项目文档（作业三）

路线三 11 组
组长：
宋定杰 学号：171250628 手机：18851132226

组员：
姓名：李辰辉 学号：171250645
姓名：梁斌 学号：171830506
姓名：陈维烨 学号：171250599

组员分工

任务 1:
陈维烨：数据爬取、清洗、整理；扩充数据，获取更多数据源；与其他组别进行数据交换
梁斌：数据库搭建；后端数据持久层搭建；数据集成；

任务 2、3:
宋定杰：服务端搭建；知识问答系统搭建
李辰辉：neo4j 数据库搭建；图数据持续集成

项目 git 地址

https://github.com/bbsngg/Data-Integration/

项目地址

图数据展示：http://ismzl.com:3000/
基于图数据的股票知识问答系统：http://ismzl.com:3000/qa.html

数据库地址

Neo4j：http://ismzl.com:7474/browser/
MySQL：http://ismzl.com:3306

数据获取与数据整理

数据源

作业三数据源分为两部分来源：一部分是其他组的数据，另一部分是来自 http://pycs.greedyai.com/的新增数据。由于更细节的基金和债券等信息需要在 tushare 上获取，而获取相应数据需要大量付费积分，因此无法将 tushare 作为新的数据源来获取数据。

获取方式

第一部分中包括另外两组的数据，直接通过联系其他组组长获取。

第二部分数据仍然通过 scrapy 进行爬取，爬取重点变为董事会相关信息，同时由于标签更加不具有特殊性，在定位数据位置时添加了对 xpath 表达式的使用。

具体而言是获取到董事会、监事会和高管面板的 tbody 标签，然后获取其下的所有 tr 元素。观察到每一行 tr 元素至少包含一位管理层成员信息，至多两位，用于存储信息的 td 标签以三个作为一组，因此直接通过索引值对特定的内容进行获取：

对于索引值 x：若 mod(x)=0，则获取字段为人名；若 mod(x)=1，则获取字段为职位；若 mod(x)=2，则获取字段为持有股数。

最后为了使得新的数据能够与原有数据产生关联，在每一行数据开头都加入了 stock_no 信息。

相较于参考 demo 的董事会信息数据，该部分还增量爬取了监事会和高管层的成员数据，也就是在公司高层成员数据方面更加完善，同时增加了数据源有提供的持股信息。对于回调函数的调用与传参处理与作业二相同。

集成流程

1. 将多个数据的数据源导出，为避免各方数据源格式不唯一的情况，我们从数据源爬取下来的数据以及从其他小组获取的数据统一抽取为 csv 格式。

2. 将所有的 csv 文件进行初步过滤，如清除无效字段后导入数据库。

3. 按照我们数据库内定义的数据字段与类型对数据进行二次过滤，将增量获取的数据分为以下几类：
   a) 重复数据：对此类数据进行筛除，抛弃无用字段或只将部分数据添加到已有数据库中，如对于已存在数据库中的实体，扩充数据只有部分字段可供使用，其余则被抛弃（不从 CSV 导入）。
   b) 命名冲突数据：对于某些实体的相同属性在不同表（CSV 文件）中可能存在不同的命名，需要对其进行分析，查找相同的属性，再做数据合并。
   c) 包含新实体类的数据：对于有新的实体类的数据，我们新建了表保存，以保证数据库表和实体的一致性。
4. 遇到困难：信息去重——按照有相同信息的字段进行去重，如股票名有出入但是根据股票代码这一主键可以去重。

关键代码

获取三个面板页面

```
boss_table = response.css("div#ml_001 tbody") #董事会成员
list_table = response.css("div#ml_002 tbody") #监事会成员
high_table = response.css("div#ml_003 tbody") #高管成员

tr_list1 = boss_table.css("tr")
tr_list2 = list_table.css("tr")
tr_list3 = high_table.css("tr")
```

