Mutma Adebayo
Akul Datta
Nisha Prasad

PaperWeb Project Proposal

Functions and Users

We plan to implement a standalone software tool/web-based application called PaperWeb. PaperWeb will be a visual representation tool of the research papers on the ArXiv website and will be hosted on GitHub Pages. The minimum viable features will include looking up a paper and being able to find:

a) The paper abstract and other data
b) All of its cited sources
c) All the papers that cite it (ranked/bigger node if that paper has more citations)

In the graph. The intended users of our tool will be researchers and professionals looking to catch up on new developments in their various fields, and potentially also students looking for papers relevant to their projects and assignments.

Significance

We need this tool because tools like it that already exist are often outdated and do not update their databases frequently enough to keep up with the rate at which papers are published. We intend to use API calls on ArXiv to regularly update our data and ensure the tool is always able to provide up-to-date information. This tool can help to change the world by enabling the researchers who do important work to find all the information and relevant papers needed to make further contributions to society in their work. It is important to address this need because expediting the research process will help the efficiency of research that could drive society forwards.

Approach

We first plan to create our project on a small sample of papers, which will be approximately 10,000 in size. We will create two SQL databases, one that stores metadata about ArXiv papers and their connections to papers that they cite, and another that will contain embedding information that will eventually help us determine similar papers.

Mutma Adebayo
Akul Datta
Nisha Prasad

We plan to use a cloud cluster to run our scripts to get the paper metadata, and we will store the data on GitHub with a SQLite file. For the embeddings, we plan to use an already existing embedding model and run it locally using Python. For our web application, we will use Streamlit. We currently plan to write all the code in Python. For displaying graphs, we will use the Networkx library. We will host the application on GitHub Pages.

We currently anticipate no risks that we'll need to mitigate.

Evaluation

We will compare the tool's features to already existing tools such as ArXivAtlas and Connected Papers. We will also select a sample of a few people in the research community at UIUC and gauge their interest in this application. We will ask them to use it and then fill out a survey to evaluate whether our tool works as intended.

Timeline and Task division

[Mar 23, 2025 Week 9]

- Create code repository **(Mutma)**
- Write scripts to get ArXiv papers via API for test dataset (e.g. 10,000 samples) **(Mutma + Nisha)**
- Initialize SQL database schema with test data **(Directed: Mutma + Undirected: Akul)**

[Mar 30, 2025 Week 10]

- Brainstorm frontend application layout **(All)**
- Create a script to get 2nd and 3rd degree connections (no ui yet) **(Mutma + Nisha)**
- Create a similar paper embedding database **(Akul)**

[Apr 6, 2025 Week 11]

- Implement working frontend prototype **(Mutma + Nisha)**
- Connect backend and frontend **(All)**
- Host project on Github Pages **(Akul)**

[Apr 13, 2025 Week 12-13]

- Add continuous updating of the graph **(Akul + Nisha)**

Mutma Adebayo
Akul Datta
Nisha Prasad

- Increase dataset size to all papers in the last few years **(Akul)**
- Implement extra features
    - Query by title, topic, date, and embeddings **(Akul)**
    - “Hot Papers” identification per major topic (citation+year based) **(Mutma)**
    - “Core paper” identification (also citation based, these are popular fundamental papers that current hot papers are based on) **(Nisha)**
    - If we have extra time: Extract scientific claims from papers and verify claims based on graph/embeddings **(All)**

[Apr 27, 2025 Week 14]

- Finish Final Writeup (4 pages) **(All)**

[May 4, 2025 Week 15]

- Create slides for project presentation **(All)**
- Practice presenting **(All)**
- Do the project presentation **(All)**