World Stock Index & Futures Prediction By Sentimental Analysis

Miles Lee 李明叡 Allen Chen 陳俊安
Tony Tsai 蔡騏丞 Michiya Chu 朱燕玲

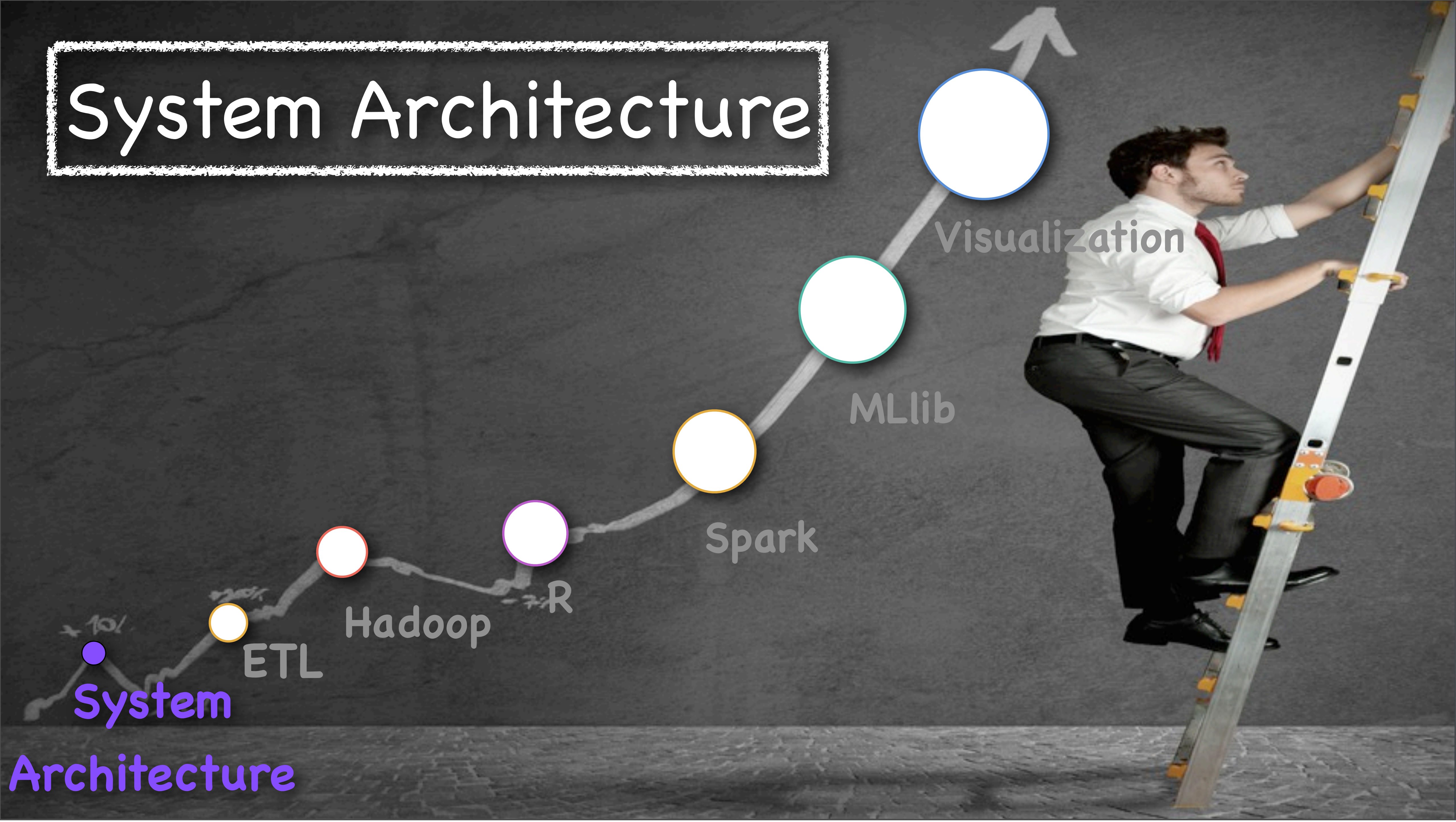
Topic

Financial markets aren't purely rational. Emotions play a large part in asset pricing. When people facing the financial markets and make their own strategy. They always think they're really rational. But is that true? We think that's impossible for a people to make a strategy without any personal emotion. How about let “market” decide their own emotion? We choose several algorithm and several training system(Hadoop & Spark) to improve our view. Let's check it !!

