**爬虫数据 作业报告**

小组成员：胡雨缦 201720210659

郑雨桐 201720210716

邓蔚垚 201720210662

我们小组选择了国内著名的财经网站——东方财富网爬取国内A股市场全部个股在2018年01月08日的资金流向数据，包括当日资金流向、3日资金流向、5日资金流向及10日资金流向，共四份数据。每份数据包含的字段为：code、name、price、change、jinge以及jingzhanbi。每份数据有3350条数据，四份数据共有13400条数据。

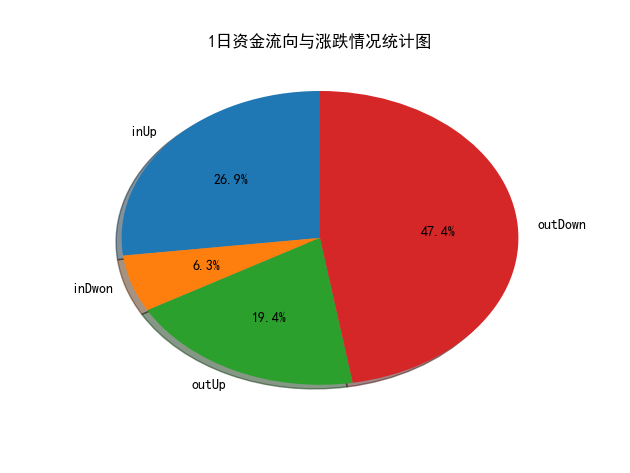
**爬虫部分：**通过urllib2库中的requests库首先爬取数据，然后通过re正则库来搜索可用数据，并以pandas的dataframe格式保存下来，最后将数据写入csv。

遇到的问题主要出现在网站的反爬上，小组此前尝试爬取的数据是豆瓣电影模块，遭遇了豆瓣的反爬，通过改headers的方式也不能完全解决问题，由于数据量偏大，所以我们重新选择了东方财富网的数据来爬取分析。

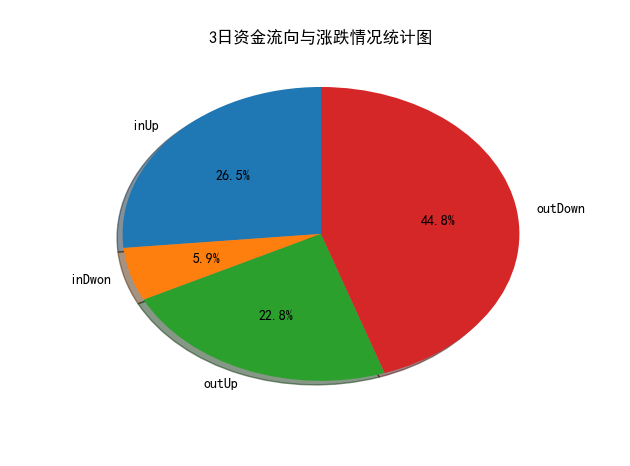
**生成可用数据库部分：**首先下载数据库MYSQL并安装。其次pip pymysql 用来连接python和数据库。接下来，将爬下来的股票数据导入数据库。用python执行数据库：创建数据库、创建数据表、用for循环依次对每个数据文件进行存储，再套嵌一个for循环迭代读取表中每行数据，依次存储。数据库部分总体安装和处理都比较顺利，并没有出现过多的问题。

**爬取数据的意义：**股票涨跌与资金流入流出关系是非常大的，在目前国内的股票市场，基本上仍然属于资金推动型的市场。资金的流入流出其实反映的就是多空双方供求关系的强弱。比如一支股票有资金流入，而且在接下来的一段时间又有资金流入加速，那往往这支股票会有一个比较理想的涨幅，反之，那下跌的概率将会比较大。为了进一步研究资金流向和股票涨幅的之间的关系，对爬取的数据进行如下分析：每份数据首先去掉缺失数据，然后将所有股票分为四类：资金流入且上涨的股票、资金流入且下跌的股票、资金流出且上涨的股票以及资金流出且下跌的股票，最后统计资金流入的股票上涨和下跌股票分别的占比一斤资金流出的股票上涨和下跌股票分别的占比。

