A Mini Project with Seminar On

**Unsupervised Clustering of Scientific Literature**

Submitted in partial fulfillment of the requirements for the award of the

**Bachelor of Technology**

in

**Department of Computer Science and Engineering** **(Artificial Intelligence and Machine Learning)**

By

1. Mr. Nikhil Garimella(21241A6623)

2. Mr. Aegyarapu Karthikeya (21241A6602)

3. Mr. Akhilesh Varma (21241A6603)

**Under the Guidance of :**

Mr. K Mallikarjun Raju

Assistant Professor, AIML Department



**DepartmentofComputerScienceandEngineering**

(**ArtificialIntelligenceandMachineLearning)**

**GOKARAJURANGARAJUINSTITUTEOFENGINEERINGAND TECHNOLOGY**

**(Approved by AICTE, Autonomous under JNTUH, Hyderabad) Bachupally, Kukatpally, Hyderabad-500090**



**GOKARAJURANGARAJUINSTITUTEOFENGINEERINGAND TECHNOLOGY**

**(Autonomous) Hyderabad-500090**

### CERTIFICATE

This is to certify that the mini-project entitled “Unsupervised Clustering of Scientific Literature” is submitted by Mr. Nikhil Garimella (21241A6623),Mr.Aegyarapu Karthikeya (21241A6602), and Mr. Akhilesh Varma (21241A6603) under the guidance of Mr. K Mallikarjun Raju, Assistant Professor, AIML Department, in partial fulfillment of the award of the degree in BACHELOR OF TECHNOLOGY in Computer Science and Engineering (Artificial Intelligence and Machine Learning) during the academic year 2023-2024.

**Internal Guide Head of the Department**

#### Mr. K Mallikarjun Raju Dr.G. Karuna

**ExternalExaminer**

### 

### ACKNOWLEDGEMENT

There are many people who helped us directly and indirectly to complete our project successfully. We would like to take this opportunity to thank one and all. First, we would like to express our deep and profound gratitude towards our guide Mr. K Mallikarjun Raju, Assistant Professor, AIML Department, for his support in the completion of our dissertation. We wish to express our sincere thanks to Dr.G. Karuna, Head of the Department, and to our Principal Dr.J.Praveen, for providing the facilities to complete the dissertation. We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

Mr. Nikhil Garimella (21241A6623)

Mr. Aegyarapu Karthikeya (21241A6602)

Mr. Akhilesh Varma (21241A6603)

## DECLARATION

We hereby declare that the mini project titled “Unsupervised Clustering of Scientific Literature” is the work done during the period from 13 March 2024 to 3 July 2024 and is submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning) from Gokaraju Rangaraju Institute of Engineering and Technology (Autonomous under Jawaharlal Nehru Technology University, Hyderabad). The results embodied in this project have not been submitted to any other university or institution for the award of any degree or diploma.

**Nikhil Garimella (21241A6623)**

**Aegyarapu Karthikeya (21241A6602)**

**Akhilesh Varma (21241A6603)**

## ABSTRACT

The proposed approach involves leveraging unsupervised machine learning techniques, including Latent Dirichlet Allocation (LDA) for topic modeling, to cluster scientific papers based on their abstracts. The primary objective is to discover underlying themes and structures within the extensive body of scientific literature, facilitating enhanced organization and comprehension of research topics. To achieve this, the research will employ traditional clustering algorithms such as K-Means, DBSCAN, and hierarchical clustering alongside LDA-based topic modeling. By integrating LDA into the clustering process, the researchers aim to uncover nuanced relationships between papers and identify cohesive clusters of documents sharing similar thematic content.

Data preprocessing involves collecting and preprocessing the abstracts of scientific papers, including tasks such as tokenization, stopword removal, and stemming/lemmatization. Traditional clustering algorithms like K-Means, DBSCAN, and hierarchical clustering will be employed to group the scientific papers. Latent Dirichlet Allocation (LDA) will be used to extract latent topics from the preprocessed abstracts. By integrating LDA-based topic modeling into the clustering process, the researchers aim to uncover nuanced relationships between papers and identify cohesive clusters of documents sharing similar thematic content.

