Data Science Canvas

Project: Research Paper Recommendation Service

Team: ScholarStream

Problem Statement

Business Case & Value Added
There is explosion of the research articles, and it is difficult to keep up with the newly introduced concepts and trending topics.

Model Selection
Embedding models are known to model a vector space that encodes real world concepts and therefore help with semantic similarity.

We will use,
TF-IDF as baseline 1,
all-minilm-l6-v2 as baseline 2

Model Requirements
- Semantic understanding.
- Domain relevance
- Scalability
- Low latency
- Input flexibility
- Low cost

Skills
What skills are needed to provide the data and model development?

- EDA
- Developing application using python framework like FASTAPI
- Vector DB
- NLP
- Machine Learning
- Containerization
- Cloud Deployment

Data Landscape

Yes data is already available from arxiv
Additional data can be generated by extracting pdf content for certain categories. But, for the purpose of this project, we will limit ourself to just the metadata (title, abstract, authors etc.)

Software & Libraries
- Numpy
- Pandas
- Sci-kit learn
- Vector db like milvus db
- LLMs (ex gemini-flash-2.5)
- Langchain
- HuggingFace
- FastAPI

Execution & Evaluation

Model Evaluation
- Precision@K
- Recall@K
- MRR
- nDCG@K

Real time monitoring:
- Latency
- Throughput
- Error rate
- Resource usage

Data Storytelling
A simple ui with detailed research paper information, recommendations and similarity scores.

Data Collection & Preparation

Data Selection & Cleansing
The data is readily available in arxiv. We filter out papers that have non ascii characters in the title or abstract.

Data Collection
No additional data is required.

Data Integration
Data integration is not required as it is from a single source

Explorative Data Analysis
We only use papers from ‘cs.ai’ category. The data is clean and doesn’t require EDA.

Adopted from: https://github.com/tomalytics/datasciencecanvas