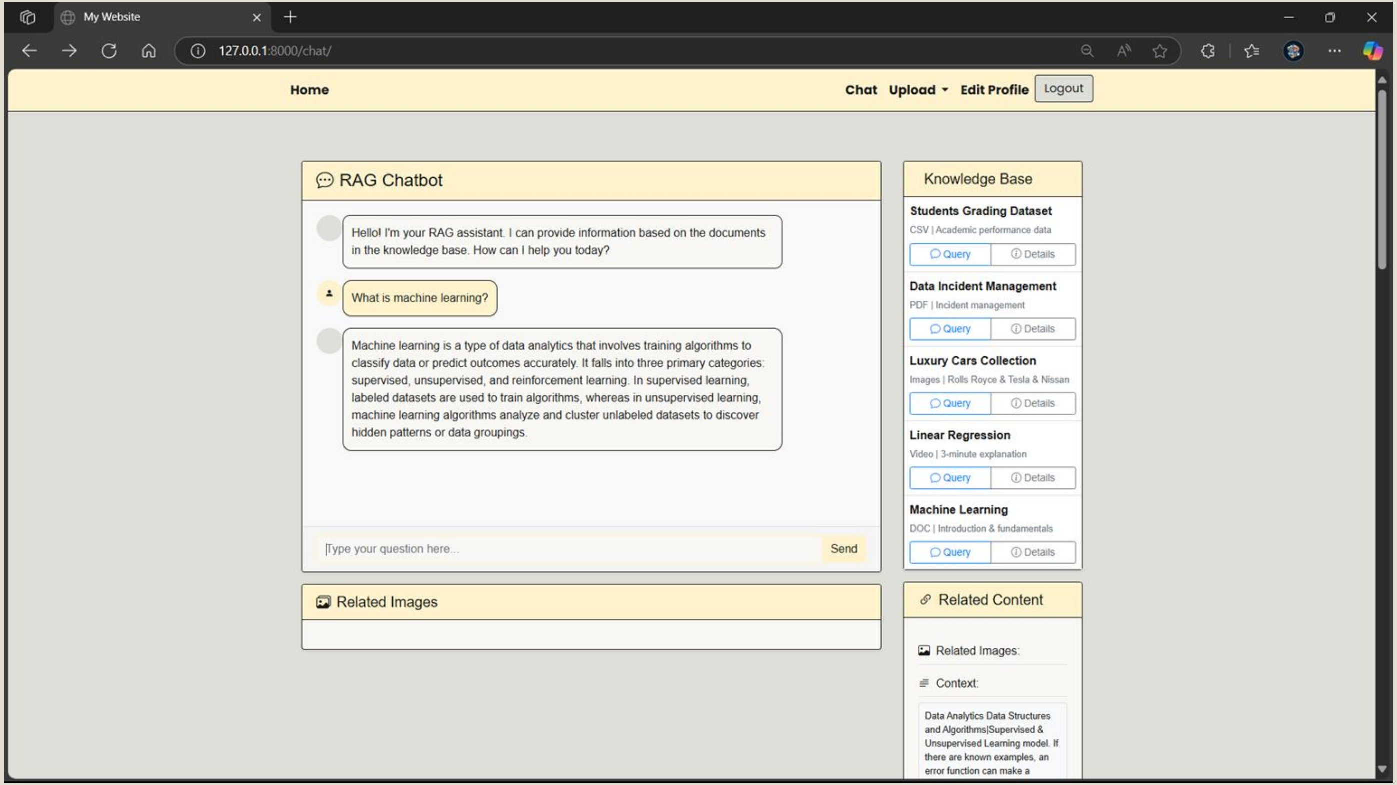


Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

give me some rolls royce cars

Based on the context provided, here are two Rolls Royce cars: 1. A white Rolls Royce convertible parked in a showroom. 2. A dark blue Rolls Royce coupe driving on a road with trees in the background.

Query from Docx Data



Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

What is machine learning?

Machine learning is a type of data analytics that involves training algorithms to classify data or predict outcomes accurately. It falls into three primary categories: supervised, unsupervised, and reinforcement learning. In supervised learning, labeled datasets are used to train algorithms, whereas in unsupervised learning, machine learning algorithms analyze and cluster unlabeled datasets to discover hidden patterns or data groupings.

Type your question here...

