

# 华东师范大学计算机科学技术系实验报告

课程名称：数据分析实践

年级：2017 级

实践作业成绩：

指导教师：兰曼

姓名：英嘉豪

作业提交日期：

实验编号：1

学号：10175102247

实践作业编号：1

---

## 1 实验名称

股票信息的采集以及数据的预处理

## 2 实验目的

掌握本地数据的读写，网络数据的采集，应用数据分析技术，对结果进行可视化，并对统计和可视化结果进行分析和讨论并对数据分析的结果保存入磁盘。

### 实验要求

1. : 从 Yahoo! Finance 下载美交所，深交所各种题材的股票（阿里，百度，京东等）
2. : 观察并分析各种题材的股票（例如保险类，新能源类，互联网相关类）的各种统计情况，趋势，相关性分析的等

## 3 实验内容

### 3.1 环境准备

#### 3.1.1 基本环境

- Windows 10

- Python 3.6.7

### 3.1.2 依赖

- pandas 0.23.4
- scipy 1.2.0
- numpy 1.15.4
- matplotlib 3.0.2

## 3.2 股票信息获取

### 3.2.1 股票信息选取

- 互联网: 百度, 阿里巴巴, 微软, 谷歌, 苹果, 拼多多, 腾讯, 亚马逊
- 银行类: 招商银行, 农业银行, 中信银行, 杭州银行, 平安银行, 郑州银行
- 新能源类: 特变电工, 岷江水电, 维克精华, 长城电工
- 医疗行业: 奥星制药, 和黄药业
- 航空运输: 达美航空, 联合大陆航空, 美国航空, 波音

### 3.2.2 获取股票工具函数

---

```
1 # 加载数据
2 Internet_list = ['BIDU', 'BABA', 'MSFT', 'GOOG', 'AAPL', 'PDD', 'TCEHY', 'AMZN']
3 # 美交所: 百度, 阿里巴巴, 微软, 谷歌, 苹果, 拼多多, 腾讯, 亚马逊
4 all_stock = [Internet_list]
5 for stock in all_stock:
6     for per in stock:
7         data = pdr.get_data_yahoo(per, start='1/1/2018')
8         data.to_csv('./mydata/' + name_dict[per] + '.csv')
```

---

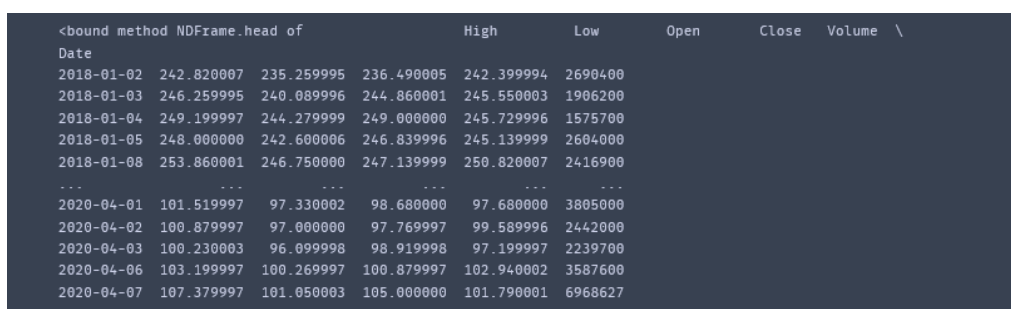
我使用 `get_data_yahoo()` 这个函数来下载股票数，第 1 个参数是股票代码，这里用了列表 `codes` 的迭代器 `ticker`，第 2 个参数是可选的起始时间，目前的默认时间是 2010 年，然而对于我们本次实验的需求，不需要这么庞大的数据量，因此我选取的是自 2018 年 1 月 1 日起的股票数据。我根据我获取的各个股票的代码抓取了股票信息，并存入本地文件夹 `mydata`，并用股票名字作为文件名字，方便下一步进行数据分析

### 3.3 数据分析函数与显示

#### 3.3.1 数据基本信息

```
1 all_stock = [Internet_list]
2 stocks = dict()
3 for stock in all_stock:
4     for per in stock:
5         data = pd.read_csv("./mydata/" + name_dict[per] + '.csv', index_col = 'Date')
6         data.index = pd.to_datetime(data.index)
7         stocks[per] = data
```

我们将本地的股票数据按照模块将其从本地文件夹 `mydata` 中读出，通过调用 `.head()` 或者 `.tail()` 函数就可以看到我们获取的数据



Date	High	Low	Open	Close	Volume
2018-01-02	242.820007	235.259995	236.490005	242.399994	2690400
2018-01-03	246.259995	240.089996	244.860001	245.550003	1906200
2018-01-04	249.199997	244.279999	249.000000	245.729996	1575700
2018-01-05	248.000000	242.600006	246.839996	245.139999	2604000
2018-01-08	253.860001	246.750000	247.139999	250.820007	2416900
...	...	...	...	...	...
2020-04-01	101.519997	97.330002	98.680000	97.680000	3805000
2020-04-02	100.879997	97.000000	97.769997	99.589996	2442000
2020-04-03	100.230003	96.099998	98.919998	97.199997	2239700
2020-04-06	103.199997	100.269997	100.879997	102.940002	3587600
2020-04-07	107.379997	101.050003	105.000000	101.790001	6968627

图 1: 数据显示

在股票信息中我们有 `High`, `Low`, `Open`, `Volume`, `Adj Close` 字段，其分别代表股票当日的最高成交价，最低成交价，开盘价，成交量和最后的调整结束价格。事实上，我们道股票的价格和成交量固然有一定规律，而且值得分析，但是大多数情况下，我们会单独对某只

股票的价格变化趋势感兴趣, 或者单独对某几只股票的成交量关系感兴趣, 很少同时涉及不同的指标。我们为了方便横向对比各个股票在这些字段的表现我们需要构建新的DataFrame。

我们在每个字段分别建立新的表格过程如下

```
1 volume = pd.DataFrame({tic: data['Volume'] for tic, data in stocks.items()})
2 high = pd.DataFrame({tic: data['High'] for tic, data in stocks.items()})
3 low = pd.DataFrame({tic: data['Low'] for tic, data in stocks.items()})
4 open = pd.DataFrame({tic: data['Open'] for tic, data in stocks.items()})
5 price = pd.DataFrame({tic: data['Adj Close'] for tic, data in stocks.items()})
```