The proposed approach promises to provide researchers with a comprehensive tool for navigating and exploring related work, fostering interdisciplinary collaboration and advancing knowledge discovery in scientific domains. The expected outcomes include improved organization and comprehension of the extensive body of scientific literature, identification of latent themes and structures within the scientific literature, enhanced interdisciplinary collaboration, and advancement of knowledge discovery in various scientific domains.

|  |  |  |
| --- | --- | --- |
|  | **Table of Contents** |  |
| **Chapter** | **TITLE** | **Page No** |
|  | **Abstract** |  |
| **I** | **Introduction** |  |
|  | 1.1Introduction to the Project Work (Rationale / Motivation) |  |
|  | 1.2 Objective of the Project |  |
|  | 1.3 Methodology adopted to satisfy the objective |  |
|  | 1.4 Architecture diagram with very brief description |  |
|  | 1.5 Organization of the Report |  |
| **II** | **Literature Survey** |  |
| **III** | **Proposed Method** |  |
|  | 3.1 Problem Statement & Objective of the Project |  |
|  | 3.2 Explanation of:   * Architecture Diagram * Modules – Connectivity Diagram * Software and Hardware Requirements |  |
|  | 3.3 Modules and its Description |  |
|  | 3.4 Requirements Engineering   * Functional and Non-Functional |  |
|  | 3.5 Analysis and Design through UML   * Class Diagram * Sequence Diagram * Use case Diagram * Activity Diagram |  |
| **IV** | **Result and Discussions** |  |
|  | 4.1 Description about Dataset |  |
|  | 4.2 Detailed explanation about the Experimental Results |  |
|  | 4.3 Significance of the Proposed Method |  |
|  |  |  |
| **V** | **Conclusion and Future Enhancements** |  |
|  | 5.1 Source code  5.2 Results (Screenshots) |  |
| **VI** | **Appendices** |  |
| **VII** | **References (Papers, Websites, Textbooks)** |  |

**list of figures**

|  |  |  |
| --- | --- | --- |
| Figure No. | Figure Name | Page No. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of tables**

|  |  |  |
| --- | --- | --- |
| Table No. | Table Name | Page No. |
|  |  |  |
|  |  |  |

**Chapter 1**

**1.Introduction**

* 1. **INTRODUCTION TO THE PROJECT WORK**

In recent decades, the realm of scientific research has undergone a profound transformation, characterized by an unprecedented explosion in the volume of scholarly output. This exponential growth, fueled by factors such as increased funding, technological advancements, and the globalization of research networks, has ushered in a new era of scientific exploration and discovery. However, alongside these advancements, researchers now face a formidable challenge: the daunting task of navigating an ever-expanding sea of information.

The traditional mechanisms for organizing and accessing scientific literature, while invaluable in their time, are struggling to keep pace with this deluge of new knowledge. The sheer magnitude and complexity of the literature pose significant obstacles to researchers seeking to stay abreast of the latest developments, identify relevant studies, and synthesize disparate findings across disciplines. As a result, there is a pressing need for innovative solutions that can effectively harness the wealth of knowledge contained within this vast corpus and facilitate the process of knowledge discovery and innovation.

Moreover, the interdisciplinary nature of many scientific endeavors further complicates matters. Breakthroughs often emerge at the intersection of multiple fields, where the cross-pollination of ideas and methodologies can lead to transformative insights. Yet, the siloed organization of scientific literature can obscure these critical linkages, hindering the very process of interdisciplinary collaboration and inhibiting the emergence of novel discoveries.

To address these challenges, our project aims to leverage the power of unsupervised machine learning techniques, specifically Latent Dirichlet Allocation (LDA) for topic modeling, coupled with established clustering algorithms. By applying these advanced computational methods to the vast corpus of scientific literature, we seek to uncover hidden patterns and relationships that may not be immediately apparent to human readers.

At the heart of our approach lies the integration of LDA-based topic modeling into the clustering process. By extracting latent topics from textual data and uncovering the underlying thematic structures, we aim to reveal the intricate tapestry of knowledge that spans disciplinary boundaries. Through the identification of cohesive clusters that share common conceptual foundations, our solution promises to provide researchers with a comprehensive tool for exploring related work, fostering interdisciplinary collaboration, and accelerating the pace of knowledge discovery.

Ultimately, our project seeks to empower researchers with a powerful toolkit that streamlines the exploration of scientific literature, enhances interdisciplinary communication, and catalyzes innovation across various domains. By providing an intuitive interface for navigating the clustered literature, we aspire to alleviate the cognitive burden of information overload, enabling researchers to focus their energies on the intellectual pursuit of advancing scientific understanding and driving progress in their respective fields.

