数据分析报告——QABasedOnStockKG

路线三 11组：宋定杰 171250628、李辰辉 171250645、梁斌 171830506、陈维烨 17125059

本项目以知识图谱为理论指导对股票数据进行分析，并最终从0到1搭建一个以股票为中心有一定规模的领域知识图谱，并以该知识图谱完成自动问答与分析服务。

项目介绍

本项目基于作业二集成的来自同花顺金融服务网http://pycs.greedyai.com/与tushare网站https://tushare.pro/的数据，以股票为核心，构建了一个包含4类规模为2.2万的知识实体，11类规模约3.6万实体关系的股票知识图谱，并依托于知识图谱数据构建了股票自动问答系统，可支持问答内容包括股票的各类属性信息与有关股票和董事的评估报告。

本项目将包括以下两部分的内容：

1. 基于金融网站数据的股票知识图谱构建
2. 基于股票知识图谱的自动问答

项目最终效果

以下两图是实际问答运行过程中的截图：

点击查看说明

输入想问的问题，点击query即刻回馈；点击clear按钮清除回答框；问答示例请点击 示例

Answer

2020-07-04 18:58:43:
****

股票名：新文化　　股票代码：300336
董事信息：郑培敏；曹军波；孙毅；万如平；王敏；杨震华；盛文蕾
所属产业：传媒娱乐
所属概念：IP变现；送转潜力
所属公司：上海新文化传媒集团股份有限公司
业务范围：影视剧的投资、制作、发行及衍生业务
资产信息：总资产 5.38亿元，每股价格 25.00元，开盘价 28.88元
经销商：海通证券股份有限公司

其他相关信息：
董事所有的股票名：良信电器；东方明珠；麦达数字；博腾股份；新文化；森马服饰
对应资产信息：7.86亿元；26.89亿元；5.79亿元；5.38亿元；31.86亿元；5.44亿元
对应业务范围：终端电器、配电电器、控制电器等三大类低压电器产品的研发、生产和销售；旅游现代服务业、媒体业务；主要从事休闲服及儿童服饰经营；包括企业互联网服务业务和电子制造服务(EMS)；影视剧的投资、制作、发行及衍生业务；医药定制研发生产服务

****

Question

新文化怎么样

Answer

2020-07-04 18:59:21: 产业钢铁行业包含的股票有：鲁银投资；红宇新材；新兴铸管；韶钢松山；海南矿业；马钢股份；安泰科技；杭钢股份；河钢股份；方大特钢；宝钢股份；鸿路钢构；金洲管道；玉龙股份；山东地矿；法 尔 胜；*ST钒钛；承德钒钛；久立特材；山东钢铁；北讯集团；广泽股份；包钢股份；邯郸钢铁；安阳钢铁；八一钢铁；杭萧钢构；大金重工；沙钢股份；三钢闽光；西宁特钢；贵绳股份；东方铁塔；*ST重钢；精工钢构；新钢股份；酒钢宏兴；中原特钢；武钢股份；常宝股份；柳钢股份；创兴资源；大西洋；*ST华菱；富煌钢构；凌钢股份；南钢股份；莱钢股份；大冶特钢；鞍钢股份；新日恒力；抚顺特钢；太钢不锈；光正集团；宏达矿业；恒星科技；金岭矿业；本钢板材；首钢股份；哈尔斯

Question

钢铁行业有哪些

项目运行方式

登入http://ismzl.com:3000/qa.html#，在Question栏目输入想问的问题，点击右下角的query按钮即可得到答案。若不熟悉提问内容，可在上方点击查看示例，提问语句的格式有较大自由度，可按照自身需求使用。

