NLP and LLMs: Building a Question-Answering Pipeline for Financial Documents

Project Overview

This project demonstrates how Natural Language Processing (NLP) and Large Language Models (LLMs) can be applied to financial research. The goal was to build a Question-Answering (QA) pipeline capable of processing large amounts of financial text data and extracting key insights to support analysts in decision-making. By testing pretrained models, evaluating performance, and scaling to more complex datasets, this project highlights both the opportunities and limitations of deploying LLMs in real-world financial analysis.

Objectives

- Build an initial QA pipeline using a single financial text document (financial_context).

- Experiment with multiple pretrained Hugging Face models and select the most suitable one.

- Evaluate performance against financial Q&A; tasks.

- Provide recommendations for scenarios where this pipeline should or should not be used.

- Bonus: Extend the pipeline to a larger set of documents from the FinQA dataset.

Data Sources

- FinQA Dataset (subset provided in the repo)
- Original FinQA Data Website: https://finqasite.github.io/index.html
- Hugging Face QA Pipeline Documentation: https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/pipelines#transformers.QuestionAnsweringPipeline

Tools and Libraries

- Python
- Hugging Face Transformers
- FinQA Dataset
- Jupyter Notebook

Deliverables

- QA Pipeline Implementation (Jupyter Notebook)
- Performance Evaluation and Recommendations
- Scalability Test with FinQA Dataset

Prerequisites

Before replicating this project, you should be able to:
- Explore NLP models in Hugging Face.
- Build and run a basic NLP pipeline.

Notes

- Results may vary depending on the model and parameters used. Always evaluate and document your outcomes.
- Please ignore the _layouts folder and config.yml file in the repo. These are required for rendering and should not be modified.