Description

This project aims to build a **Content-Based News Recommendation System** that suggests news articles to users based on the **similarity between news content and user interests**. The system builds a user profile from their explicitly stated preferences (e.g., preferred categories, topics, or keywords) and recommends articles with **similar content** using techniques like **TF-IDF** and **cosine similarity**.

We will use the **MIND (Microsoft News Dataset)**, a large-scale dataset from Microsoft News that includes detailed article metadata such as titles, abstracts, categories, and subcategories—ideal for content-based modeling.

Requirements

Functional Requirements:

- Load and preprocess the MIND dataset.
- Extract meaningful features from article content (e.g., title, abstract) using TF-IDF.
- Construct a user profile vector based on selected categories, keywords, or preferred articles.
- Compute cosine similarity between the user profile and news articles.
- Recommend top-N most similar articles to the user.

Technical Requirements:

Programming Language: Python 3.x

Libraries:

- pandas, numpy
- scikit-learn (for TF-IDF and similarity computation)
- NLTK or spaCy (for text preprocessing, optional)
- matplotlib / seaborn (for optional visualizations)

Dataset: Microsoft News Dataset (MIND):

https://www.kaggle.com/datasets/arashnic/mind-news-dataset/data

---

Deliverables

README.md: Project description, setup, and usage instructions.

data/: Folder containing the MIND dataset files.

- `news.csv`: Metadata for news articles.

notebooks/:

- `01_data_preprocessing.ipynb`: Load and clean the dataset; extract text features.
- `02_user_profile_construction.ipynb`: Build user profile vector from explicit input.
- `03_content_similarity.ipynb`: Compute article similarities using TF-IDF and cosine similarity.
- `04_ranking_and_recommendation.ipynb`: Rank and display recommendations.

results/: Folder for sample outputs, visualizations, or screenshots.

- `sample_recommendations.csv`
- `user_feedback_notes.txt`

requirements.txt: List of Python dependencies.