这样我们就可以方便的在某一维度进行股票信息的横向对比。调用`.describe()`我们就可以实现对表格的五数概述分析, 当然也可以添加自己的统计量, 例如定义极差、四分位差等。

	BIDU	BABA	MSFT	GOOG	AAPL	PDD	TCEHY	AMZN
count	570.000000	570.000000	570.000000	570.000000	570.000000	428.000000	570.000000	570.000000
mean	173.519755	179.272404	119.735883	1173.201454	206.737909	27.737874	46.247213	1741.593701
std	56.816894	19.990360	24.875017	112.370002	43.359674	7.327749	5.643444	181.422857
min	83.620003	130.600006	82.121544	976.219971	139.753540	17.150000	32.059837	1189.010010
25%	116.790001	166.377506	100.134264	1091.304993	173.431591	21.212500	42.060428	1628.132538
50%	168.084999	178.970001	110.913696	1155.930054	197.096191	25.130000	45.635849	1769.039978
75%	227.217495	190.487499	137.239510	1220.987488	220.926579	34.647499	50.060491	1870.314972
max	284.070007	230.479996	188.185989	1526.689941	327.200012	43.529999	60.662487	2170.219971

图 2: 对成交价的五数概述

在每个字段的表格中我们也可以方便的进行各个股票相关性的分析, 以及某只股票历史波动率和对数收益率的分析。

```
1 # 对 price 表中各个股票相关性分析
2 price.corr()
3 # 对百度股票成交价格的历史波动, 对数收益, 平均移动均值的计算
4 df = pd.DataFrame(price['BIDU'], index = price.index)
5 df['Return'] = np.log(df['BIDU']/df['BIDU'].shift(1))
6 df['42d'] = df['BIDU'].rolling(window = 42, center = False).mean()
```

```
7 df['252d'] = df['BIDU'].rolling(window = 252, center = False).mean()  
8 df['Mov_Vol'] = df['Return'].rolling(window = 252, center = False).mean()*math.sqrt(252)
```

## 3.4 数据可视化

### 3.4.1 折线图

直接对某只股票的对象调用`plot()`方法，该代码画出了对应各个股票的各项数据的折线图

```
1 price.plot(figsize = (16,8))
```

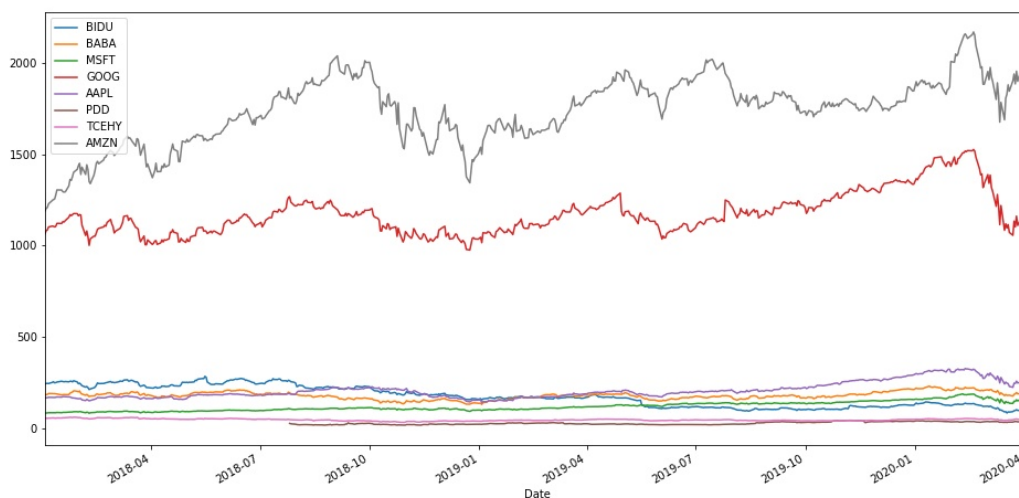


图 3: 互联网股票的价格波动折线图

### 3.4.2 盒图

盒图是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法，它也可以粗略地看出数据是否具有有对称性，分布的分散程度等信息，特别可以用于对几个样本的比较。

```
1 price.plot.box(figsize = (16,8))
```

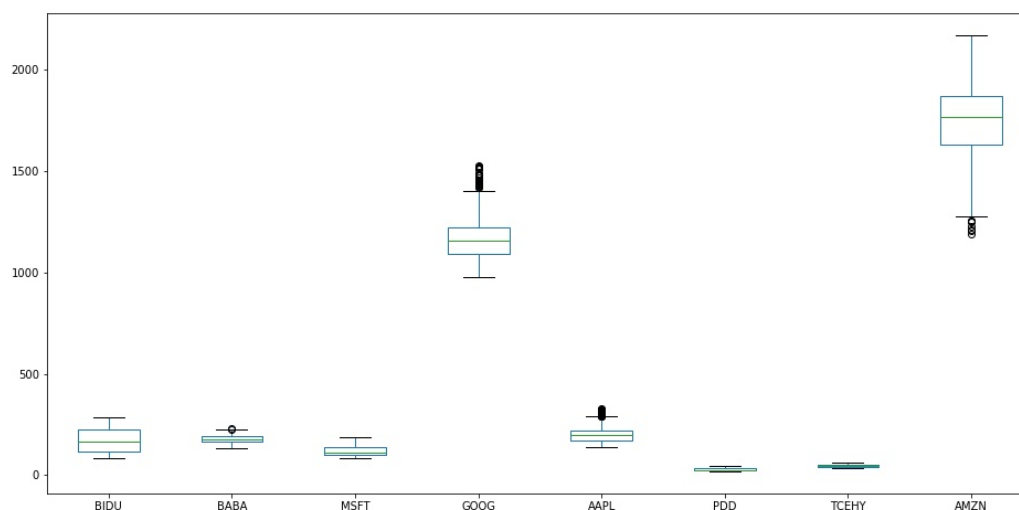


图 4: 互联网股票的价格

## 4 实验结果及分析

### 4.1 各类题材股票的分析

#### 4.1.1 互联网类总体趋势

图 5 的纵轴为Adj close, 即已调整或者复权后的收盘价, 能比较真实反映股票的表现。折线图的可视化使我们直观地看到了股票价格的变化, 从图 5 可以看出, 亚马逊和谷歌的股价遥遥领先其他互联网公司; 从图 6 纵轴为Volume, 即成交量可以看出股票交易量的排名与股票价格的排名似乎没有显然的关系, 在股价上, 亚马逊和谷歌遥遥领先, 而在成交量上, 百度和苹果占据大头。

