Project Report

Problem definition:
Cluster research papers produced by all the researchers at Ashoka University into topics based on their content and compare them with researcher profiles. This could help you discover the main research themes and trends at Ashoka University, as well as identify potential collaborators or mentors across disciplines.

ML Problem Type:

Topic Modelling:

Topic modeling is a type of unsupervised learning technique that helps in discovering hidden topics or themes in a corpus of text data. It is a widely used approach in natural language processing (NLP) for analyzing large volumes of text data. In this project, we will use topic modeling to cluster research papers produced by all the researchers at Ashoka University based on their content. Latent Dirichlet Allocation (LDA) algorithm can be used, which models topics as probability distributions over words and assumes that each document is a mixture of these topics.

Recommendation system:

Recommend potential research collaborators according to past co-authors. Collaborative Filtering algorithm can be used, which uses the past behavior of users (in this case, researchers) to make recommendations.

Dataset:

CSV File 1: *papers.csv*
Data about all the research papers produced by researchers at Ashoka.
The dataset includes the following features:

researcher_id, paper_title, paper_id, publication_date, journal, publication_type, authors, article_url

CSV File 2: *researchers.csv*
Data about all the researchers at Ashoka with their researcher id.
The data set includes the following features:

Name, researcher_id, photo_url, description,scopus_url, website_url,emails, publication_count, department, position, journal_count

Note: Data from https://publications.ashoka.edu.in/. *scraper.py* is included in the submission.