Overview

This script scrapes data from the Reddit tech subreddit, including post titles, authors, timestamps. It then processes the scraped data, conduct clustering to the posts and stores the data in a database.

Usage

1. Put the `cluster.py` and `main.py` under the same directory.
2. Run the `main.py` script like:

```
python main.py [interval_in_minutes]
```

3. Follow the instructions printed in the terminal.

Requirements

- See in requirements.txt

Installation

1. Install Python 3.x from python.org.
2. Install all the required Python packages using pip:

```
pip install -r requirements.txt
```

3. Install geckodriver for Firefox. You can download it from the geckodriver releases page and add it to your system's PATH.

Configuration

Ensure you have a compatible version of geckodriver installed for Firefox. Adjust the executable_path parameter in the scrape function if necessary. Customize the number of posts to scrape and other parameters as needed.

Files

1. main.py: Main script to run the data scraping and processing.
2. cluster.py: Script that contains the functions related to clustering. The functions are called in main.py.
3. cluster_notebook: A notebook that shows details of how the clustering part works, and a visualization image of the clustering results.

4. README.md: This file providing instructions and information about the project.

Credits

1. Wendy & Lorenzo: Clustering of the posts.
2. Jade: Integrate the clustering part with the automation and database.