介绍详细方案

一、股票知识图谱构建

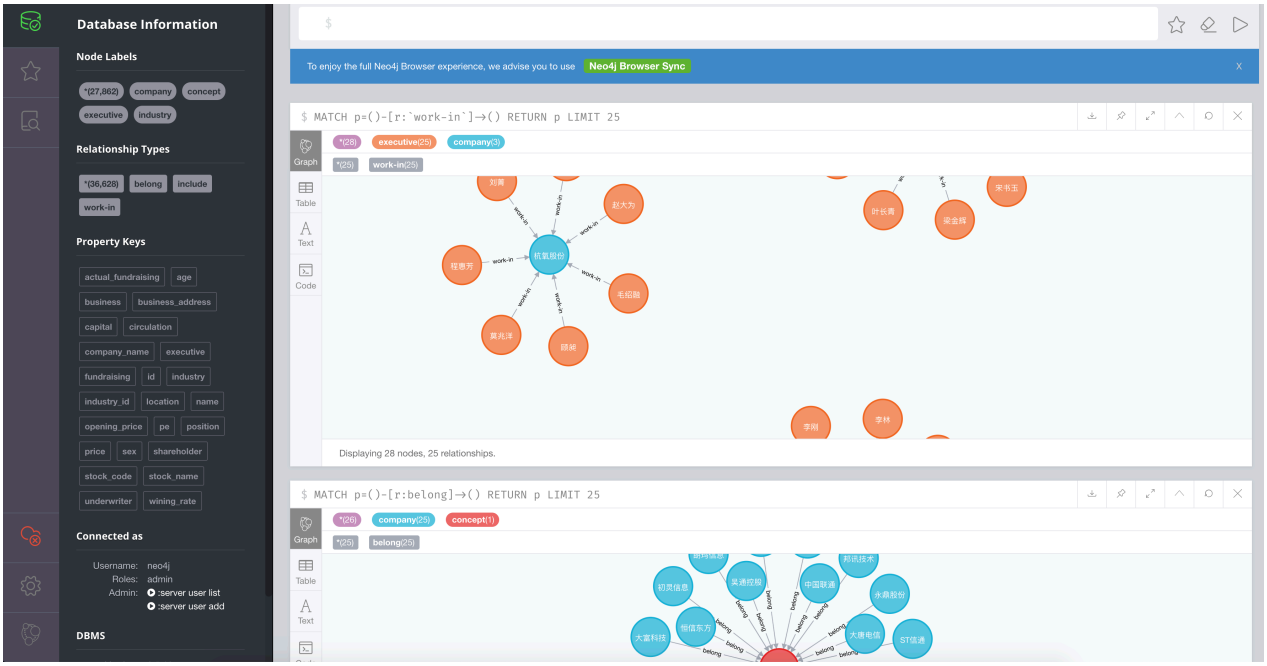
1.1 知识图谱构建脚本目录

爬虫代码文件存放位置：

https://github.com/bbsngg/Data-Integration/tree/master/SecuritiesKnowlegeGraph

1.2 股票领域知识图谱规模

1.2.1 neo4j图数据库存储规模



1.2.2 知识图谱实体类型

实体类型	中文含义	实体数量	举例
Company	股票	2,876	黄山B股
Concept	概念	162	赛马概念
Industry	产业	49	钢铁行业
Executive	董事	24,775	宋杰
Total	总计	27,862	2.7万级

1.2.3 知识图谱实体关系类型

实体关系类型	中文含义	关系数量	举例
belong	属于	8,978	<黄山Ｂ股,属于,赛马概念>
include	包含于	2,875	<黄山Ｂ股,包含于,钢铁行业>
work-in	作为董事于	24,775	<宋杰,作为董事于,黄山Ｂ股>
Total	总计	36,628	3.6万级