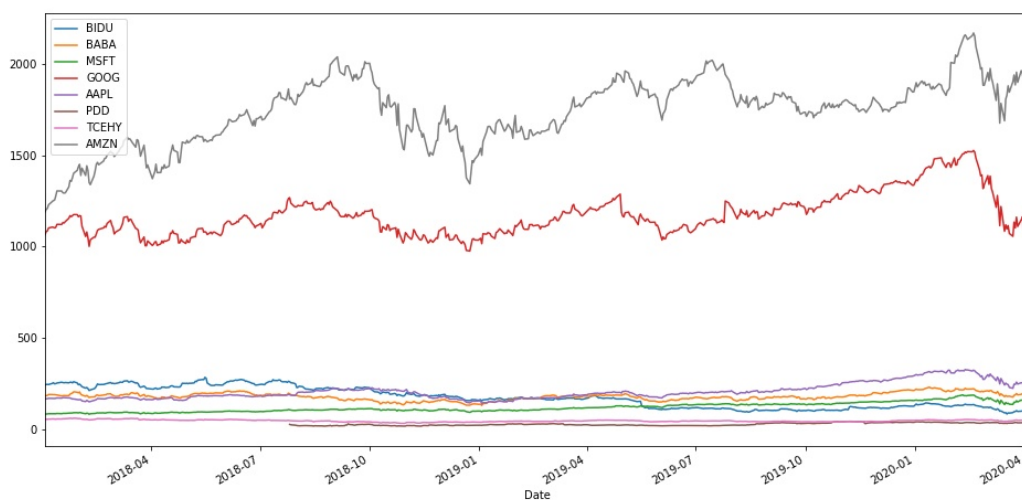


图 5: 互联网类股票的收盘价

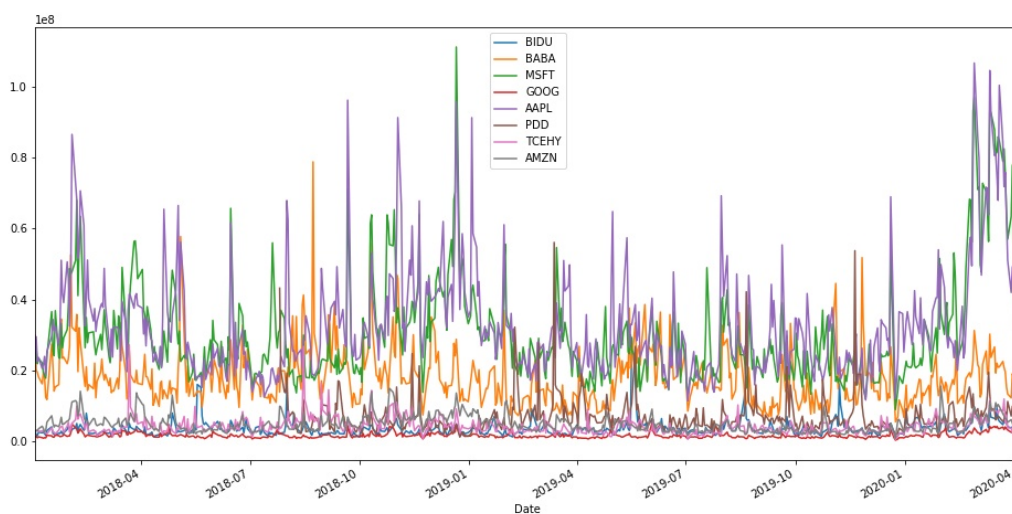


图 6: 互联网类股票的成交量

上市时间上，百度，谷歌，微软，苹果均在初期就已上市，阿里巴巴则 14 年中旬才上市，拼多多上市时间为 18 年中旬，是新兴的公司。因而可以看出股票价格的差距还是很显著的。从 2018 年到 2020 年互联网股票的价格总体呈波动状态，除了在 2018 年 10 月份美

股市场在牛市后迎来了持续暴跌, 堪称黑色星期三。为了看出更明显的股票变化趋势, 我将一些具有代表特征的互联网股票重新做从 2009 年开始重新做了爬取见图 7。

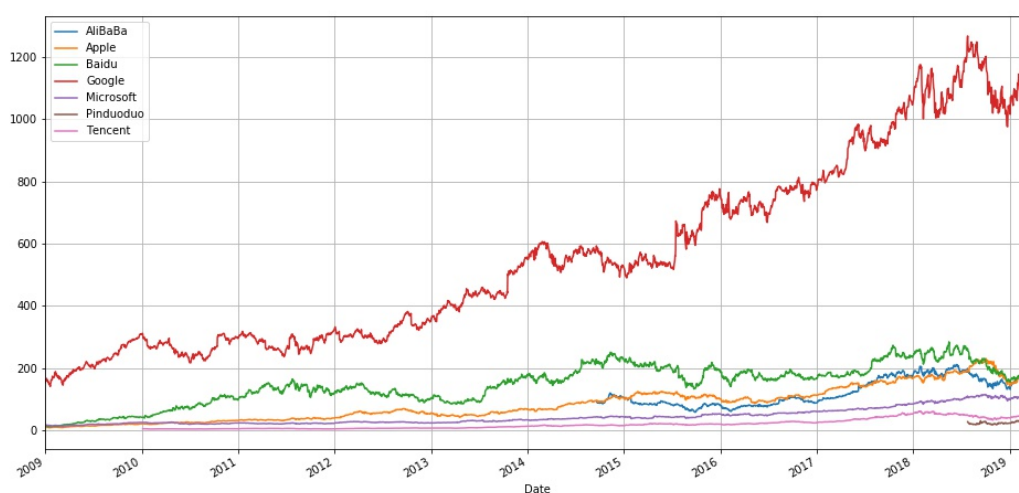


图 7: 2009 年到 2020 年互联网类股票的成交量

由图 7 可见从 2009 年至今互联网类的股票整体呈上升趋势, 在 18 年末, 互联网行业遇冷, 但 19 年又开始回温, 但是在 2020 年三月遭遇了罕见的三级熔断股票价格整体下降。

