**个人代码：**

**/home/luowang/data/financial\ reports/Remote\_call/Remote\_call/**

注： 1) 如果所用机器以前还没连接过所用服务器的，则要手动的 ssh 连接一遍才行

2) 如果要修改所使用的服务器，请对各 \*.sh 上的 machine\_list 进行修改，以及一些 \*.py 中的 machine\_N 进行修改

各服务器上执行指令：

bash cmd.sh 或者 bash su-cmd.sh，两者区别在于权限

# 注意，bash 自动调用的 shell 我没修改，注意回退时按 Ctrl + Backspace 就好

如果是安装文件，看 po\_test/adapt\_cmd.sh，在 expect "root" 那句改一下安装指令。不过服务器环境不一，有些机器对某些工具可能难以安装

跑代码：

1、删除服务器端上的数据

1) bash cmd.sh 或 su-cmd.sh

2) rm -r -f path ( 如 /home/lijunjie/lw/data/282\_tagged\_files/ )

2、传数据

sudo python post\_data.py source\_path\_in\_local target\_path\_in\_machines

如 sudo python post\_data.py /home/lijunjie/lw/data/282\_tagged\_files/ /home/

lijunjie/lw/data/

# 注意，服务器上的数据应与本地的路径保持一致，这样后面分配任务才好做。这也是为什么我在 172.31.238.165 上也创建一个 /home/lijunjie/lw/ 的原因

3、调用各服务器进行处理

python adapt\_process.py command\_in\_each\_machine target\_data\_path

如 python adapt-process.py "python /home/lijunjie/lw/code/wy\_wk5\_tmp/

3txt\_tag\_init.py /home/lijunjie/lw/data/282\_tagged\_files/ /home/lijunjie/

lw/data/282\_tagged\_init\_file/" /home/lijunjie/lw/data/282\_tagged\_files/

4、从各服务器中拉数据

修改 merge\_data.py 中的 suffix/ ①后，sudo python merge\_data.py 即可

如 从 result/ 拉数据到本地所需文件夹 sudo cp -r /home/luowang/data/financial\

reports/Remote\_call/Remote\_call/result/.../ /home/luowang/data/financial\

reports/result\_merge/

① 表示处理后的数据存在服务器上 /home/lijunjie/lw/data/ 中的哪个文件夹下

5、从本地(165机器)的 Remote\_call 下的 result 中，复制数据到指定地方

如 sudo cp -r /home/luowang/data/financial\ reports/Remote\_call/

Remote\_call/result/....../ /home/lijunjie/lw/data/

6、(可选) 删除 result/ 下面的文件

**信息文件：**

**/home/luowang/data/financial\ reports/result\_merge/**

filter\_files：应该是年报TXT吧

282\_tagged\_init\_file：过滤掉标点符等的年报TXT

282\_tagged\_files：tree\_tagger 标识后的文件

split\_sentence：filter\_files 中的文件分句后文本

282\_txt\_freq：原始词频信息，存在 txt 中

282\_excel\_freq：把年报中的词频信息存在 excel 中

282\_Count：记录三个词典下各情感维度的词频统计信息

282\_emotional\_freq：记录了三个词典所有情感维度下的词频信息，存在 excel 中

282\_Summary：汇总了词典情感维度下的词频信息和股价变化情况(Label，+1、-1、0)。前者缺失值以 '0' 表示，后者缺失值以 '#N/A' 表示

282\_dict\_summary：freq\_summary.xls 记录了三个词典各情感维度下，某股票某年的词频信息；其他两个 pkl 则只是将这部分信息分为 Annual 和 Interim 存储，具体格式为 freq\_info[ stock ][ year ][ dimension ] = freq

282\_cov\_summary：三个词典构成的情感维度下的词频与股价变化的协方差(没什么参考价值)

282\_tmp\_vec：根据 282\_cov\_summary/ 下的协方差信息，针对第7、30天选出的股票词频信息。按叶的要求调整的过的格式。具体为：freq\_info[ stock ][ year ][ dimension + 'Label'\_i ] = freq.

# dimension + 'Label'\_i：各词典情感维度 dimension 以及 'Label0' - 'Label4'

282\_vec\_pkl：emotionalDim\_\* —— 根据 282\_cov\_summary/ ，针对某天的股价协方差信息选出的股票信息①；freq\_McDonald\_count2.pkl —— 根据词典 LoughranMcDonald\_MasterDictionary\_2014 得到的，以一个单词单独作为维度形成的向量信息，文件挺大的，需要降维处理；label\_info.pkl 根据股价变化得出来的 Label 信息，具体格式为 label\_info[ stock ][ year ][ 0-4 ] = label

① 针对 Annual/ 的；简单的筛选标准为，nagetive 维度下为负相关，positive 维度下为正相关；有 \_100\_ 的为放松标准筛选出的前 100 个股票，没有的则为严格筛选，通常仅 70+ 个

**/home/lijunjie/lw/**

部署在各服务器上的数据的样式。基本上都是王毅的代码，部署时会有修改，数据可以参见 yyy/wk5/

修改后的代码文件，请参见 Remote\_call/code/wy\_wk5\_tmp/ 与 /home/lijunjie/lw/code 处。建议用 diff 指令查看