Project Directory

spiders : a distribution spider system controlled by Celery

|---crawl\_worker : a scrapy project to get ccass data

|---other\_worker : other scrapy projects could be added

|---queue : save queue data to link Celery and Scrapy

|---tasks : celery tasks and settings

|---tools : some tools to deal with data

---crawl\_manager.py : celery task distributor

Docker Environment

1. MySQL

$ sudo docker run --name mysqllib –v ~/mysqllib:/var/lib/mysql -e MYSQL\_ROOT\_PASSWORD=toor -d mysql:5.6

2. Workers

$ sudo docker run --name ccass\_1 -it -v ~/Documents:/home/work --link mysqllib:mysql crawl\_worker:0.1

$ sudo docker run --name ccass\_2 -it -v ~/Documents:/home/work --link mysqllib:mysql crawl\_worker:0.1

$ sudo docker run --name ccass\_3 -it -v ~/Documents:/home/work --link mysqllib:mysql crawl\_worker:0.1

3. Manager

$ sudo docker run -it --name crawl-manager -v ~/Documents:/home/work --link mysqllib:mysql python:3.6

(Note: pip install celery)

Execute Crawl with Celery Tasks

1. MySQL

$ sudo docker exec -it ccass-mysql bash

# mysql – u root –p

Mysql> 1) distinct, 2) if not exists …

2. Workers – wait and run celery tasks

$ sudo docker exec -it ccass\_1 bash

# cd /home/work/spiders

# celery --loglevel=info -A tasks.scrapy\_task worker --max-tasks-per-child 1

---

Repeat ccass\_2, ccass\_3

3. Manager – distribute tasks

$ sudo docker exec -it crawl\_manager bash

# cd /home/work/spiders

# python crawl\_manager.py -i "00006" -d 60

Monitor

docker <http://172.17.0.1:9000/#/dashboard>

mq <http://67.209.179.247:15672/#/>