Intelligent Research Paper Summarization and Citation Network System

Nivedita Nair
nivedita.nair@sjsu.edu
018184777

Shanmukha Manoj Kakani
shanmukhamanoj.kakani@sjsu.edu
018195645

Kalyani Chitre
kalyani.chitre@sjsu.edu
017622917

Selected Project Option: 1 - Automatic Data Mining for Large PDF Files

Project Overview: Keeping up with research is harder than ever. With thousands of new papers published daily across different fields, it's nearly impossible to manually sift through them all and find the most relevant work. Our project aims to solve this problem by building an intelligent system that can automatically summarize research papers, extract key citations, and organize them into an easy-to-navigate structure.

Key Objectives:

1. **Summarization of Research Papers:** Automatically extract and condense key points from research papers using a hybrid approach combining extractive and abstractive summarization techniques.
2. **Citation and Reference Analysis:** Extract references and citations from papers to construct an interactive citation network that visualizes research relationships.
3. **Smart Semantic Search & Question Answering (QA):** Implement a semantic search engine and QA system to enable users to query across multiple papers using natural language, surpassing basic keyword searches.
4. **Visualization of Research Connections:** Generate topic maps and interactive citation graphs to reveal research trends and connections.

Additional Exploration (Time Permitting):

Comparative Evaluation - Assess the performance and usability of the system as a general-purpose paper analysis tool versus a domain-specific solution.

Citation Recommendations: Develop an AI-powered system to suggest relevant citations for researchers drafting new papers.

Datasets We'll Use:

- **Semantic Scholar Open Research Corpus** (SSORC)
- **ArXiv Papers** (focused on AI, Computer Science, etc.)
- **PubMed Articles** (for Biomedical and Life Sciences research)

Tools & Technologies:

- **PDF Processing**: PyMuPDF (to extract text from research papers)
- **NLP & Summarization**: HuggingFace Transformers, LlamaIndex, LangChain
- **Graph Analysis & Visualization**: NetworkX, Plotly Dash
- **Semantic Search & AI models**: Vector embeddings, retrieval-augmented generation (RAG)

Expected Impact: This system will significantly reduce the time required for researchers to locate and understand relevant papers. By automating summarization, citation analysis, and enabling intuitive research exploration, we hope to make academic research more accessible, efficient, and insightful.

Related References:

- Semantic Scholar Open Research Corpus
- HuggingFace Transformers
- LangChain
- PyMuPDF
- NetworkX
- [2104.03057] Enhancing Scientific Papers Summarization with Citation Graph
- opendatalab/PDF-Extract-Kit: A Comprehensive Toolkit for High-Quality PDF Content Extraction