Clustering of research paper dataset

Structure of dataset:

No. of research papers/ rows = 100000

Fields in the Dataset:

- id: A unique identifier for each paper.
- title: The title of the research paper.
- authors: The list of authors involved in the paper.
- venue: The journal or venue where the paper was published.
- year: The year when the paper was published.
- n_citation: The number of citations received by the paper.
- references: A list of paper IDs that are cited by the current paper.
- abstract: The abstract of the paper.

Example:

- "id": "013ea675-bb58-42f8-a423-f5534546b2b1",
- "title": "Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors",
- "authors": ["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"],
- "venue": "Journal of Computational Chemistry",
- "year": 2017,
- "n_citation": 0,
- "references": ["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"],
- "abstract": "This paper studies …"

Cite: https://www.aminer.cn/citation

Clustering Task: Group similar research papers together based on title and abstract separately

Report on Preprocessing Tasks

This report outlines the preprocessing steps performed on a dataset containing research paper titles and abstracts. The goal of these tasks was to clean and standardize the text data, making it suitable for further analysis or machine learning tasks.

1. Data Cleaning

1.1 Handling Missing Data

- The dataset was cleaned by removing any rows with missing values in either the 'title' or 'abstract' columns using the dropna() function.

1.2 Text Cleaning

- A clean_text() function was defined to perform the following operations:
  - Convert all text to lowercase
  - Remove special characters, numbers, and punctuation using regular expressions
-
- This function was applied to both the 'title' and 'abstract' columns, creating new columns 'title_clean' and 'abstract_clean'.

2. Tokenization

- The cleaned text in 'title_clean' and 'abstract_clean' was tokenized using the word_tokenize() function from NLTK.
- This process split the text into individual words or tokens.
- The results were stored in new columns 'title_tokens' and 'abstract_tokens'.

3. Stop Word Removal

- A set of English stop words was created using NLTK's stopwords corpus.
- Stop words were removed from both the title and abstract tokens.
- This step helps in reducing noise and focusing on more meaningful words in the text.

4. Lemmatization

- The WordNet Lemmatizer from NLTK was used to lemmatize the tokens.
- Lemmatization reduces words to their base or dictionary form, which helps in standardizing the text and reducing vocabulary size.
- This process was applied to both title and abstract tokens.

5. Reconstruction of Processed Text

- After all the above steps, the processed tokens were joined back into strings.
- New columns 'processed_title' and 'processed_abstract' were created, containing the fully processed text.

Report on TF-IDF Vectorization and SVD Dimensionality Reduction

This report details the process of converting the preprocessed text data into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, followed by dimensionality reduction using SVD (Singular Value Decomposition).

1. TF-IDF Vectorization

Two separate TfidfVectorizer instances were created, one for titles and one for abstracts, with the following parameters:

- min_df=3: Terms appearing in less than 3 documents were ignored.
- max_features=10000: Limited the vocabulary to the top 10,000 terms.
- ngram_range=(1, 2): Considered both unigrams and bigrams.
- stop_words='english': Used the built-in English stop words list.

The vectorizers were then applied to the processed titles and abstracts:

```
tfidf_title = tfidf_vectorizer_title.fit_transform(df['processed_title'])
tfidf_abstract = tfidf_vectorizer_abstract.fit_transform(df['processed_abstract'])
```

This step converted the text data into sparse matrices where each row represents a document (title or abstract) and each column represents a term in the vocabulary.

2. Dimensionality Reduction with SVD

To reduce the high-dimensional TF-IDF matrices to a more manageable size, Truncated SVD (also known as LSA - Latent Semantic Analysis) was applied:

```
svd_title = TruncatedSVD(n_components=30, random_state=42)
reduced_title_features = svd_title.fit_transform(tfidf_title)

svd_abstract = TruncatedSVD(n_components=30, random_state=42)
reduced_abstract_features = svd_abstract.fit_transform(tfidf_abstract)
```

Key points:

- n_components=30: The number of dimensions was reduced to 30 for both titles and abstracts.
- random_state=42: Ensures reproducibility of the results.

