The Research of Stock Data Analysis

105033467 Shi-Yi Tan
Department of Industrial Engineering and Engineering Management
National Tsing Hua University, Taiwan

Word count：2547 words

Abstract. The stock market always has the wealth and risk. Make great fortune overnight always attract people especially young gays to follow and ignore the risk. The young guys wanted to be rich often labored at stock market and end with losing. Since that the stock price affected by very many unpredictable factors which make it impossible to predict the stock price. So how to predict the change trend of stock price effectively and make the right decision becomes very important research topic. Our research object is to find an effective method to get the data from the internet and mining the data which is “too big” to make you confuse. And then we will build a model basic on some stock technical analysis tools to simulate stock change trend. In the end we will use the analytics model to analysis the processed data and find the optimal solution to give you some advice to make the decision.

Keywords. Big data, Data Mining, Stock data analysis, Stock change trend model

Introduction

Stock plays an important role in human’s economic life, even more so in the digital age. With the advance of science and technology, people could easily trade stock on personal or cellphone at anywhere anytime even sometime on work or on duty. So how to get the stock information and make the decision become more and more important. But the stock price affected by very many unpredictable factors which make it impossible to predict the stock price. Sometime we joke that the information from analysis is not as right as information from inside the company. As a result, get the stock data and make data analysis became a fatal issue. This issue will be direct and effective economic benefits. It addresses the problem we use Python to get the data from the web. Further, store the data as a .txt document which can easily communicate with Matlab. Finally use Matlab to build a model to analysis the data and get the optimal solution. Basic on the solution we can the volatility of stocks which can help us to make decision.

Therefore, the aim of this study is to establish an effective method to give you some advice for trading stock. The paper is organized as follows. In chapter 1, we discuss the model of affect, the motivation and problem background for our research. Chapter 2 illustrates the methodology and the framework of this study. Experimental analysis are discussed in chapter 3. Conclusions and potential research issues for future study are given in chapter 4.

1. Motivation and problem background

With the development of network technology, data is ubiquitous in people's life. If the last generation we call it Internet age, the next age will belong Internet of Thing belong data. With data now becoming the primary source of knowledge, success and revenue, the need to store, analysis, view and understand it is crucial to any business regardless of size especially in today’s competitive market. Especially in stock market data is particularly important, since stock is that thing basic on lots of data itself. But without processing, data always data. From the data, we know the past, data help us determine what happened, why it happened and how often it happened, and data help us understanding the future, data drive predictions about what will happen next and whether it’s a trend. They can be used to develop worst-and best-case scenarios and outcomes, and determine the optimum action based on multiple scenarios (Chun-I P. Chen et al., 1996).

To sum up, what we want to stress is what is the data and why is the data so important. But how to get data, even though we live around with data, but the data you create is not belong to you, it’s belong to Google, Baidu, Facebook and so on expect yourself. The first problem you should face up to it’s that how can you get the data and store it. This step about process & store data we often call it “Big data”. Based on volume, variety and velocity, we may require Big data & related technologies. Big data & technology helps in reducing cost in processing volume of data and also making it feasible to do a few typically analysis. If you own the data, a feasible data, how can you use the data, what the insights the data tell you. This step about analyze & generate insights we call it “Data mining”. So you can see every problem we try to deal with, is a very popular topic, the process we handle this problem is meaningful. We can sum up the process as the graph show:

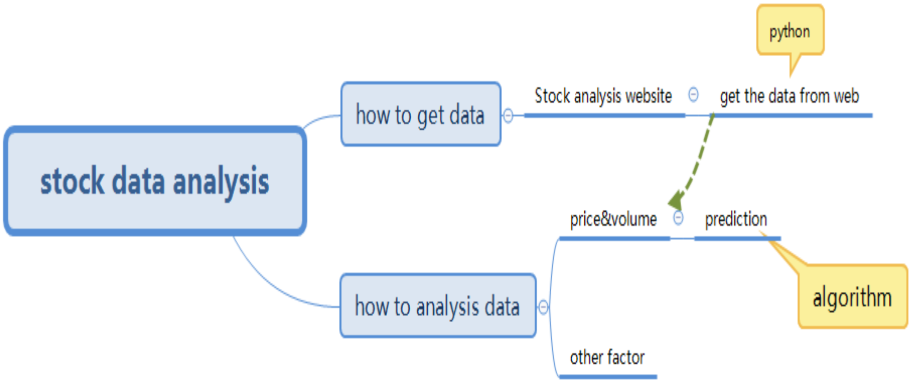


Figure 1. The structure of analysis to the problem

The process focu on this two problem: get data & analysis data. For getting the data, we should use python to get the data which represent price and volume from web. Through the model, by using algorithm we can predict the price. Since, many other factors can affect the price, so some warning must attend. To get the best decision, maybe we need use some algorithm to optimize. We talk detail method at chapter 2.

2. Method and analysis process

The aim of this paper is to establish an intelligent stock data analysis system, which could select appropriate data and use the data in a prediction model which object function is about the possibility of the stock increase to give you some advice to help you make decision since making decision is expensive.