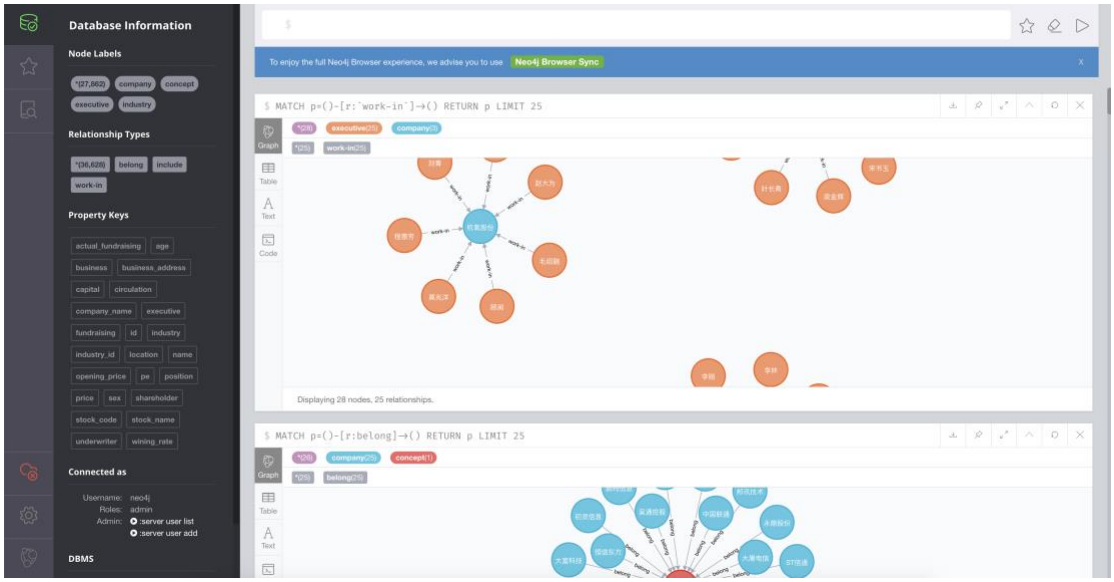
获取字段信息

```
person_1 = td_list[0].css("a::text").get()
position_1 = td_list[1].xpath("./text()").extract()[0]
stocknum_1 = td_list[2].css("span.to_tip::text").get()
if stocknum_1 is None or len(stocknum_1)<=2:
    stocknum_1 = 'NULL'
with open(r'G:\scpy\MyDtScpy\MyDtScpy\data.txt', 'a') as bossFile:
    row = stock_no + '\t' + meeting + '\t' + person_1 + '\t' + position_1 + '\t' + stocknum_1
    bossFile.write(row)
    bossFile.write('\n')
    bossFile.close()
    print("yield one: " + row)
```

