Analyzing Research Trends in Computer Science Using the DBLP Dataset

Olutoba "Toba" Sanyaolu (2135924), Diamond Oladeji (2164567), Diego Coronado (2303693),
Data Science I Project Proposal
November 07, 2025

Contents

1 · Project Motivation

In computer science, research evolves rapidly across subfields such as artificial intelligence, data mining, cybersecurity, and software engineering. Understanding how these areas cluster together and how they have evolved over time can provide insights into emerging topics, collaboration trends, and the overall landscape of the discipline.

This project aims to analyze the DBLP dataset to identify and visualize clusters of research topics and examine how they change over time. Such an analysis could reveal which fields have grown in popularity, which have declined, and how different areas are interrelated.

2 · Dataset Description

The DBLP dataset, provided via AMiner, includes over 3 million research papers and 25 million citation relationships up to October 2017. Each record contains details such as: - Paper ID - Title - Authors - Venue (journal or conference) - Year - Number of Citations - Abstract - References

This dataset is suitable for both text-based and temporal analyses. Because it contains titles and abstracts, we can derive meaningful TF-IDF features that capture topic-related terms. The presence of year and venue attributes allows us to perform time-based trend visualizations and study the growth of different research areas.

3 · Initial Exploration

Our preliminary review of the dataset structure indicates:

- Multiple JSON files (`dblp-ref-0.json` to `dblp-ref-3.json`) must be merged for full analysis.
- The data contains missing fields and inconsistencies that will require cleaning (e.g., papers without abstracts or venues).

- Titles and abstracts contain rich textual information that can be vectorized for clustering and keyword frequency analysis.
- The number of papers increases dramatically after 2000, suggesting trends may shift over time.

4 · Planned Tasks

4.1 · Data Preprocessing & Feature Generation

- Merge and clean the four JSON files.
- Remove entries with missing or incomplete key attributes (title, year, abstract).
- Sample a manageable subset (e.g., 50,000–100,000 papers) for computational efficiency.
- Apply TF-IDF vectorization on titles and abstracts to represent textual content.
- Perform dimensionality reduction (e.g. PCA) to enable later visualization and reduce sparsity.

4.2 · Topic Clustering

- Use K-Means or DBSCAN to cluster papers into groups representing potential research fields.
- Identify top keywords in each cluster to interpret research themes (e.g., “neural,” “database,” “graph”).

4.3 · Temporal Trend Analysis

- Analyze how topic clusters evolve over time.
- Compute the yearly frequency of papers per cluster and visualize topic trends (e.g. growth of “machine learning” post-2010).
- Plot line graphs showing the rise, decline, or stability of clusters, and annotate shifts and transitions.

4.4 · Predictive Modeling (Classification)

- Train supervised models to classify papers into venues or topic labels based on textual and metadata features.
- Compare algorithms such as Logistic Regression, Random Forests, and LightGBM using 5-fold CV.
- Evaluate models with accuracy, macro/micro F1-scores, and confusion matrices to evaluate prediction performance.

4.5 · Network Analysis

- For the citation network, we will construct a graph of papers and citations, applying PageRank, centrality measures, and community detection to identify influential papers and collaboration clusters.

5 · Methods

- Programming Language: Python
- Key Libraries: pandas, json, scikit-learn, matplotlib, seaborn, wordcloud
- Analytical Techniques:
  - TF-IDF Vectorization (max_features=5000, stop_words=‘english’)
  - PCA for visualization (2D/3D scatterplots)
  - Clustering with K-Means or DBSCAN
  - Year-based frequency and trend graphs

- Word clouds for cluster keyword interpretation

6 · Expected Outcomes

- Visualizations of topic clusters (2D scatterplots, word clouds).
- Trend plots showing growth and decline of topics by year.
- Insights on which venues and authors dominate certain fields.
- Discussion of the most influential or rapidly evolving topics in computer science research.

Deliverables will include: - Figures and tables (topic clusters, trends, distributions). - Analytical discussion connecting patterns to real-world research trends.

7 · Anticipated Challenges

- Handling large data volumes efficiently (potential need for sub-sampling).
- Choosing optimal parameters for TF-IDF and clustering (balancing interpretability and accuracy).
- Dealing with noisy or incomplete textual data.
- Ensuring visualizations are both meaningful and reproducible.

8 · Contribution and Significance

This project contributes by combining topic modeling and temporal analysis on a large, real-world research dataset. Unlike simple EDA, our work aims to visualize research evolution and uncover meaningful topic clusters. The findings can benefit academic institutions, researchers, and policymakers seeking to understand the growth of emerging computer science fields.

9 · Collaboration Plan

Task #	Task Name	Assigned To	Notes
1	Data Preprocessing & Feature Generation	Toba	Merge JSON files, clean data, apply TF-IDF & PCA
2	Topic Clustering	Diego	Cluster papers by topic using K-Means/DBSCAN and identify key research themes
3	Temporal Trend Analysis	Diamond	Visualize topic growth, decline, and stability over time
4	Predictive Modeling (Classification)	All	Train and evaluate models (LogReg, Random Forest, LightGBM)
5	Network Analysis	All	Construct and analyze citation graphs (PageRank, centrality, communities)