**1.2 OBJECTIVE OF THE PROJECT**

This framework utilizes unsupervised machine learning techniques to cluster scientific papers based on their abstracts, enabling researchers to navigate the ever-growing volume of literature more effectively.

**1. Data Preprocessing:**

* **Data Collection:** Gather a comprehensive dataset of scientific papers and their abstracts from reliable sources like online repositories or digital libraries.
* **Cleaning and Preprocessing:** Clean the abstracts by removing irrelevant information such as stop words (common words like "the" or "a"), punctuation, and formatting symbols. Apply stemming or lemmatization to reduce words to their root form.

**2. Topic Modeling with Latent Dirichlet Allocation (LDA):**

* **Train the LDA Model:** Train the LDA model on the preprocessed abstracts. Specify the desired number of topics (k) that best represent the underlying themes within the dataset.
* **Topic Identification:** The LDA model will identify k latent topics that represent the core thematic clusters within the abstracts. Each topic will be characterized by a set of keywords with associated probabilities, indicating their importance within that topic.

**3. Clustering Algorithm Evaluation and Selection:**

* **Candidate Algorithms:** Evaluate various unsupervised clustering algorithms like K-Means, DBSCAN, Hierarchical clustering, or Spectral clustering.
* **Evaluation Metrics:** Compare the performance of each algorithm using metrics like Silhouette Score, Calinski-Harabasz Index, or Davies-Bouldin Index. These metrics assess the quality of cluster separation and cohesion.
* **Selection:** Based on the evaluation results, choose the clustering algorithm that demonstrates the best performance in organizing the papers based on their thematic similarities identified by LDA.

**4. Interactive User Interface:**

* **Visualization:** Develop a user-friendly interface that allows researchers to visualize the clustered papers. Consider techniques like interactive scatter plots where each paper is represented by a point, and its position reflects its thematic similarity to other papers. Colors or shapes can represent different clusters.
* **Topic Exploration:** Allow users to explore the identified topics in detail. Display the keywords associated with each topic and highlight the most representative papers within each cluster.
* **Search Functionality:** Implement a search bar where users can enter keywords or topics to find relevant papers within the clustered landscape.

**5. User Efficiency Assessment:**

* **User Study Design:** Conduct a user study with a group of researchers to evaluate the impact of the system on their efficiency in accessing and comprehending relevant literature.
* **Data Collection:** Before and after using the system, researchers can answer questions related to time spent searching for relevant literature, the perceived difficulty of understanding the literature landscape, and their overall satisfaction with the search process.
* **Analysis:** Analyze the collected data to assess if the system improves researchers' efficiency in finding relevant papers and their overall comprehension of the literature.

**Additional Considerations:**

* **Scalability:** Design the framework to handle large datasets of scientific papers efficiently. Explore distributed computing techniques if necessary.
* **Explainability:** Develop mechanisms to explain the rationale behind the clustering results. This can be achieved by highlighting the keywords associated with each cluster and showcasing a few representative papers from each cluster.
* **Continuous Improvement:** Continuously monitor user feedback and explore advanced techniques like sentiment analysis or author expertise modeling to further enhance the framework's capabilities.

By implementing this framework, researchers can leverage the power of unsupervised machine learning to navigate the vast ocean of scientific literature. Extracting underlying themes through LDA and organizing papers into meaningful clusters can significantly improve the efficiency and effectiveness of literature searches, ultimately accelerating scientific progress.

**1.3 METHODOLOGY ADOPTED TO SATISFY THE OBJECTIVE**

**1. Data Acquisition and Preprocessing:**

* **Data Sources:**
  + Identify and access reliable sources of scientific papers and abstracts, such as online repositories (PubMed Central, arXiv) or digital libraries (ScienceDirect, IEEE Xplore).
  + Consider partnering with specific research institutions or scientific societies to access domain-specific datasets.
* **Data Cleaning and Preprocessing:**
  + Implement text cleaning techniques to remove irrelevant information from abstracts:
    - Eliminate stop words (common words like "the" or "a").
    - Remove punctuation and formatting symbols.exclamation
  + Apply stemming or lemmatization to reduce words to their root form (e.g., "running" and "ran" become "run").
  + Perform named entity recognition (optional): Identify and categorize named entities like authors, institutions, or chemicals to potentially include them as additional features for clustering (if relevant).