Knowledge Base

Students Grading Dataset
CSV | Academic performance data

Data Incident Management
PDF | Incident management

Luxury Cars Collection
Images | Rolls Royce & Tesla & Nissan

Linear Regression
Video | 3-minute explanation

Machine Learning
DOC | Introduction & fundamentals

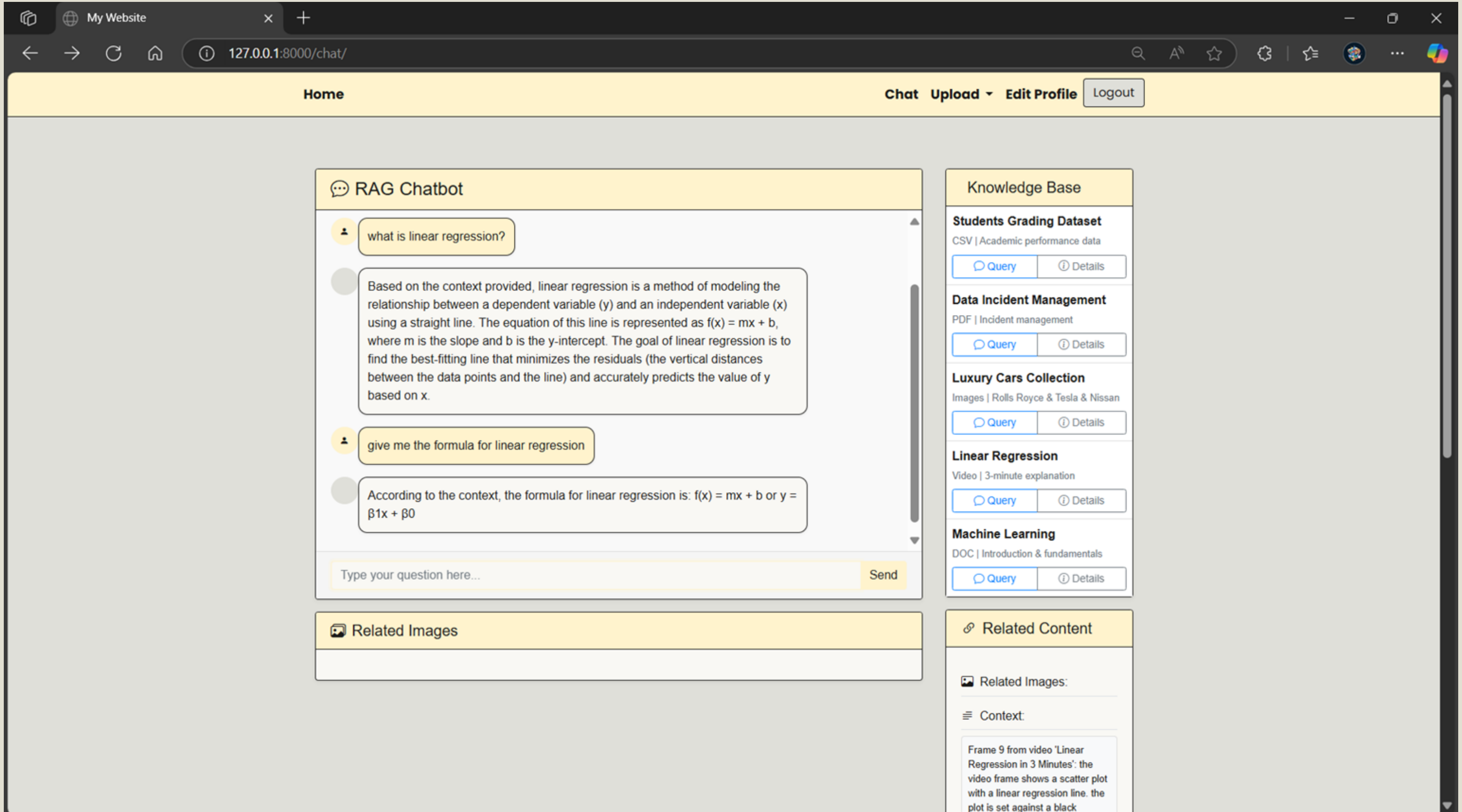
Related Content

Related Images:

Context:

Data Analytics Data Structures and Algorithms|Supervised & Unsupervised Learning model. If there are known examples, an error function can make a

Query from Video Data



what is linear regression?

