Mapademic

Final Project for CAPP 122[1]
Winter 2025

Allen Wu (songting) Peiyu Chen (peiyuch)
Shiyao Wang (shiyao611) Yue Pan (pany17)

1 Abstract

Mapademic is an interactive platform designed to chart the global distribution and temporal dynamics of academic research by integrating bibliometric analysis with geospatial visualization to address strategic issues in education and industry. The system is built on a Python technology stack, utilizing the Scopus API for bibliometric data retrieval and Natural Earth for geospatial boundary analysis. It employs Lasso regression analysis to identify significant shifts in research trends and regional focal points, while interactive heatmaps (generated using Plotly, coloring based on CRDI.) and keyword-driven word clouds dynamically illustrate research density and conceptual hotspots. A front-end interface developed with Streamlit enables users to submit custom queries and filter results by time frame. Motivated by the pursuit of accessible knowledge, this project aims to address two core questions: (1) For students and researchers, in which specific regions is global research concentrated? (2) For industry stakeholders, which geographic areas offer the most favorable conditions for academic-industrial collaboration?

2 Data

2.1 Data Sources

2.1.1 Academic Data

We used the Scopus API to collect information about papers and the administrative affiliations of the institutions to which they belong.

2.1.2 Geospatial Data

We used Natural Earth's over 4,500 internal administrative divisions of countries Shapefile data to construct the administrative divisions polygon for mapping.

2.2 Mathematical Model (Research Density)

Our model trys to evaluate the research impact of a geographical region based on three key indicators: $Total\ Paper\ Density$, $Total\ Citation\ Density$, and $Global\ Quality\ Coefficient$. These indicators collectively contribute to the Comprehensive Research Density Index (CRDI).

$$CRDI = \frac{1}{3}\frac{\sum_{t=0}^{T} Paper\ Count_t}{Area} + \frac{1}{3}\frac{\sum_{t=0}^{T} Citation\ Count_t}{Area} + \frac{1}{3}\frac{\sum_{t=0}^{T} Paper\ Count_t}{\sum_{t=0}^{T} Paper\ Count_t + \epsilon}$$

[1] All code for this project will be open-sourced on GitHub under the MIT License.