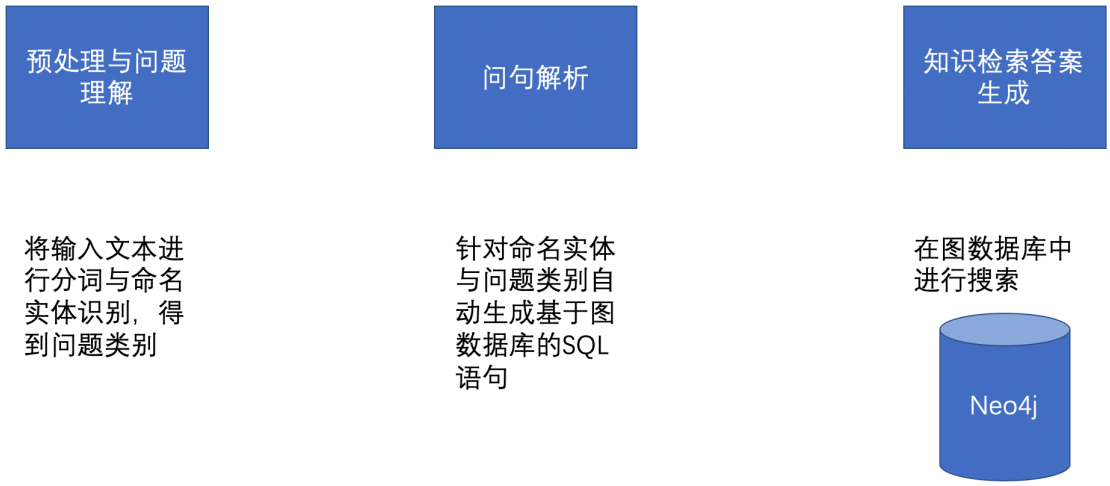
1.2.4 知识图谱属性类型

字段	类型	含义	举例
id	bigint	主键id	主键
stock_code	bigint	股票代码	键
company_name	varchar	公司名称	
executive	varchar	董事长	
location	varchar	公司地址	
industry_id	bigint	行业id	
industry	varchar	行业名称	
business	varchar	主营业务	
shareholder	varchar	股东	
capital	varchar	注册资本	
business_address	varchar	办公地点	
circulation	varchar	发行量	
price	varchar	发行价	
pe	varchar	市盈率	
fundraising	varchar	预计募资	
opening_price	varchar	开盘价	
wining_rate	varchar	中签率	
actual_fundraising	varchar	实际募资	
underwriter	varchar	主承销商	
stock_name	varchar	股票名称	

二、基于股票知识图谱的自动问答

2.1 技术架构

知识问答系统的总架构分为四个大模块：预处理与问题理解、问句解析、知识检索答案生成。在每个大模块里包含算法和处理逻辑。



2.2 脚本结构

2.2.1 问句类型分类脚本(question_classifier.py)

对文本进行分词与命名实体识别，识别出与股票概念相关的关键词，并基于关键词对问题进行20类问题分类。支持问答类别在2.3详细展开。

```
model init finished ......
input an question:钢铁行业有哪些?
{'args': {'钢铁行业': ['industry']}, 'question_types': ['industry_company']}
input an question:
```

2.2.2 问句解析脚本(question_parser.py)

根据分类脚本对问题的分类结果与关键词提取，解析问句，将问句转化为Neo4j图数据库的Cypher语句。

```
[{'question_type': 'company_corp', 'sql': ["MATCH (m:company) where m.stock_name = '黄山B股' return m.stock_name, m.company_name"]}]
```

2.2.3 知识搜索脚本(answer_search.py)

链接服务器中的Neo4j数据库，传入Cypher语句进行知识查询，得到结果。

```
小宋：股票黄山Ｂ股的董事成员包括：孙峻；黄世稳；陶平；章德辉；高舜礼；迟武；郭永清；陈俊；裴斌
股票黄山Ｂ股的主要股东为：黄山旅游集团有限公司
```

2.2.4 问答程序脚本(chatbot_graph.py)

负责启动服务器中的问答程序。

```
model init finished ......
用户:黄山Ｂ股的股东是谁
{'args': {'黄山Ｂ股': ['company']}, 'question_types': ['company_executive', 'company_shareholder']}
[{'question_type': 'company_executive', 'sql': ["MATCH (m:executive)-[r:`work-in`]->(n:company) where
[{'n.stock_name': '黄山Ｂ股', 'm.name': '郭永清'}, {'n.stock_name': '黄山Ｂ股', 'm.name': '陈俊'}, {'n.st
[{'m.stock_name': '黄山Ｂ股', 'm.shareholder': '黄山旅游集团有限公司'}]
小宋：股票黄山Ｂ股的董事成员包括：孙峻；黄世稳；陶平；章德辉；高舜礼；迟武；郭永清；陈俊；裴斌
股票黄山Ｂ股的主要股东为：黄山旅游集团有限公司
```