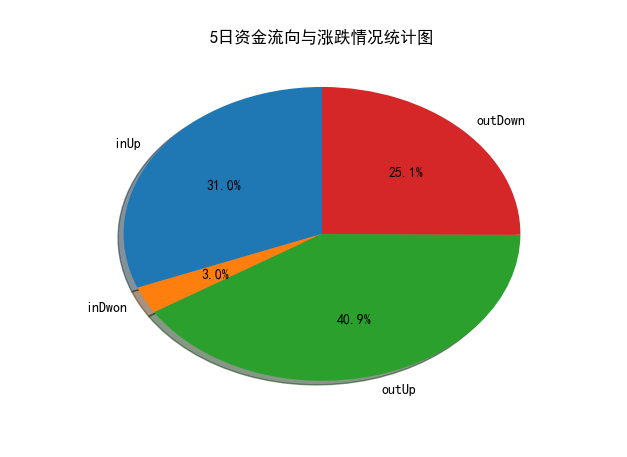
根据上述思路编写Python程序处理数据后得到如下四张饼状图，分别如下图1-图4所示。初步观察这四张图，资金流入时股票多数是上涨的，资金流出时股票多数是下跌的，基本符合我们开始的设想的结论：资金流入，股票易上涨；资金流出，股票易下跌。



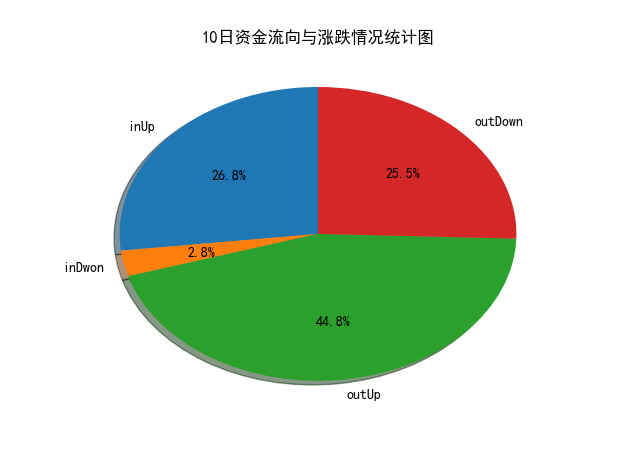
**图1 1日资金流向与股票涨跌统计**



**图2 3日资金流向与股票涨跌统计**



**图3 5日资金流向与股票涨跌统计**



**图4 10日资金流向与股票涨跌统计**

为了进一步分析1日、3日、5日以及10日资金流向哪一时间段内的资金流向与涨跌情况更符合上述规律，我们对四种时间段内的情况分别进行了以下如1日资金流向与涨跌情况统计的计算。

1日资金流向与涨跌情况统计：

1. 资金流入且上涨的股票占资金流入股票的百分比为：100\*26.9/(26.9+6.3) = 81.02%;
2. 资金流入且下跌的股票占资金流入股票的百分比为：100\*（1-0.8102） = 18.98%；
3. 资金流出且上涨的股票占资金流出股票的百分比为：100\*19.4/(19.4+47.4) = 29.04%;
4. 资金流出且下跌的股票占资金流出股票的百分比为：100\*（1-0.2904） = 70.96%；

得到如表1所示的不同时间段股票资金流向与涨跌情况占比对比表。仔细观察表格发现，当资金流入时候，时间段越长，越符合上述规律；当资金流出的时候，时间段越短，越符合上述规律。

**表1 不同时间段股票资金理想与涨跌情况占比对比表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | inUp | inDown | outUp | outDown |
| 1日 | 81.10% | 18.90% | 29.10% | 70.90% |
| 3日 | 81.90% | 18.10% | 33.73% | 66.27% |
| 5日 | 91.22% | 8.78% | 61.99% | 38.00% |
| 10日 | 90.45% | 9.55% | 63.73% | 36.27% |

根据上面的分析，我们对不同时间段内资金流向和股票涨跌情况之间的规律有了一定认知，在后续的交易操作中，我们可以根据这个规律，指导我们的投资行为。