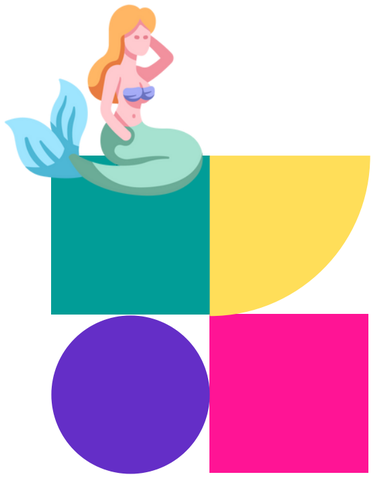


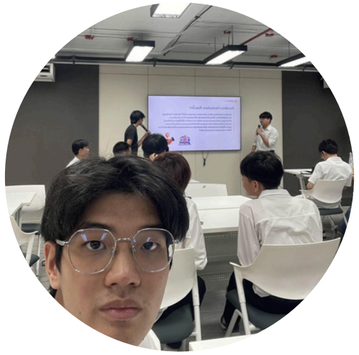
Little Mermaid

Introduction to Data Science and Data Engineering

Member

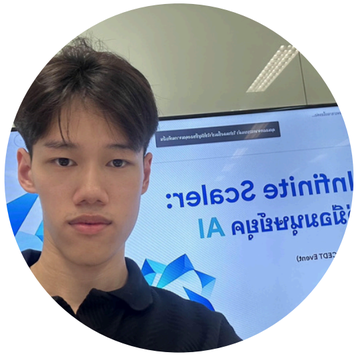
Pattaradanai Thanomsittikul

6633185721



Supanat Thanaphonpho

6633249221

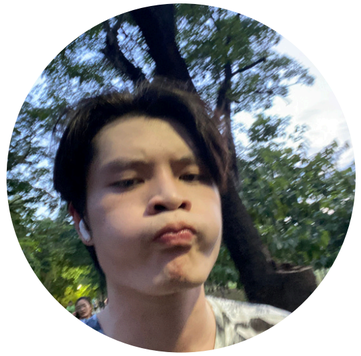


Naphat Chartwanchai

6633059321

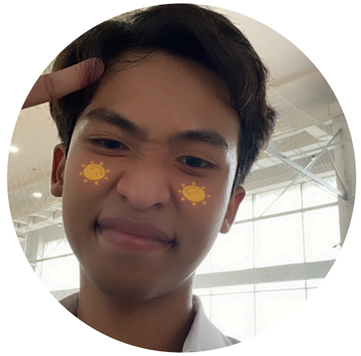
Phavarisa Pitavaratorn

6633181121



Jirayu Khunrak

6633036921



Peeratuch Khammeesak

6633175421

Objectives

To analyze engineering research data , identifying key trends, collaboration, and emerging topics, while presenting actionable insights through a comprehensive and visually engaging pipeline.

Data Used

Raw Data (JSON)

primary data requirement provided data from Scopus

Scopus API

data from Scopus API from 2015-2024

Web Scraping

Perform web scarping from Arxiv

Data Collected

Title

The name or headline of the research paper.

Abstract

A concise summary of the research.

Author

Names of authors who contributed to the research.

Aggregation Type

The type or method used for grouping or categorizing this paper.

Publisher

The organization or entity responsible for publishing the research.

Publication Date

The date when the paper was published.

Institutions

The academic or research institutions affiliated with the authors.

Keywords

A set of terms that capture the main topics of the research.

Data Pipeline

Raw Data (JSON)

Scopus API

Web Scraping

kafka

PySpark

pandas

- Analysis
- AI/ML

Data Engineering

Data from

- **Scopus API**: getting Title, Abstract, Authors, Aggregation type, Publisher, Publication Date, Institutions, and Keywords collected to CSV file
- **Web Scraping** from Arxiv getting Title, Abstract, Authors, and Publication Date collected to CSV file
- **Raw Data**

Data Cleansing

Collect all the data and aggregate it using Kafka running on Docker. The Kafka Producer will create topics for each year to organize the data for each year into a single file. Afterward, the data will be consumed through a Kafka Consumer, producing output files separated by year. These files will then be used for further processing.

Data Scientist

Topic Modeling	to uncover the latent (hidden) thematic structure in a large collection of unstructured text data and aims to group words into coherent topics and assign those topics to documents.
Recommend System	A system designed to allow users to input sentences or words and find the most relevant or suitable research articles. This system leverages NLTK (Natural Language Toolkit) to process text and identify similarities or connections between the input and a database of research articles.
Publication Trend	A system designed to analyze and predict trends in research publication over the years, categorized by quarterly periods.

Data Analysis

Coauthor Network	Visualize the collaboration network of researchers by displaying connections between authors based on their joint publications. This network representation highlights key collaborators, research clusters, and influential authors within the field.
Topic Analysis	Display trending research topics for each year by leveraging machine learning techniques to classify and extract themes from research publications. This analysis provides insights into the evolution of research interests and emerging areas of study.
Keyword Analysis	Compare the most frequently used keywords year by year to identify shifts in focus areas and terminology within the research domain. This analysis highlights trending keywords and provides a deeper understanding of the research landscape.

Dashboard &
Results Showcase

Conclusion

From the analysis conducted in our project, several key insights and future trends have emerged. One notable finding is that the highest number of publications consistently occurs in the first quarter (Q1) of each year, and this trend is expected to continue in the foreseeable future. Over the past decade, the top five most frequently used keywords have been COVID-19, Machine Learning, Artificial Intelligence, Educational Innovation, and Higher Education. In terms of emerging topics for 2024, the focus appears to be on AI and ML applications in industry, sustainability, environmental concerns, and agriculture, with aggregation systems also being a key area of interest. Additionally, journals remain the most popular type of publication among researchers. These trends provide valuable insights into the evolving research landscape, guiding future research directions aligned with current and emerging topics.

Thank You