2.3 支持问答类型

问句类型	中文含义	问句举例
company_concept	股票概念	古井贡酒的概念有哪些?
concept_company	已知概念找股票	赛马概念有哪些?
company_industry	股票产业	黄山Ｂ股的所属产业是什么？ 古井贡酒的所属产业是什么?
industry_company	已知产业找股票	钢铁行业有哪些?
company_executive	股票董事	黄山Ｂ股的董事有哪些？ 古井贡酒的董事有哪些?
executive_company	已知董事找股票	宋杰持有哪些股票?
company_corp	股票公司	黄山Ｂ股的公司是什么?
company_province	股票省份	黄山Ｂ股的所在地?
company_code	股票代码	黄山Ｂ股的代码是多少?
company_business	股票公司服务	黄山Ｂ股的业务内容?
company_shareholder	股票股东	黄山Ｂ股的股东是谁?
company_capital	股票资产	黄山Ｂ股的资产有多少?
company_price	股票单价	黄山Ｂ股的单价是多少?
company_openprice	股票开盘价	黄山Ｂ股的开盘价是多少?
company_underwriter	股票承销商	黄山Ｂ股的承销商是谁?
company_risk	股票风险评估	黄山Ｂ股怎么样?
executive_risk	董事风险评估	评估一下宋杰风险?

问答结果展示

用户:古井贡酒的概念有哪些?
小宋： 古井贡酒所属的概念包括：融资融券；含B股

****
用户:赛马概念有哪些?

小宋：概念赛马概念包含的股票有：平潭发展；新潮能源；中弘股份；中牧股份；栖霞建设；亚通股份；海航创新；日发精机；华联综超；罗 牛 山；珠江实业；武汉控股；海澜之家；新 华 都

****

用户:古井贡酒的所属产业是什么？？

小宋：古井贡酒所属的产业为：酿酒行业

****

用户:钢铁行业有哪些?

小宋：产业钢铁行业包含的股票有：玉龙股份；*ST钒钛；*ST重钢；马钢股份；酒钢宏兴；久立特材；本钢板材；杭萧钢构；海南矿业；西宁特钢；安阳钢铁；精工钢构；哈尔斯；北讯集团；鞍钢股份；
沙钢股份；贵绳股份；邯郸钢铁；莱钢股份；太钢不锈；恒星科技；安泰科技；武钢股份；包钢股份；新兴铸管；*ST华菱；山东地矿；南钢股份；凌钢股份；山东钢铁；鸿路钢构；八一钢铁；韶钢松山；
中原特钢；宏达矿业；光正集团；金岭矿业；法 尔 胜；新日恒力；方大特钢；三钢闽光；抚顺特钢；柳钢股份；富煌钢构；首钢股份；东方铁塔；大西洋；创兴资源；承德钒钛；大金重工；广泽股份；
新钢股份；鲁银投资；红宇新材；河钢股份；宝钢股份；常宝股份；金洲管道；大冶特钢；杭钢股份

****

用户:古井贡酒的董事有哪些?

小宋：股票古井贡酒的董事成员包括：梁金辉；宋书玉；李培辉；叶长青；闫立军；周庆伍；杜杰；许鹏；王高

****

用户:宋杰持有哪些股票?

小宋：宋杰作为董事成员的股票有：京东方Ｂ；京东方Ａ

****

用户:黄山Ｂ股的公司是什么?

小宋：股票黄山Ｂ股的公司为：黄山旅游发展股份有限公司

****

用户:黄山Ｂ股的所在地?

小宋：股票黄山Ｂ股所在的省份为：安徽省

****

用户:黄山Ｂ股的代码是多少?

小宋：股票黄山Ｂ股的股票代码为：900942

****

用户:黄山Ｂ股的业务内容?

小宋：股票黄山Ｂ股的主要业务内容为：酒店业务、索道业务、园林开发业务、旅游服务业务、商品房销售

****

用户:黄山Ｂ股的股东是谁?

小宋：股票黄山Ｂ股的董事成员包括：郭永清；黄世稳；迟武；陈俊；孙峻；裴斌；陶平；章德辉；高舜礼

股票黄山Ｂ股的主要股东为：黄山旅游集团有限公司

****

用户:黄山Ｂ股的资产有多少?