The methodology in this paper is divided into two parts. Phase I “Big data”, use Python to get the data from the web, and process the data and make it feasible to do a few typically analysis.

Phase II is about “data mining”, since we get data we should use some technology to analysis data to establish a model to give the prediction.

2.1 Phase I: Get “big data” from the web

Data mining is a series of process that extracting interesting information from database, and processing, transforming, mining and evaluating. Data mining is a part of knowledge discovery in databases (KDD) (Fayyad et al., 1996). So the first thing we need to do is how to get data. The process sum up as the graph show:

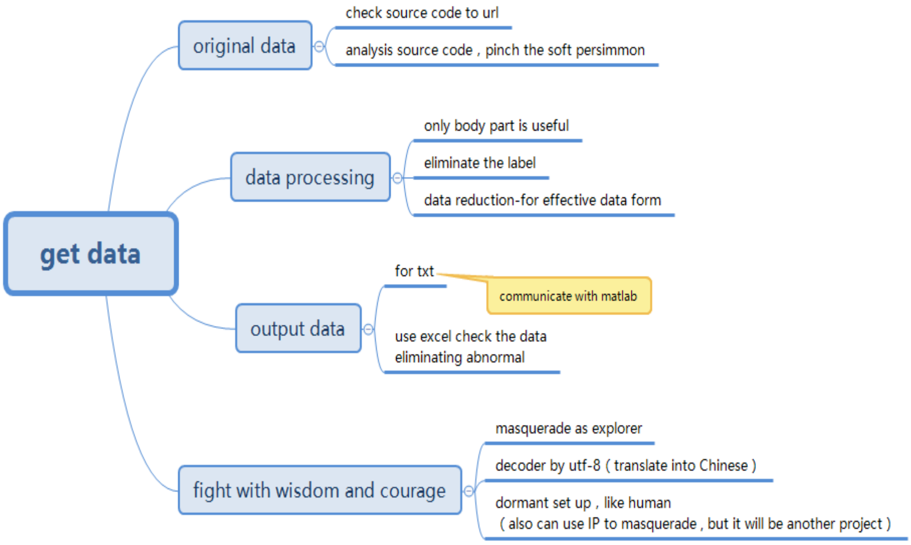

get data
original data
check source code to url
analysis source code , pinch the soft persimmon
data processing
only body part is useful
eliminate the label
data reduction-for effective data form
output data
for txt
communicate with matlab
use excel check the data
eliminating abnormal
fight with wisdom and courage
masquerade as explorer
decoder by utf-8 (translate into Chinese)
dormant set up , like human
(also can use IP to masquerade , but it will be another project)

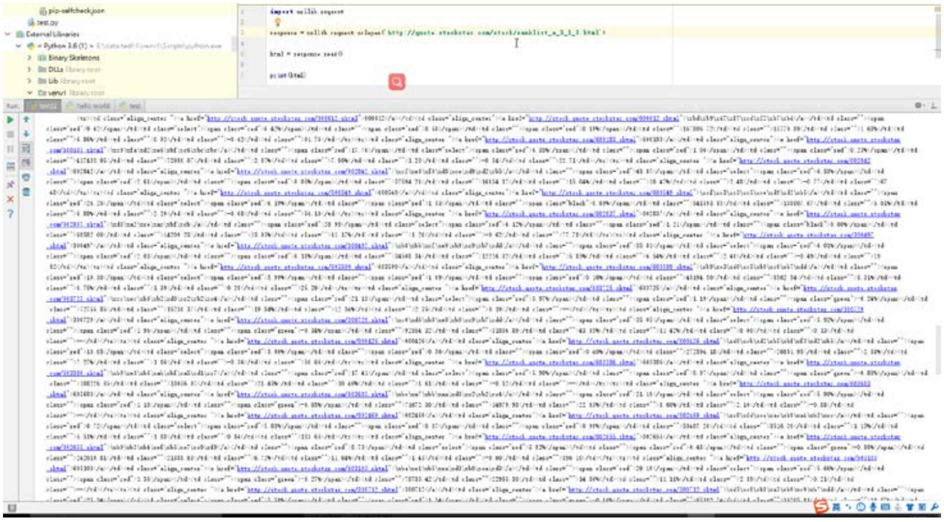
Figure 2. The process of part I get “big data”

Step 1: Original data

Trading stock is a important part in people’s economical life, and there are so many web give the real-time stock information. We can check the source code of the web, usually the form of the source code is .HTML or .CSS, from the source code information we get the URL of the web. There are some tips here, since there are so many web we can pinch the soft persimmom.

The simplest way to get the data from the web only need few row code, as show:

```
import urllib.request

respose =
urllib.request.get("http://quote.stockstar.com/stock/ranklist_a_3_1_3.html")

html =respose.read()

print(html)
```



```
import urllib.request

response = urllib.request.urlopen('http://quote.stockstar.com/stock/ranklist_a_3_1_3.html')

html = response.read()

print(html)
```

Figure 3. Result of simplest way to get data from web

Step 2: Data process

By using the simplest way to get data from web which was shown as Figure 3 contains too much redundant information and the form of the data is ASCII, machine language, we need pick up the useful information and decode the ASCII to the human language (ASCII → UTF - 8). A complete example is given in chapter 3.

