crawl

ItemsPipeline

+ Sort itmes into Profile and Posts

+ Assign Items to mongo database
+ Save Items to local Json file

Scheduler

+ URL scheduler

+ Add URL to queue

Scrapy Crawl hsinWeiboSpider

Internet

+ Weibo data

+ Request HTTP sent
+ Retrun Json response

Spider

+ Start URL
+ Request
+ Parse

+ URL queue
+ Request profile
+ Request weibo
+ Parse all pages

Middleware

+ Fake human

+ User agents changed
+ Cookies changed
+ Auto delay setted

Advantages:

use scrapy as framework plus mongodb as database

1. Generally the crawl works fine. For each user can get all the data of profile and posts. 10mins for 15k+ data

2. Crawling from mobile websites with help of fiddler helps user accounts more difficult to be baned

3. A user agents changing and cookies changing system

4. Mongo database for data management and filter data.

Disadvantages:

1. Scrapy is too heavy
2. Without Redis distributed system, crawling is not fast
3. The most important feature(get profile and post for specific person) is realized but it is not very extendable.