实验报告

姓名：张港港 学号：2017212011

## 背景

通过爬取1000家上市公司近五年股票交易数据，并分析同时期的股票公告，对不同公司在交易数据方面同涨同跌等相关性进行聚类分析，对公司公告等文本内容进行分析，研究其与公司近期交易情况的关系。

## 思路

在对不同公司在交易数据方面同涨同跌等相关性进行聚类分析时，我们需要找到两个变量特征，从而可以对这变量特征进行聚类，我选择了每日的涨跌幅和换手率来进行聚类，在此之前我首先对这两个特征进行标准化处理，然后在进行聚类分析

## 实验部分

### 数据与文字爬取

#### 准备库

import requests 爬取网页

from lxml import etree 解析网页

import re 正则提取

from sklearn.cluster import k\_means,KMeans 导入数据聚类k\_means算法

import matplotlib.pyplot as plt

import pandas as pd 导入pandas库，数据读取，预处理

import numpy as np 导入numpy库，数据预处理

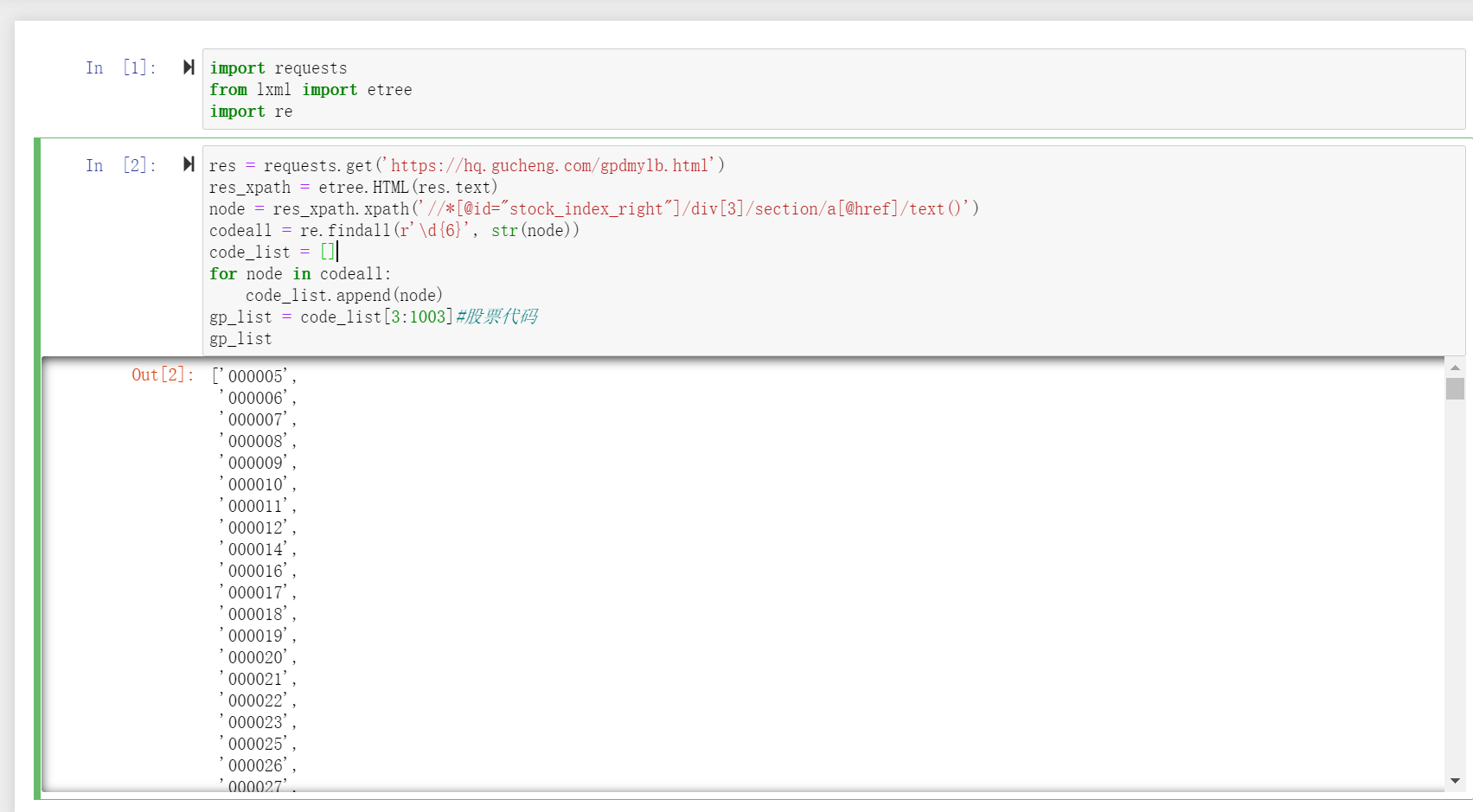
import os 读取文件夹下的子文件

from bs4 import BeautifulSoup 正则提取

import urllib.request 爬取网页

from pandas.core.frame import DataFrame 导入数据框

