Maria Castro, Seyla Wickramasinghe, Salma Yousry
Deliverable 1

Dataset Choice

We chose the ArXiv dataset because it's cleanly structured, large, and matches the goal of building an article recommender. It provides all the essential metadata for both content-based and collaborative filtering approaches. It is continuously updated making it suitable for a reliable and scalable recommendation system.

Methodology

Data Preprocessing

- The dataset chosen is feasible because it's clean, large, and in json format.
- The most useful information is the title, the abstract, and the categories
- Data preprocessing: Convert text to lowercase, remove special characters, text normalization (saw this online, removes common words)
- The goal is to recommend similar/related research papers based on user input features.

Machine learning model

- The proposed machine learning model is the KNN regression because it is a supervised model that will predict what research paper the user would like to read based on input features, such as title, abstract and keywords.
- Alternative models: Linear Regression could predict how likely a user is to like a paper based on features
- KNN Regression Pros: Easy to implement. KNN Regression Cons: Slow because it computes distance to all samples

Evaluation Metric

- Precision@k: the proportion of recommended items in the top-k recommendations of the system
- Recall@k: the proportion of relevant items found in the top-k recommendations

Application

Our web application will allow users to discover research papers related to a topic of interest based on keyword input. Users will enter text or keywords describing the topic, question, or concept they want to explore (for example, "neural networks" or single-cell data"). The input will be provided through a simple text box on the landing page. By default, the user will receive the top 10 most related papers.The user will receive a list of recommended papers, each displayed with its **title** and **DOI**. The results will be presented in a scrollable interface with clickable links to access the papers directly.