**2. Topic Modeling with Latent Dirichlet Allocation (LDA):**

* **Train the LDA Model:**
  + Utilize a machine learning library like scikit-learn (Python) or Gensim (Python) to implement LDA.
  + Experiment with different numbers of topics (k) to determine the optimal granularity that captures the thematic structure of the abstracts without overfitting.
  + Evaluate perplexity, a measure of model fit, to choose the optimal k value.exclamation
* **Topic Interpretation:**
  + Analyze the keywords associated with each identified topic. These keywords represent the core themes within the cluster.
  + Consider visualization techniques like word clouds to depict the most prominent keywords within each topic.

**3. Clustering Algorithm Evaluation and Selection:**

* **Candidate Algorithms:**
  + Choose a set of unsupervised clustering algorithms to evaluate:
    - K-Means: A centroid-based algorithm requiring pre-specified cluster number (k).expand\_more
    - DBSCAN: A density-based algorithm that can identify clusters of arbitrary shapes without pre-defining k.
    - Hierarchical clustering: A bottom-up or top-down approach that creates a hierarchy of clusters based on similarities.
    - Spectral clustering: Utilizes spectral properties of a similarity matrix to group data points.expand\_more
* **Evaluation Metrics:**
  + Implement evaluation metrics like:
    - Silhouette Score: Measures the average distance between a point and its closest cluster compared to the distance to points in neighboring clusters.
    - Calinski-Harabasz Index: Compares the variance within clusters to the between-cluster variance.expand\_more
    - Davies-Bouldin Index: Compares the ratio of the within-cluster scatter to the between-cluster separation.expand\_more
* **Selection:**
  + Evaluate the performance of each clustering algorithm on a representative sample of the data using the chosen metrics.
  + Select the algorithm that demonstrates the best performance in terms of cluster separation (low within-cluster variance) and cohesion (high between-cluster variance) based on the thematic similarities captured by LDA.

**4. Interactive User Interface Design:**

* **Visualization:**
  + Develop an interactive interface using web development frameworks like Flask (Python) or Django (Python) to create a user-friendly experience.
  + Utilize libraries like D3.js (JavaScript) to create visualizations for exploring the clustered papers.
  + Implement scatter plots where each point represents a paper, positioned based on its thematic similarity to other papers according to the clustering algorithm. Color or shape coding can differentiate clusters.
  + Allow users to zoom in and hover over points to view paper details (title, authors, abstract).exclamation
* **Topic Exploration:**
  + Provide a dedicated section to explore the identified topics.
  + Display the keywords associated with each topic and highlight the most representative papers within each cluster. Users can click on keywords to filter the displayed papers within the chosen cluster.
* **Search Functionality:**
  + Implement a search bar where users can enter keywords or topics to retrieve relevant papers within the clustered landscape.
  + Leverage the LDA topic model to identify papers whose abstracts share semantic similarity with the user's search query.

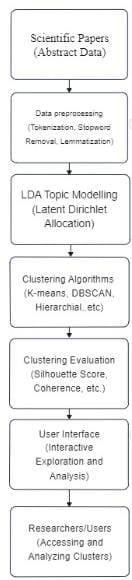
**5. User Efficiency Assessment:**

* **User Study Design:**
  + Recruit a representative group of researchers from various scientific disciplines to participate in a user study.exclamation
  + Divide participants into two groups: a control group using traditional search methods and an experimental group using the developed framework.
* **Data Collection:**
  + Before and after using the system, administer surveys to both groups:
    - Time spent searching for relevant literature.
    - Perceived difficulty of understanding the literature landscape.
    - Overall satisfaction with the search process.exclamation

Consider researching further to make sure the statement is credible.

thumbs\_up\_down

**1.4 ARCHITECTURE DIAGRAM WITH VERY BRIEF DESCRIPTION**



* The first step involves identifying and collecting relevant scientific papers. This could involve searching through online databases or digital libraries.
* The next step involves data preprocessing which includes tasks like tokenization (breaking text into individual words or phrases), removing stop words (common words like "the" or "a" that don't carry much meaning), and lemmatization (reducing words to their base form). Essentially, this step cleans the text data for further analysis.
* Then, the process moves on to topic modeling using Latent Dirichlet Allocation (LDA). LDA is a statistical technique that can identify hidden thematic structures within a collection of documents. In the context of scientific papers, it can reveal the underlying topics that unify seemingly disparate research articles.
* The final step shows the identification of the main topics. After the LDA model is trained on the preprocessed text data, it outputs a set of latent topics along with the keywords that are most representative of each topic.