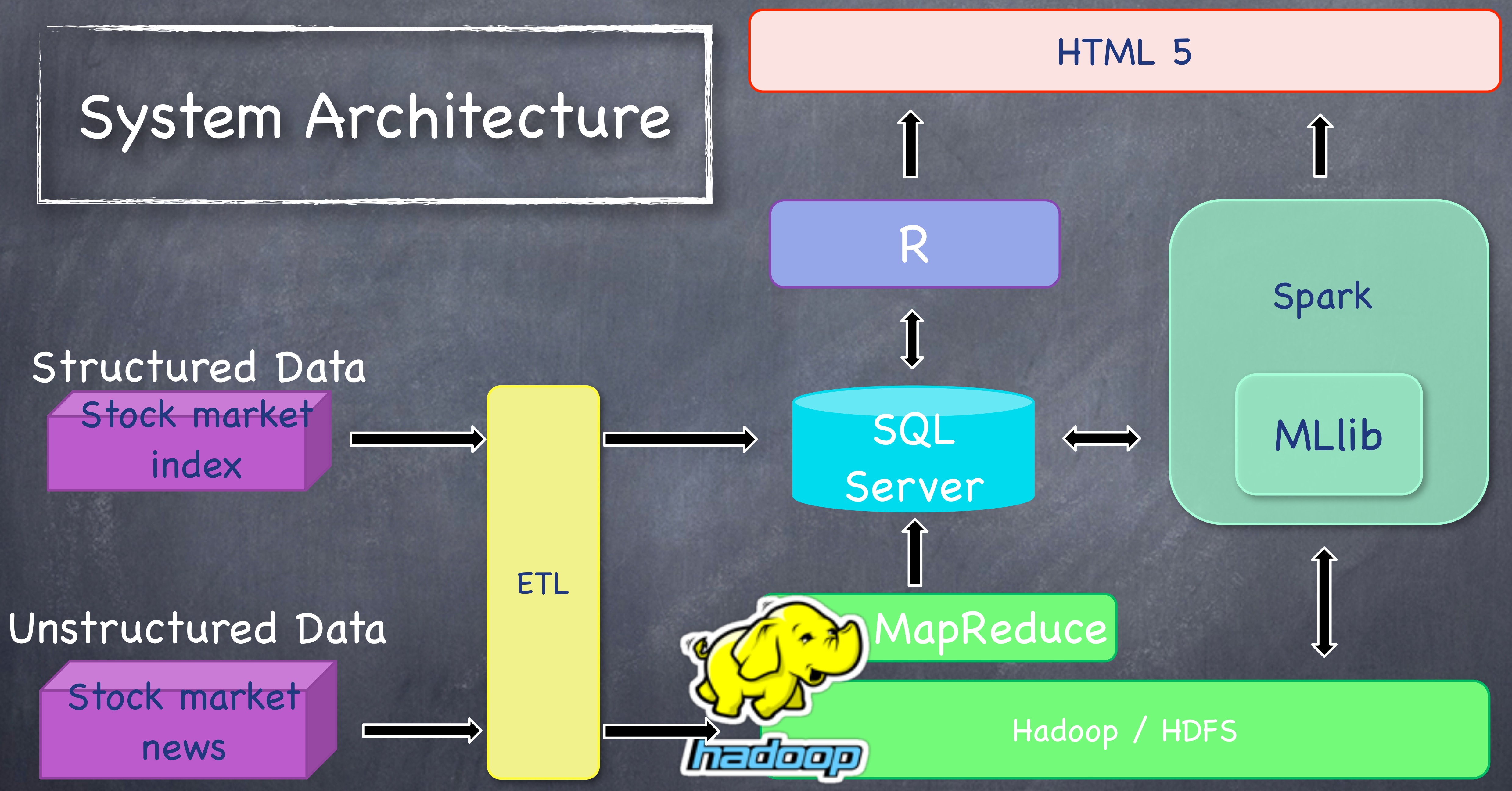
Job Allocation

Miles Lee	Tony Tsai	Allen Chen	Michiya Chu
1.Architecture design 2.Hadoop&Spark deployment 3.MapReduce 4.Spark code 5.Modeling 6.Visualization 7.SQL	1.ETL 2.Hadoop deployment 3.MapReduce 4.Visualization 5.R-Modeling	1.ETL 2.Powerpoint design 3.JDBC	1.Visualization 2.Data study 3.SQL