数据管理

Neo4j

负责构建和维护知识图谱，便于开展后期应用，如知识问答系统。



MySQL

负责维护数据之间的关系，便于后端对数据进行计算挖掘与前端展示。
股票：

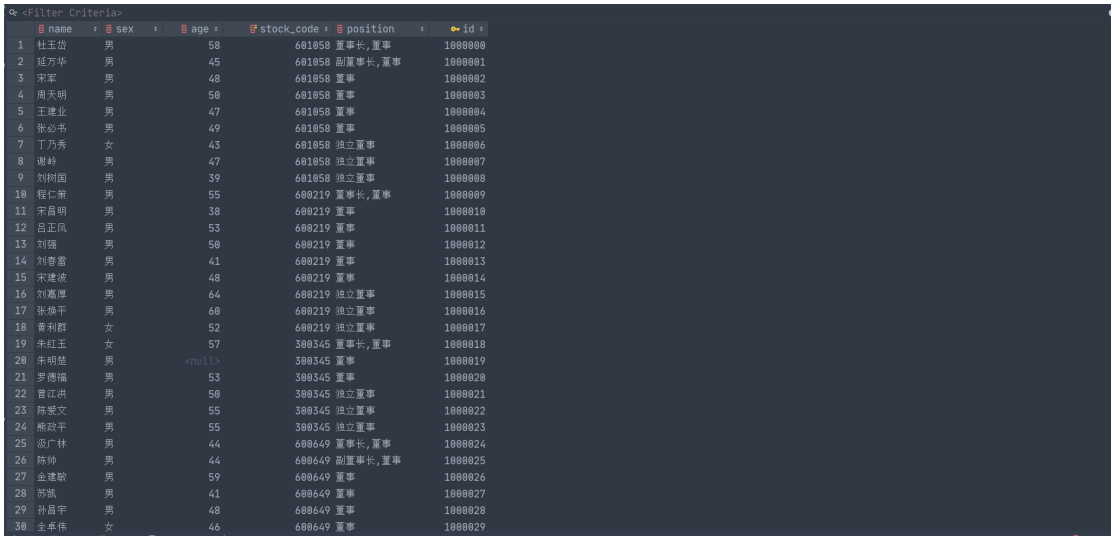
	id	stock_code	company_name	executive	location	industry_id	industry	business	shareholder
1	2877	600082	天津海泰科技发展股份有限	宋克新	天津市	100000032	综合 – 综合	提供租赁和商务服务业。	天津海泰控股集团有限公司
2	2878	600208	新湖中宝股份有限公司	林俊波	浙江省	100000032	房地产 – 房地产	房地产开发和销售	浙江新湖集团股份有限公司
3	2879	596	安徽古井贡酒股份有限公司	梁金辉	安徽省	100000027	食品饮料 – 饮料	白酒的生产与销售、包装材料的生产与销售、农副产品的加工与生产、(	安徽古井集团有限责任公司
4	2880	2430	杭州杭氧股份有限公司	蒋明	浙江省	100000003	机械设备 – 专用设	大中型空分设备、石化设备的生产和销售和工业气体的生产和销售	杭州制氧机集团有限公司
5	2881	603083	上海剑桥科技股份有限公司	Gerald G Wong	上海市	100000028	信息设备 – 通信设	基于合作模式（主要为JDM和ODM模式）进行家庭、企业及工业应用类I	Cambridge Industries Compa
6	2882	600538	北海国发海洋生物产业股份	潘利斌	广西壮族自治区	100000042	医药生物 – 中药	医药制造及医药流通产业、农药产业，酒店和环保	朱蓉娟
7	2883	2654	深圳万润科技股份有限公司	李志江	广东省	100000038	电子 – 光学光电子	从事LED应用与照明产品、LED光电元器件的研发、设计、生产、销售；	李志江、罗小艳、李驰
8	2884	613	海南大东海旅游中心股份有	袁小平	海南省	100000043	餐饮旅游 – 酒店旅	旅游服务业中的住宿和餐饮	罗牛山股份有限公司
9	2885	601616	上海广电电气(集团)股份有	侯松容	上海市	100000029	机械设备 – 电气设	高低压输配电成套设备、各类元器件及零配件的生产销售	新余昊杰投资管理有限公司
10	2886	601011	宝泰隆新材料股份有限公司	焦云	黑龙江省	100000021	采掘 – 煤炭开采	焦炭和化工产品	黑龙江宝泰隆集团有限公司
11	2887	600630	上海龙头(集团)股份有限公	王卫民	上海市	100000014	纺织服装 – 服装鞋	品牌销售与国际贸易	上海纺织(集团)有限公司
12	2888	300288	贵阳朗玛信息技术股份有限	王伟	贵州省	100000036	信息服务 – 计算机	计算机技术及软件开发、销售；计算机硬件及耗材销售；计算机网络互联	王伟
13	2889	600061	国投资本股份有限公司	叶柏寿	上海市	100000002	金融服务 – 证券	提供证券经纪、投资咨询、资产管理及相关证券金融服务。	国家开发投资集团有限公司
14	2890	900919	上海绿庭投资控股集团股份	龙炼	上海市	100000034	金融服务 – 保险业	投资管理和资产管理	绿庭(香港)有限公司
15	2891	300242	广东佳兆业佳云科技股份有	郑毅	广东省	100000038	信息服务 – 传媒	移动互联网营销业务	深圳市一号仓佳速网络有限公司
16	2892	600493	福建凤竹纺织科技股份有限	陈逊清	福建省	100000014	纺织服装 – 纺织制	针织织造、染整加工、漂染筒子色纱、鞋业生产	陈逊清
17	2893	300251	北京光线传媒股份有限公司	王长田	北京市	100000038	信息服务 – 传媒	栏目制作与广告、演艺活动、影视剧的投资及发行业务。	上海光线投资控股有限公司
18	2894	150	宜华健康医疗股份有限公司	陈奕民	广东省	100000015	医药生物 – 医疗服	健康医疗服务	宜华企业(集团)有限公司
19	2895	2475	立讯精密工业股份有限公司	王来春	广东省	100000038	电子 – 电子制造	连接器产品研发、生产和销售	立讯有限公司
20	2896	757	四川德钧机电股份有限公司	颜广彤	四川省	100000040	交运设备 – 汽车零	从事汽车发动机曲轴的研发、制造与销售	天津市德钧机电汽车贸易有限公司
21	2897	2483	江苏润邦重工股份有限公司	吴建	江苏省	100000003	机械设备 – 专用设	物料搬运业务(主要包括起重装备、立体停车设备等产品)以及海工与船	南通威望实业有限公司
22	2898	507	珠海港股份有限公司	欧辉生	广东省	100000008	交通运输 – 港口服	港口及其配套设施的项目投资；电力项目投资；玻璃纤维制品项目投资；	珠海港控股集团有限公司
23	2899	600086	东方金钰股份有限公司	赵宇	湖北省	100000037	轻工制造 – 家用轻	珠宝玉石，贵、铂金及镶嵌饰品的生产、销售	云南兴龙实业有限公司
24	2900	600638	上海新黄浦置业股份有限公	程齐鸣	上海市	100000032	房地产 – 房地产	房地产开发与销售	上海新华闻投资有限公司
25	2901	600292	国家电投集团远达环保股份	郑武生	重庆市	100000026	公用事业 – 环保工	环保、节能和水务	国家电力投资集团有限公司
26	2902	2353	烟台杰瑞石油服务集团股份	孙伟杰	山东省	100000003	机械设备 – 专用设	油田专用设备制造，设备维修改造及配件销售，油田工程技术服务。	孙伟杰、王坤晓、刘贞峰
27	2903	852	中石化石油机械股份有限公	袁建强	湖北省	100000003	机械设备 – 专用设	制造、销售石油钻采设备等	中国石油化工集团公司
28	2904	600500	中化国际(控股)股份有限公	NULL	上海市	100000045	化工 – 化学制品	化工原料、精细化工、农用化工、塑料、橡胶制品等的进出口、内销贸	中国中化股份有限公司
29	2905	2536	河南省西峡汽车水泵股份有	孙耀志	河南省	100000040	交运设备 – 汽车零	汽车零部件及其机械产品的研发、制造、销售。	河南省宛西控股股份有限公司