#### 4.1.2 互联网类股票相关性分析

我们以阿里巴巴为例子来看看价格和成交量之间的关系, 其相关性图见图 8



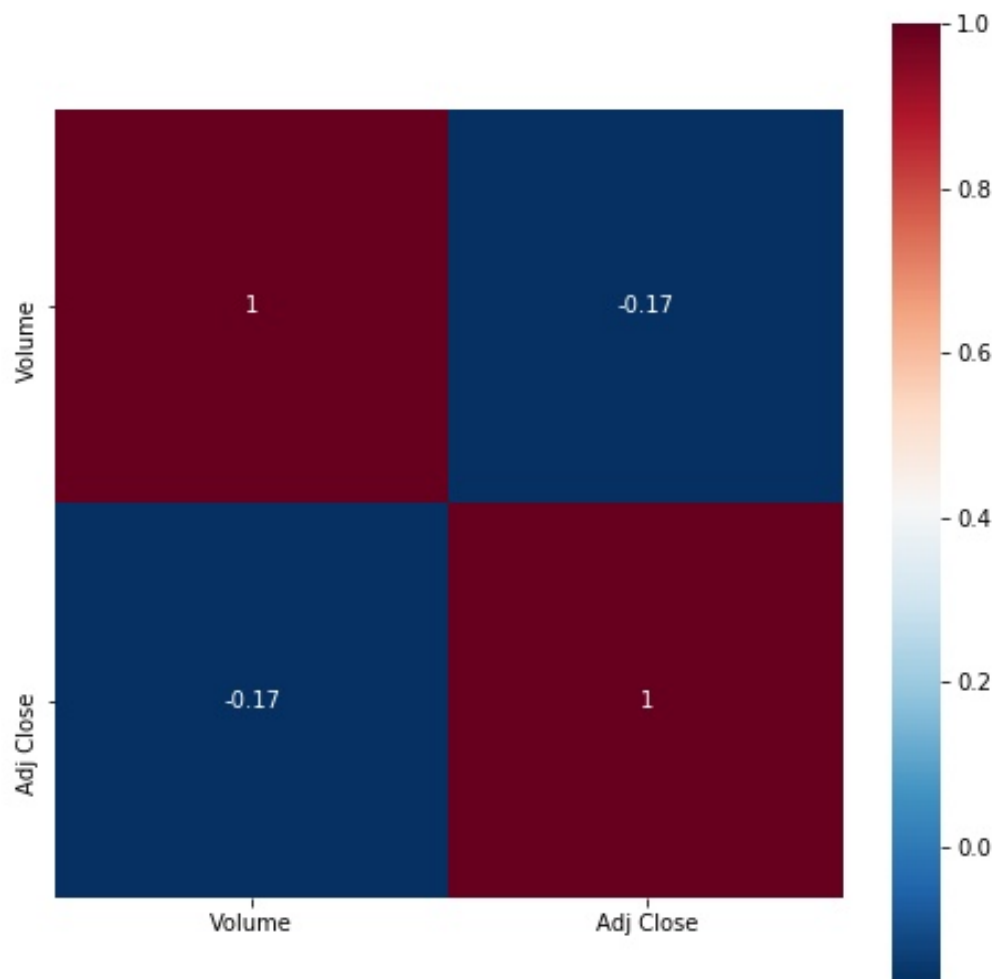


图 8: 股票成交量和价格相关性

同一股票的成交量与股价之间的关系，从生活常识感知股价的上涨，往往伴随着成交量的增加。事实上由我们的计算结果可以看出相关系数趋近于 0。后查阅相关资料得知成交量对股价的影响并不是一成不变的，需要结合具体情况，进行分析。如：

1、在温和的情况下“量价齐升”是指股价的上涨，往往伴随着成交量的增加。成交量是股价的基础，在温和情况下，成交量平稳发展，不断的增加，刺激股价逐步上升，呈多头排列

趋势。

2、在持续上涨的情况下“放量滞涨”是指成交量巨大，但是股价并没有上涨，出现停滞不前的现象。这种情况一般出现在持续上涨的阶段，其中放量并不是意味着买入增加，而是持有该股票的投资者，大量的抛出手中的股票，这可能导致后市股价大幅下跌。

