README

提交物说明

文件	说明
weibo_crawl.py	爬取微博数据的爬虫代码文件
weibo_check.py	爬取并验证微博数据的代码文件
table.py	爬取新浪财经表格数据的爬虫文件
table_check.py	爬取并验证新浪财经表格数据的代码文件
weibo	此文件夹中存储爬取的微博数据，子文件分别对应某个用户
weibo_log.txt	此文件存储验证微博数据时产生的错误信息(不存在则因为无错误信息)
stock	此文件存储爬取的新浪财经上的数据
stock/上市公司业绩公告.csv	存储爬取的新浪财经上上市公司业绩公告expected result
stock/融资融券数据.csv	存储爬取的新浪财经上融资融券数据expected result
stock/基金历史信息.csv	存储爬取的新浪财经上基金的历史信息 expected result
stock/上市公司业绩公告log.txt	此文件存储验证新浪财经表格数据时产生的错误信息(不存在则因为无错误信息)

运行环境

- Python 3.6
- Pycharm 2018.2.1
- JRE: 1.8.0_152-release-1248-b8 amd64
- JVM: OpenJDK 64-Bit Server VM by JetBrain s.r.o

代码说明

- 验证器采用边爬取数据边对比expected_result的方式进行验证，所以不会产生actual_result的文件，而是产生对比结果的log文件。

- 微博数据的爬取需要添加cookie信息。代码中的cookie信息为组员的微博cookie信息。weibo_crawl.py爬取了内容数据和文件数据。
- 新浪财经表格为动态渲染。table.py爬取了表格数据。

数据说明

微博数据

数据	说明
微博id	某个微博的id(来自新浪微博url)
微博正文	微博的正文内容
原始图片	原创微博的图片，数组形式，每一个元素包含了图片的url，类型和大小。
被转发微博原始图片	转发的微博的图片，数组形式，每一个元素包含了图片的url，类型和大小
是否为原创微博	True表示是原创微博，False表示不是原创微博
发布位置	微博发布位置
发布时间	微博发布时间
发布工具	发送设备的说明
点赞数	微博的赞数
转发数	微博的转发数
评论数	微博的评论数

其中，微博id，微博正文，原始图片，被转发微博原始图片，是否为原创微博，发布位置，发布时间，发布工具为验证器验证的对象。

新浪财经基金数据

爬取的全部基金的名称和代码：

http://vip.stock.finance.sina.com.cn/fund_center/index.html#jzkfall

后通过网页上基金的链接爬取基金的历史信息：

数据	说明
日期	2019/02/19到2019/12/24期间的数据
单位净值	某日零时统计的单位净值信息
累计净值	某日零时统计的累计净值信息

数据	说明
基金代码	基金代码
基金名称	基金名称

验证：

验证时没有发现问题

新浪财经每日融资融券数据

例如2019-12-12日，新浪财经的融资融券数据记录页：

http://vip.stock.finance.sina.com.cn/q/go.php/vInvestConsult/kind/rzrq/index.phtml?tradedate=2019-12-12

<<上一周 | 12月9日(周一) | 12月10日(周二) | 12月11日(周三) | 12月12日(周四) | 12月13日(周五) | 下一周>>

融资融券交易总量

市场	本日融资余额(元)	本日融资买入额(元)	本日融资偿还额(元)	本日融券余量金额(元)
沪市	467,717,152,298	14,755,841,841	14,783,664,475	0
深市	383,142,874,751	24,298,140,216	0	2,356,355,740

融资融券交易明细

序号	股票代码	股票名称	融资			融券				
			余额(元)	买入额(元)	偿还额(元)	余量金额(元)	余量(股)	卖出量(股)	偿还量(股)	融券余额(元)
1	000001	平安银行	3,560,126,545.00	51,169,966.00	--	24,280,043.00	1,556,413	116,300	--	3,584,406,588.00
2	000002	万科A	3,059,425,963.00	115,756,890.00	--	35,016,800.00	1,250,600	131,400	--	3,094,442,763.00
3	000006	深振业A	596,091,963.00	5,358,464.00	--	609,042.00	117,803	3,000	--	596,701,005.00
4	000008	神州高铁	369,954,540.00	10,181,703.00	--	77,826.00	21,800	0	--	370,032,366.00
5	000009	中国宝安	1,143,747,788.00	10,367,750.00	--	1,648,088.00	358,280	3,000	--	1,145,395,876.00
6	000012	南玻A	429,236,657.00	7,224,956.00	--	673,030.00	150,230	48,200	--	429,909,687.00
7	000016	深康佳A	135,836,507.00	18,972,789.00	--	183,920.00	44,000	10,000	--	136,020,427.00
8	000021	深科技	524,142,081.00	65,128,211.00	--	4,554,225.00	432,500	72,300	--	528,696,306.00
9	000025	特力A	60,463,607.00	1,872,192.00	--	4,393,956.00	217,200	7,000	--	64,857,563.00
		深圳能								