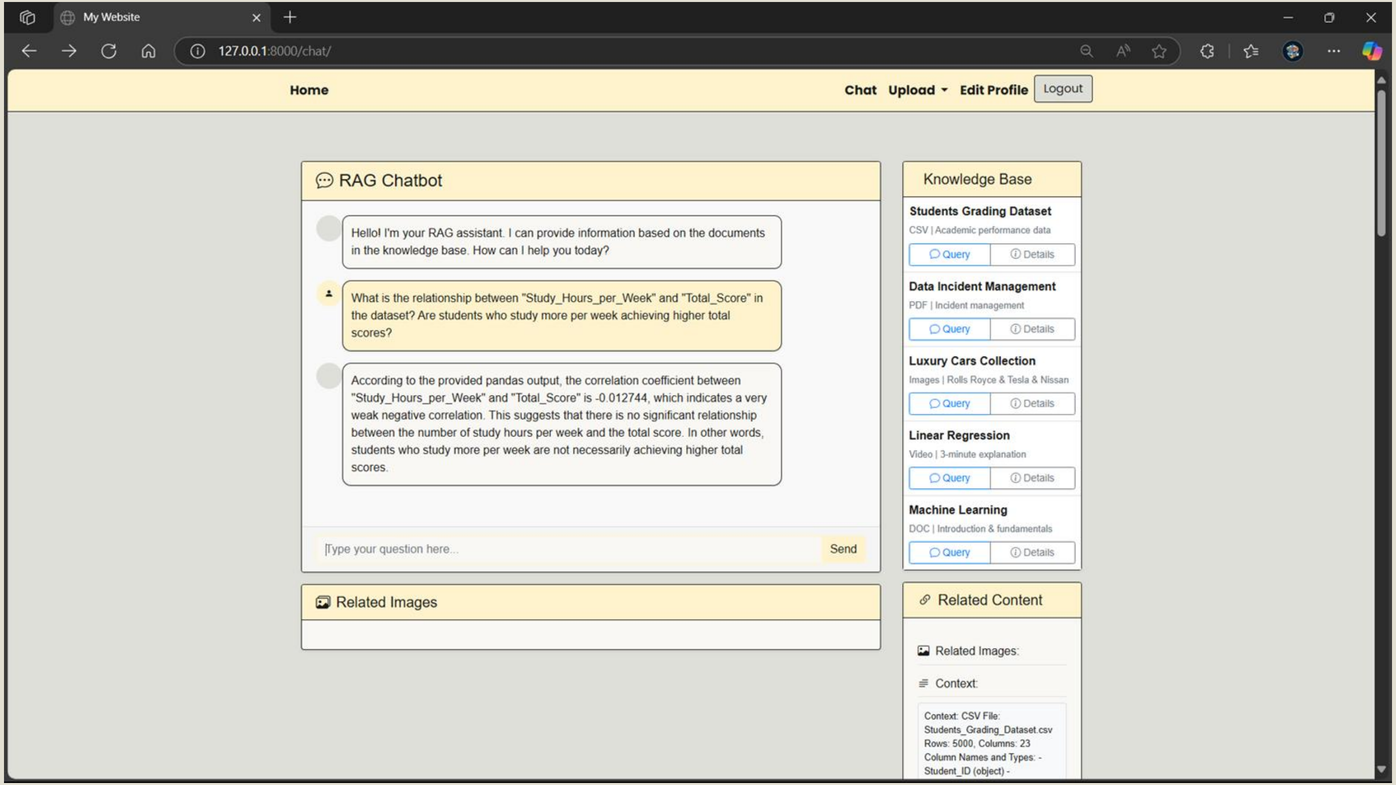
Based on the context provided, linear regression is a method of modeling the relationship between a dependent variable (y) and an independent variable (x) using a straight line. The equation of this line is represented as f(x) = mx + b, where m is the slope and b is the y-intercept. The goal of linear regression is to find the best-fitting line that minimizes the residuals (the vertical distances between the data points and the line) and accurately predicts the value of y based on x.

give me the formula for linear regression

According to the context, the formula for linear regression is: f(x) = mx + b or y = β1x + β0

Frame 9 from video 'Linear Regression in 3 Minutes': the video frame shows a scatter plot with a linear regression line. the plot is set against a black

Query from CSV Data



Hello! I'm your RAG assistant. I can provide information based on the documents in the knowledge base. How can I help you today?

What is the relationship between "Study_Hours_per_Week" and "Total_Score" in the dataset? Are students who study more per week achieving higher total scores?

According to the provided pandas output, the correlation coefficient between "Study_Hours_per_Week" and "Total_Score" is -0.012744, which indicates a very weak negative correlation. This suggests that there is no significant relationship between the number of study hours per week and the total score. In other words, students who study more per week are not necessarily achieving higher total scores.

Students Grading Dataset
CSV | Academic performance data

Data Incident Management
PDF | Incident management

Luxury Cars Collection
Images | Rolls Royce & Tesla & Nissan

Linear Regression
Video | 3-minute explanation

Machine Learning
DOC | Introduction & fundamentals

Context: CSV File: Students_Grading_Dataset.csv Rows: 5000, Columns: 23 Column Names and Types: - Student_ID (object) -

Thank you