Step 3: Output data

When we have lots of data it will still data only we output it and make it feasible to do a few typically analysis. So we output the data as a .TXT document which can communicate with Matlab. Before load data in Matlab, we need use a EXCEL for double check to eliminate abnormal data. Prevent trouble before it happens, will save lots of computing resources.

Step 4: Fight with wisdom and courage

Some web forbid Python code read the information and get data. That means if you want to download the data from the web, you should masquerade as explorer and link the web like human being. Even more the Python code should contain a dormant function, because of one can’t link the same web so often, so if you want you Python get data from web automatically and not to be found you using a Python code, you should need some dormant more like human behavior not that so automatically. There are some skills, which means you should get the data with wisdom and courage. But Illegal ACTS are strictly prohibited.

2.2. Phase II: Analysis the data “data mining”

Since we get data we should use some technology to analysis data to establish a model to give the prediction. Before “data mining”, we should check the data one more time, as the old saying, a good begin is half done. Then we can input the data to the analysis model, a simple representation of reality that helps us to understand how something works. At last, use some stock analysis tools to show the trend of the stock price and give a optimal solution, to help you to make a decision. The detail process is show as a graph:

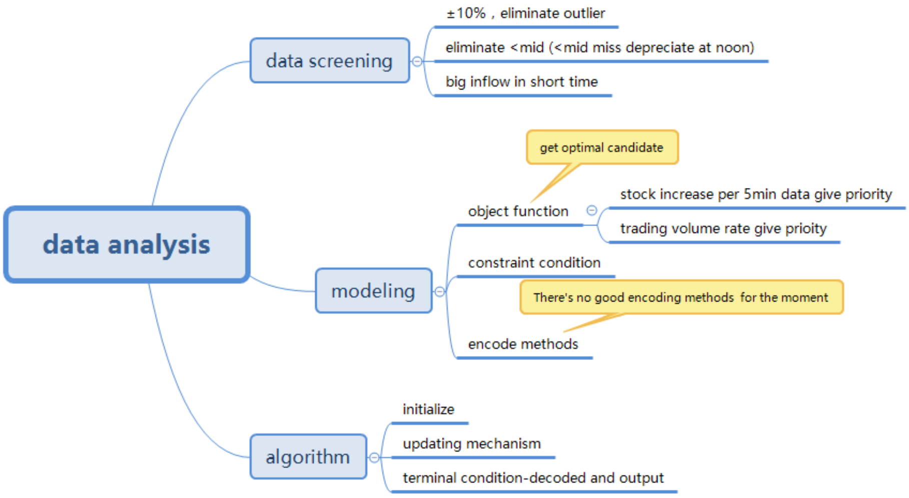

data analysis
data screening
±10%，eliminate outlier
eliminate <mid (<mid miss depreciate at noon)
big inflow in short time
modeling
get optimal candidate
object function
stock increase per 5min data give priority
trading volume rate give prioity
constraint condition
There's no good encoding methods for the moment
encode methods
algorithm
initialize
updating mechanism
terminal condition-decoded and output

Figure 4. Process of data analysis

Step 1: Data screening

Stock market at mainland China limit the increase and decrease range of the stock (±10%) expect some stock reenter the market which obey another row. So we eliminate outlier. And we statistics the frequency of each stock and make the mid number as a filtrate, if the frequency < mid which means the stock may depreciate at afternoon. Such stock usually decreases tomorrow.

Step 2: Modeling

At first, in address the problem, two methods were brought forward. One focus on the stock itself use algorithm to calculate each price possibility of the stock, get the optional price the stock tomorrow will be and the price range, so we can know the buy point and sale point. But mainland China stock market is T+1 mechanism of exchange, so this way has some defect can’t fit the market rule well. Another method is treat the whole stock market as a pool, we dig the most possible increase stock and buy it, sale tomorrow. Both method is show below:

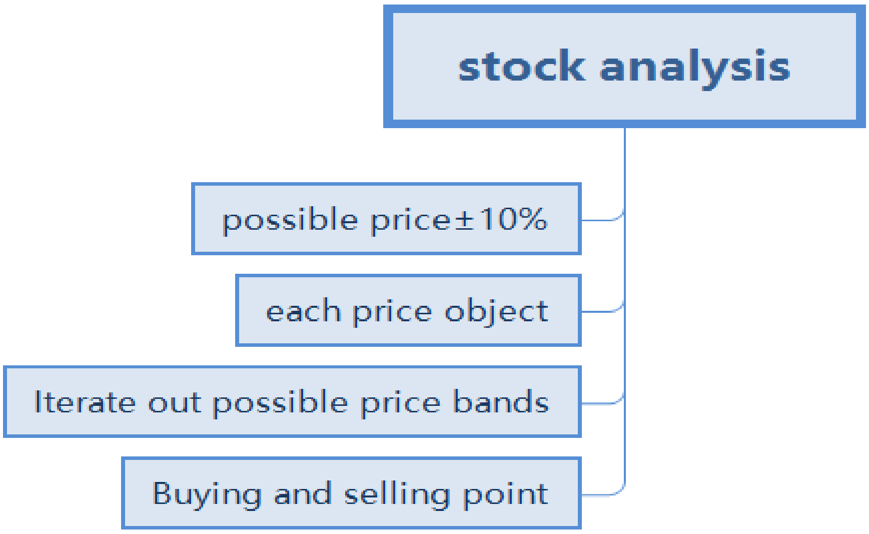


Figure 5. Stock data analysis method which focus on stock itself

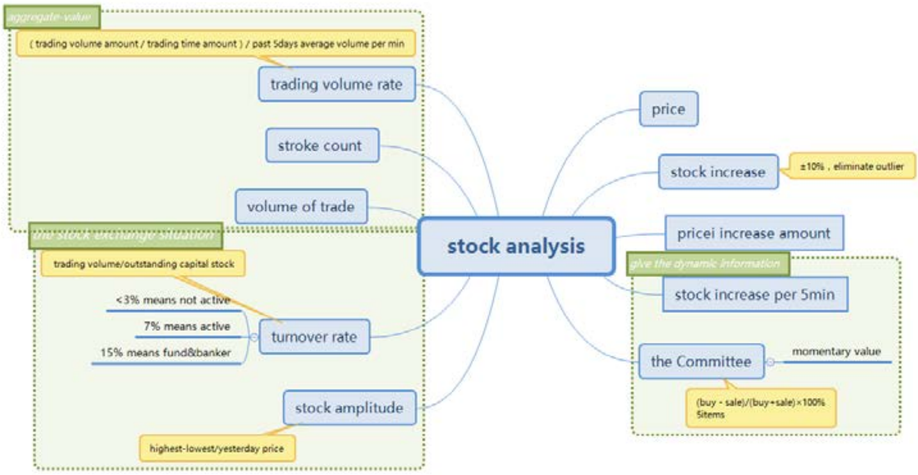


Figure 6. Stock data analysis method which treat the market as a pool to dig the "good" stock

The detail process would be illustrated in Chapter 3.

Step 3: Algorithm

Basic on method TWO, we build a model with the object function about increase possibility, the object function and constraints are shown below table:

$$
\begin{aligned}
\text{object: } z = {} & w_1 * rand_1 * \text{stock increase per 5min} + \\
& w_2 * rand_2 * \text{turnover rate} + \\
& w_3 * rand_3 * [\text{stroke count } / (\text{volume of trade} * \text{stock price})] - \\
& rand_4 * \text{stock amplitude} + \\
& rand_5 * \text{price} - \text{earnings ratio/stock increase}
\end{aligned} \qquad 2\text{-}1
$$

$$
turnover\ rate > 0.3 \qquad 2\text{-}2
$$

$$
stock\ increase \in (-0.1, 0.1) \qquad 2\text{-}3
$$

3. Case Study

3.1 Phase I: get “big data”

3.1.1 Step 1 original data

By check the source code of the stock web which provide timely and effective stock information, we choose the web “stock star” and get the URL, the source code of the “stock star” as shown below:

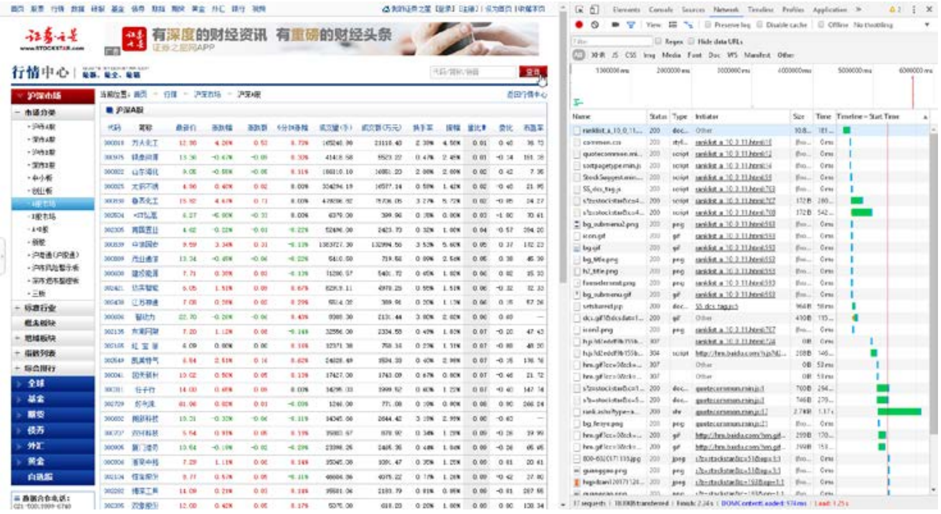
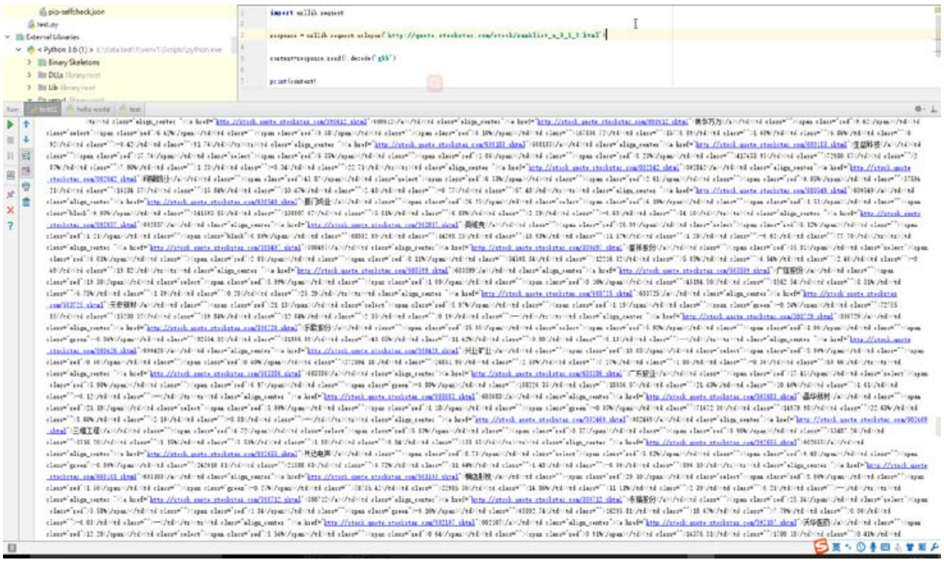


Figure 7. Source code of web “stock star”

3.1.2 Step 2 data process

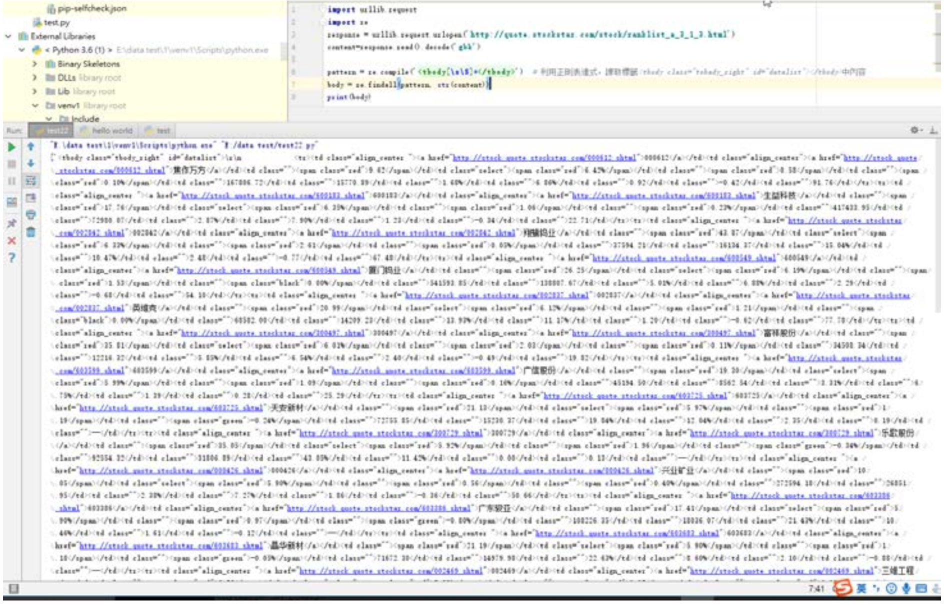
The URL of the web is http://quote.stockstar.com/stock/ranklist_a_3_1_1.html, and then we request the data form the server (which is shown as figure 3). And we need decode the data for store, the result is shown below:



```
import urllib.request

response = urllib.request.urlopen('http://quote.stockstar.com/stock/ranklist_a_3_1_1.html')

content=response.read().decode('gbk')

print(content)
```