概念：

<Filter Criteria>

	name	id
1	3D打印	10000030
2	4G概念	10000041
3	5G概念	10000047
4	IPV6概念	10000036
5	IP变现	10000015
6	O2O模式	10000118
7	QFII重仓	10000076
8	ST板块	10000009
9	三沙概念	10000083
10	三网融合	10000147
11	上海本地	10000017
12	上海自贸	10000055
13	业绩预升	10000032
14	业绩预降	10000082
15	东亚自贸	10000072
16	丝绸之路	10000104
17	云计算	10000043
18	互联金融	10000140
19	京津冀	10000095
20	低碳经济	10000067
21	体育概念	10000069
22	保险重仓	10000013
23	保障房	10000102
24	信息安全	10000160
25	信托重仓	10000080
26	充电桩	10000130
27	免疫治疗	10000129
28	养老概念	10000024
29	内贸规划	10000157
30	军工航天	10000021

董事：



	name	sex	age	stock_code	position	id
1	杜玉岱	男	58	601058	董事长,董事	1000000
2	延万华	男	45	601058	副董事长,董事	1000001
3	宋军	男	48	601058	董事	1000002
4	周天明	男	50	601058	董事	1000003
5	王建业	男	47	601058	董事	1000004
6	张必书	男	49	601058	董事	1000005
7	丁乃秀	女	43	601058	独立董事	1000006
8	谢岭	男	47	601058	独立董事	1000007
9	刘树国	男	39	601058	独立董事	1000008
10	程仁策	男	55	600219	董事长,董事	1000009
11	宋昌明	男	38	600219	董事	1000010
12	吕正风	男	53	600219	董事	1000011
13	刘强	男	50	600219	董事	1000012
14	刘春蕾	男	41	600219	董事	1000013
15	宋建波	男	48	600219	董事	1000014
16	刘嘉厚	男	64	600219	独立董事	1000015
17	张焕平	男	60	600219	独立董事	1000016
18	曹利群	女	52	600219	独立董事	1000017
19	朱红玉	女	57	300345	董事长,董事	1000018
20	朱明楚	男	<null>	300345	董事	1000019
21	罗德福	男	53	300345	董事	1000020
22	曾江洪	男	50	300345	独立董事	1000021
23	陈爱文	男	55	300345	独立董事	1000022
24	熊政平	男	55	300345	独立董事	1000023
25	邵广林	男	44	600649	董事长,董事	1000024
26	陈帅	男	44	600649	副董事长,董事	1000025
27	金建敏	男	59	600649	董事	1000026
28	苏凯	男	41	600649	董事	1000027
29	孙昌宇	男	48	600649	董事	1000028
30	全卓伟	女	46	600649	董事	1000029

产业:

	name	id
1	交通运输	100000008
2	仪器仪表	100000007
3	传媒娱乐	100000030
4	供水供气	100000017
5	公路桥梁	100000001
6	其它行业	100000034
7	农林牧渔	100000010
8	农药化肥	100000042
9	化工行业	100000031
10	化纤行业	100000002
11	医疗器械	100000015
12	印刷包装	100000047
13	发电设备	100000018
14	商业百货	100000013
15	塑料制品	100000012
16	家具行业	100000041
17	家电行业	100000048
18	建筑建材	100000011
19	开发区	100000035
20	房地产	100000032
21	摩托车	100000046
22	有色金属	100000016
23	服装鞋类	100000037
24	机械行业	100000003
25	次新股	100000028
26	水泥行业	100000044
27	汽车制造	100000040
28	煤炭行业	100000021
29	物资外贸	100000045
30	环保行业	100000026