我们再从互联网各个股票相关性着手进行分析，各个股票相关性表如图 9 所示，其中颜色越接近红色其相关性就越高。

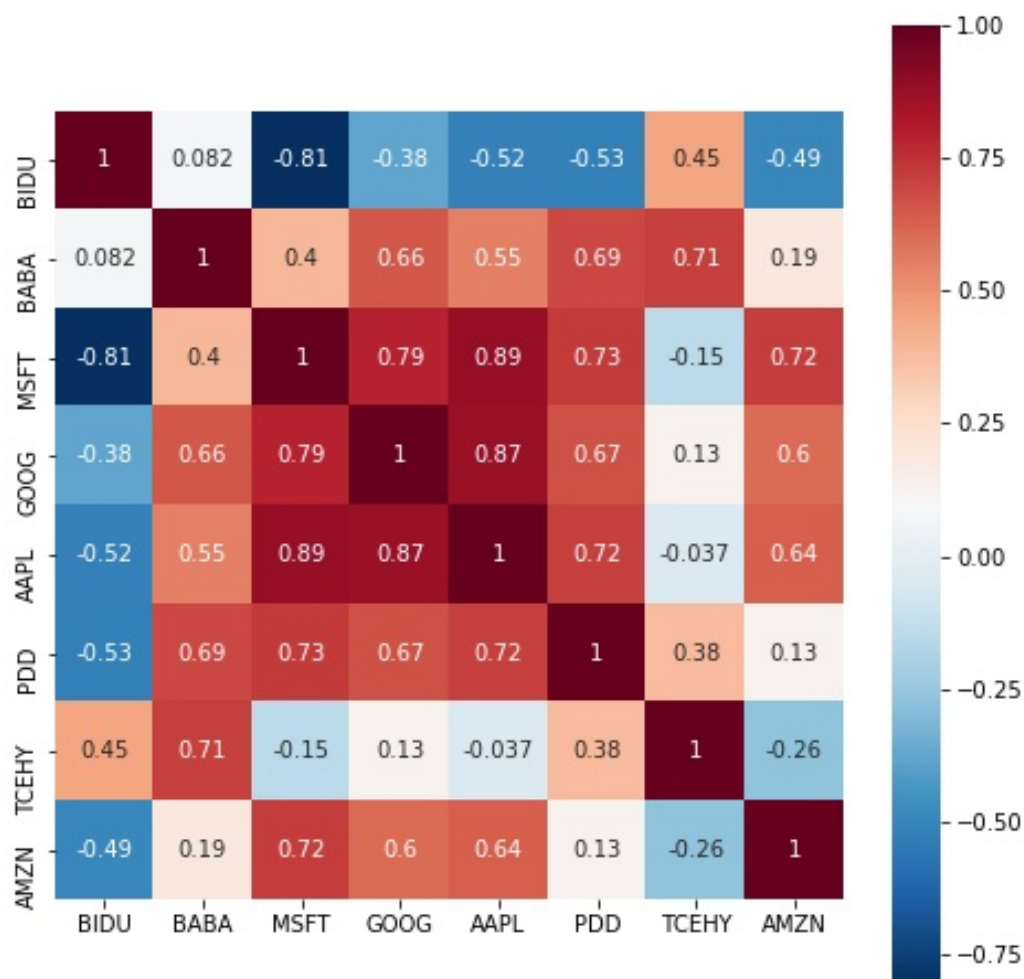


图 9: 互联网股票相关性

可以看出各个股票相关性还是比较强的，值得注意的是蓝色方框大多集中在百度股票中，也就说明百度股票和股票和其他股票相关性较差，这也是神奇的一点。百度与其他公司的相关系数不那么明显，可能更大的原因是百度的股价经常受到负面新闻的印象，波动比较大。

#### 4.1.3 国内互联网单只股票分析

我们以百度为例我们来看他的一些具体数据，我们求出他的对数收益率，历史波动率和平均移动值。

```
1 df = pd.DataFrame(price['BIDU'], index = price.index)
2 df['Return'] = np.log(df['BIDU']/df['BIDU'].shift(1))
3 df['42d'] = df['BIDU'].rolling(window = 42, center = False).mean()
4 df['252d'] = df['BIDU'].rolling(window = 252, center = False).mean()
5 df['Mov_Vol'] = df['Return'].rolling(window = 252, center = False).mean()*math.sqrt(252)
6 df[['BIDU', 'Return', 'Mov_Vol']].plot(subplots = True, figsize = (16,8))
7 df[['BIDU', '42d', '252d']].plot(figsize = (16,8))
```

