**Reddit Scraper**

**Description**

* Scrape Reddit posts.
* Find the news content linked in each post.
* Extract keywords from each news item.
* Store posts and keywords in mysql database.

**Setup**

pip install praw newspaper3k selenium beautifulsoup4 pandas spacy

python3 -m spacy download en\_core\_web\_sm

Ensure MySQL is running and the user credentials provided in the script match your MySQL setup. Ensure selenium webdriver browser and path is correct. (Need to change praw API and MySQL credentials to yours in the scrape\_reddit.py, webdriver browser and path to yours in helper.py)

**Running the script**

python3 scrape\_reddit.py 20

Will scrape the first 20 new reddit posts and get full news content from the url in each post.

**User Options:**

Follow the on-screen prompts to interact with the system. Specify the number of posts to scrape.

**Note:**

Reuters uses CAPTCHA to block bot activities such as web scraping. If both local machine and virtual machine fail using selenium, then there is no way to retrieve news from <https://www.reuters.com/>.

**Author:**

Yi Sun, Ting-Han Chen, Jingyao Zhou