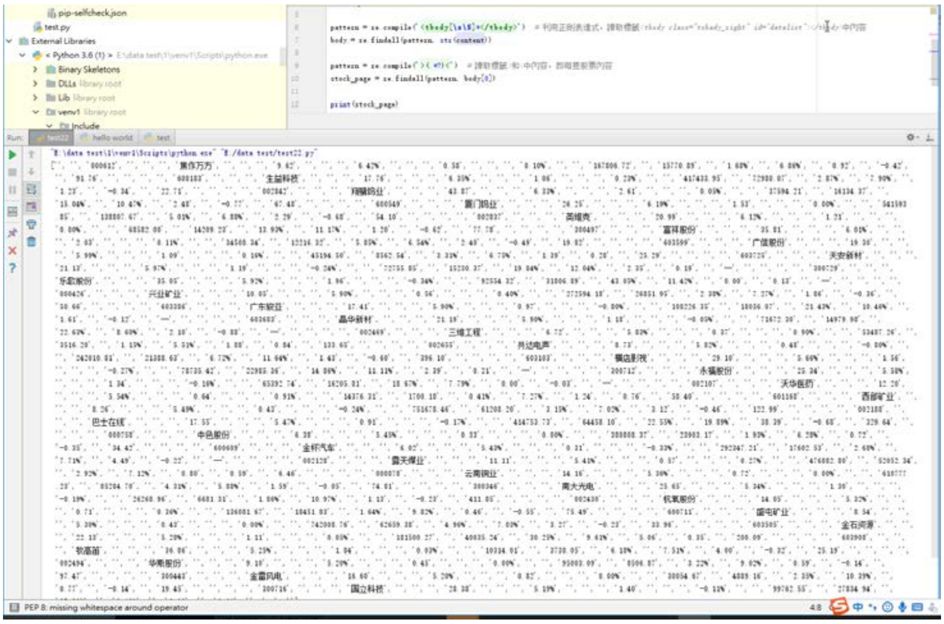
Figure 8. Result of decode

The data we need is between the two tags, so we use the function .re to data screen



```
import urllib.request
import re
response = urllib.request.urlopen('http://quote.stockstar.com/stock/ranklist_a_3_1_1.html')
content=response.read().decode('gbk')

pattern = re.compile('<tbody[\s\S]*</tbody>')  # 利用正则表达式，提取标签 <tbody class="tbody_right" id="datalist"></tbody> 中内容
body = re.findall(pattern, str(content))
print(body)
```

```
"E:\data test\1\venv1\Scripts\python.exe" "E:/data test/test22.py"
['<tbody class="tbody_right" id="datalist">\r\n            <tr><td class="align_center "><a href="http://stock.quote.stockstar.com/000612.shtml">000612</a></td><td class="align_center"><a href="http://stock.quote.stockstar.com/000612.shtml">焦作万方</a></td><td class=""><span class="red">9.62</span></td><td class="select"><span class="red">6.42%</span></td><td class=""><span class="red">0.58</span></td><td class=""><span class="red">0.10%</span></td><td class="">167006.72</td><td class="">15770.89</td><td class="">1.68%</td><td class="">6.06%</td><td class="">0.92</td><td class="">-0.42</td><td class="">91.76</td></tr><tr><td class="align_center "><a href="http://stock.quote.stockstar.com/600183.shtml">600183</a></td><td class="align_center"><a href="http://stock.quote.stockstar.com/600183.shtml">生益科技</a></td><td class=""><span class="red">17.76</span></td><td class="select"><span class="red">6.35%</span></td><td class=""><span class="red">1.06</span></td><td class=""><span class="red">0.23%</span></td><td class="">417433.95</td><td class="">72900.07</td><td class="">2.87%</td><td class="">7.90%</td><td class="">1.23</td><td class="">-0.34</td><td class="">22.71</td></tr><tr><td class="align_center "><a href="http://stock.quote.stockstar.com/002842.shtml">002842</a></td><td class="align_center"><a href="http://stock.quote.stockstar.com/002842.shtml">翔鹭钨业</a></td><td class=""><span class="red">43.87</span></td><td class="select"><span class="red">6.33%</span></td><td class=""><span class="red">2.61</span></td><td class=""><span class="red">0.05%</span></td><td class="">37594.21</td><td class="">16134.37</td><td class="">15.04%</td><td class="">10.47%</td><td class="">2.48</td><td class="">-0.77</td><td class="">67.48</td></tr><tr><td class="align_center "><a href="http://stock.quote.stockstar.com/600549.shtml">600549</a></td>
```



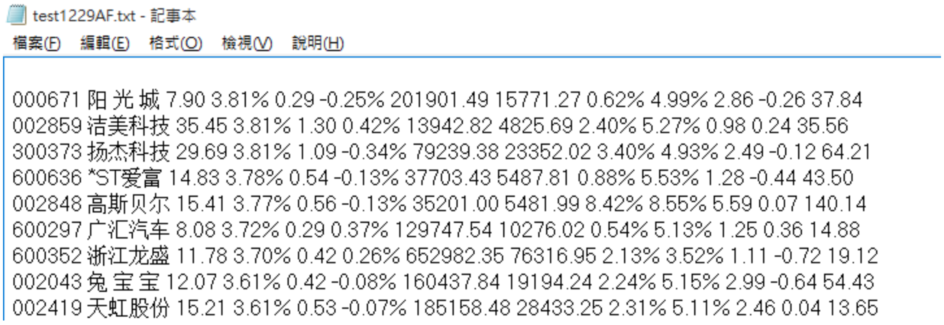
```
pattern = re.compile('<tbody[\s\S]*</tbody>')  # 利用正則表達式，讀取標籤<tbody class="tbody_right" id="datalist"></tbody>中內容
body = re.findall(pattern, str(content))

pattern = re.compile('>(.*?)<')  # 讀取標籤<td>和</td>中內容，即每筆股票內容
stock_page = re.findall(pattern, body[0])

print(stock_page)
```

```
"E:\data test\1\venv1\Scripts\python.exe" "E:/data test/test22.py"
```

