**ITC Financial Analysis with AI Scraping & LLM Integration**

**🎯 Objective**

To analyze ITC Ltd.'s revenue trends, profitability, and financial performance by:

* Automatically searching and extracting official financial documents
* Embedding content into a searchable format
* Using a Language Model (LLM) to answer financial questions interactively

**1. Requirements**

**🔹 Data Extraction**

* Use Tavily AI to search and retrieve ITC’s official:
  + Annual Reports for 2023 and 2024
  + Investor presentations
  + Stock price data
* Limit results to sources from ITC’s official domain.
* Extract relevant PDF URLs and download them for processing.

**🔹 Data Processing**

* Load and split PDF documents into readable chunks.
* Convert document chunks into vector embeddings using Gemini-based embedding model.
* Store the embeddings in a vector database (FAISS) for retrieval.

**🔹 LLM Integration**

* Use Google Gemini via LangChain to create a retrieval-based QA system.
* Send user queries along with relevant document context to Gemini.
* Use retrieved content to generate factual, grounded answers.

**🔹 Output Interface**

* Build a Streamlit app (optional) to:
  + Accept natural language financial questions
  + Display LLM-generated answers
  + Include citations pointing to specific pages/sections of the report

**2. Deliverables**

**📁 GitHub Repository**

The project repository includes:

* Scripts for web search, downloading, and parsing documents
* Logic for embedding and storing document data
* LLM integration for querying answers
* A requirements.txt file listing all dependencies
* A detailed README.md with setup instructions

**🖥️ Streamlit App**

* **Provides a user-friendly interface to:**
  + **Type financial queries**
  + **View answers generated by the LLM**
  + **See source references pulled from ITC documents**

**3. Sample Code Structure**

* itc-financial-analysis/
* ├── tavily\_search/
* ├── pdf\_loader/
* ├── vectorstore/
* ├── llm\_chain/
* ├── app.py
* ├── requirements.txt
* └── README.md