SANDHIYA C V

Data Science | Machine Learning | Artificial Intelligence

☎ 7871019893 ✉ sandhiyagiri07@gmail.com in Sandhiya C V SandhiyaGiri

An enthusiastic data science student proficient in in Python, machine learning and experience in large language models (LLMs). Seeking opportunities in a startup to contribute to real-time problems in data-driven environments.

EXPERIENCE

Data Science Intern **Halliburton** **April 2024 -May 2024**

- Developed a RAG system for querying data from multiple PDF documents, employing **Neo4j's Graph database** for structured storage.
- Integrated vector search using **ember-v1**(Dense vector) & **BM25 encoder** (Sparse vector) and graph retrieval using **Cypher query** for efficient data retrieval.
- Established relationships between text chunks and entities, as well as between entities, enhancing data organization and analysis.
- Implemented a **reranker mechanism** to refine queries and extract top K relevant text chunks.
- Deployed the model using **Streamlit** as a local host web application for POC.
- **Tech stack :** Neo4j, Cypher, Reranker.

PROJECTS

QA Retrieval system

- Developed a Q&A retrieval system that utilizes **open source language models** for querying unstructured data from multiple PDF documents.
- Used sentence transformer **- all-mini-lm-l6-v2** as an embedding model and stored the embeddings in **FAISS** for efficient **similarity search** and **clustering**.
- Integrated **Mistral-7b-instruct-v0.1** for final response generation for the user query.
- **Tech stack :** RAG, LLM.

Sentiment Analysis

- Developed a sentiment analysis model utilizing **Naive Bayes classifier** to evaluate Movie review.
- Trained the model with IMDB 50k dataset and achieved an accuracy rate of **85.7%** through preprocessing techniques.
- Serialized the trained model and reused it across a different dataset scraped from web.
- **Tech stack :** NLP, Serialization, Data Preprocessing.

Succeeding word predictor

- Programmed a **trigram text generator** using the Wikipedia API to retrieve data, that produces up to 100 words based on a two-word seed phrase. The model utilizes the 're' module for data preprocessing, with a maximum sentence length of 15 words.
- **Tech Stack :** Wikipedia API, Regular expression, Tokenization.

SKILLS

- **Programming Languages** : Python , C/ C++.
- **Frame works** : Streamlit, FastAPI.
- **Libraries** : Pandas, Numpy,Langchain, Regex, OpenCV, NLTK, Scikit-learn,Pytorch.
- **Tools** : Visual Studio Code, Jupyter Notebook, GitHub.
- **Data Techniques** : Data Preprocessing , Dimensionality reduction.
- **Supervised Learning, Unsupervised Learning** : Linear Regression, Logistic Regression,Naive Bayes, Neural Networks , ID3 Algorithm, PCA-SVD.

EDUCATION

Bachelor of Engineering in Electronics and Communication **CGPA : 8.55***
Government College of Technology, Coimbatore. 2025