Figure 9. Pick up the data

3.1.3 Step 3 output data

When we have lots of data it will still data only we output it and make it feasible to do a few typically analysis. So we output the data as a .TXT document which can communicate with Matlab. the results are shown as below:



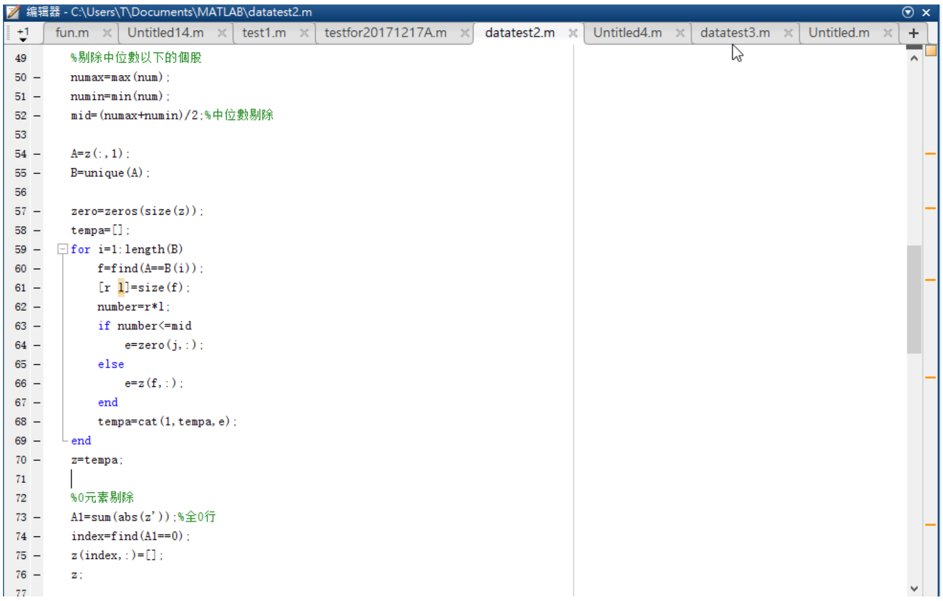
```
000671 阳 光 城 7.90 3.81% 0.29 -0.25% 201901.49 15771.27 0.62% 4.99% 2.86 -0.26 37.84
002859 洁美科技 35.45 3.81% 1.30 0.42% 13942.82 4825.69 2.40% 5.27% 0.98 0.24 35.56
300373 扬杰科技 29.69 3.81% 1.09 -0.34% 79239.38 23352.02 3.40% 4.93% 2.49 -0.12 64.21
600636 *ST爱富 14.83 3.78% 0.54 -0.13% 37703.43 5487.81 0.88% 5.53% 1.28 -0.44 43.50
002848 高斯贝尔 15.41 3.77% 0.56 -0.13% 35201.00 5481.99 8.42% 8.55% 5.59 0.07 140.14
600297 广汇汽车 8.08 3.72% 0.29 0.37% 129747.54 10276.02 0.54% 5.13% 1.25 0.36 14.88
600352 浙江龙盛 11.78 3.70% 0.42 0.26% 652982.35 76316.95 2.13% 3.52% 1.11 -0.72 19.12
002043 兔 宝 宝 12.07 3.61% 0.42 -0.08% 160437.84 19194.24 2.24% 5.15% 2.99 -0.64 54.43
002419 天虹股份 15.21 3.61% 0.53 -0.07% 185158.48 28433.25 2.31% 5.11% 2.46 0.04 13.65
```

Figure 10. Output data

3.2 Phase I: “data mining”

3.2.1 Step 1 data process

Even we have a .TXT document, when we input the data into Matlab, we should have some skill (such as mid number, change the data form and so on) to process the data, to reduce the computation.