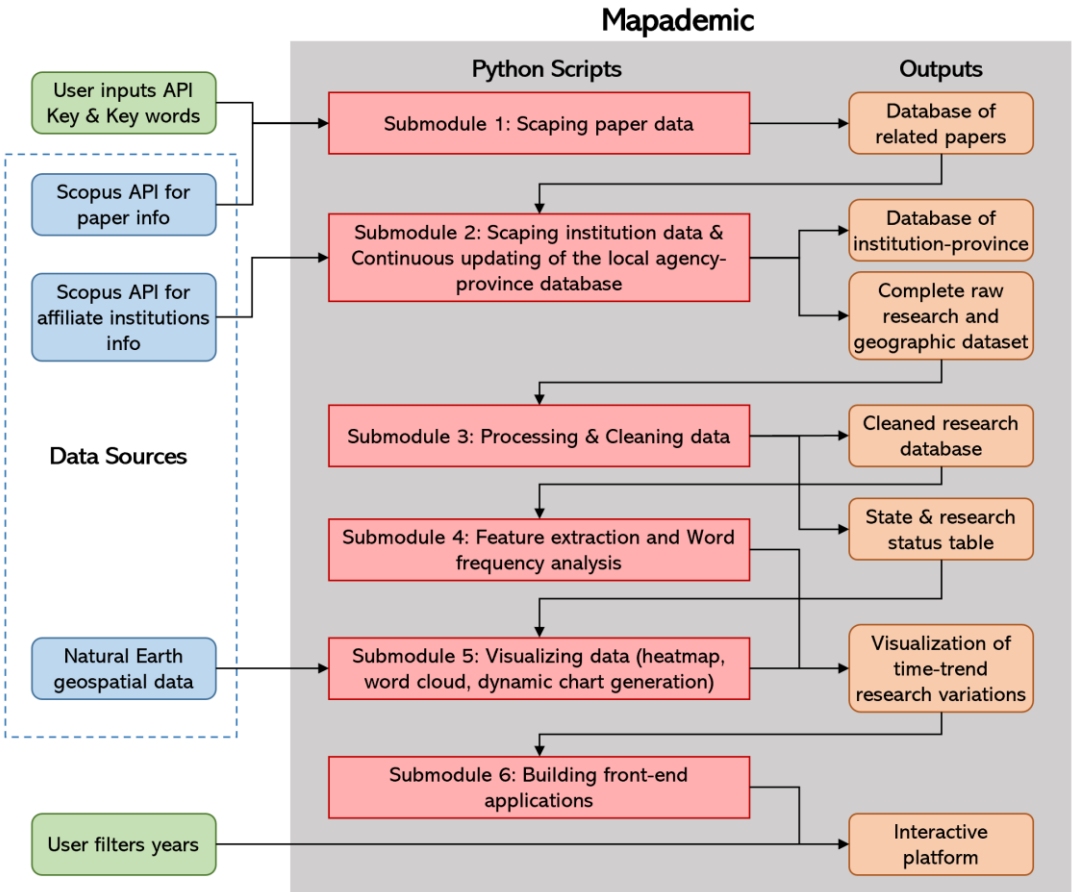
2.3 Data Flow

User inputs API Key and keywords → API calls to obtain keyword-related papers and author data → Matching and preliminary processing of author's institutional information → Cleaning, normalization, and feature extraction of collected data → Visualization

2.4 Cautions

- Please be aware of the stability of your API Key, the original version of the code set the limit of thesis calls to 100 per year for protection purposes, and the team is not aware of its call limit.
- Matching administrative district data from different sources is the biggest difficulty encountered in this project to map research information into geospatial, if you find anomalous matches during the process, please report them to us using the issue in GitHub.

3 Application Structure



4 Team responsibilities

Allen Wu	**API Calling**	• Deliver search results including author names, affiliated institutions, geographic data, and abstracts • Optimized API usage by reducing redundant calls • Addressed limitations in geographic details by handling institution IDs and mapping affiliations to state names
	Data access and processing	• Created and continuously refined a proprietary database by mapping affiliations to state names, enhancing the overall data pipeline

Peiyu Chen	**Data access and processing**	• Convert and clean Natural Earth's provincial shapefile into geojson data
	Visualization	• Heat mapped of research distribution by year and integration by timeline • Applied parallel computing to reduce plotting time
	Front-end construction	• Applied Streamlit caching to reduce data reads • Optimized the interface UI and enabled year filtering of charts
	Project Management	• Documentations and overall design • Built git branching and repo structures, managed project versions
Shiyao Wang	**Data processing**	• Data cleaning for abstract/title text data; Mapped API paper dataframe with geojson data • Constructed a research density index and employed a Lasso regression model to examine the influence of various factors on citation counts • Calculated word frequency, top cited institutions, top cited state/provinces by year
	Visualization	• Analyzed the dynamic trends in word frequency over time, generated annual word cloud visualizations, and plotted Lasso regression coefficients for all academic features
Yue Pan	**Front-end construction**	• Gather user inputs and invoke the keyword_search.py script to fetch paper metadata • Parse and standardize JSON data using cleaning scripts • Generate a heat map and allow users to view images like Top Features and Word Cloud

5 Conclusions and Reflections

During its inception, Mapademic was envisioned as an interactive search platform that would visualize the evolution of global research distribution and researcher mobility, thereby highlighting geographic variations in research concentration within specific fields. Based on this vision, we coined the name “Mapademic” by combining “Map” and “Academic.”

During implementation, limitations in database resources and the complexity of researcher information prevented the realization of the researcher mobility feature. Nonetheless, we maintain that this objective holds significant value and remain committed to pursuing it in the future.

Revisiting the project’s original intent—to elucidate the evolution of academic distribution and knowledge progression—we re-evaluated its relevance to individual researchers. Consequently, we refocused our efforts to address two key questions: (1) Where can I study and engage with relevant research? and (2) What are the emerging trends in my areas of interest? To this end, we enhanced the project with features that display the temporal evolution of top features, word clouds, and word frequency data.

Finally, we uphold the belief that knowledge is a public asset. In alignment with our slogan, “Unfold the Map of Discovery,” Mapademic will continue to strive toward becoming a global visualization platform for academic mobility and knowledge evolution.