基于爬虫的股票交易活跃度研究

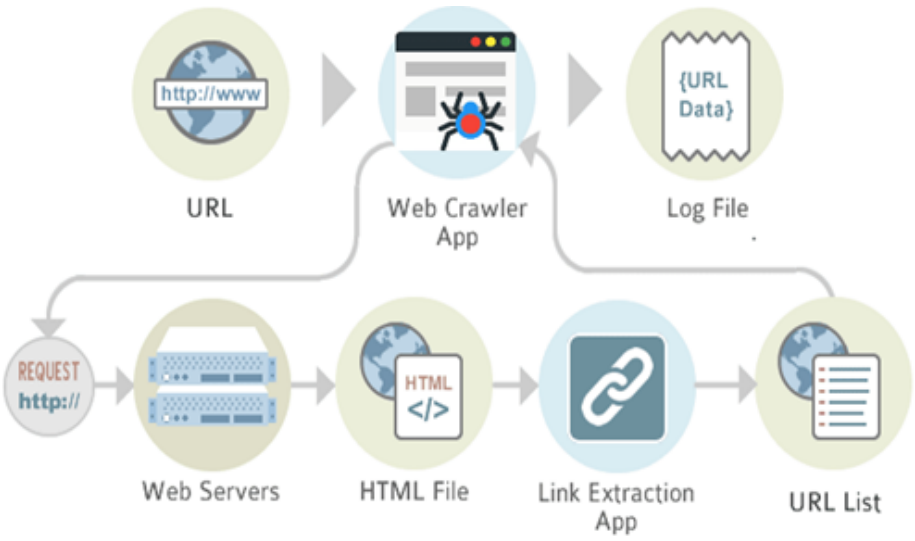
引言

我们解决了通过结构化数据选择交易活跃的股票，接下来是解决非结构化数据的问题，这里的非结构化数据就是指网络上的论坛讨论、网络新闻等。一个人访问所有网站，并做记录肯定是不切实际的。这就需要网络爬虫技术。网络爬虫，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

Python学习网络爬虫主要分四大的版块：明确目标，抓取，分析，存储

1. 明确目标 (要知道你准备在哪个范围或者网站去搜索)
2. 抓取 (将所有的网站的内容全部爬下来)
3. 分析(去掉对我们没用处的数据)
4. 处理数据（按照我们想要的方式存储和使用）

以下是一个网络爬虫的示意图



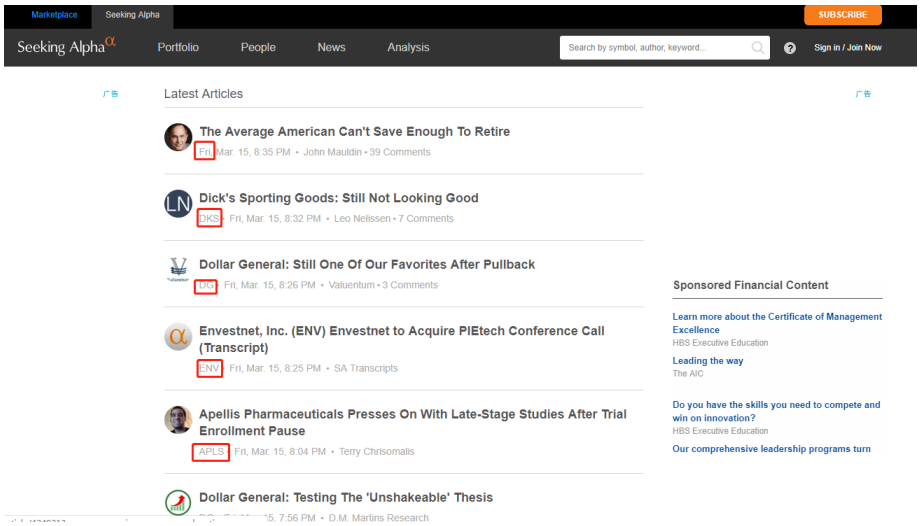
可以看出，网络爬虫主要模仿了人浏览网站的过程，然后在网页中，根据一定规则，选择所需的内容下载下来。

方法

我们主要利用python中的‘BeautifulSoup’、‘requests’、‘nltk’ 等“package”来进行网络爬虫。

首先，选择需要爬取的网站。这在里选取“seekingalpha.com”，Seeking Alpha是一家投资研究网站，网站基于自由投稿人对金融市场的讨论，网站成立于2004年，目前已有13,240位分析师，行业分析师的撰稿人，每月有120,000个新评论，报道了过12,780家来自70个国家的公司，总文章量达到691,441篇。拥有4

