Level 5 实验报告

PB20111686 黄瑞轩

选题：基金吧舆论情感走势与基金净值走势的相关检验

背景：随着支付宝、东方财富网购买基金功能的完善，越来越多的大学生加入了炒基金的行列。和传统炒股、炒基金模式不同的是，现在购买基金的 APP 往往提供"评论区"功能，用户可以发帖、回帖、分享自己的见解。Level 5 中，我将综合利用爬虫、情感分析和其他图形库技术，对东方财富网基金吧内舆论进行情感走势与基金净值走势的相关检验。

Step1：分析基金吧网页结构

基金吧 url：`http://guba.eastmoney.com/list,of{ID}{_{Page}}.html`，其中`{ID}`表示基金代码，如果不是第一页，则需要加上`{_{Page}}`，`{Page}`表示当前页号。

最大页号 xpath：`//*[@id="articlelistnew"]/div[82]/span/span/span/span`

标题 xpath 构成：

- 第一个帖子的 xpath：`//*[@id="articlelistnew"]/div[2]/span[3]/a/text()`
- 最后一个帖子的 xpath：`//*[@id="articlelistnew"]/div[81]/span[3]/a/text()`

很容易看出遍历帖子所需要的变量。

发帖时间 xpath：`//*[@id="articlelistnew"]/div[2]/span[5]`，这里的 2 处为帖子位置。

Step2：分析基金净值网页结构

爬基金净值的网页 url：`https://www.dayfund.cn/fundvalue/{ID}.html`，其中`{ID}`表示基金代码。

日期 xpath：`//*[@id="his_nav_table"]/tbody/tr[1]/td[1]`

增长率 xpath：`//*[@id="his_nav_table"]/tbody/tr[1]/td[4]`

Step3：封装基金吧网页、基金净值网页对象

先简单将翻译功能封装成`translate`模块，这里不赘述。

基金吧网页支持如下方法：

```
class talk_bar:
    url = ''
    max_page = 0

    def __init__(self, id):
        # 按基金编码来初始化url，初始化翻到第一页
        # 获取最大页数

    def change_page(self, page):
        # 翻到基金吧第page页

    def get_title(self, num):
        # 获取第num个标题内容和时间
        # 返回值是一个二元素列表，第0个元素是时间，第1个元素是内容
```

基金净值网页支持如下方法：

```
class value:
    url = ''

    def __init__(self, id):
        # 按基金编码来初始化url

    def change_page(self, page):
        # 翻到第page页

    def get_rate(self, num):
        # 获取表格第num行的内容
        # 返回值是一个二元素列表，第0个元素是时间，第1个元素是当日净值增长率
```

这里列出的也是**伪代码**。实际情况中因为对每一页的访问量大，并且 API 翻译需要时间，因此可以对每一页的 html 页面进行缓存，以减少 `request` 次数，这里没有列出具体方法，实现详情见代码。

这样，只需要调用

```
import url

test = url.talk_bar('161725')
test2 = url.value('161725')

for i in range(1, 61):
    pr2 = test.get_title(str(i))
    pr = test2.get_rate(str(i))
    if pr2 is not None:
        print(pr2)
    # if pr is not None:
    #     print(pr)
```