K-Means Clustering Analysis Report

This report provides an overview of the K-means clustering analysis conducted on a dataset of research paper titles and abstracts. The analysis includes clustering results, evaluation metrics, visualization, and interpretation of findings.

1. Clustering Process

K-means clustering was applied separately to the SVD-reduced features of titles and abstracts. The initial number of clusters was set to 5 for both datasets.

Clustering Results for k=5:

- Title Clustering:
  - Silhouette Score: 0.1984
  - Davies-Bouldin Index: 1.8911
- Abstract Clustering:
  - Silhouette Score: 0.0677
  - Davies-Bouldin Index: 2.9981

2. Cluster Themes

Title Clusters:

1. Cluster 0: Focus on design and implementation, with terms like "design," "implementation," and "architecture."
2. Cluster 1: Emphasis on network-related topics, including "network," "wireless," and "sensor network."
3. Cluster 2: Methodological focus with terms such as "method," "optimization," and "analysis."
4. Cluster 3: Applications and models, featuring words like "application," "model," and "data."
5. Cluster 4: General terms with a mix of "using," "model," and "algorithm."

3. Evaluation Across Different Numbers of Clusters

Title Clustering:

- Optimal performance observed at 5 clusters (Silhouette: 0.1984, Davies-Bouldin: 1.8911) and at 8 clusters (Silhouette: 0.1957, Davies-Bouldin: 1.9552).
- Silhouette scores ranged from 0.1613 to 0.1984.
- Davies-Bouldin indices ranged from 1.7229 to 2.2703.

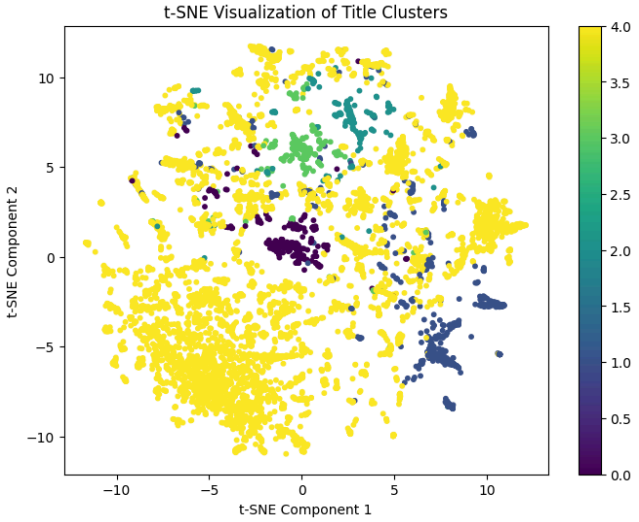
Abstract Clustering:

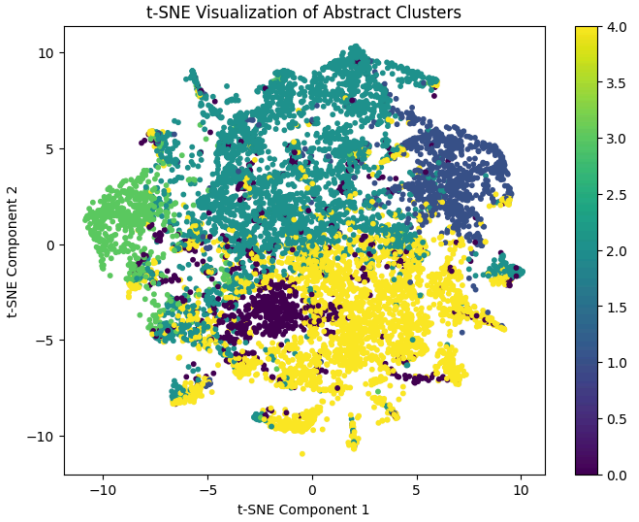
- Performance improved with more clusters, best at 10 clusters (Silhouette: 0.0813, Davies-Bouldin: 2.4484).
- Silhouette scores ranged from 0.0508 to 0.0813.
- Davies-Bouldin indices ranged from 2.4484 to 4.2505.