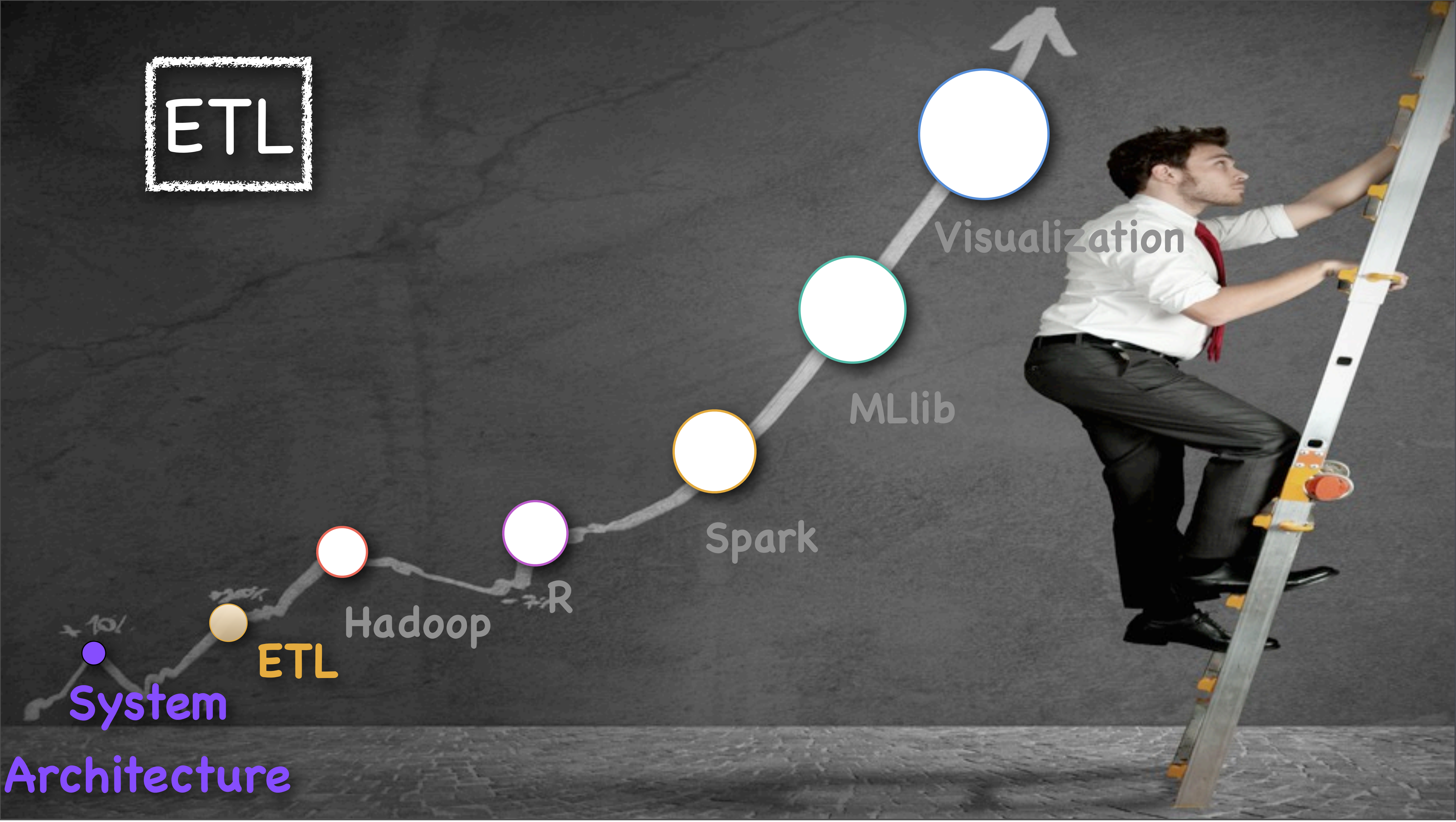
System Architecture



System Architecture



ETL



System Architecture

ETL

Hadoop

R

Spark

MLlib

Visualization

Data Come From

結構化資料

StockQ.org

資料訊息: 國際股市指數

資料時間: 2007 - 2014(即時資料)

