# Exploring Research Trends: A Data-Driven Approach to Scientific Discovery and Topic Evolution

**Shashwat Shrivastava**

## Abstract

This project uses research paper trend analysis and topic modeling to investigate automated scientific discovery in NLP. Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP) modeling, and time series forecasting of topic prevalence using ARIMA are the three main approaches used to identify underlying themes in scientific literature. The project provides insights into the evolution of scientific topics over time, identifies related clusters of research, and predicts future trends based on historical data. The methods used are intended to provide a scalable method for automating scientific trend analysis and speeding up knowledge discovery.

## 1 Introduction

In the field of Natural Language Processing (NLP), the necessity of automated tools to support scientific discovery has become more widely acknowledged. The difficulty of determining and monitoring the development of research ideas throughout time is addressed by this project. The project seeks to identify significant patterns and trends that can guide future study by examining massive collections of scientific publications. The research focuses on three approaches: ARIMA for time series forecasting, HDP for hierarchical clustering, and LDA for topic modeling. The concept of automated scientific discovery in NLP is advanced by these methods, which when combined allow for a thorough approach to topic discovery, clustering, and prediction.

## 2 Methodology

### 2.1 Data Source

The dataset has been taken from Kaggle. The original dataset contains a total of 1,000,000 records with the following columns: Abstract, Authors, n_citation, References, Title, Venue, Year, and ID. For this project, I used a subset of the dataset containing 1500 records, with 500 records for each year (2015–2017).

### 2.2 Text Preprocessing

One text column is created by combining the title and abstract columns. The text is cleaned using the preprocessing function `preprocess_text` by:

- Removing numbers and punctuation.
- Converting all words to lowercase.
- Using ENGLISH_STOP_WORDS from scikit-learn to eliminate stopwords.

### 2.3 Feature Extraction

The cleaned text is converted into a document-term matrix (DTM) using the CountVectorizer, which represents the text as a sparse matrix of word counts. The term frequencies are then transformed into TF-IDF values using a TfidfTransformer, prioritizing less common but more significant words.

### 2.4 Concepts

#### 2.4.1 Latent Dirichlet Allocation (LDA)

LDA was used to model topics. The processed text was transformed into a document-term matrix using TF-IDF. By identifying 300 topics, LDA highlighted the recurring themes in scientific writing. Figure1 shows the top 10 keywords for each topic.

#### 2.4.2 Hierarchical Dirichlet Process (HDP)

HDP was used in conjunction with LDA for hierarchical clustering. A dendrogram visualizes the relationships between topics. Figure 2 shows the hierarchical structure of topics.

#### 2.4.3 Time Series Forecasting using ARIMA

ARIMA was used to predict the future prevalence of the top 10 topics. Figure 8 shows the predicted trends for the next 3 years (2018-2020).

## 3 Results

### 3.1 Latent Dirichlet Allocation (LDA)

- Figure 1 shows the top 10 keywords for each category.

- The topic coherence score (figure2) for the 300 topics was 0.1189. This low score indicates that some topics are less coherent.Despite the relatively low score, it indicates that some themes may be less coherent, meaning that the most popular words for such topics may not always be in line with a distinct, well-organized theme

- Figure 3 shows the distribution of topics across research papers.

- Figure 4 shows the evolution of topics over time (2015-2017).

### 3.2 Hierarchical Dirichlet Process (HDP)

The hierarchical clustering revealed different clusters of topics. Figure 5 shows the topic clusters generated by the HDP model. Different clusters of study subjects within the corpus were identified by the hierarchical topic modeling technique. The distance metric (FIGURE7) on the y-axis of the dendrogram visualization shows a distinct hierarchical structure with different levels of topic similarity. With distances ranging from 0 to 3.0, the analysis found multiple large subject clusters that suggested both more specific subtopics and more general thematic groups. There are two main levels of organization visible in the structure (figure 7): There is a significant break at distance 3.0, suggesting that the study themes are fundamentally divided. At distances of 1.0 to 1.5, several smaller clusters appear, signifying more closely related study subtopics.

• Coherence Analysis The cluster coherence scores reveal (figure6): • Most clusters (5-10) show relatively low coherence (below 0.07) • Three clusters (1, 2, and 3) demonstrate notably higher coherence • Cluster 1 significantly outperforms others with its 0.958 coherence score This structure suggests that while most topics are loosely related, there exists one very cohesive cluster of topics (Cluster 1) that represents a well-defined research area within the corpus.

### 3.3 Time Series Forecasting using ARIMA

The ARIMA model predicted the trends for the top 10 topics over the next 3 years. Figure 8 shows the prediction results for the top 2 topics.

## 4 Key Insights

- Topic modeling and hierarchical clustering helped find linked subjects and monitor new research trends.

- TF-IDF helped improve the evaluation of topic quality.

- ARIMA time series forecasting demonstrated potential in predicting future research trends.

## 5 Conclusion

This study effectively illustrated the potential for automated scientific discovery using topic modeling, hierarchical clustering, and time series forecasting. The methods successfully predicted future research trends and provided insights into the evolution of research topics. Future work will explore refining the forecasting models and investigating additional clustering and trend analysis techniques.

