



本节知识点概述

爬取拉勾职位信息

数据库安装

MongoDB安装

可视化工具

基础使用

实战环节

分析请求

动手写爬虫



DC学院
class.pkbigdata.com

造数



MongoDB安装

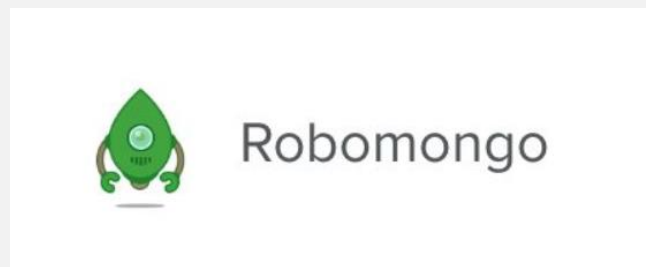
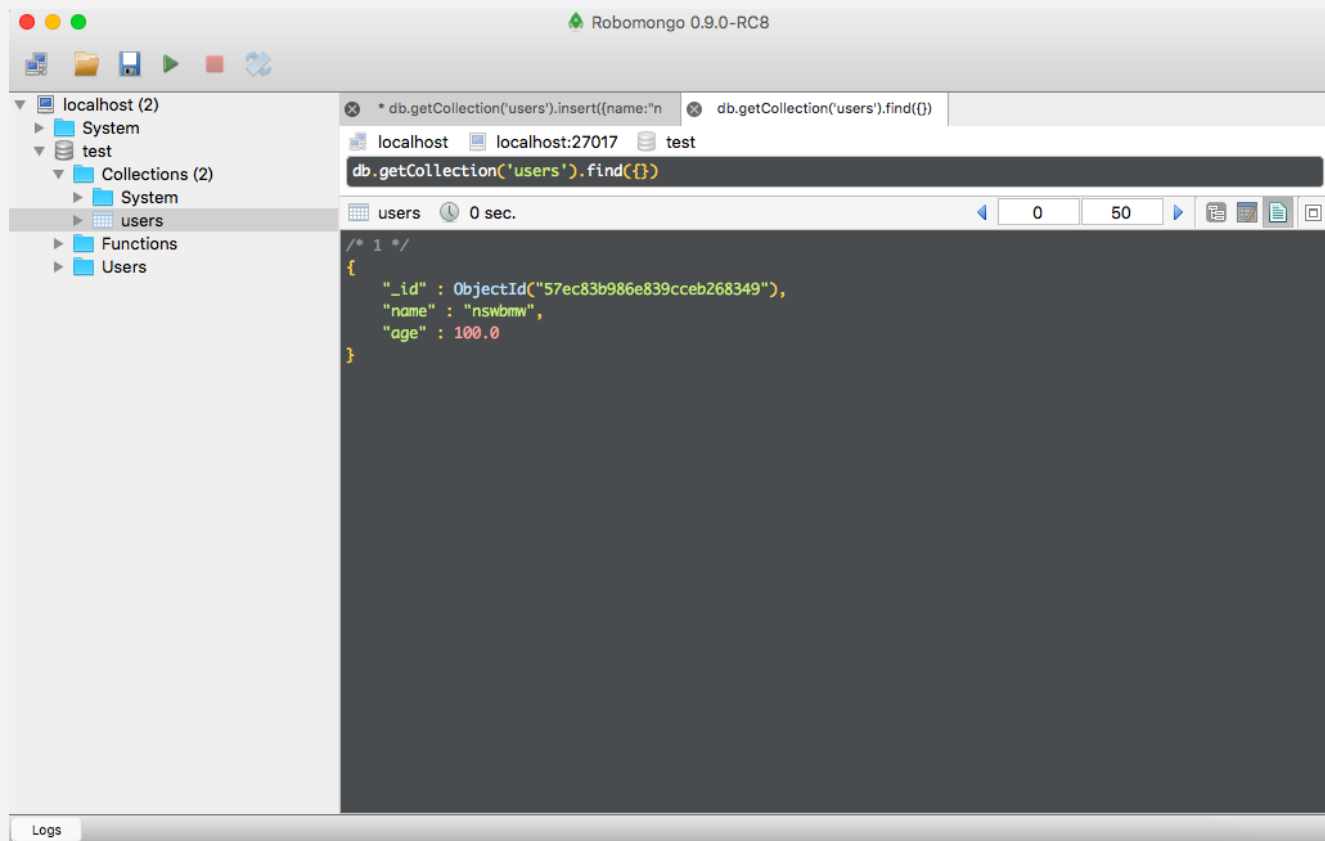
MongoDB 将数据存储为一个文档，数据结构由键值(key=>value)对组成
安装链接：

```
{  
  name: "sue",  
  age: 26,  
  status: "A",  
  groups: [ "news", "sports" ]  
}
```

