# Mapping the Academic Landscape

## Creating a Knowledge Graph of Professors and Research Areas at IIT Kharagpur

*by Aditya Chawla(19CH3AI12), Apurva Anand(19EE10008), Pauras Meher(19ME3AI21), Saurabh Mishra(19IE3AI20)*

## Introduction

The project aims to create a knowledge graph of professors and their research areas at IIT Kharagpur. The project team used data scraping, preprocessing, and clustering techniques to gather and organise the data, and then generated a CSV file and an ontology with Protégé for further analysis and visualisation.

## Project Objective

The objective of the project is to create a comprehensive knowledge graph that maps the academic landscape of IIT Kharagpur, providing insights into the research areas and expertise of professors in various departments. This will enable students to efficiently locate and identify suitable professors to undertake research projects within their respective areas of interest.

## Methodology

### Data Scraping

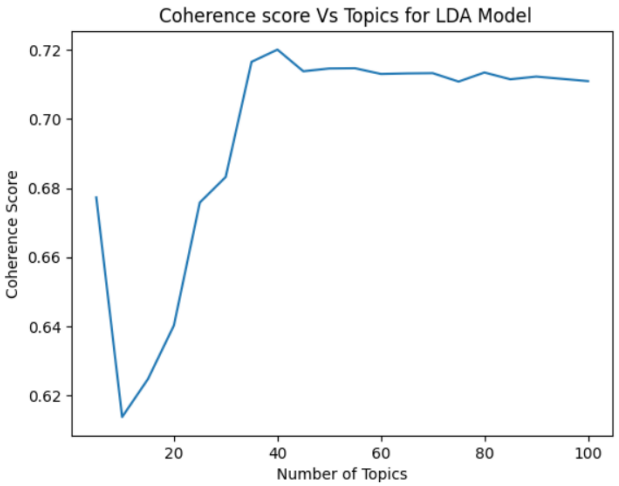
We scraped data from the IIT Kharagpur website [1] for each department, including the names and research areas of the faculty members. We used the requests and BeautifulSoup modules to extract the necessary information from the HTML content of the web pages.
**Link to code -** DataScrape

### Clustering

After data scraping, we obtained data for 390 professors and 1545 research topics. Owing to such a huge number of research topics, we decided to cluster related topics into broad research areas. The data was preprocessed to remove punctuation, convert all characters to lowercase, remove stop words, and create bigrams and trigrams to capture frequently occurring word combinations with a minimum count of 5 and a threshold of 10. Lemmatization was also performed to reduce words to their base form.

#### Models used

1. *LDA (Latent Dirichlet Allocation)*
   - Applied LDA for topic modelling on the preprocessed data.
   - LDA is a generative probabilistic model that assigns topics to documents and words to topics.
   - Plotted the number of topics vs coherence score and selected the optimal number of topics for which the coherence score was maximum.




- From the graph the optimal number of clusters is 35.
- However, the visualised topics did not have good clusters due to less data and lack of context from outside data.

### *2. BERT with K Means Clustering*

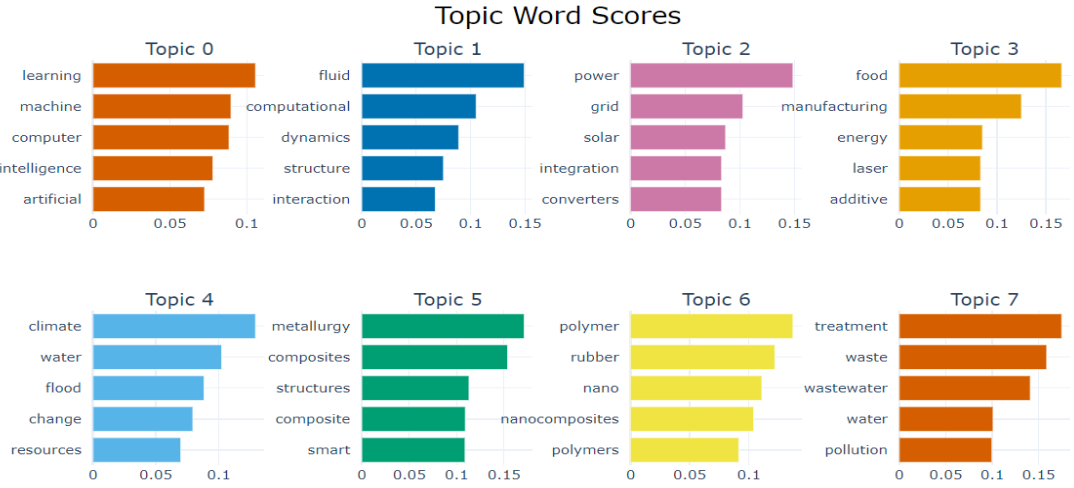
- Implemented BERT with K Means Clustering for topic modelling.
- However, the results were not very satisfactory as the algorithm was clustering outliers into the topics, leading to poor clustering results.

