README for Reddit Data Scraping Scripts

Introduction
This document provides detailed information about two Python scripts developed for data collection and preprocessing from the Reddit platform, focusing on the 'tech' subreddit. The scripts are designed to automate the process of fetching and cleaning data for analysis purposes.

1. fetch_data_from_tech.py

The 'fetch_data_from_tech.py' script is designed to interact with the Reddit API via the 'praw' library to collect posts from the 'tech' subreddit. The script allows for customization of search queries to target specific content within the subreddit. The collected data includes post titles, text content, unique post IDs, post scores, total number of comments, and URLs. This data is compiled into a pandas DataFrame and subsequently exported to a CSV file named 'Posts.csv', facilitating easy storage and further analysis.

In addition to fetching and preprocessing data, the 'fetch_data_from_tech.py' script has been enhanced to include functionality for connecting to a MySQL database and storing the processed Reddit posts data. This extension leverages the 'mysql-connector-python' library to establish a connection to a MySQL database, where a dedicated table is used to store post details such as titles, content, post IDs, scores, total comments, and URLs. To utilize this feature, we have also ensured that a MySQL database is set up with the appropriate schema and that the necessary connection details (host, user, password, database name) are correctly configured within the script.

2. data_preprocessing_lab_4.py

The 'data_preprocessing_lab_4.py' script takes the raw data collected by 'fetch_data_from_tech.py' and applies a series of preprocessing steps to prepare it for analysis. These steps include removing duplicate entries, filling or removing missing values, extracting domain names from post URLs, and analyzing post titles to extract keywords and topics using Natural Language Processing (NLP) techniques. The script utilizes libraries such as 'pandas' for data manipulation, 'nltk' for natural language tasks, and 'spaCy' for advanced NLP processing. The outcome is a cleaned and structured dataset ready for analytical tasks.

Requirements
Running these scripts requires a Python 3.x environment with specific libraries installed, including 'pandas' for data manipulation, 'praw' for Reddit API interaction, 'nltk' and 'spaCy' for natural language processing tasks, and 'BeautifulSoup' for HTML parsing in data preprocessing. Installation of these libraries can be done via pip. Additionally, users must have valid Reddit API credentials configured to use 'praw' for data collection.

Usage

To use these scripts effectively, start by executing 'fetch_data_from_tech.py' to collect the desired data from the 'tech' subreddit. After the data collection is complete, and the 'Posts.csv' file is generated, run 'data_preprocessing_lab_4.py' to clean and preprocess the collected data. It's crucial to ensure that all dependencies are installed and that the environment is properly set up with the necessary API credentials before running the scripts.

Team name: HealthEdBot

Cici Chang (8706354957) Daoyang Li (7168949829) Yifan Yang(8386626967)