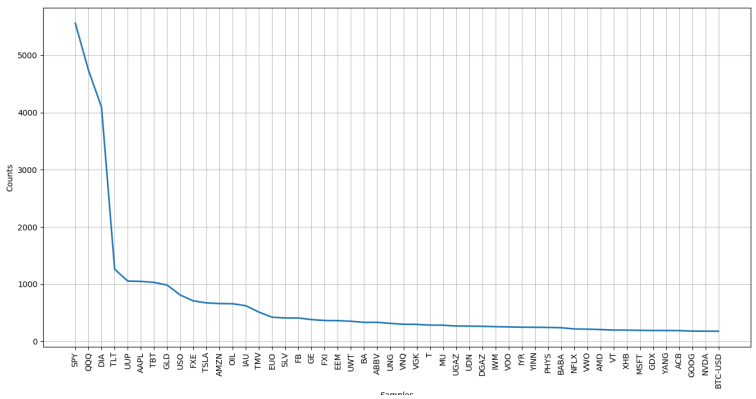
百万注册用户，14,147个博客。公司目前有150个员工。每月会吸引900万的访客。

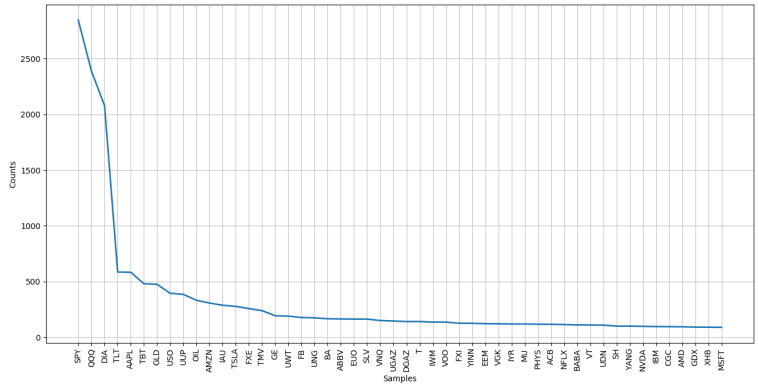


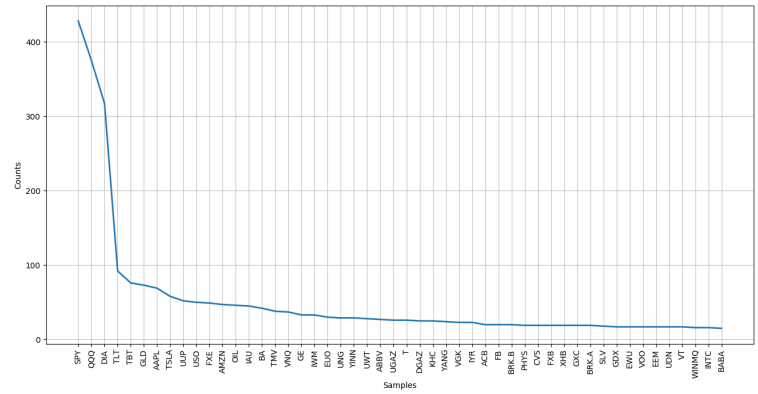
可以看出，在seekingalpha中analysis模块，有很多评论，并在题目下方有讨论股票的“symbol”。于是，我们很容易想到，通过爬虫获取讨论的“symbol”，计算每只股票出现的次数，将其存入“CSV”文件中，以衡量股票的交易活跃度。爬虫代码见附录。

结果

我分别爬取了前一百页（一个月）、五百页（半年）、一千页（一年）的analysis模块。取各前五十名得到以下表格，可见前四名，均为指数或ETF，这也在情理之中，我们需要在前五十名中找到符合已选择的板块的股票。







RANK	NAME(100)	COUNT	NAME(500)	COUNT	NAME(1000)	COUNT
1	SPY	428	SPY	2846	SPY	5562
2	QQQ	375	QQQ	2389	QQQ	4744
3	DIA	317	DIA	2078	DIA	4092
4	TLT	92	TLT	585	TLT	1261
5	TBT	76	AAPL	583	UUP	1053
6	GLD	73	TBT	480	AAPL	1048
7	AAPL	69	GLD	475	TBT	1031
8	TSLA	58	USO	395	GLD	983
9	UUP	52	UUP	385	USO	809
10	USO	50	OIL	332	FXE	706
11	FXE	49	AMZN	307	TSLA	672
12	AMZN	47	IAU	287	AMZN	660
13	OIL	46	TSLA	277	OIL	657
14	IAU	45	FXE	257	IAU	622
15	BA	42	TMV	239	TMV	510
16	TMV	38	GE	193	EUO	421
17	VNQ	37	UWT	190	SLV	409
18	GE	33	FB	177	FB	408