#### 爬取1000家公司的的股票代码



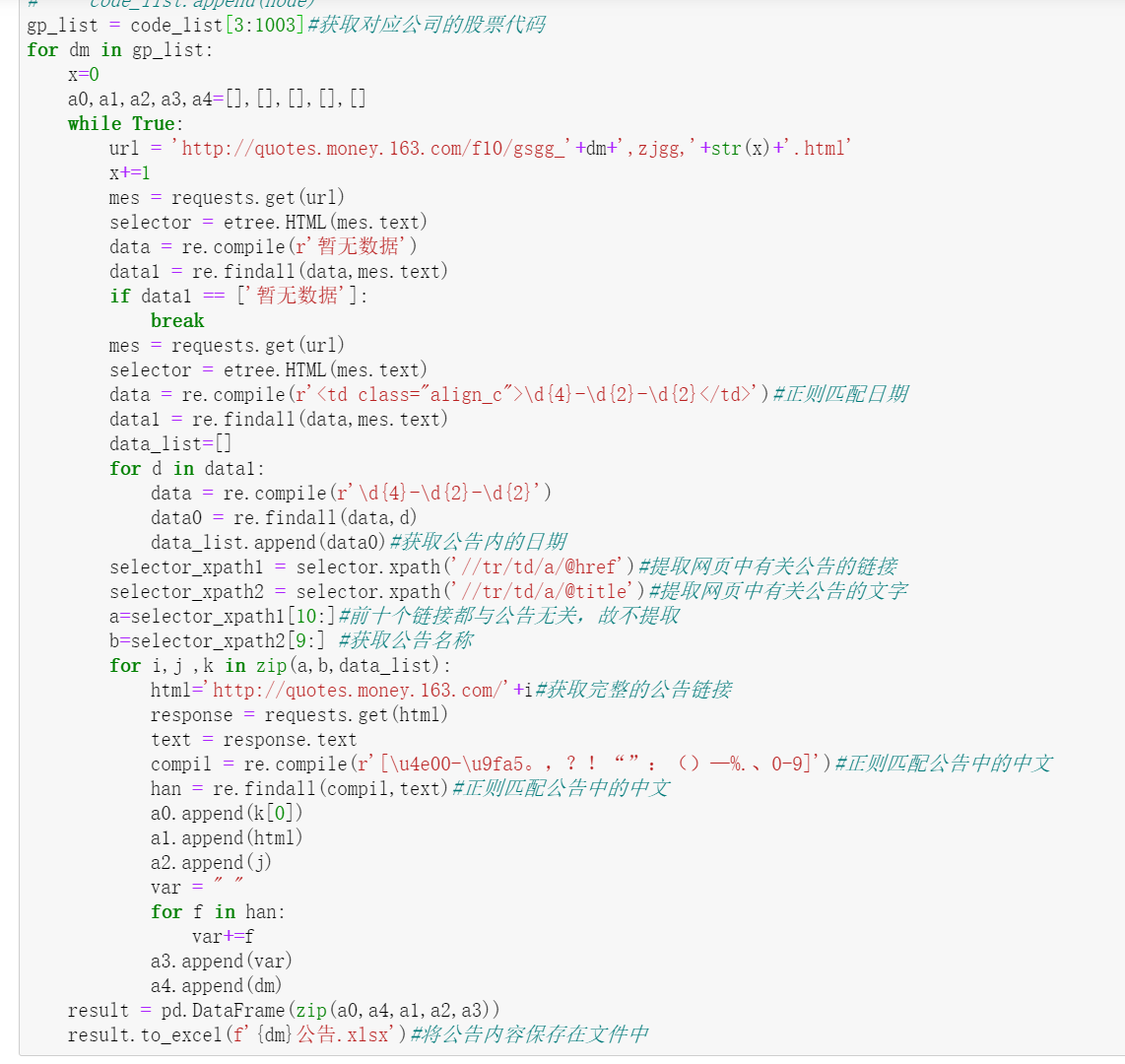
#### 爬取1000家公司过去五年的交易数据

通过网易财经接口可以获取股票的历史数据，我只需要设置起始日期和终止日期就可以，另外，通过观察下载链接，我发现以6和3开头的股票代码code前都有0，以0开头的股票代码code前都有1，由于我爬取的股票都是0开头的，所以不考虑前一种情况。



#### 爬取公告新闻

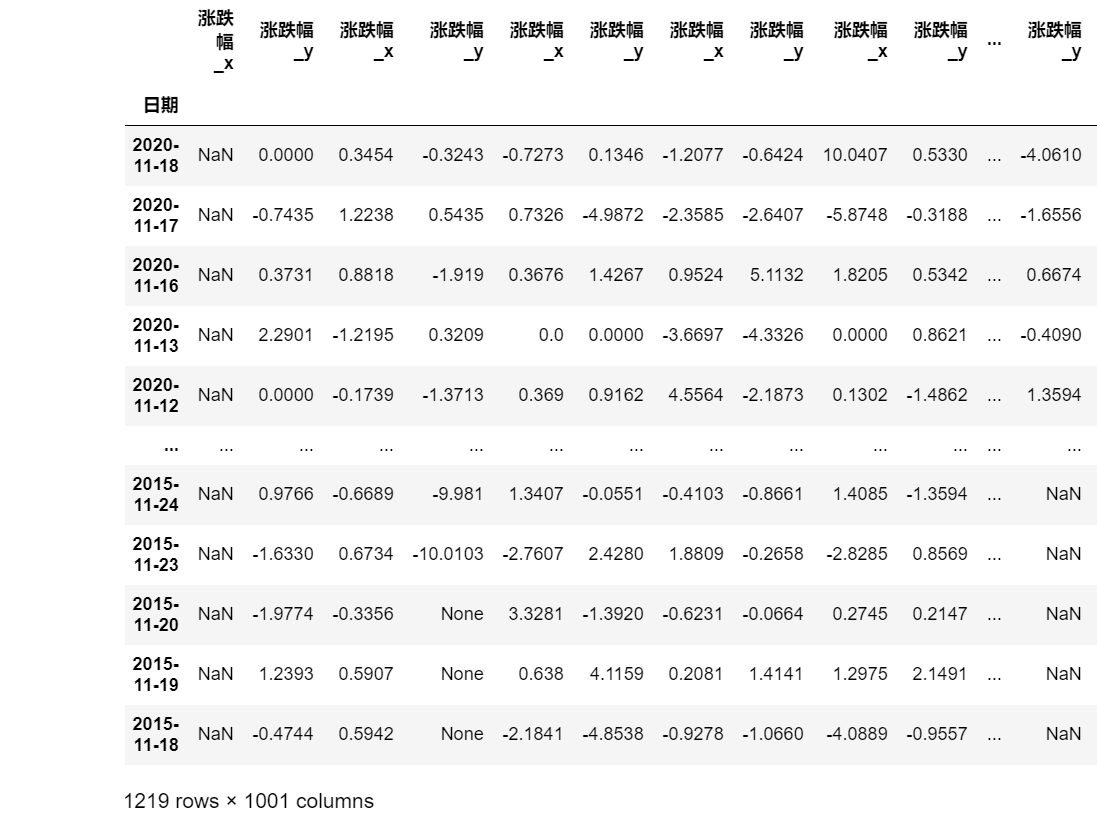
通过上面获取的股票代码列表，对对应的股票公告页面进行解析，用正则表达式和xpath，我提取出了公告发布的日期、股票代码、对应的链接、公告名称和公告内容。



