Organizing Research Papers Using Text Embeddings and Clustering

Henry Roeth

CPSC 40100.00 – Integrated Research Component II

Spring 2025

Project Overview

This project is about creating a tool that helps organize research papers into groups based on their topics. The idea is to analyze the abstracts of research papers, represent their content using text embeddings, and use clustering to automatically sort the papers into meaningful categories. The end result will organize these papers into folders, making it easier to find related work.

The tool will be built using Python and libraries like PyMuPDF (for extracting text), sentence-transformers (for generating embeddings), and scikit-learn (for clustering). The process will be fully automated and should work on any folder of PDFs.

Interdisciplinary Approach

This project brings together skills and ideas from different areas:

- **NLP (Natural Language Processing):** To extract text from research papers and turn it into meaningful data using embeddings.

- **Machine Learning:** To use clustering methods (like K-means) for grouping similar papers.

- **Programming and Tools:** To write code that integrates all the parts and creates a working tool.

I'll combine these areas to solve a real problem—managing and organizing research papers.

Why This Project?

I chose this project because I find it hard to keep my own research papers organized, especially as my collection grows. This is a chance to solve a problem I've personally experienced while also learning more about natural language processing and machine learning. I'm excited about the idea of creating something practical and useful.

Plan and Goals

The project will have three stages to make sure it stays manageable:

Stage 1: Minimal Goals

- Extract the abstract text from research papers using PyMuPDF.
- Convert the abstracts into embeddings using sentence-transformers.
- Use K-means clustering to group the papers.
- Automatically move the papers into folders based on clusters.

Stage 2: Expected Completion

- Add a method to figure out the best number of clusters, like the elbow method.
- Allow users to pick a folder to organize and set the number of clusters through a command-line interface.
- Test the tool with real research papers and write about the results in the final report.

Stage 3: Stretch Goals (If I Have Time)

- Build a simple graphical interface to make the tool easier to use.
- Add visualizations to show how the papers are grouped (e.g., using PCA or t-SNE).
- Expand the tool to work with other file types like Word documents.

Updated Timeline

- **Weeks 3-4:** Extract text, create embeddings, and implement clustering. Organize papers into folders.
- **Weeks 5-6:** Add features like finding the best number of clusters. Test the tool and document results. Start writing the final paper.
- **Week 7:** Polish the tool and finalize clustering. Write about limitations and improvements in the paper.
- **Week 8:** Finalize the research paper and prepare the presentation slides.
- **Week 9:** Deliver the presentation.

Conclusion

This project will not only help me organize research papers but also teach me valuable skills in NLP, machine learning, and software development. By breaking the work into stages, I can focus on creating a useful and functional tool while still leaving room for additional features if I have extra time. I'm excited to explore these technologies and see how this project can make research easier.