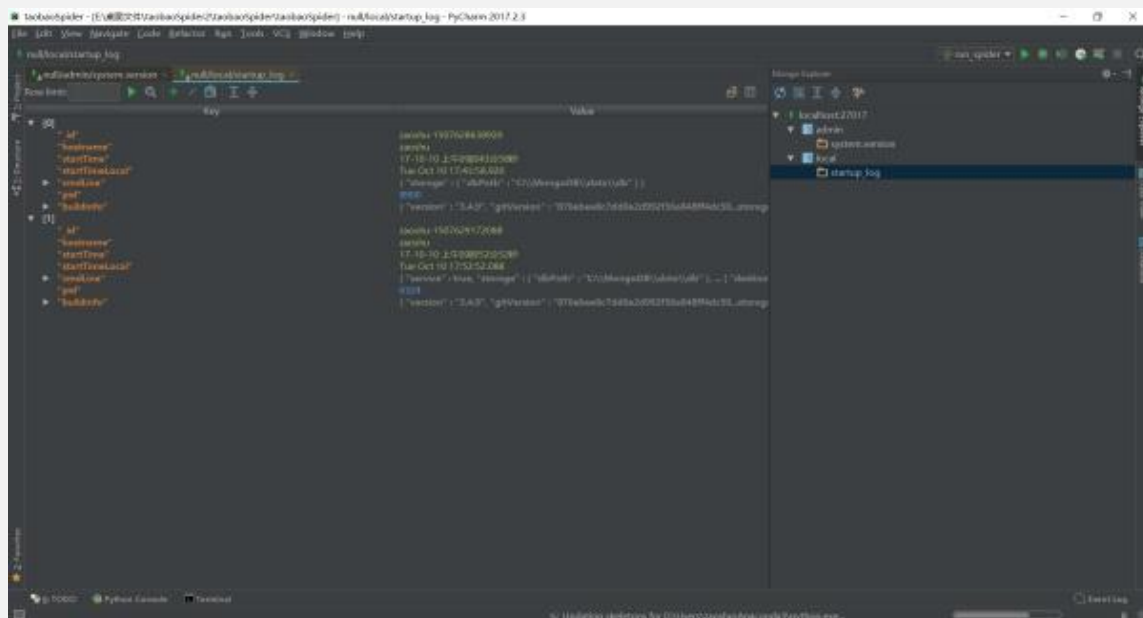
← field: value
← field: value
← field: value
← field: value



可视化工具



PyCharm插件



1、先安装第三方库

```
$ pip install pymongo
```

2、安装PyCharm的MongoDB插件

File——Plugins——mongo



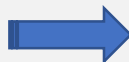
基础使用

```
#!/usr/bin/env python
# -*- coding:utf-8 -*-

from pymongo import MongoClient

client = MongoClient()
db = client.test #连接test数据库, 没有则自动创建
my_set = db.set #使用set集合, 没有则自动创建

my_set.insert({"name": "yyy", "age": 18})
```



db.getCollection('set').find({})

