Codes Documentation

ANDREW LI

11/22/2019

Description

daily_scraping.py is a multi-processes and multi-threads enabled program for scraping the inforation from Eastmoney forum periodically. This program will be activated twice a day on (1:00 PM and 2:30 PM) and stay silent in background for the rest of the day.

1:00 PM

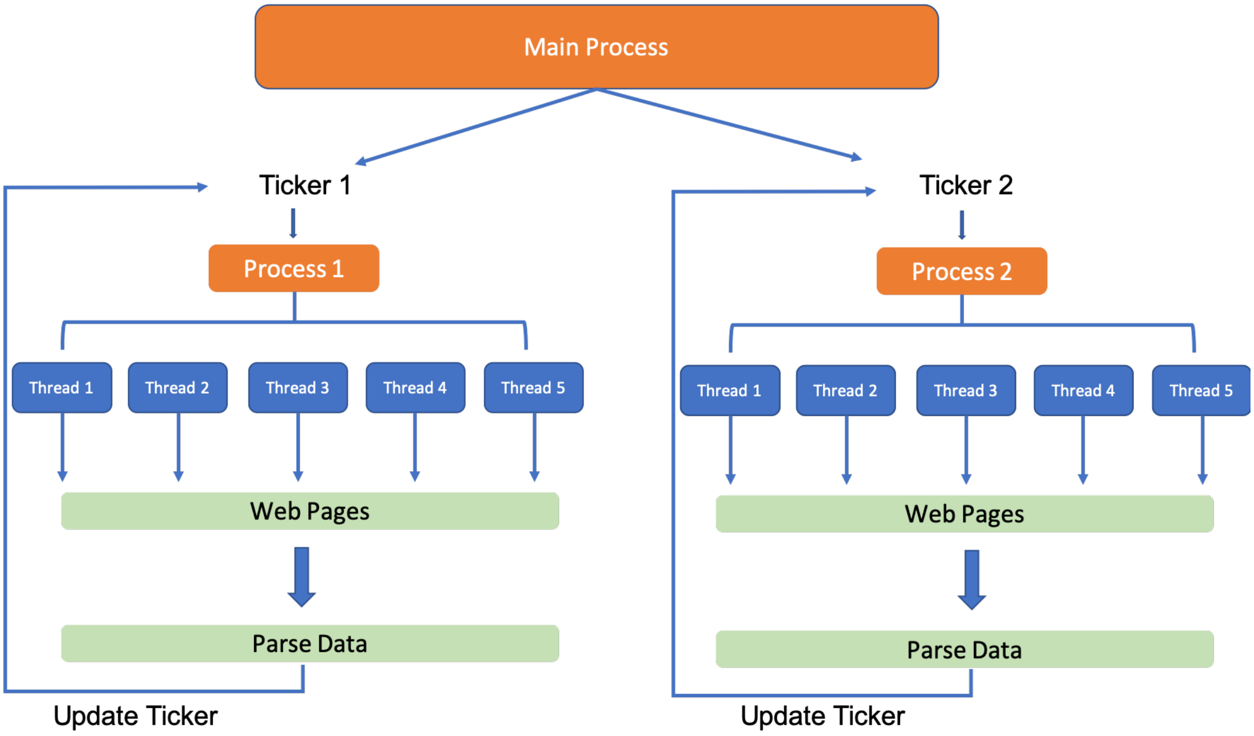
> The main purpose of this period is to update the database by adding the latest data.
>
> The program will scrape the 300 tickers from CSI300. For each ticker, 15 pages will be requested so total $300 * 15 = 4500$ pages will be scraped. The program will update the historical data in the database by using latest information.
>
> **Time complexity**: 10 - 15 minutes (depending on the number of total post of that day)
>
> **Concurrency**: 2 Processes and 5 Threads are used, which means that each process is able to send 5 requests to the remote sever at a single time point. After acquiring 15 pages information, the program will take 2 - 3 seconds to parse and reformat the information followed by downloading the rest of tickers.

2:30 PM

> The main purpose of this period is to update the database and create the daily summary table.
>
> The program will scrape the 300 tickers page by page during this period. We will compare the timestamp of scraped page with our historical data to decide whether more pages are needed to be scraped or not.
>
> **Time complexity**: 3 - 5 minutes (depending on the number of post of that day)
>
> **Concurrency**: 4 Processes and 1 Thread are used, which means that each process can only send 1 request to the remote sever at a single time point.

Work Flow

As shown below, the main process will distribute the tickers to sub-processes and each sub-process will create threads to request web page and wait for response independently. After having the complete information, the data will be formatted by each sub-process.



Execution

1. install python3.6 through here
2. run `pip install -r requirements.txt` to install required dependencies
3. run `python -W ignore daily_scraping.py`