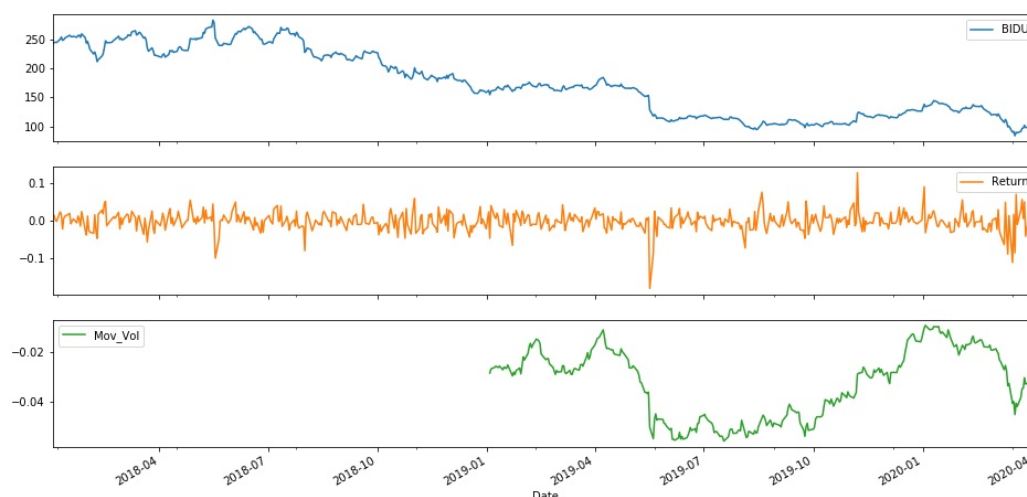


图 10: 百度股价，对数收益和历史波动率

这里比较难过的是没有从图中很清楚的看到所谓的：‘杠杆效应假设’市场下跌时历史移

动波动率倾向于升高,而在市场上涨时波动率下降。我更换了窗口大小和股票依然没有很好的效果。

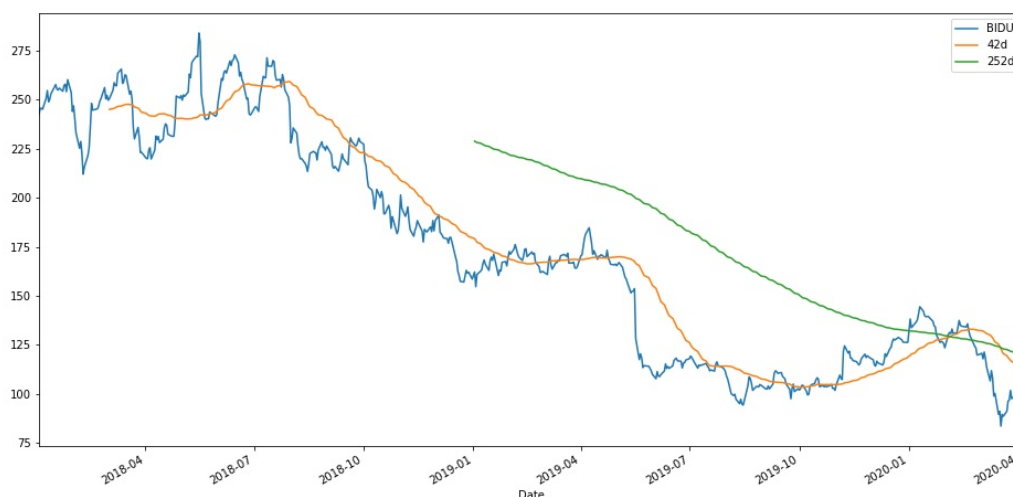


图 11: 平均移动值

可以看出自从 2018 年 8 月百度股票开始走下跌趋势,步入互联网 2018 年寒冬,2020 稍有好转但是紧接着美股熔断,又再次跌入谷底。曾几何时,作为中文互联网巨头,百度在 BAT 中都排在第一位,如今百度却被腾讯、阿里远远甩开,甚至出现了亏损。百度都做错了什么,才让自己从神坛跌落?

### 一、高层动荡

据不完全统计,从 2007 年到 2019 年间,百度至少有十多位副总裁、二十多位高管相继离职。从 2018 年 5 月 18 日百度“二把手”陆奇宣布卸任 COO 至今,百度一直没有摆脱高管的剧烈动荡,至少有 7 位高管先后离开。百度高层动荡,首席科学家吴恩达、总裁陆奇、搜索业务总裁向海龙等相继离职,张亚勤也在去年申请退休。高管离职更多地应该理解为公司内部战略摇摆,矛盾激化,业务方向不能达成共识。如果方向明确,可能像阿里一样进行高管轮岗,就能解决问题。

### 二、豪赌 AI, 可能是一个美丽的错误

百度提出 All in AI 已经很久了。但这个豪赌,可能是一个美丽的错误。百度在人工智能领域的布局和持续发力好像也有些成效,但和投入远远不成正比。作为比较,谷歌都卖出

了世界最强的人工智能企业波士顿动力公司，因为只有投入，没有产出。一直以来致力于科研的波士顿动力公司始终只能推出一些实验室产品和 YouTube 上的爆款视频。而百度在 AI 上的研发投入，每年都有上百亿元人民币。这种没有收入的巨额投入，谷歌都无法承受，何况百度的利润只是谷歌的零头。

### 三、搜索上的封闭化

去年一篇网文《搜索引擎百度已死》批评百度搜索中第一屏有太多百家号链接，这涉及到百度搜索的内容来源问题。自从阿里巴巴、腾讯微信等禁止百度搜索爬取其内容，国内各互联网企业就开始封闭其系统，百度搜索里能搜到的东西就越来越少。

### 四、内容优势荡然无存

在百度的业务布局中，内容本来应该是极其重要的关键一环。而且由于搜索被其他巨头们限制，内容对百度就更加重要。但是百度认识到这个问题还是太晚了，内容领域的今日头条竟然在短短几年内从创业公司成长为体量庞大的巨头。去年今日头条最新一轮的融资，估值已经达到 750 亿美元，甚至已经远超百度市值了。当谷歌还在中国的时候，百度是依靠百度贴吧、百度百科、百度音乐、百度知道、百度文库等内容优势领先谷歌的，当时百度的内容在中国互联网企业里肯定是第一位的。但是如今百度贴吧衰落、百度音乐早已让位腾讯音乐、网易云音乐、百度知道不如知乎、悟空问答。百度当年辉煌的内容优势荡然无存。即使在去年意识到信息流可以称为重要的收入来源后，不断加码百家号与好看视频，现在的百度内容端仍然落后于今日头条与腾讯，甚至也不如阿里巴巴。

#### 4.1.4 能源类

最近几年，新能源工业成了国家优先发展的新兴战略产业之一。新能源股票集万千宠爱于一身，市场持续看好，热点不断，在国家大力发展新能源的政策下，新能源股票也得到很大的关注。在中国，现在虽然风力发电设备的生产已经初具规模，但太阳能光伏技术缺乏整体配套能力，新型蓄电池还没有达到规模化生产的阶段，新能源汽车也只是刚进入市场。总体来说，中国的新能源产业，规模还不小，尚没有进入可以整体盈利的阶段。