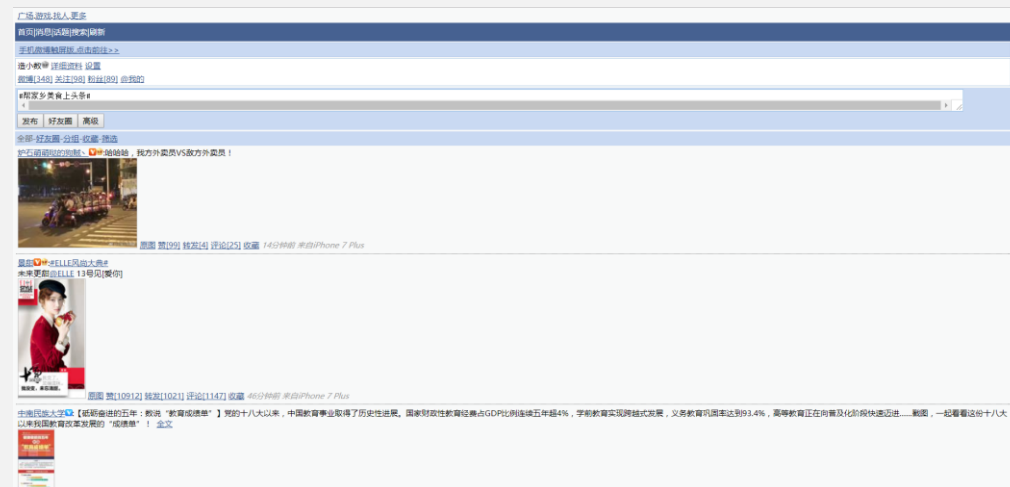
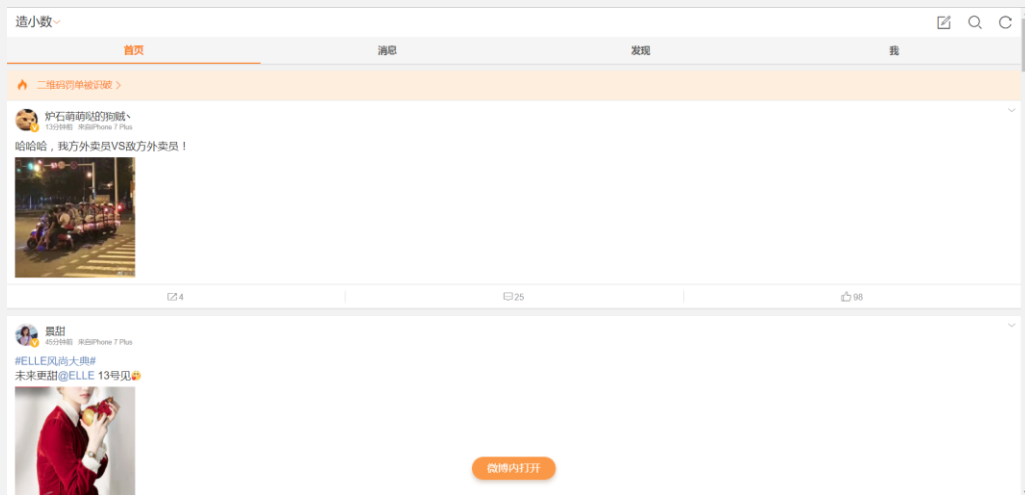
| Key | Value | Type |
|-----------------------------------|---------------------------------------|----------|
| (1) ObjectId("59dca411b385c7...") | { 3 fields } | Object |
| _id | ObjectId("59dca411b385c711cc897b...") | ObjectId |
| name | yyy | String |
| age | 18 | Int32 |



实战环节

<https://m.weibo.cn/>

<https://weibo.cn/>





实战环节

找工作-互联网招聘求职网-拉勾网

排序方式: 默认 最新 月薪: 不限 工作性质: 不限

爬虫工程师 [东城区] 09:53发布

10k-20k 经验3-5年 / 本科

软件开发 Java 算法

同牛科技 金融 / 成长型(A轮)

"五险一金,带薪年假,年度"

爬虫工程师 [海淀区] 2017-09-27

20k-30k 经验1-3年 / 本科

后端开发 分布式

Face++ 移动互联网,硬件 / 成熟型

"技术驱动,极客氛围,大牛"

爬虫工程师-政企事业部 [中关村] 15:44发布

10k-20k 经验1-3年 / 本科

中级 python

普林科技 数据服务 / 成长型(A轮)

"平台好,发展空间大"

爬虫工程师 [学院路] 14:12发布

10k-20k 经验1-3年 / 本科

中级 初级 Python Java

阿博茨科技 移动互联网 / 成长型(A轮)

