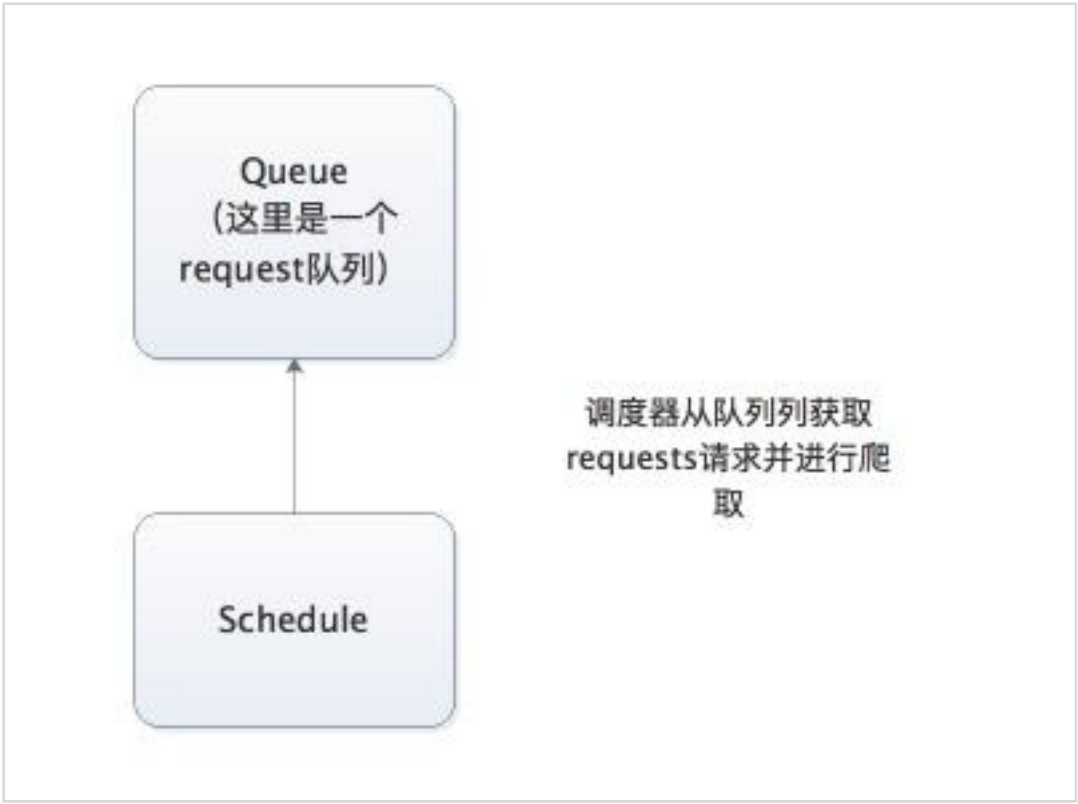
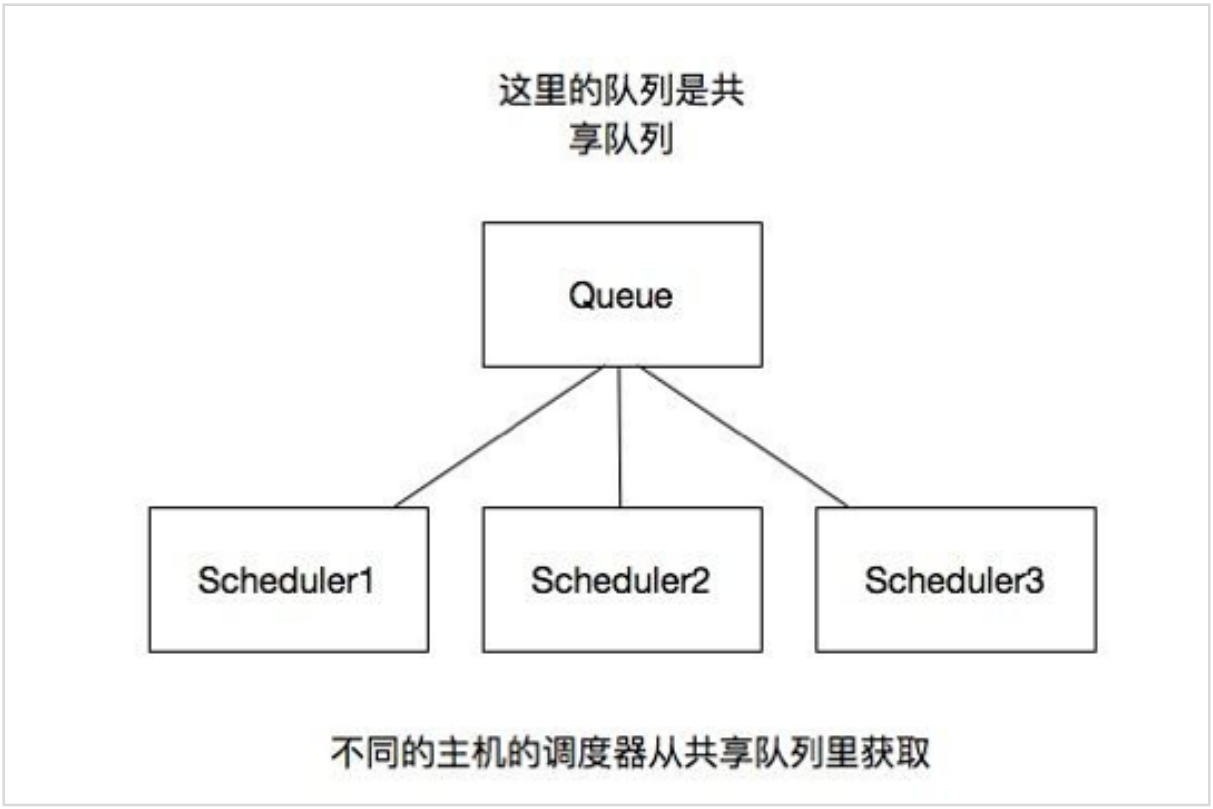
爬虫 day10 分布式爬虫

分布式爬虫是将多台主机组合起来，共同完成爬取任务

单机爬虫



分布式爬虫



爬取队列：使用Redis列表或有序集合 (默认是redis有序集合)
去重：使用Redis集合保存Request的指纹，提供重复过滤
中断续爬：调度器从Redis队列中取上次没有爬的继续爬取

使用分布式
安装scrapy-redis

```
pip install scrapy-redis
```

mongodb

```
show dbs;
use yaoqi
db.getCollectionNames()
db.yaoqis.count()
db.yaoqis.find()
```

redis
有序集合，在集合的基础上，为每元素排序；元素的排序需要根据另外一个值来进行比较，所以，对于有序集合，每一个元素有两个值，即：值和分数，分数专门用来做排序。

```
exists comic:requests

 exists comic:dupefilter

ZRANGE comic:requests 0 -1 withscores zcount
comic:requests 0 -1
```

headers 要一样 爬虫名要一样

连接redis redis-cli -h ip -p port

redis-cli -h 38.138.138.138 -p 6379