RANK	NAME(100)	COUNT	NAME(500)	COUNT	NAME(1000)	COUNT
19	IWM	33	UNG	174	GE	380
20	EUO	30	BA	166	FXI	364
21	UNG	29	ABBV	164	EEM	362
22	YINN	29	EUO	163	UWT	351
23	UWT	28	SLV	163	BA	332
24	ABBV	27	VNQ	150	ABBV	332
25	UGAZ	26	UGAZ	146	UNG	314
26	T	26	DGAZ	141	VNQ	298
27	DGAZ	25	T	141	VGK	296
28	KHC	25	IWM	136	T	284
29	YANG	24	VOO	136	MU	282
30	VGK	23	FXI	126	UGAZ	268
31	IYR	23	YINN	125	UDN	266
32	ACB	20	EEM	122	DGAZ	263
33	FB	20	VGK	121	IWM	256
34	BRK.B	20	IYR	119	VOO	252
35	PHYS	19	MU	119	IYR	247
36	CVS	19	PHYS	117	YINN	246
37	FXB	19	ACB	116	PHYS	243
38	XHB	19	NFLX	114	BABA	238
39	GXC	19	BABA	111	NFLX	217
40	BRK.A	19	VT	110	VWO	213
41	SLV	18	UDN	109	AMD	206
42	GDX	17	SH	100	VT	197
43	EWU	17	YANG	100	XHB	196
44	VOO	17	NVDA	98	MSFT	192
45	EEM	17	IBM	96	GDX	189
46	UDN	17	CGC	95	YANG	189
47	VT	17	AMD	94	ACB	187
48	WINMQ	16	GDX	91	GOOG	179
49	INTC	16	XHB	90	NVDA	177
50	BABA	15	MSFT	89	BTC-USD	177

之后，我们综合来看三个时间段，各个股票的讨论活跃度的变化。

NAME	RANK(100)	RANK(500)	RANK(1000)
SPY	1	1	1
QQQ	2	2	2

NAME	RANK(100)	RANK(500)	RANK(1000)
DIA	3	3	3
TLT	4	4	4
TBT	5	6	7
GLD	6	7	8
AAPL	7	5	6
TSLA	8	13	11
UUP	9	9	5
USO	10	8	9
FXE	11	14	10
AMZN	12	11	12
OIL	13	10	13
IAU	14	12	14
BA	15	20	23
TMV	16	15	15
VNQ	17	24	26
GE	18	16	19
IWM	19	28	33
EUO	20	22	16
UNG	21	19	25
YINN	22	31	36
UWT	23	17	22
ABBV	24	21	24
UGAZ	25	25	30
T	26	27	28
DGAZ	27	26	32
KHC	28	-	-
YANG	29	43	46
VGK	30	33	27
IYR	31	34	35
ACB	32	37	47
FB	33	18	18
BRK.B	34	-	-
PHYS	35	36	37
CVS	36	-	-
FXB	37	-	-
XHB	38	49	43
GXC	39	-	-

NAME	RANK(100)	RANK(500)	RANK(1000)
BRK.A	40	-	-
SLV	41	23	17
GDX	42	48	45
EWU	43	-	-
VOO	44	29	34
EEM	45	32	21
UDN	46	41	31
VT	47	40	42
WINMQ	48	-	-
INTC	49	-	-
BABA	50	39	38
FXI	-	30	20
MU	-	35	29
NFLX	-	38	39
SH	-	42	-
NVDA	-	44	49
IBM	-	45	-
CGC	-	46	-
AMD	-	47	41
MSFT	-	50	44
VWO	-	-	40
GOOG	-	-	48
BTC-USD	-	-	50

我们通过每列做差分析，得出在这一年时间里，YANG、ACB、IWM、YINN、VNQ、BA、UGAZ、DGAZ、XHB、UNG、IYR、TSLA、T、PHYS、GE、KHC、BRK.B、CVS、FXB、GXC、BRK.A、EWU、WINMQ、INTC等24只股票热度是具有上升趋势的。因此，就非结构化数据分析来看，我们选择了以上股票。

附录

```
import requests
from bs4 import BeautifulSoup
import re
import nltk
import matplotlib.pyplot as plt
import time
import random
import csv
```

```
word=[]




for i in range(100):
    url = "https://seekingalpha.com/latest-articles?
page="+str(i)
    headers = [
        "Mozilla/5.0 (Windows NT 6.3; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95
Safari/537.36",
        "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_2)
AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/35.0.1916.153 Safari/537.36",
        "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0)
Gecko/20100101 Firefox/30.0"
        "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_2)
AppleWebKit/537.75.14 (KHTML, like Gecko) Version/7.0.3
Safari/537.75.14",
        "Mozilla/5.0 (compatible; MSIE 10.0; Windows NT
6.2; Win64; x64; Trident/6.0)"]


    randdom_header = random.choice(headers)
    headers = {'User-Agent': randdom_header}
    page = requests.get(url,headers = headers)
    time.sleep(8)
    #print(randdom_header)
    print(page)
    soup = BeautifulSoup(page.text, "html.parser")

    # title=soup.find_all('a',{'class':"a-title"})

    symbol=soup.find_all('a',
{'href':re.compile('/symbol/*')})


    #print(symbol)
    #print(title)
    # for i in title:
    #     print(i.get_text())


    for s in symbol:
        #print(i.get_text())
        word.append(s.get_text())
```

```
freq = nltk.FreqDist(word)
out = open('data100.csv','a', newline='')
csv_write = csv.writer(out,dialect='excel')
for key,val in freq.items():
    #print (str(key) + ':' + str(val))
    data=[key,val]
    csv_write.writerow(data)
freq.plot(50, cumulative=False)
plt.show()
```