小宋：股票黄山Ｂ股的资产有：7.47亿元

****

用户:黄山Ｂ股的单价是多少?

小宋：股票黄山Ｂ股的每股价格为：0.26元

****

用户:黄山Ｂ股的开盘价是多少?

小宋：股票黄山Ｂ股的开盘价为：0.29元

****

用户:黄山Ｂ股的承销商是谁?

小宋：股票黄山Ｂ股的经销商为：国元证券股份有限公司

****

用户:我可以买什么股票?

小宋：您好，我是小宋金融智能助理，希望可以帮到您。如果没答上来，可联系1025149919@qq.com。祝您赚到大钱!

****

用户:宋杰怎么样

小宋：

董事姓名：宋杰　　年龄：51　　性别：男

股票信息：股票名 京东方Ａ　股票代码 725

所属产业：电子器件

所属概念：融资融券；含B股

所属公司：京东方科技集团股份有限公司

业务范围：显示器件业务、智慧系统产品业务、智慧健康服务业务

资产信息：总资产 347.98亿元，每股价格 16.80元，开盘价 19.32元

经销商：南方证券有限公司

其他相关信息:
董事所有的股票名:中国铁建;云南白药;高鸿股份;华夏银行;爱施德;中国联通;京东方A;京东方B
对应资产信息:310.28亿元;9.08亿元;10.41亿元;347.98亿元;12.39亿元;135.8亿元;128.23亿元
对应业务范围:吸收存款;发放贷款;办理国内外结算;发行金融债券;同业拆借等业务;建筑工程承包、勘察设计及咨询、物流贸易、工业制造、房地产开发;在中国境内提供移动和固网语音及相关增值服务、宽带及其他互联网相关服务、信息通信技术服务以及商务及数据通信服务GSM移动语音、WCDMA移动语音、TD-LTE/LTEFDD移动语音,以及相关增值服务以下称为"移动业务"。除移动业务外,上述其他业务以下统称为"固网业务"。;国内商业,移动通讯、电子产品及相关配套产品的购销与代理,供应链管理等;显示器件业务、智慧系统产品业务、智慧健康服务业务;制造和销售以云南白药系列产品和天然植物药系列产品为主的各类药品。;企业信息化业务、信息服务业务、IT销售业务

****
用户:新文化值得投资吗
小宋:
股票名:新文化 股票代码:300336
董事信息:孙毅;万如平;曹军波;杨震华;王敏;郑培敏;盛文蕾
所属产业:传媒娱乐
所属概念:送转潜力;IP变现
所属公司:上海新文化传媒集团股份有限公司
业务范围:影视剧的投资、制作、发行及衍生业务
资产信息:总资产 5.38亿元,每股价格 25.00元,开盘价 28.88元
经销商:海通证券股份有限公司

其他相关信息:
董事所有的股票名:新文化;良信电器;博腾股份;麦达数字;森马服饰;东方明珠
对应资产信息:26.89亿元;5.44亿元;5.79亿元;5.38亿元;31.86亿元;7.86亿元
对应业务范围:主要从事休闲服及儿童服饰经营;包括企业互联网服务业务和电子制造服务(EMS);医药定制研发生产服务;终端电器、配电电器、控制电器等三大类低压电器产品的研发、生产和销售;影视剧的投资、制作、发行及衍生业务;旅游现代服务业、媒体业务

****

总结

1. 本项目基于同花顺金融服务网http://pycs.greedyai.com/与tushare网站的数据,并增量集成了来自其他项目组的真实数据,并以股票为核心,构建了一个包含4类规模为2.2万的知识实体,11类规模约3.6万实体关系的股票知识图谱,并依托于知识图谱数据构建了股票自动问答系统。
2. 本项目以neo4j作为存储,经过实际测试,数据导入时间可在2小时内完成,问答系统启动可在5秒内完成,超过90%的问题可在2秒内获得结果,可以较好地满足用户对问答系统的需求。
3. 本项目可以以脚本的形式快速运行,可快速部署到多端应用,如本例可以较容易地部署到HTTP网络环境中(见chatbot_server.py文件)。