4. Visualization

t-SNE Plots:

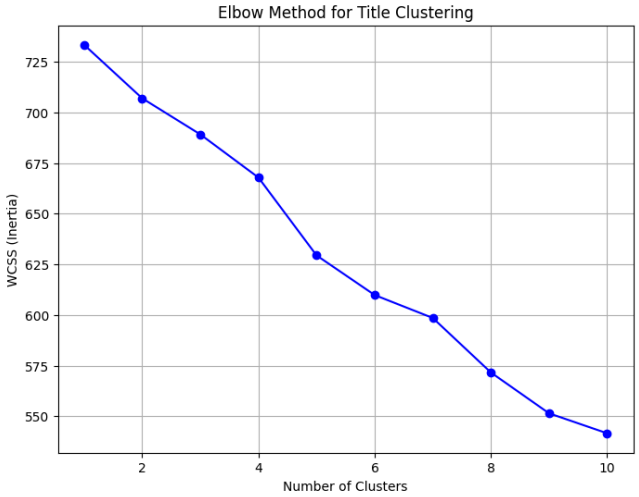
- Visualizations for both title and abstract clusters were generated using t-SNE, highlighting the distribution and separation of clusters in a reduced dimensional space.

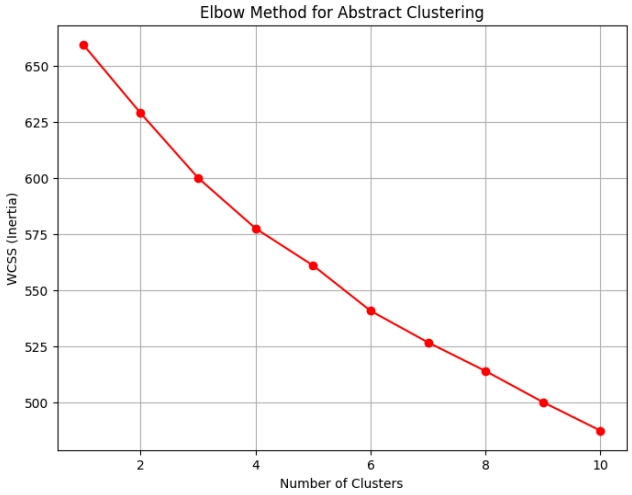




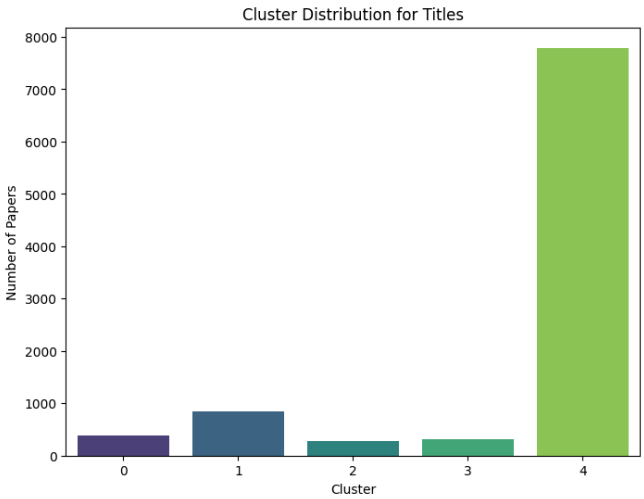
Elbow Method:

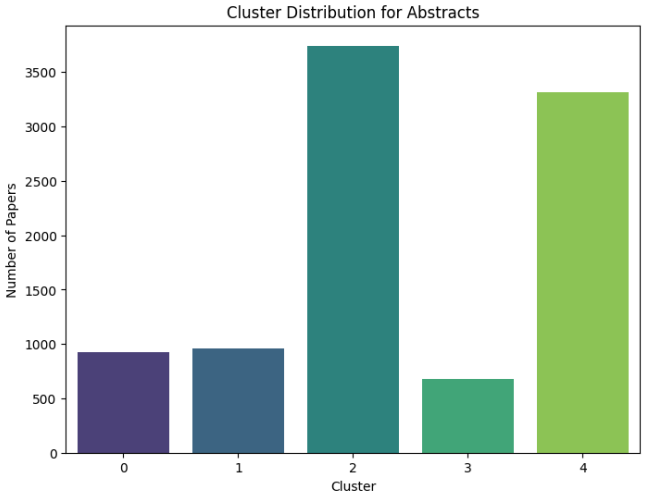
- Elbow plots were created for both title and abstract clustering to help determine the optimal number of clusters by showing the WCSS (Within-Cluster Sum of Squares) for different numbers of clusters.

