提交截止： **2018年7月7日，11：59 PM**。对于之后提交的作业，不提供任何的反馈，并不会加入到课程总分当中。对于按时提交的项目，我们会在 **10个工作日** 内给与反馈。

项目目标： 利用网络上公开的数据来搭建一个小型的证券知识图谱

数据源： 本项目需要用到两种数据源：一种是公司董事的信息，另一种是股票的行业以及概念信息。

- 公司董事的信息： 这部分数据包含在"exe_member.zip"压缩文件中，里面的每一个文件是以“XXXXXX.html”命名，其中XXXXXX是股票代码。这部分数据是由同花顺的网页爬取而来的。比如对于“ 600007.html”，这部分内容来自于 http://stockpage.10jqka.com.cn/600007/company/#manager

-股票行业以及概念信息：这部分信息也可以通过网上公开的信息得到。在这里，我们使用Tushare工具来获得，详细细节见之后具体的任务部分。

任务1： 从网页中抽取董事会的信息。（20%）

在我们给定的html文件中，需要对每一个股票/公司抽取董事会成员的信息，这部分信息包括董事会成员“**名字**”、“**职务**”、“**性别**”、“**年龄**” 共四个字段。首先，姓名和职务的字段来自于：

在这里总共有12位董事成员的信息，都需要抽取出来。 另外，性别和年龄字段也可以从下附图里抽取出来：

最后，生成一个 **“executive_prep.csv”**文件，格式如下：

高管姓名、性别、年龄、股票代码、职位

朴明志， 男， 51， 60007， 董事长、董事

高燕， 女， 60， 600007， 执行董事

刘永政， 男， 50， 600008， 董事长、 董事

.....

任务2： 获取股票行业和概念的信息。（10%）

对于这部分信息，我们可以利用Tushare工具来获取，官网为 http://tushare.org/index.html，并可以从官网下载Tushare工具包。下载完之后，在python里即可以调用股票行业和概念信息。

通过以下的代码即可以获得股票行业信息，并把返回的信息直接存储在**“stock_industry_prep.csv”**文件里。http://tushare.org/classifying.html#id2

```
import tushare as ts
df = ts.get_industry_classified()
// TODO 保存成"stock_industry_prep.csv"
```

类似的，可以通过以下的代码即可以获得股票概念信息，并把它们存储在 **“stock_concept_prep.csv”**文件里。

```
df = ts.get_concept_classified()
//  TODO 保存成“stock_concept_prep.csv”
```

任务3： 设计知识图谱（20%）

设计一个这样的图谱：

- 创建“人”实体，这个人拥有姓名、年龄、姓名

- 创建“公司”实体，除了股票代码，还有股票名称

- 创建“概念”实体，每个概念都有概念名

- 创建“行业”实体，每个行业都有行业名

- 给“公司”实体添加“ST”的标记，这个由LABEL来实现

- 创建“人”和“公司”的关系，这个关系有董事长、执行董事等等

- 创建“公司”和“概念”的关系

- 创建“公司”和“行业”的关系

把设计图存储为“**design.png**”文件。 注：实体名字和关系名字需要易懂，对于上述的要求，并不一定存在唯一的设计，只要能够覆盖上面这些要求即可以。“ST"标记是用来刻画一个股票严重亏损的状态，这个可以从给定的股票名字前缀来判断。

任务4： 创建可以导入Neo4j的.csv文件（20%）

在前两个任务里，我们已经分别生成了 “executive_prep.csv”, "stock_industry_prep.csv", "stock_concept_prep.csv"，但这个文件不能直接导入到Neo4j数据库。所以需要做一些处理，并生成能够直接导入Neo4j的.csv格式。

我们需要生成这几个文件： **“executive.csv”, "stock.csv", "concept.csv", "industry.csv", "executive_stock.csv", "stock_industry.csv", "stock_concept.csv"。**

对于格式的要求，请参考：https://neo4j.com/docs/operations-manual/current/tutorial/import-tool/

任务5： 利用上面的csv文件生成数据库（0%）

bin/neo4j-admin import --nodes executive.csv --nodes stock.csv --nodes concept.csv --nodes industry.csv --relationships executive_stock.csv --relationships stock_industry.csv --relationships stock_concept.csv

这个命令会把所有的数据导入到Neo4j中，数据默认存放在 graph.db 文件夹里。如果graph.db文件夹之前已经有数据存在，则可以选择先删除再执行命令。

把Neo4j服务重启之后，就可以通过 localhost:7474 观察到知识图谱了。

任务6： 利用已经构建好的知识图谱，并通过编写cypher语句回答以下几个问题。（20%）

(1) 有多少个公司目前是属于 “ST”类型的?

(2) “600519” 公司的所有独立董事人员中，有多少人同时也担任别的公司的独立董事职位？

(3) 有多少公司既属于环保行业，又有外资背景？

(4) 对于有锂电池概念的所有公司，独立董事中女性人员比例是多少?

请提供对应的cypher语句以及答案，并把结果写在 “**result.txt**”

任务7： 构建人的实体时，需要考虑重名情况，那这个问题具体怎么解决？（10%）

把解决思路简单写在**“result.txt”**文件中。

作业提交方式：把任务1到任务7的所有的结果放在一个文件夹里，把文件夹命名为 “[用户ID]_项目1.zip", 并把文档提交。里面需要包含的文件有：“executive_prep.csv”，“stock_industry_prep.csv”，“stock_concept_prep.csv”，“executive.csv”, "stock.csv", "concept.csv", "industry.csv", "executive_stock.csv", "stock_industry.csv", "stock_concept.csv", "result.txt", "design.png"

把[用户ID]_项目1.zip 文件发送到 submit@greedyai.com 邮箱，设置邮箱的标题为： 项目1_用户ID