## A Appendix

This appendix includes supplementary material and additional details on the methodology, figures, and results that support the analysis in the main sections.

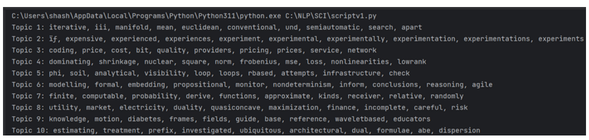


Figure 1: Top 10 keywords for each category (LDA).

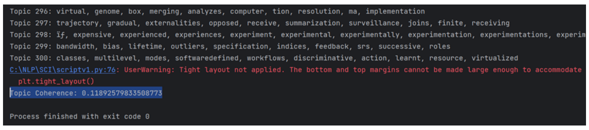


Figure 2: Coherence Score for the LDA model..

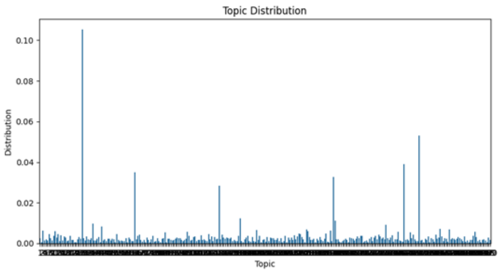


Figure 3: Topic distribution of research papers.

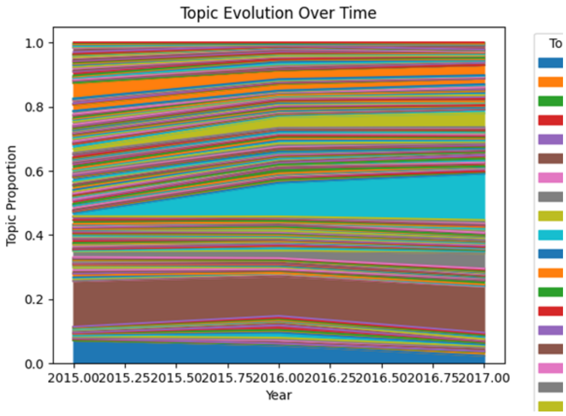


Figure 4: Evolution of research topics over time.

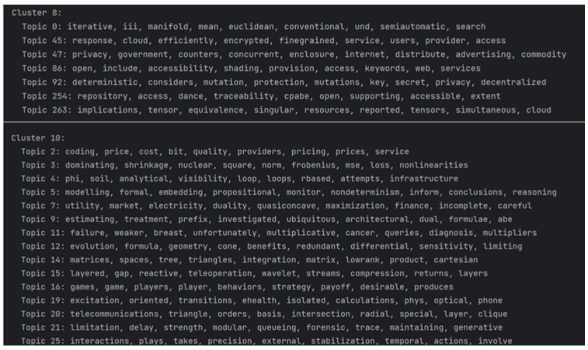


Figure 5: Clusters of research papers generated by HDP model.

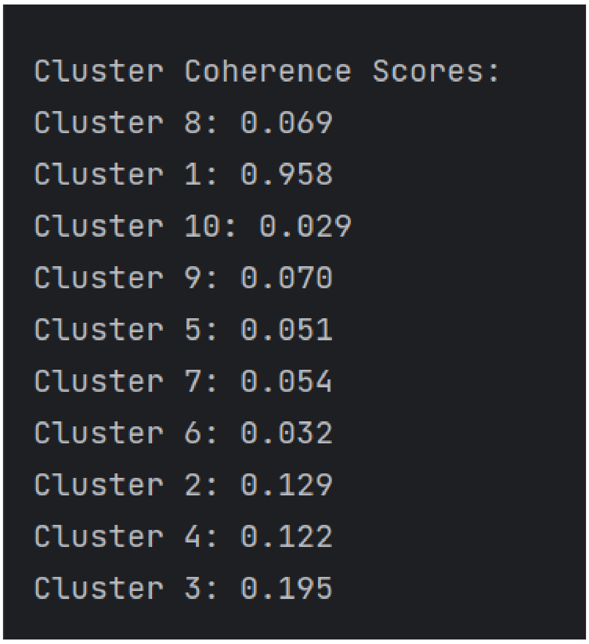


Figure 6: Coherence score of different clusters of research papers generated by HDP model.

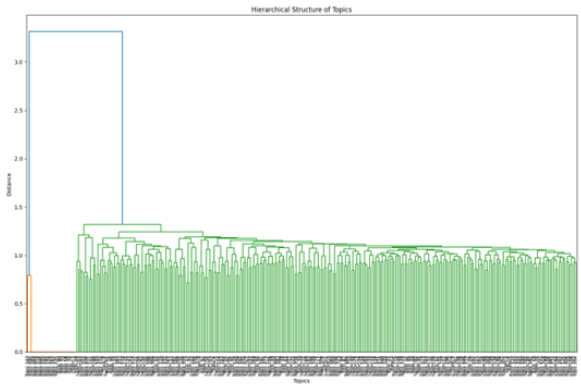


Figure 7: Dendogram showing the hierarchical structure of the topics.


Figure 8: Results of time series forecasting using ARIMA model.