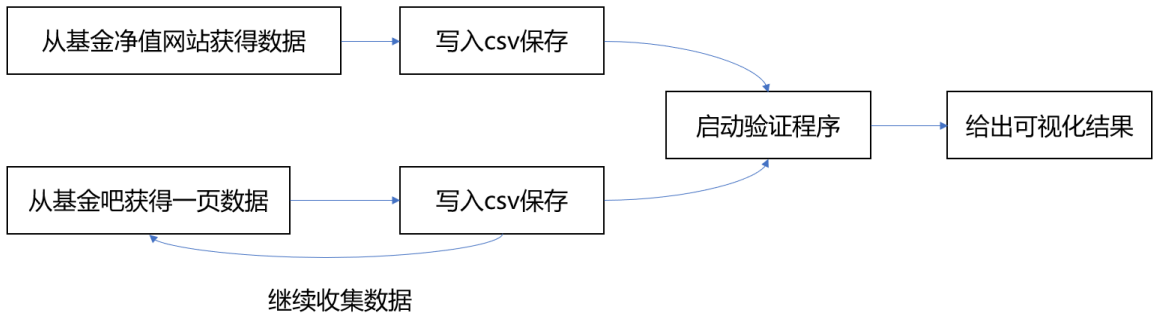
就可以相当方便的给出我们需要的数据。

```
['06-06 06:15', '用行动赢取“苹果华为”礼包']
['06-05 17:21', '【讲故事瓜分2w】我和我的新能源故事']
['06-03 11:51', '【30000惊喜好礼】鑫元中短债夏日派对']
['06-06 11:17', '为什么说这两个板块有望迎来爆发行情？']
['06-06 11:04', '策略：确定了，明天A股走势推演已出炉，请尽快查收！']
['06-06 11:03', '重磅美股纳指大跌超2%，下周A股何去何从？（附策略）']
['06-06 11:00', '策略|端午节后A股会涨吗？直接告诉你']
['06-06 10:42', '市场反弹后感受也好了，而不买会踏空吗？告诉你布局逆']
['06-06 11:30', '0606猴哥点评：科创量价齐升再度大涨近3%，锂电板块持']
['06-06 11:30', '白酒今天百分百大跌[大笑][大笑][大笑][大笑][大笑][']
['06-06 11:29', '给大家的一封信：']
['06-06 11:28', '$白酒基金LOF(SZ161725)$跑不赢大盘还']
['06-06 11:27', '观点鲜明：2022年6月6日，各板块基金，今天的操作建议']
['06-06 11:27', '新能源带涨创业板，踏空的朋友要追吗？']
['06-06 11:27', '拍断大腿，触底反弹又是新能源，你们别买让我来[兴奋]']
['06-06 11:26', '不要和趋势作对，趋势向下买易亏。趋势向上卖易飞，顺']
```

```
['2022-06-02', '-0.77%']
['2022-06-01', '-0.53%']
['2022-05-31', '3.21%']
['2022-05-30', '3.29%']
['2022-05-27', '0.52%']
['2022-05-26', '0.04%']
['2022-05-25', '-0.52%']
['2022-05-24', '-1.78%']
['2022-05-23', '-1.52%']
['2022-05-20', '3.96%']
['2022-05-19', '-0.91%']
['2022-05-18', '-0.87%']
['2022-05-17', '0.87%']
['2022-05-16', '-0.82%']
['2022-05-13', '-0.36%']
['2022-05-12', '1.01%']
```

Step4：各种库模块封装以及数据流传送设计

数据流传送图：



可见写入和验证程序的读取是比较固定的模式，把这个操作封装成库（与 numpy 及 matplotlib 等库结合）。

数据命名规范：

- 基金净值：`{ID}_jz_{create_time}.csv`，其中 ID 为基金代号，create_time 是一个 `yyyy-mm-dd` 形状的导出时间。
- 基金吧数据：`{ID}_tk_{page}_{create_time}`，其中 ID 为基金代号，page 是导出时此页为第几页，create_time 是一个 `yyyy-mm-dd` 形状的导出时间。

这样，用户就可以无需看到复杂的内部结构，直接初始化（这里以 *招商中证白酒指数分级基金：161725* 为例）

```
from time import sleep
import mycsv

test = mycsv.collector('161725')

test.get_jz()
sleep(5)
test.get_tk('1')
```