非結構化資料

Google news

FINANCIAL TIMES

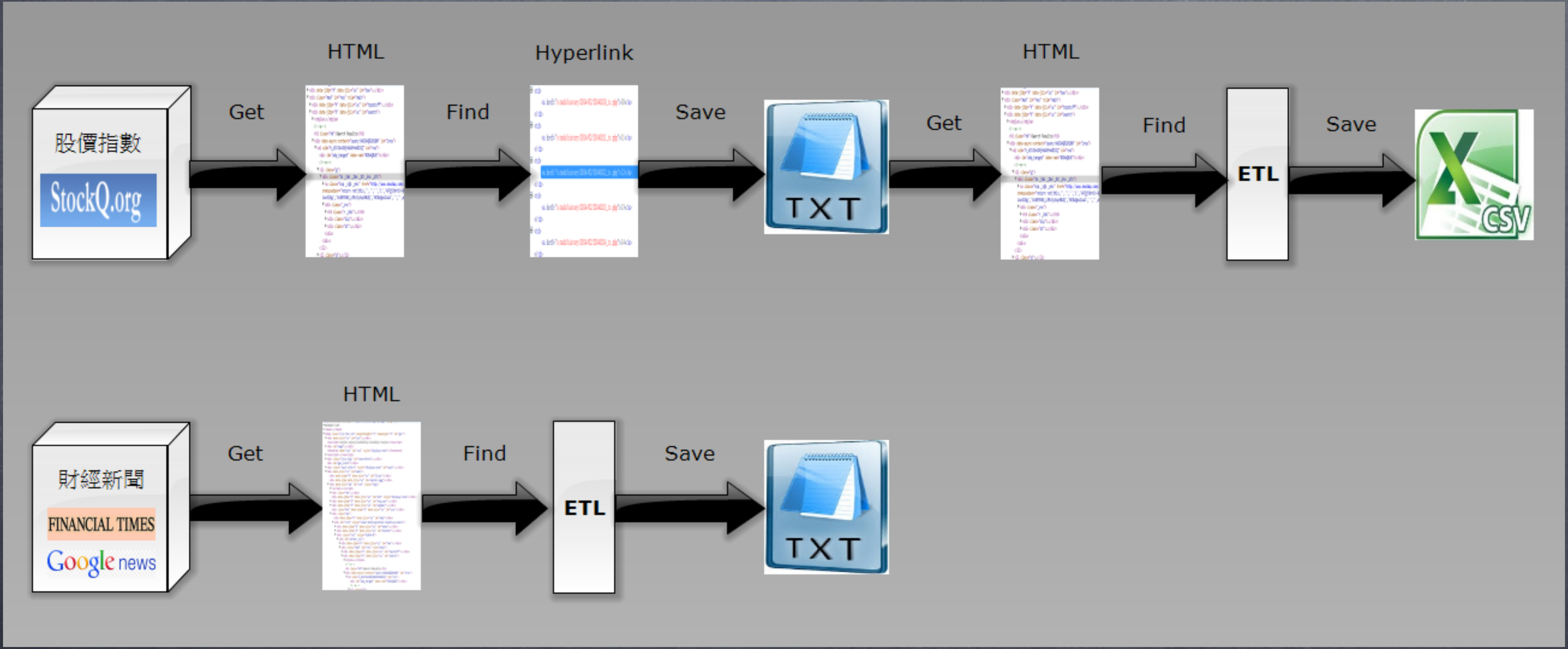
資料訊息: 國際主要股市新聞摘要 (即時)

資料訊息: 國際主要股市新聞摘要 (歷史)