"团队年轻,人工智能,扁平"

爬虫工程师 [朝阳区] 1天前发布

13k-25k 经验1-3年 / 本科

超对称 移动互联网,企业服务 / 成

我要反馈

Network

Hide data URLs All XHR JS CSS Img Media Font Doc WS Manifest Other

| Name | Status | Type | Initiator | Size | Time | Waterfall |
|-----------------------------------|--------|------|------------------|---------|--------|-----------|
| positionAjax.json?city=%E5%8C% | 200 | xhr | vendor_e3ddee... | 19.4 KB | 345 ms | |
| companyAjax.json?city=%E5%8C% | 200 | xhr | vendor_e3ddee... | 624 B | 317 ms | |
| approve.json?companylds=14166... | 200 | xhr | vendor_e3ddee... | 548 B | 438 ms | |
| oss.html?u=/jobs/list_%E7%88%A... | 200 | xhr | oss.js:1 | 213 B | 19 ms | |

4 / 95 requests | 20.7 KB / 181 KB transferred | Finish: 2.68 s | Load: 2.15 s

寻找请求



DC学院
class.pkbigdata.com

造数



动手写爬虫

```
lagou_spider.py
1 # -*- coding:utf-8 -*-
2
3 import requests
4 from pymongo import MongoClient
5 import time
6 from fake_useragent import UserAgent
7
8
9 ua = UserAgent()
10
11 client = MongoClient()
12 db=client.test
13 lagou = db.lagou
14
15 headers = {
16     'User-Agent':ua.random ,
17
18 正在爬取第 17 页的数据...
19 正在爬取第 18 页的数据...
20 正在爬取第 19 页的数据...
21 正在爬取第 20 页的数据...
22 正在爬取第 21 页的数据...
23 正在爬取第 22 页的数据...
24 正在爬取第 23 页的数据...
25 正在爬取第 24 页的数据...
26 正在爬取第 25 页的数据...
27 正在爬取第 26 页的数据...
28 正在爬取第 27 页的数据...
29 正在爬取第 28 页的数据...
30 正在爬取第 29 页的数据...
31 正在爬取第 30 页的数据...
[Finished in 98.1s]
```



```
[0]
{
  "id": "59dca636b385c712f8ce9445",
  "companyId": "1561",
  "positionId": "3682270",
  "industryField": "移动互联网, 硬件",
  "education": "本科",
  "workYear": "3-5年",
  "city": "北京",
  "positionAdvant": "人工智能, 黑科技, 大牛多",
  "createTime": "2017-10-10 14:49:37",
  "salary": "20k-40k",
  "positionName": "高级Python开发工程师",
  "companySize": "500-2000人",
  "companyShortN": "Face++",
  "companyLogo": "i/image/M00/8A/62/CgqKkVh1n0eACORfAAQPSvw5sM235.jpg",
  "financeStage": "成熟型(C轮)",
  "jobNature": "全职",
  "approve": "1",
  "companyLabelL": ["科技大牛公司", "自助三餐", "年终多薪", "超长带薪年假"],
  "district": "海淀区",
  "positionLables": ["专家", "Java"],
  "industryLables": [],
  "publisherId": "39658",
  "businessZones": "null",
  "score": "0",
  "companyFullNa": "北京旷视科技有限公司",
  "adWord": "0",
  "imState": "today",
  "lastLogin": "1507628118000",
  "explain": "null",
  "plus": "null",
  "pcShow": "0",
  "appShow": "0",
  "deliver": "0",
  "gradeDescriptio": "null",
  "promotionScore": "null",
  "firstType": "开发/测试/运维类",
  "secondType": "后端开发",
  "isSchoolJob": "0",
  "companySize": "14-49人"
}
```



DC学院
class.pkbigdata.com

造数



完成作业

- 在电脑上安装MongoDB
- 尝试爬取多个职位数据
- 改进爬虫，爬取详情页数据



DC学院
class.pkbigdata.com

造数

爬虫工程师

更多数据科学课程，上DC学院：class.pkbigdata.com



关注 DataCastle



关注造数



造数