- 

- 

Cluster Distribution:

- Bar plots illustrated the distribution of papers across different clusters for both titles and abstracts.
- 



Cluster Similarity:

- A heatmap visualized the cosine similarity between title cluster centers, indicating varying degrees of similarity between clusters.

5. Key Findings and Interpretations

1. Cluster Quality: Title clustering is more well-defined than abstract clustering, as indicated by higher silhouette scores and lower Davies-Bouldin indices.
2. Optimal Number of Clusters:
    - Titles: Using either 5 or 8 clusters is recommended based on performance metrics.
    - Abstracts: Increasing beyond 10 clusters might yield better results but should be balanced against interpretability.
3. Title vs Abstract Clustering: Title clustering consistently outperforms abstract clustering across all metrics and cluster numbers.
4. Cluster Themes: The identified themes in title clusters provide a high-level categorization of research papers, potentially useful for organizing large collections of academic literature.

5. Cluster Distribution: Some variation in cluster sizes is observed, but specific quantification is not possible without detailed distribution data.
6. Cluster Similarity: The heatmap reveals distinct relationships between title clusters based on cosine similarity.

Conclusions and Recommendations

1. Title clustering appears more effective for quick categorization of papers compared to abstract clustering.
2. For title clustering, using either 5 or 8 clusters is recommended based on performance metrics.
3. Abstract clustering may benefit from exploring higher numbers of clusters, balancing performance gains against interpretability.
4. Further analysis could include investigating higher cluster numbers for abstracts, exploring reasons for poor abstract clustering performance, and applying alternative clustering algorithms better suited to high-dimensional text data.
5. The identified cluster themes could be utilized to create a broad categorization system for research papers, aiding in the organization of large academic literature collections.

Agglomerative Hierarchical Clustering Analysis Report

This report presents the results of Agglomerative Hierarchical Clustering applied to a dataset of research paper titles and abstracts. The analysis includes clustering results, evaluation metrics, visualizations, and interpretation of findings.

1. Clustering Process

Agglomerative Hierarchical Clustering was applied separately to the SVD-reduced features of titles and abstracts. The number of clusters was set to 5, using the Ward linkage method.
Clustering Results:

- Title Clustering:
    - Silhouette Score: 0.17784253422285404
    - Davies-Bouldin Index: 1.6003855947616414
- Abstract Clustering:
    - Silhouette Score: 0.06067622721409091
    - Davies-Bouldin Index: 3.411757466172543

2. Cluster Themes

Title Clusters:

1. Cluster 0: General systems and algorithms (system, using, based, model, analysis)
2. Cluster 1: Networks and wireless systems (network, neural, ad hoc, wireless, mobile)
3. Cluster 2: Sensor networks and protocols (sensor, network, wireless, protocol)
4. Cluster 3: Design and implementation (design, system, analysis, implementation)
5. Cluster 4: Image processing and analysis (image, using, based, segmentation, analysis)