就可以获得下面的结果：

	A	B
1	date	rate
2	2022/6/2	-0.77%
3	2022/6/1	-0.53%
4	2022/5/31	3.21%
5	2022/5/30	3.29%
6	2022/5/27	0.52%
7	2022/5/26	0.04%
8	2022/5/25	-0.52%
9	2022/5/24	-1.78%
10	2022/5/23	-1.52%
11	2022/5/20	3.96%
12	2022/5/19	-0.91%
13	2022/5/18	-0.87%
14	2022/5/17	0.87%
15	2022/5/16	-0.82%
16	2022/5/13	-0.36%
17	2022/5/12	1.01%
18	2022/5/11	1.40%
19	2022/5/10	1.55%
20	2022/5/9	-2.11%
21	2022/5/6	-3.80%
22	2022/5/5	0.17%
23	2022/4/29	0.55%
24	2022/4/28	2.66%
25	2022/4/27	2.53%

	A	B
1	date	title
2	2022/6/6 6:15	用行动赢取“苹果华为”礼包
3	2022/6/6 14:33	【30000惊喜好礼】鑫元中短债夏日派对
4	2022/6/5 17:21	【讲故事瓜分2w】我和我的新能源故事
5	2022/6/6 16:17	策略|端午节后A股会涨吗？直接告诉你
6	2022/6/6 16:16	刚要入睡，证券市场又传重磅利好，6月准备迎接大行情
7	2022/6/6 16:05	重磅美股纳指大跌超2%，下周A股何去何从？（附策略）
8	2022/6/6 16:00	A股行情普涨突破3200点，应证提示小心踏空
9	2022/6/6 14:21	6.05复盘：A股估值处于历史低位，白酒军工新能源怎么看？
10	2022/6/6 16:36	周末消息面总结解读和下周A股操作思路
11	2022/6/6 16:36	A股出现逆转信号，牛市的要来了吗，下半年准备扬眉吐气
12	2022/6/6 16:33	百万基金实盘：小高理财6月6日投资分享，成功预判节后上涨行情
13	2022/6/6 16:33	生活中不可避免的思维盲点——归因谬误（上）
14	2022/6/6 16:32	复盘：确定了，A股开始加速上涨，千万要踏准节奏！
15	2022/6/6 16:32	继续优化仓位结构，A股形成向上突破的趋势雏形
16	2022/6/6 16:29	我用的闲钱，不挣十几点不会走，现在又不是高位[呲牙][呲牙]
17	2022/6/6 16:29	周线级别金叉已经形成，大盘波段反弹已经开启
18	2022/6/6 16:28	A股基金收评|A股6月上涨预测成功 下一步我们这样做
19	2022/6/6 16:26	假期过后大盘又迎来开门红，沪指突破3200点后该如何走
20	2022/6/6 16:22	今天想必大家是开心的，又是一个吃大肉的行情，特别是新能源和
21	2022/6/6 16:22	收评|市场如期迎来“报复性”反弹，可从防守逐步转向进攻？市场行
22	2022/6/6 16:19	不要和趋势作对，趋势向下买易亏。趋势向上卖易飞，顺势而为利
23	2022/6/6 16:17	百万实盘：大头6月6日投资分享，科创板有望迎来牛市？
24	2022/6/6 16:16	盈亏日记丨今天基金盈利接近1.8万，节后有望延续反弹
25	2022/6/6 16:16	真正穿越牛熊的中长期优质基金！（附详细名单）

拿到数据，就可以来做数据检验了。

> 这里不把翻译后的数据放进来，是因为翻译的结果要消耗大量请求，极易发生错误，本人认为保存原始数据的操作越安全越好。

Step5：数据检验设计

> 依然以 *161725.of* 为例，封装一个读取数据的 csv 行为对象以及 matplotlib 画图对象，这里不展开。
>
> 由于基金的涨跌牵涉到的方面较为复杂，单从**舆论情感**一方面量化可能不完善，因此本项目先对“**涨**”、“**跌**”和“**平**”（定义为当天涨跌幅在0.2%以内）做三态的分析，以 `simple` 为接口标记。最后会实现一个量化的接口，但并不保证其准确性（超出了Level 5的范围）。