数据	说明
日期	2011/06/01到2019/12/29期间的每日数据
序号	某股票当日记录序号

数据	说明
股票代码	某日零时统计的累计净值信息
股票名称	某股票的股票名称
融资-余额（元）	融资栏下的余额（元）
融资-买入额（元）	融资栏下的买入额（元）
融资-偿还额（元）	融资栏下的偿还额（元）
融券-余量金额（元）	融券栏下的余量金额（元）
融券余量（股）	融券栏下的余量（股）
融券-卖出股（股）	融券栏下的卖出股（股）
融券-偿还量（股）	融券栏下的偿还量（股）
融券-融券金额（元）	融券栏下的融券金额（元）

验证：

验证时没有发现问题

新浪财经上市公司业绩公告

如下链接，为某一季度上市公司的公告列表

http://finance.sina.com.cn/realstock/income_statement/2016-03-31/issued_pdate_ac_10.html

股票代码	股票名称	批露日期↑	每股收益(元)	营业收入		净利润		每股净资产(元)	净资产收益率(%)	每股现金流量(元)	毛利率(%)	分配方案	明细	PDF报告
				营业收入(万元)	同比(%)	净利润(万元)	同比(%)							
601179	中国西电	2016-04-28	0.05	243160.97	-10.12	23253.38	43.06	0.00	1.25	0.00	33.54	不分配	明细	查看
601699	潞安环能	2016-04-28	0.01	246506.77	-10.00	2115.56	-70.38	0.00	0.12	0.00	16.56	不分配	明细	查看
601766	中国中车	2016-04-28	0.07	4033097.10	72.45	198777.70	104.56	0.00	2.03	0.00	22.51	不分配	明细	查看
000710	天兴仪表	2016-04-28	-0.01	5262.26	13.04	-87.80	84.11	0.00	-0.83	0.00	6.65	不分配	明细	查看
000707	双环科技	2016-04-28	0.01	77068.35	1.23	502.37	-8.34	0.00	0.35	0.00	18.59	不分配	明细	查看
000705	浙江震元	2016-04-28	0.02	60219.65	11.59	827.83	21.57	0.00	0.65	0.00	12.67	不分配	明细	查看
000702	正虹科技	2016-04-28	-0.03	23189.35	-20.73	-792.53	-443.31	0.00	-1.76	0.00	8.80	不分配	明细	查看
000701	厦门信达	2016-04-28	0.01	695853.66	36.50	2935.35	-22.95	0.00	0.13	0.00	3.61	不分配	明细	查看
000698	沈阳化工	2016-04-28	0.11	199339.76	64.41	8667.62	251.56	0.00	2.21	0.00	7.63	不分配	明细	查看
000692	惠天热电	2016-04-28	0.18	82368.70	-0.82	9456.19	-30.61	0.00	6.64	0.00	22.06	不分配	明细	查看
000652	泰达股份	2016-04-28	0.00	238830.71	329.00	238.62	-30.10	0.00	0.08	0.00	2.70	不分配	明细	查看
601177	杭齿前进	2016-04-28	0.02	33668.39	-14.08	935.42	84.86	0.00	0.58	0.00	18.49	不分配	明细	查看
601166	兴业银行	2016-04-28	0.82	4091700.00	19.62	1570000.00	6.15	15.95	5.31	-4.19	49.29	不分配	明细	查看
601158	重庆水务	2016-04-28	0.06	97003.73	9.89	30443.96	-4.66	0.00	2.21	0.00	45.42	不分配	明细	查看
601137	博威合金	2016-04-28	0.12	64117.87	-8.15	2547.03	10.13	0.00	1.24	0.00	12.07	不分配	明细	查看

其中最后一栏为公告pdf下载链接，通过url可获得对应的文件元数据。

表格包含的数据如下：

股票代码	**股票名称**	**披露日期**
每股收益（元）	营业收入（万元）	营业收入同比（%）
净利润（万元）	净利润同比（%）	每股净资产（元）
净资产收益率（%）	每股现金流（元）	毛利率（%）
分配方案	明细	PDF报告
文件大小	文件类型	文件最后修改时间

Reference

[1]https://github.com/dataabc/weiboSpider