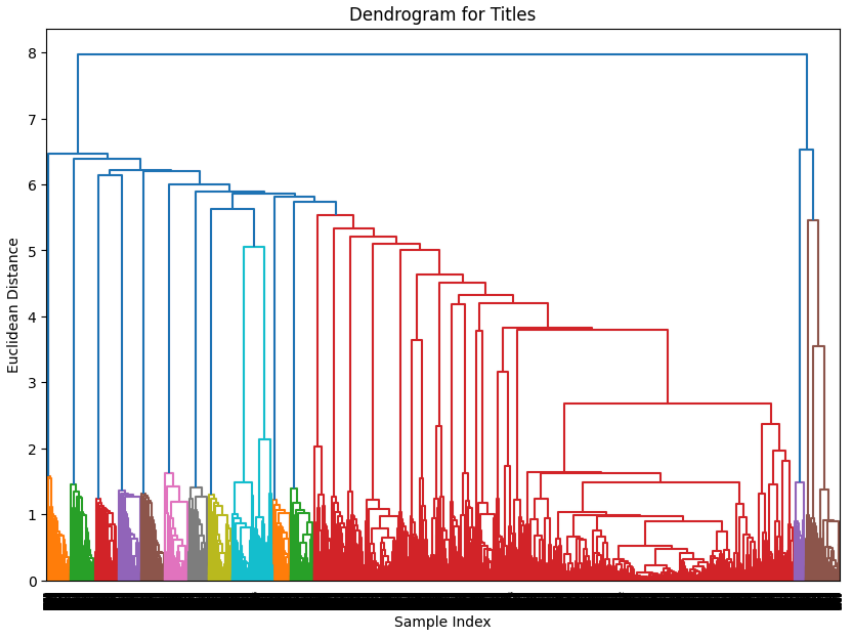
Abstract Clusters:

1. Cluster 0: General research topics (system, model, data, paper, result)
2. Cluster 1: Algorithms and methods (algorithm, method, system, problem, paper)
3. Cluster 2: Image and video processing (image, method, object, algorithm, video)
4. Cluster 3: Communication systems (channel, system, scheme, code, performance)
5. Cluster 4: Network and protocol research (network, node, algorithm, protocol, paper)

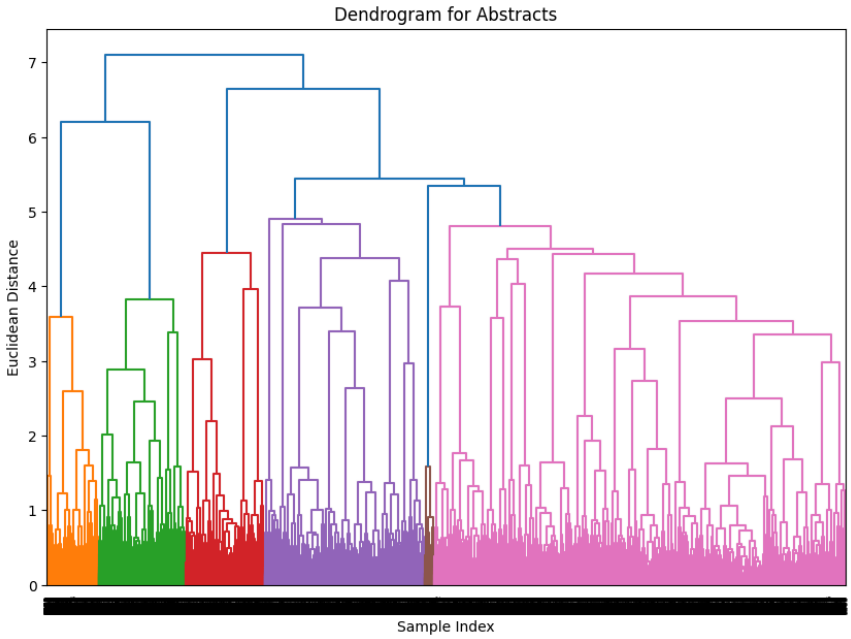
3. Visualization

Dendrograms:

- Dendrograms for both title and abstract clustering were generated, showing the hierarchical structure of the clusters.