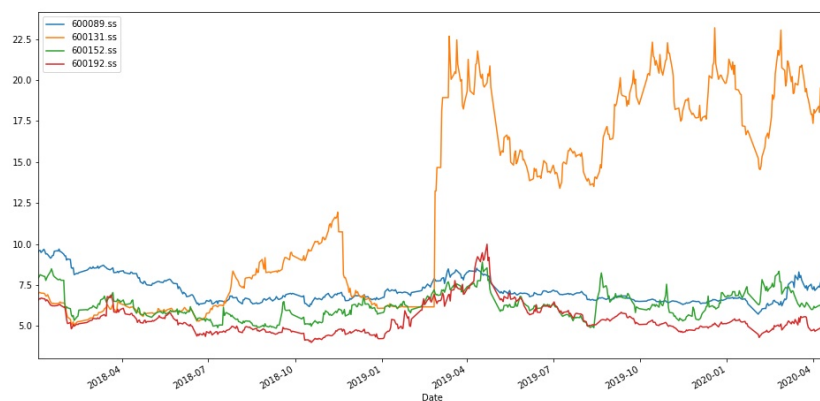


图 12: 能源类信息

所选择的五种能源类股票小除了岷江水电外均波动性较小，总体来说还是处于平稳上升阶段。各类股票的相关性如图 13

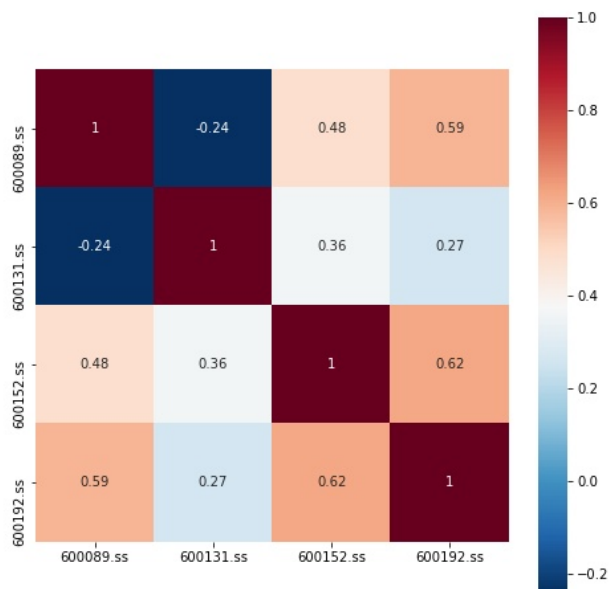


图 13: 能源类信息

可以看出各类股票的相关性还是很差的

1. 这是因为不同的公司采用的技术可能不一样，有的石基于太阳能，有的是基于风能，还有的比如像特变电工是基于多种技术的。
2. 不同公司生产的产品应用的地区不同岷江水电是参股西藏华冠科技涉足太阳能产业而特变电工参股新疆新能源从事太阳能光伏组件制造。我们下面看一下发展较好的岷江水电



图 14: 岷江水电历史波动

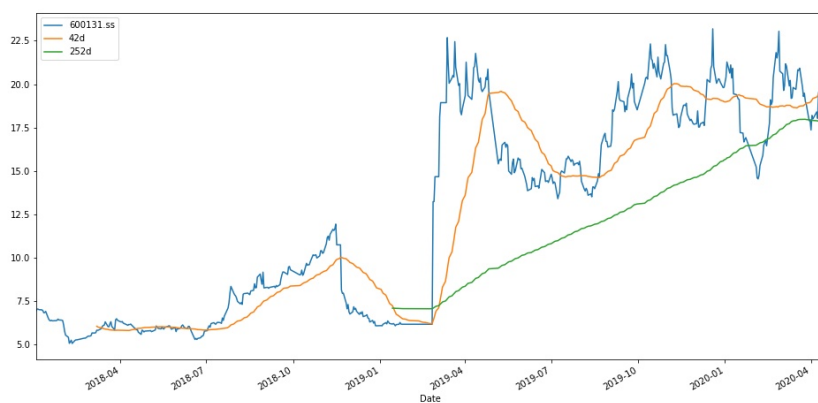


图 15: 岷江水电移动平均值

岷江水电呈现一直上升趋势，受益于国家的改革，岷江水电的盈利大幅提升，导致股价也大涨。

#### 4.1.5 医疗类

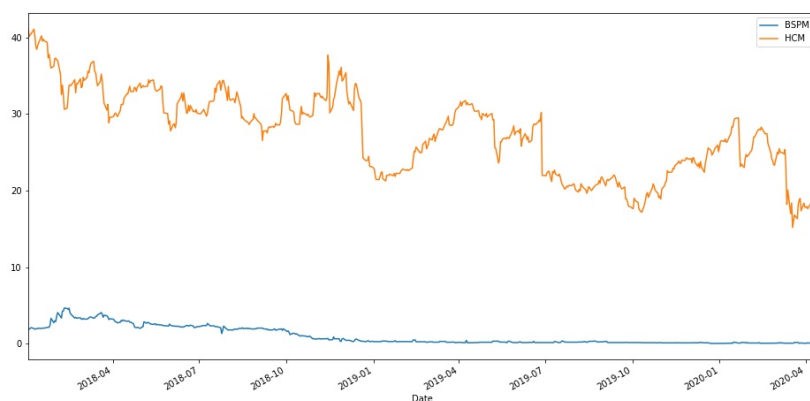


图 16: 能源类信息

选取医疗类股票就是想看看在新冠病毒爆发期间, 和收到熔断影响医疗股的表现怎么样子, 我选区的和黄医药可以看到在疫情爆发期间迎来了发展的上升趋势, 但是可能收到熔断影响又下跌了一段。总的来说因为疫情爆发了医疗行业和口罩行业都赢来了发展的'春天'。在刚爆发阶段奥星制药的涨幅超过 26%。和黄药业、泰和诚医的涨幅分别达到 3.8% 及 4.7%。科兴生物及中国医药控股有限公司的涨幅也都超过 3%。

