arXiv Recommendation System

Overview

The project revolves around enhancing arXiv, a popular research platform, by developing a recommendation system. This system aims to suggest five related articles to users based on their interests, thereby refining arXiv's search results and aiding in inspiration and research discovery.

Approach

- **Data Collection**: Utilizing the arXiv package in Python, the team collected a dataset of 20,000 articles relevant to a specified term. This dataset, structured and stored in a CSV file, was fetched using the arXiv API, ensuring relevance.
- **Variables and Features**: To determine relevance to a user, a custom approach was employed. Each article's congruence to the user's interests was quantified using cosine similarity. This method ranked the top 5 related articles for a given user.
- **Modeling Techniques**: The project employed Natural Language Processing (NLP) technologies. The team tokenized and vectorized the fetched articles, then calculated and ranked their cosine similarities based on titles and abstracts. Efficiency was enhanced through the implementation of the heapq package and Inverse Document Frequency Weighting (TFIDF) in vectorization.

Results

- **Accuracy and Strengths**: While the system's accuracy is subjective and not easily measurable, its efficiency is a significant strength. It quickly generates recommendations and uses TFIDF for effective vectorization of academic articles.
- **Weaknesses**: Limitations include reliance on a single generating term, slow dataset updates, and potential improvements in similarity calculations (considering categories and author relationships).

Future Work

- Improve the diversity of the data acquisition process
- Assign weights to the categories of articles to refine the result.
- Enhance the user preference by leveraging the user’s advisor’s and coworkers’ data
- Upgrade the vectorizer by using pre-trained models like Word2Vec and Doc2Vec