首先，收集感兴趣对象的基金净值数据以及基金吧舆论数据：

```
test = mycsv.collector('161725')

stage = 'get data'

# 获取数据
if stage == 'get data':
    for i in range(1, 400):
        test.get_tk(str(i*5))
        sleep(random.randint(10,20))
        print(f"Round {i} success, start next trying...")

    test.get_jz()
```

然后，根据舆论数据按日期降序排列的特点，将所有的内容翻译，交给 Transformers 判断情感，结果存在 `./log` 文件中。

```
test2 = mycsv.analysier()

stage = 'have got data'

if stage == 'have got data':
    test2.trans_tk()
```

log - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
{'06-02': array([-0.99806362, -0.93423724, -0.92530268,  0.95604026, -0.55491638,
       -0.80474508, -0.99473226,  0.69771445, -0.99921787,  0.99986649,
       -0.99941754, -0.53760475, -0.99968958, -0.99977452,  0.99925071,
       -0.98912209, -0.98204523, -0.99820793, -0.99975413, -0.98715091,
        0.99240959, -0.58778149, -0.99557424,  0.97196668, -0.99975222,
       -0.99540102,  0.96218228, -0.99980575, -0.53306615,  0.98607999,
        0.91166162, -0.98210686, -0.99828774, -0.96648741, -0.94496667,
       -0.99810022, -0.57657206,  0.99716824, -0.98068994, -0.78222823,
        0.99816209, -0.99933738, -0.99756348, -0.9994185 , -0.60034353,
       -0.99829072,  0.99624121, -0.9661116 ,  0.99960929,  0.99837315,
        0.96620917, -0.99793184, -0.99979275, -0.97409534, -0.99868888,
       -0.98492843, -0.99968159, -0.99977154,  0.99932539,  0.99716324,
       -0.97509891,  0.99879837, -0.98806107, -0.99250519,  0.99081224,
       -0.99285746, -0.99567151,  0.99968994, -0.96480525,  0.94046789,
        0.99924123,  0.68203712, -0.99844462, -0.99896812, -0.98749357,
        0.99710542, -0.9952845 , -0.54321063,  0.99887437]), '06-01': array([ 0.99927324, -0.9987561
-0.99268299,  0.99831933,
       -0.92646605, -0.99516845, -0.99345607, -0.99744898,  0.99916875,
       -0.98780721,  0.99666601,  0.99852955,  0.97272599,  0.88992876,
        0.99941623,  0.99950218,  0.98182839, -0.99666154, -0.99683726,
       -0.98240924,  0.99679321, -0.99668282,  0.93092138, -0.98887873,
        0.96222454, -0.93143624,  0.99716502, -0.99660194,  0.99727684,
        0.952393  ,  0.99979061, -0.93310726, -0.8240962 ,  0.87633634,
       -0.54004192,  0.9998523 , -0.7949242 , -0.91077566, -0.98943335,
       -0.99911612, -0.9972921 ,  0.99937028, -0.89728421, -0.98278922,
       -0.99691725, -0.98133838,  0.99670261, -0.97324872,  0.99871564
```

之后，根据返回的 dict 字典，计算各情感值得分的平均。对结果进行正负及当天游程检验，得结果

```
id:161725
当天的符合度是 0.5204081632653061
负游程检验的符合度是 0.5204081632653061
正游程检验的符合度是 0.6428571428571429
```

结论：基金吧舆论对基金涨跌有影响，但是这种影响是正向的，即舆论倾向受净值的影响比净值受舆论的影响更显著。

level 5 项目文件（code 文件夹中）

- analysis.py （情感分析封装）
- log （分析结果）
- main.py （主函数）
- mycsv.py （读写 csv 操作封装）
- translate.py （翻译功能封装）
- url.py （爬虫模块封装）

注：按照所述步骤，将 `161725` 换成其他的基金代码就可以检验其他基金数据。所获得的

Total : 5 files, 358 codes, 38 comments, 84 blanks, all 480 lines