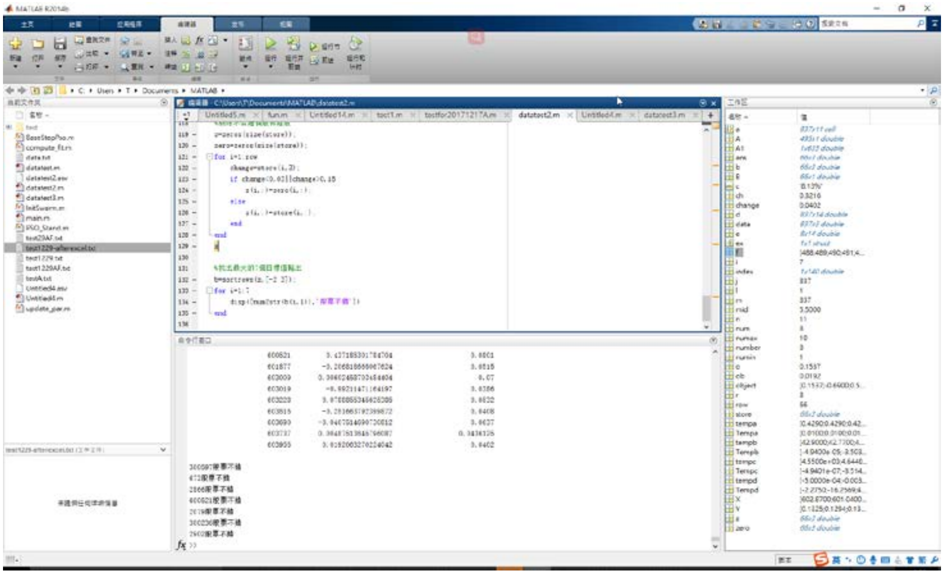


```
%剔除中位數以下的個股
numax=max(num);
numin=min(num);
mid=(numax+numin)/2;%中位數剔除

A=z(:,1);
B=unique(A);

zero=zeros(size(z));
tempa=[];
for i=1:length(B)
    f=find(A==B(i));
    [r l]=size(f);
    number=r*l;
    if number<=mid
        e=zero(j,:);
    else
        e=z(f,:);
    end
    tempa=cat(1,tempa,e);
end
z=tempa;

%0元素剔除
A1=sum(abs(z'));%全0行
index=find(A1==0);
z(index,:)=[];
z;
```

Figure 11. Data process

3.2.2 Step 2 optimal solution

We build a model which treat the whole stock market as a pool, we dig the most possible increase stock and buy it, sale tomorrow. We pick up the optimal solution by use the object function and limit by the constraints we have mentioned. A simple result is shown below:



```
z=zeros(size(store));
zero=zeros(size(store));
for i=1:row
    change=store(i,2);
    if change<0.02||change>0.15
        z(i,:)=zero(i,:);
    else
        z(i,:)=store(i,:);
    end
end
z

%找出最大的7個目標值輸出
b=sortrows(z,[-2 2]);
for i=1:7
    disp([num2str(b(i,1)),'股票不錯'])
end
```

```
          600521     0.427185301784704          0.0601
          601877    -0.206818666067624          0.0515
          603009     0.386624687024544            0.07
          603019    -0.99211471164197           0.0386
          603223     0.970006534462630          0.0622
          603515    -0.231663792399872          0.0408
          603690    -0.946781469907?0012        0.0627
          603737     0.364878138467960?7     0.0438126
          603955     0.0152063270224042         0.0402

300097股票不錯
672股票不錯
2866股票不錯
600621股票不錯
2079股票不錯
300236股票不錯
2902股票不錯
>>
```

Figure 12. Test result

4. Conclusions and future work

This research aims at building a model to analysis the data and get the optimal solution. Basic on the solution we can the volatility of stocks which can help us to make decision. In address this problem we depart it into two parts: get data & analysis data. For getting the data, we should use python to get the data which represent price and volume from web. Through the model, by using algorithm we can predict the price. Since, many other factors can affect the price, so some warning must attend. To get the best decision, maybe we need use some algorithm to optimize. We propose detail process in order to promote the process when we solve the other problem.

But there are still many shortcomings that are not being considered, such as “Black Swan event”, “trading psychology”. Also the object function is not as real as the stock market. In the future, our research hopes to extend to more detail and reliable to build a effective prediction model. Also, we imagine that we can use the mind from block chain technology. If more and more people use the similar system which parent model is our model and share the data and solution with each other, then we can really predict the stock very detail. Because of that we believe the solution increase and share out will make sure the result. But most important is the mind behind the process.

References

[1] Pentaho Business Analytics, 2012. <http://www.pentaho.com/explore/pentaho-business-analytics/>.

[2] Deng Cai, Xiaofei He, Jiawei HanSrda: an efficient algorithm for large-scale discriminant analysis IEEE Trans. Knowl. Data Eng., 20 (1) (2008), pp. 1-12 View Record in Scopus

[3] Hsu, Y-W. and Chiu, M-C. "Investigating the Relationship between Therapeutic Music and Emotion: A Pilot Study on Healthcare Services," International Conference on Concurrent Engineering 2014 (CE 2014), Sept. 8-11, Beijing, China, 2014

[4] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences 275 (2014): 314-347

[5] What is big data? https://www.oracle.com/big-data/index.html

[6] 曹海燕. 网页爬虫系统的设计[J]. 中国科技博览, 2015(5):226-226.

[7] 刘寿臣. 网页爬虫技术的关键技术研究探索[J]. 电脑知识与技术, 2016, 12(6X):16-17.

[8] 郝以珍. 基于页面分析的网络爬虫系统的设计与实现[D]. 硕士学位论文]. 武汉: 华中科技大学, 2012.

[9] Python 開發簡單爬蟲 1 課程介紹 https://www.youtube.com/watch?v=lkrpGhSfKRk&list=PLO5e_-yXpYLAYi9W9n4FukZJR_fEHqwtt

[10] 什麼是網路爬蟲？ https://www.youtube.com/watch?v=ceUhb2-gYOU&list=PLohb4k71XnPaQRTvKW4Uii1oq-JPGpwWF

.