Project Proposal: Research Paper Clustering System for Topic Discovery and Literature Organization

By Team Armah, AI Saturdays Lagos Cohort 9 - Flipped

Problem Statement

The rapid growth of academic publications across disciplines has made it increasingly difficult for researchers, students, and policymakers to stay updated with emerging knowledge. Thousands of new research papers are published daily across multiple platforms such as arXiv, IEEE Xplore, PubMed, and Springer. This overwhelming volume leads to information overload, redundant research, and difficulty in identifying related works across domains.

Traditional search engines rely heavily on keyword matching and citation networks, which often fail to capture the semantic relationships between papers. Two papers may address similar problems using different terminology and thus remain disconnected in searches. This makes the literature review process time-consuming and inefficient.

To address this challenge, this project proposes a machine learning–based system that automatically clusters research papers based on their topics and semantic similarity. By using text representation methods such as TF-IDF and dimensionality reduction with clustering algorithms like K-Means or Hierarchical Clustering, the system aims to group related papers together—making it easier for researchers to explore and analyze scientific domains.

Existing Solutions

- **Semantic Scholar** – AI-powered academic search engine that provides topic-based recommendations.
- **Google Scholar** – Keyword-based retrieval system with limited semantic clustering.
- **Connected Papers** – Graph-based visualization of related works.

Gap: Most existing platforms are closed-source, domain-specific, or depend on deep learning architectures that require large computational resources. There is a lack of lightweight, transparent, and

customizable traditional machine learning tools that students and early-stage researchers can deploy locally to organize and explore literature in their field of interest..

Objectives

1. To develop a lightweight and interpretable research paper clustering system using traditional ML techniques.
2. To apply text preprocessing and vectorization methods (e.g., TF-IDF, CountVectorizer) for document representation.
3. To implement and compare clustering algorithms (K-Means, Agglomerative Clustering, DBSCAN) for topic grouping.
4. To visualize dataset imbalance through clear, interpretable charts
5. To visualize the clusters using dimensionality reduction techniques (PCA, t-SNE) for intuitive interpretation.
6. To provide interactive outputs that allow users to explore and analyze grouped research topics easily.

Proposed Dataset

1. **arXiv Academic Papers Dataset** (available on Kaggle): Includes metadata such as title, abstract, authors, and subject categories
2. **Custom Dataset:** Scrap or access them from publicly available API (arXiv, IEEE Xplore, PubMed)

Proposed Methodology

1. **Data Input** – Users upload folder containing research papers or research paper datasets containing titles, abstracts, keywords, and metadata (CSV or JSON format) Users upload CSV/tabular datasets.
2. **Preprocessing** – Perform text cleaning (remove stop words, punctuation, and special symbols), tokenization, and lemmatization. Extract key text fields such as title and abstract for representation.

3. **Feature Extraction** – Calculate class imbalance, diversity ratios, missingness, and governance-related metricsConvert text into numerical form using traditional Natural Language Processing techniques:
    - TF-IDF Vectorization – to capture word importance relative to the corpus.
    - CountVectorizer – for frequency-based representation.
    - Optional: Apply Latent Semantic Analysis (LSA) to capture hidden topic dimensions.
4. **Clustering Analysis** – Generate histograms, pie charts, and bar charts for interpretabilityApply unsupervised learning algorithms to group semantically similar papers:
    - K-Means – baseline clustering approach for general topic grouping.
    - Hierarchical Clustering – to visualize topic relationships in a dendrogram.
    - DBSCAN – for density-based clustering to detect niche topics or outliers.
5. **Modeling** – Baseline ML models (Naive Bayes, Logistic Regression) and neural network classifier, evaluate performance, and perform hyperparameter tuning to optimize results.
6. **Visualization** – Use **PCA** or **t-SNE** for dimensionality reduction and plot research paper clusters in 2D/3D. Visualize topic distributions, most frequent keywords, and inter-cluster similarities using bar and word cloud charts.
7. **Evaluation** – Assess clustering performance with internal metrics such as Silhouette Score, Davies-Bouldin Index, and manual topic coherence inspection.
8. **Reporting** –Generate summaries per cluster, highlighting dominant themes, top keywords, and representative papers. Export reports as CSV, JSON, or PDF for research use.
9. **Deployment** – Build an interactive Streamlit web app that allows users to upload datasets, view topic clusters, explore related works, and download cluster reports.

Model Development Plan

We plan to test multiple clustering algorithms and representations:

- **K-Means:** Efficient for large datasets; provides clear cluster centroids for interpretation.
- **Agglomerative (Hierarchical) Clustering:** Useful for visualizing relationships between subtopics.
- **DBSCAN:** Detects isolated or niche topics that don't fit mainstream clusters.

- **Latent Semantic Analysis (LSA):** Captures underlying semantic structures to improve topic quality.

Deployment Plan

- **Prototype:** Streamlit-based or Jupyter Notebook interface for easy local experimentation.
- **Public Deployment:** Streamlit Cloud for public accessibility.

Expected Outcome

1. **Efficient Literature Organization**
    - Researchers can quickly identify topic clusters and discover related works, saving substantial time in literature review.
2. **Improved Knowledge Discovery**
    - The system highlights cross-domain connections between papers that traditional keyword search might miss.
3. **Transparency and Interpretability**
    - Uses traditional, explainable ML methods (TF-IDF, K-Means, LSA) to ensure clarity and ease of understanding for non-technical users.
4. **Scalability and Accessibility**
    - Lightweight and deployable locally or on the cloud (Streamlit Cloud), requiring minimal computational resources.
5. **Reusable Research Tool**
    - Designed as an open, customizable framework that students and researchers can adapt to their own datasets and domains.
6. **Data-Driven Insights**
    - Generates interpretable visual and quantitative analytics for publication trends, cluster composition, and thematic distributions.

Community Impact

The development of this research document clustering AI system holds significant potential for the academic and research community. By automating the organization, categorization, and retrieval of scholarly articles, the system enhances accessibility to knowledge and supports faster, more accurate literature discovery within the desired category. This contributes to more efficient research processes, fosters collaboration among scholars, and empowers students and professionals to stay informed about current developments in their specific topic. Beyond academia, the project promotes knowledge equity by making research outputs more discoverable to communities and institutions with limited access to premium databases, ultimately advancing innovation and societal progress through informed research.

Team Members

1. Latifat Olabisi Idris
2. Abdusshakur Olarewaju Olabisi
3. Nunsi Shiaki

Acknowledgement

This project was inspired by a strong desire to apply knowledge gained from AI Saturday trainings to a practical and impactful initiative. It was driven by a passion to support the academic community and researchers through innovative tools that simplify access to knowledge and enhance research efficiency. The work also reflects a deep motivation for continuous learning and growth, serving as both a contribution to the field and a personal journey of exploration in applying Artificial Intelligence to real-world academic challenges.

References

1. Connected Papers, Connected Papers | Find and explore academic papers
2. Dataset: arXiv Academic Papers Dataset
3. Google Scholars, Google Scholar
4. Semantic Scholar, Semantic Scholar | AI-Powered Research Tool