#### 数据切割

导入下载的股票交易数据，将日期作为索引，挑选出表示涨跌幅的那一列，此时，数据框每一列分别表示对应股票在过去五年每天的涨跌幅





将得到的数据框添加列名，然后下载到文件夹，可以查看文件名为：预处理后股票五年间每天的涨跌幅.xlsx



**同样的步骤我们可以得到预处理后股票五年间每天的换手率**

#### 数据标准化和预处理

导入之前切割下载好的数据，并分别做标准化处理



由于有的公司在爬取的时间来退市或者上市时间不满五年，所以会存在缺失值，接下来删除含有缺失值的所有列，然后行列转置，设置列名。



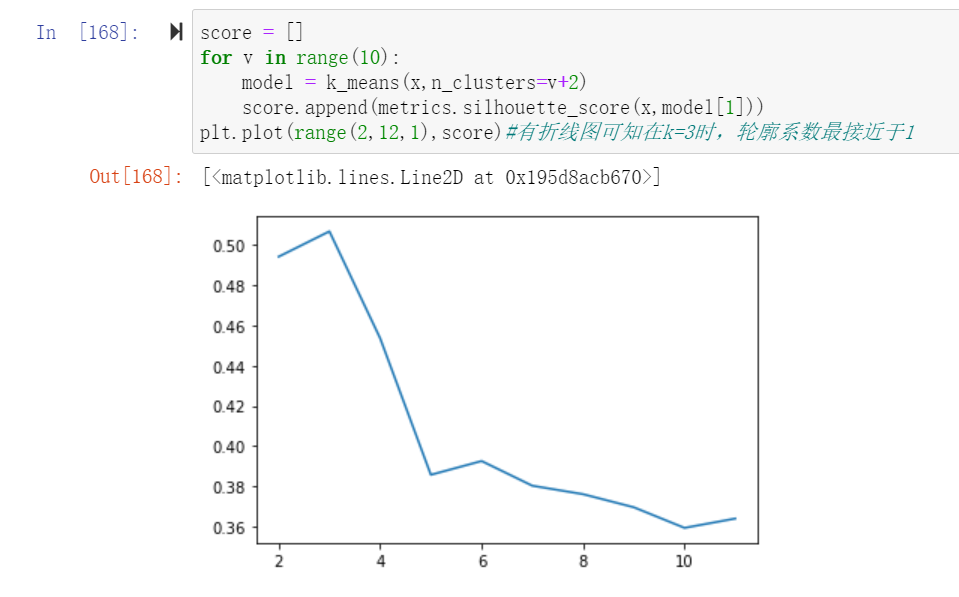
### 数据分析

#### 数据聚类与可视化

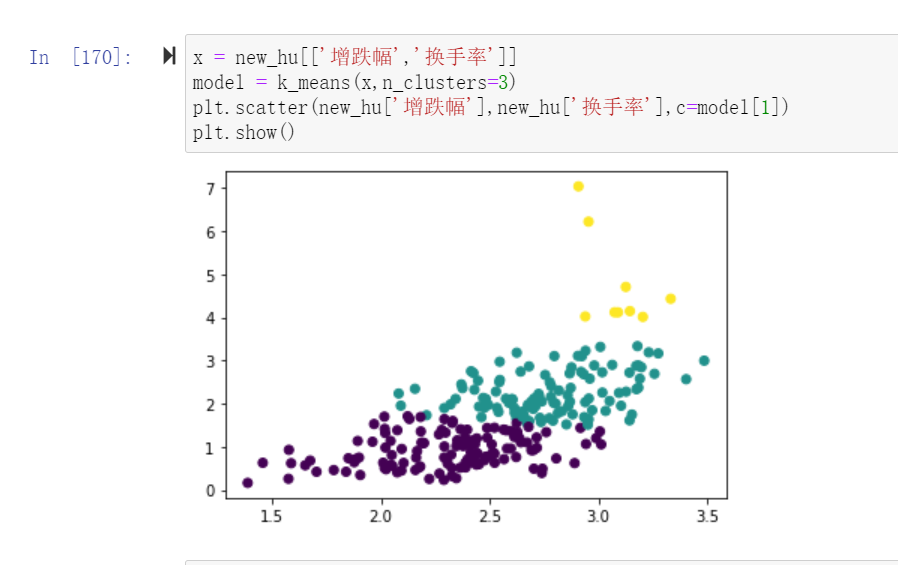
##### 查看两个特征的散点图，初步估计聚类的k值

##### 根据轮廓系数，确定k值

有折线图我们可以看出当k=3时，轮廓系数最接近于1，再结合上面的折线图，我们可以确定k值为3

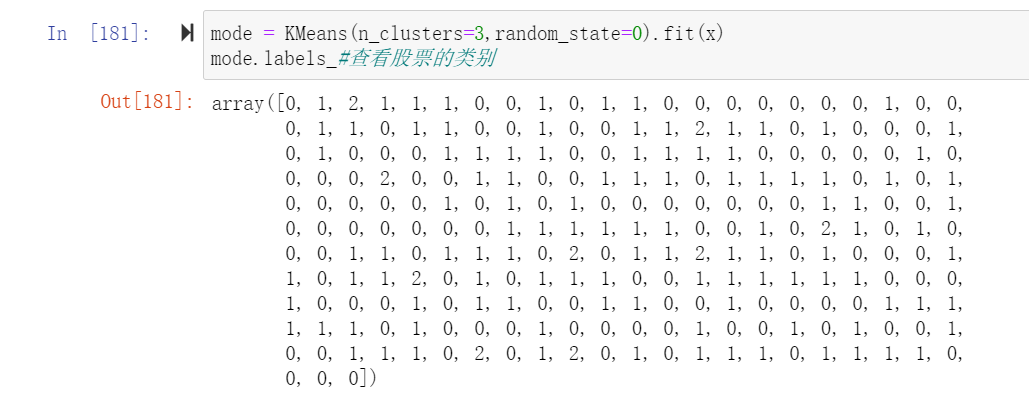


##### 数据聚类与可视化



##### 对股票类别进行分类并下载

根据前面的k值，我们可以知道股票被分为3类，我们可以查看股票的类别



根据股票类别，查看其对应的代码，最后确定股票名称和所属的行业，最后的输出结果可以看文件：股票聚类后的类别结果.docx





结果分析：

