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

项目文档（作业二）

# 组员分工

陈维烨：数据爬取，清洗，整理

梁斌：数据库搭建，后端数据持久层搭建，数据集成

宋定杰：服务端搭建

李辰辉：前端搭建

# 项目git地址

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

# 数据获取与数据整理

## 数据源

<http://pycs.greedyai.com/>：股票的公司基本信息以及发行信息数据源

<tushare库>：股票行业及概念信息数据源

## 获取方式

通过Scrapy框架爬取股票的公司基本信息以及发行信息，通过直接调用tushare库的API获取股票行业及概念信息。

Scrapy爬虫部分，首先在首页 http://pycs.greedyai.com/ 获取所提供的所有股票代码信息所在的网页链接，并将链接存储于文本文件中。

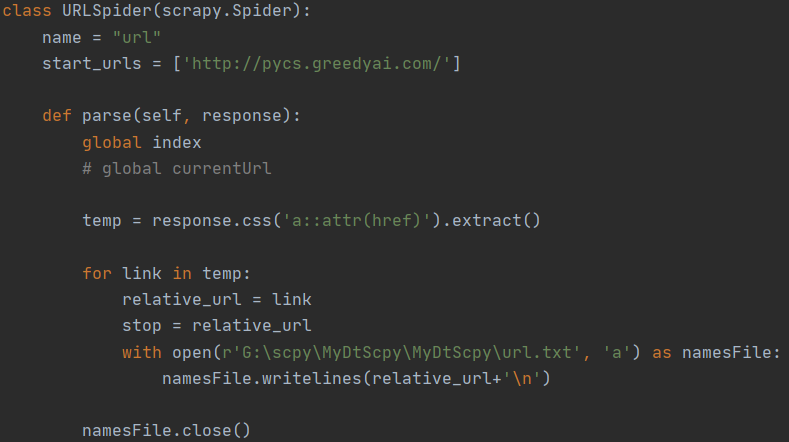
爬取具体数据之前，先从第一部分获取的链接文本文件中读取所有的url，存放于临时数组中，在调用回调函数需要传入下一个爬取的页面链接时，就可以通过全局索引，从临时数组中获取下一个爬取页面的url并传递给回调函数。

在股票信息页面，观察到大部分的信息都以表格形式展现，因此可以获取页面所有的td标签，通过遍历判断的形式获取所需的信息。对于所有获取到的信息，为了保证一只股票相关所有字段信息均位于一行内，需要进行去空白格式化处理。

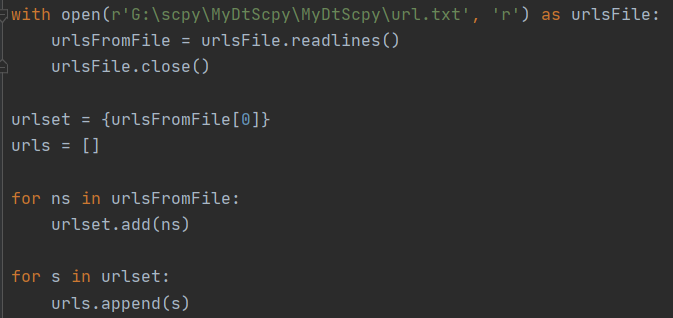
最后，由于部分股票的部分信息在页面上显示为空或"-"，因此在最终导入数据库前进行格式化预处理，将所有无效信息字段值替换为字符串"NULL"，至此完成数据获取。

## 关键代码

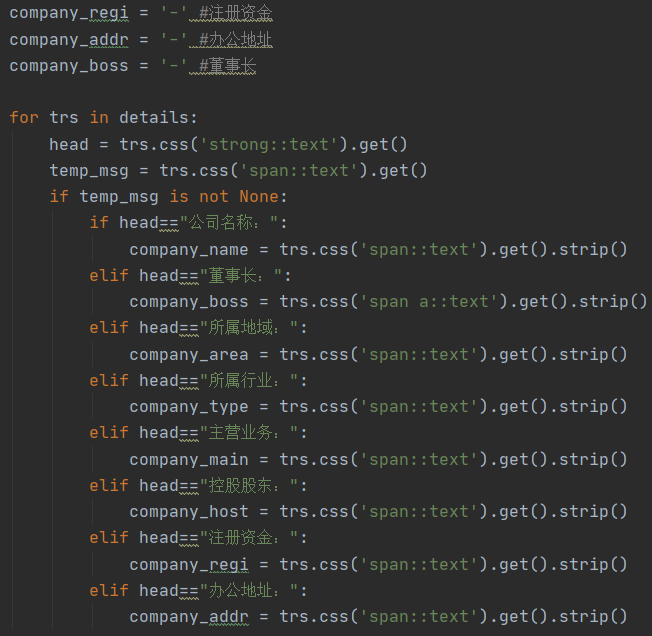
爬取所有股票信息所在页面的url



读取所有url至临时数组



爬取所需股票字段（遍历td标签）





获取下一个爬取页面url并调用回调函数