### *3. BERTopic with Hierarchical DBSCAN Clustering Model*

- Utilised BERTopic algorithm for topic modelling with hierarchical DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering model.
- The BERTopic algorithm involves three stages: embedding the textual data, clustering documents, and creating topic representation using class-based TF-IDF and Maximal Marginal Relevance to improve word coherence.

## Topics Visualization

- Visualised the topics for better understanding of the clusters.




**Frequency of sub-topics in each broader research area**

| | Topic | Count |
|---|---|---|
| 2 | Advanced Battery Management Systems for Electric Vehicles | 9 |
| 3 | Artificial Intelligence and Machine Learning | 66 |
| 4 | Computational Fluid Dynamics and Fluid-Structure Interaction | 81 |
| 5 | Renewable Energy Systems and Integration | 68 |
| 6 | Advanced Manufacturing and Food Processing | 38 |
| 7 | Integrated Water Resources Management and Climate Change Adaptation in River Basins | 71 |
| 8 | Materials Science and Engineering | 117 |
| 9 | Materials, Composites and Polymers | 101 |
| 10 | Water Management and Environmental Engineering | 44 |
| 11 | Communication Networks and Systems | 63 |
| 12 | Urban and Regional Planning | 39 |

Link to the complete final topics: Table of Topics

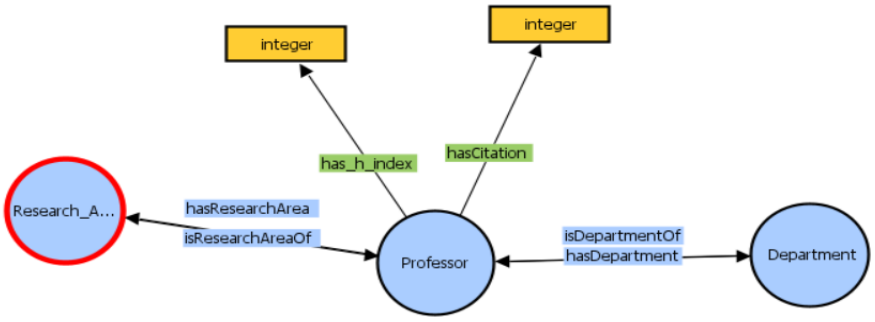
## CSV File Generation

- Input : Generated two dictionaries. The structure of the first one looks like this - `{Department: {Professor Name: [List of research topics]}}`
- The second dictionary contains the name of broad research areas (generated through ChatGPT) and the subtopics which fall into them.
- Using the above two dictionaries, we created the dataframe containing Name, Department and broad research area of the professors.

**Links to the code & final CSV file generated - Code, CSV file**

## Ontology Design

- Once the CSV file is generated, we used the Cellfie tool to create ontology from that.
- We used three **classes** - Professor, Department and Research_Area. We also declared four **object properties** - hasDepartment, hasResearchArea, isDepartmentOf and isResearchAreaOf. We have two **data properties** as well - has_h_index and hasCitation. Below is the visualisation of the ontology -




**Link to Youtube video - https://youtu.be/MMKfNg-jZus**

**Link to Github Repository - https://github.com/Chawlaji8781/KMSWT-Term-Project**

## References

[1] - http://www.iitkgp.ac.in/department/{dep}/faculties