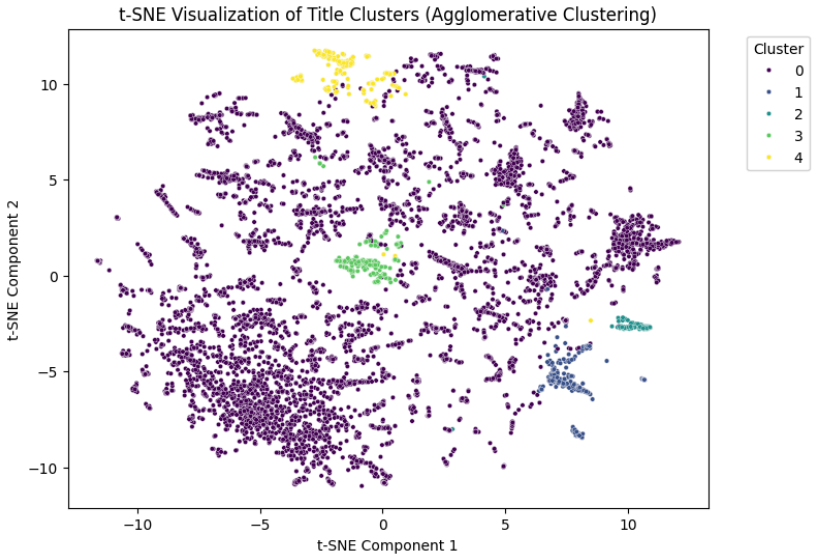


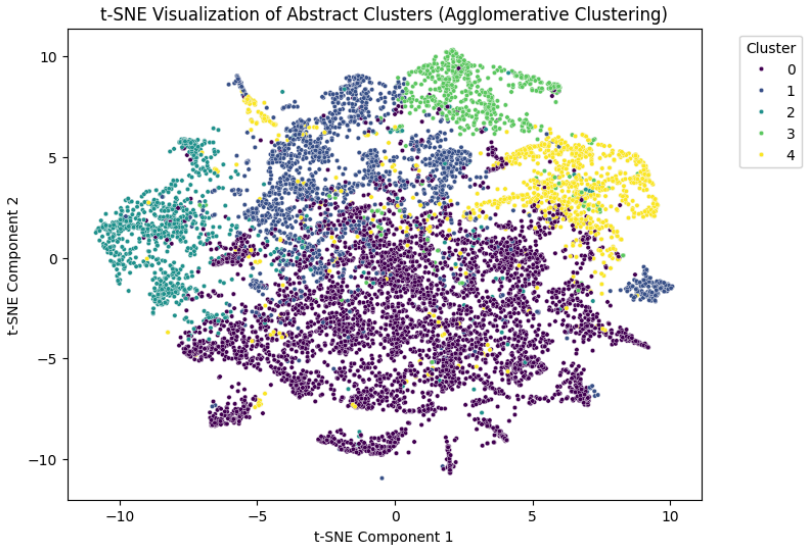
-



t-SNE Plots:

- t-SNE visualizations for both title and abstract clusters were created, illustrating the distribution of papers in a 2D space.





-

4. Evaluation of Different Linkage Methods

Various linkage methods were evaluated for title clustering:

1. Ward: Silhouette Score: 0.17784253422285404, Davies-Bouldin Index: 1.6003855947616414
2. Complete: Silhouette Score: 0.15983909881066063, Davies-Bouldin Index: 1.0016043646036805
3. Average: Silhouette Score: 0.5622112932339514, Davies-Bouldin Index: 0.6411216498647743
4. Single: Silhouette Score: 0.5408364281223241, Davies-Bouldin Index: 0.29063539642117164

5. Key Findings and Interpretations

1. Cluster Quality: Title clustering shows better-defined clusters compared to abstract clustering, as indicated by higher silhouette scores and lower Davies-Bouldin indices.
2. Linkage Methods: Average and Single linkage methods produced the highest silhouette scores and lowest Davies-Bouldin indices, suggesting they may be more suitable for this dataset than the Ward method initially used.
3. Cluster Themes: The identified themes in title and abstract clusters provide a high-level categorization of research papers, revealing distinct research areas and methodologies.
4. Visualization: The t-SNE plots show some separation between clusters, but there is also significant overlap, especially for abstract clusters.

6. Conclusions and Recommendations

1. Agglomerative clustering appears more effective for titles than for abstracts, suggesting titles may be more suitable for quick categorization of papers.
2. Further exploration of different linkage methods, particularly Average and Single linkage, could potentially improve clustering results.
3. The identified cluster themes could be utilized to create a broad categorization system for research papers, aiding in the organization of large academic literature collections.

4. Given the overlap observed in t-SNE visualizations, especially for abstracts, it may be worth exploring other clustering algorithms or feature extraction methods to achieve better separation.