通过最后的输出类别的结果，我们可以看到在3个类别中的股票，所处的行业不存在大的差异性，每一个类别都含有各种行业，，这也从侧面反映了股市大盘具有趋同性，在股市大盘上涨下跌中个股几乎也跟随上涨或者下跌。

## 总结

### 遇到的问题：

1. Pandas读取文件时，excel里面股票数字代码前面的0自动缺失

解决方法：在读取文件时加上converters函数，将数据转换为字符串converters={列名：str}

1. 正则表达式运用不够娴熟，导致无法从网页页面中匹配出需要的数据

解决方法：查资料、剔除不需要的数据（就比如在爬取股票的公告时，正则匹配出的链接有的不是公告链接，但是每一次爬取前九条链接都不是，索引我只挑选第九条之后的链接）

1. 由于第一次爬取数据没有把后面数据分析时需要的数据都爬全，导致后来又多次重复爬取，再加上代码不够简洁，循环中又嵌套着循环，导致爬取时间过长。
2. 爬取股票公告时，由于给出的网站历史公告数据不完整，导致爬取的公告数据不够代表股票五年的波动。

### 收获：

1. 将课堂上学的知识真正的运用于实践中，从中明白了自己存在的不足，对python基础语法不够熟练，导致写的代码不够简洁，接下来需要补短板。
2. 熟练掌握了数据爬取、分析的过程，了解了常用的库和语法
3. 在数据分析过程中，了解了常用的数据聚类算法，对数据的预处理和。
4. 对股票的交易数据有了新的理解，我们可以从历史的交易数据中预测未来
5. 在数据分析中，首先需要有一个好的思路，明白后面的数据分析需要什么，这样才能在刚开始爬取原始数据时知道要爬取什么数据。