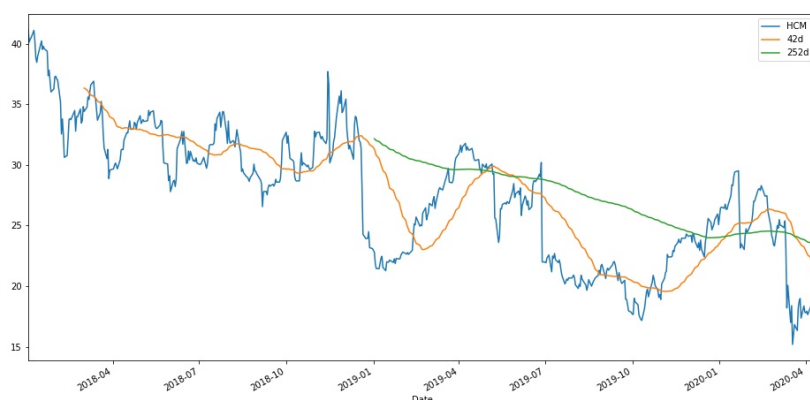


图 17: 和黄医药历史波动



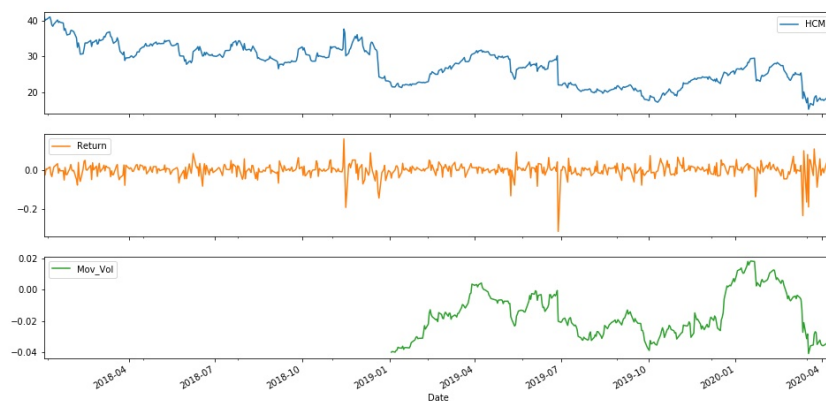


图 18: 和黄医药移动平均

#### 4.1.6 航天交通类

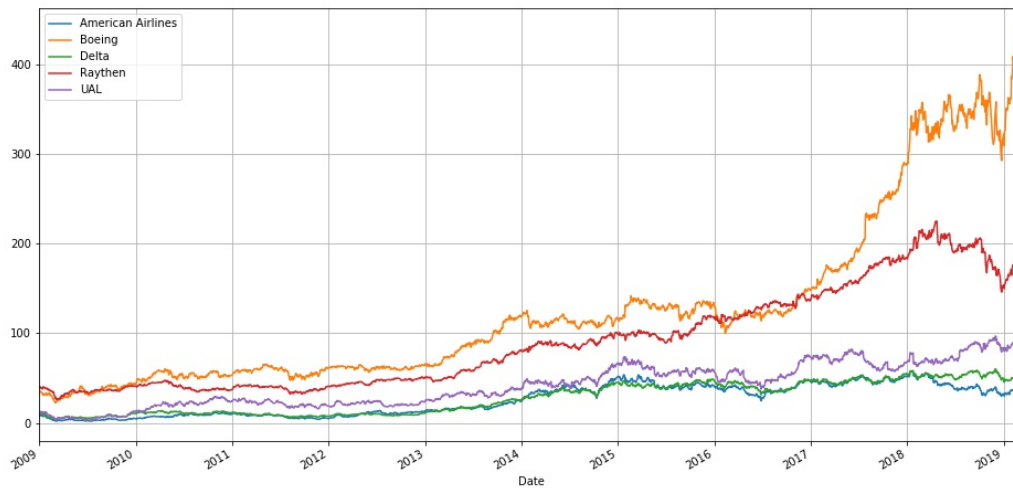


图 19: 航天类

航空类股票整体较稳定。波音和雷神从 2016 年开始呈现较明显上升状态，2018-2019 波动较大。达美航空和美联航总体呈现波动稳定状态。美国航空 2018-2019 呈现下降状态。波音发展势头最猛，但也波动较大。其次是雷神。达美、美联航、美国航空市值较低。分析原

因, 2017 年开始, 美国波音公司 (NYSE:BA) 下属各部门在手的在未交货订单, 累计金额已经超过 4820 亿美元。这为该公司股票提供了良好的增长机会。而从美国国内及国际航空公司和各国政府, 与该公司签订的订单将继续增多的势头不难看出, 波音公司有强劲的发展后劲。波音公司的收入是与其飞机是否能按合同约定时间交货息息相关的, 这个交货率与则与公司的生产率紧密相连。为此, 波音公司目前也在致力于提高生产效率。波音公司也确实完成了此目标, 所以发展较好, 股票涨势明显。而雷神公司, 有四个主要部门: 情报, 信息和服务 (占销售额的 22%), 综合防御系统 (23%), 太空和机载系统 (23%) 以及导弹系统 (30%)。最后两个是真正的核心专长, 都与战争的变化面相关, 其中空中霸权和远距离攻击的能力对于保持优势至关重要。2018 年 11 月和 2019 年 3 月, 有两次波音飞机坠机事件。随之, 波音和雷神都有两次较明显的股票下跌。其他航空类股票在这个时间点也有两次较不明显下跌。未来趋势, 有两个主要因素将会一直驱动波音公司的收入向好发展。首先, 波音公司提高生产率的计划将会提高交货率, 这将直接反映在每年的收入会增加。其次, 全球经济和航空业的发展继续支撑波音公司的市场订单, 这也说明波音公司的未来收益将会继续增长。但这必须建立在波音公司对坠机事件有较好的反省, 并依此对公司策略进行整改, 将安全性可靠性放在首位。巴菲特也借航空类股票下跌机会对达美航空等航空类股票进行买入, 所以均看到航空类股票未来的发展。

## 5 问题讨论

1. 对股票极不了解, 花费一些时间去了解股票, 及其分析指标。但是总觉得自己分析不是很到位, 分析的方向比较片, 有时单单停留在数据和图片的表面。
2. 股票题材和数量极多, 不易选择, 而且不同股票可能在多个不同交易所上市, 也难以找出他们之前的联系。
3. 股票价格变换的背后一定有相应的事件或者原因, 单单从可视化的折线图, 盒图结果中难以发觉其中原因。需要去查阅相关资料, 但是有时候却也找不到合理的解释。
4. 杠杆效应假设 ‘市场下跌时历史移动波动率倾向于升高, 而在市场上涨时波动率下降。’ 但我在图中并没有很好很清楚的看出来这个规律, 可能是我理解得不够深刻。

## 6 结论

1. 互联网股票虽在 2018 年下半年出现整体下跌，但从 2018 年底开始回升。整体处于波动上升趋势但是在 2020 年又遭遇了罕见的熔断导致股票价格又有所下降。在回升阶段 Google 市值最大且增长趋势较好。微软，苹果，阿里巴巴，腾讯均处于上升趋势。百度增长趋势不太理想，出现反复下跌，并渐渐被其他互联网股票赶超。

2. 股票成交量和股价没有显著的相关性。

3. 航空类股票从 2018 年开始出现部分下跌和停滞。2018 年和 2019 年发生的坠机事件对航空类股票产生较大影响，尤其是美国波音公司，出现两次明显的断崖式下跌。整体而言，航空类公司若能抓住机会进行整改、创新和公司形象公关，会在未来有较好发展。

4. 新能源总体发展趋势缓慢，趋于平滑。但是也有少数黑马发展较迅速，随着政府支持力度加大我认为很多新能源企业的盈利能够大幅提升，股价会逐渐上升。

5. 不是所有的股价变动都能找到对应的事件，有的可能是天灾也有可能是人为。

6. 随着新冠爆发，部分医疗企业和口罩防护生产企业迎来了”春天“

7. 美股熔断发生导致外国股票价格波动加剧。