資料時間: 2007 - 2014

新聞數量: 約35000篇

ETL Framework



Python ETL

Python

- 使用Python 來做網路資訊擷取的工具
- 語法簡潔及對於網頁擷取有提供完整的Library，可有效率的開發程式 。
- 對於文字的處理上較為方便與直覺

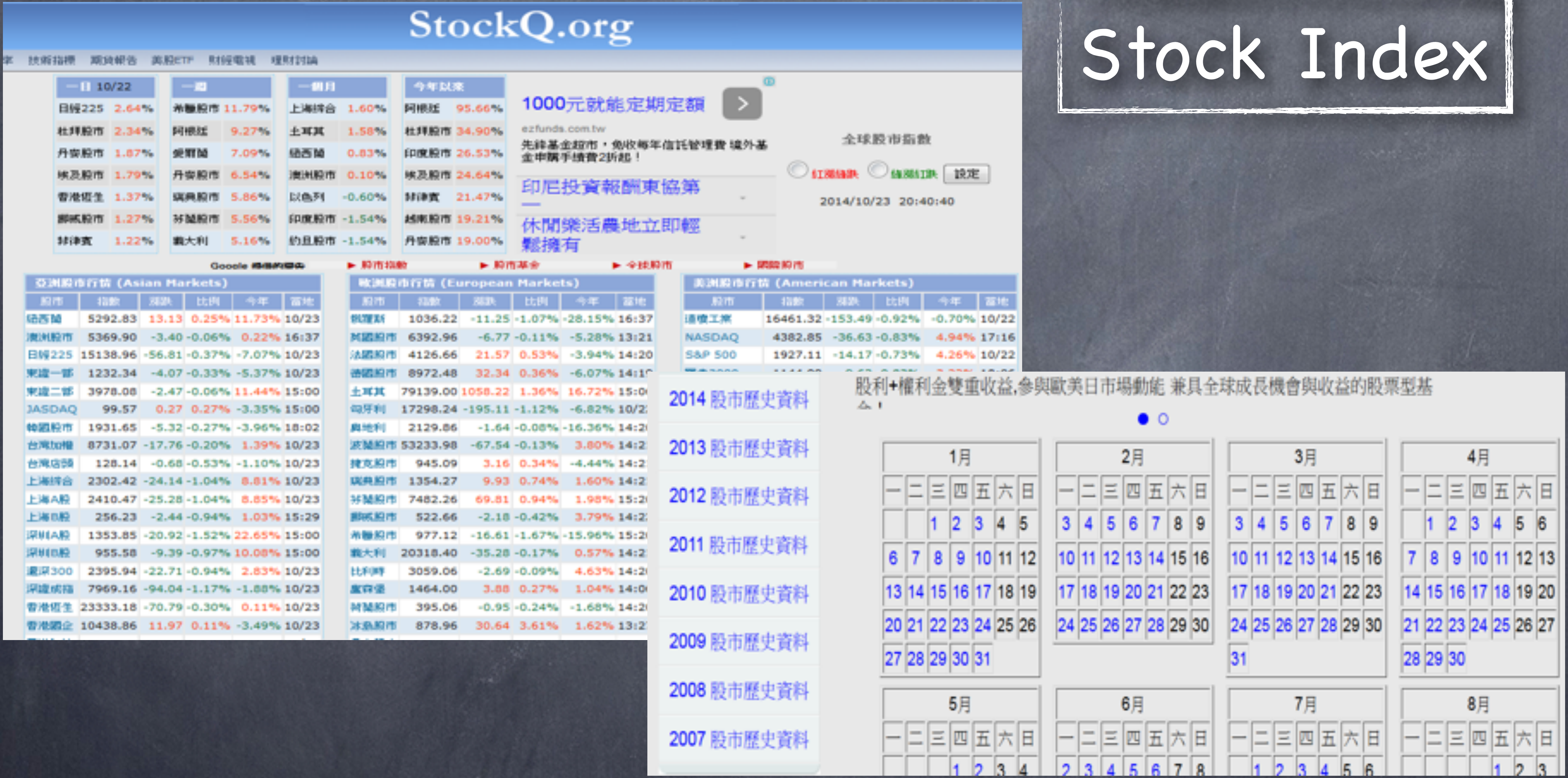
股市指數

- 主要收集各國股市、原物料、證券、債券、匯率、期貨等每日指數

股市資料的處理

- 每一檔標的名稱為檔案名稱
- 每一個檔案名稱裡有2007年以後的股價資料
- 儲存檔案為.CSV格式以方便匯入RDBMS

Stock Index



StockQ.org

率 技術指標 期貨報告 美股ETF 財經電視 理財討論

一日 10/22	
日經225	2.64%
杜拜股市	2.34%
丹麥股市	1.87%
埃及股市	1.79%
香港恆生	1.37%
挪威股市	1.27%
菲律賓	1.22%

一週	
希臘股市	11.79%
阿根廷	9.27%
愛爾蘭	7.09%
丹麥股市	6.54%
瑞典股市	5.86%
芬蘭股市	5.56%
義大利	5.16%

一個月	
上海綜合	1.60%
土耳其	1.58%
紐西蘭	0.83%
澳洲股市	0.10%
以色列	-0.60%
印度股市	-1.54%
約旦股市	-1.54%

今年以來	
阿根廷	95.66%
杜拜股市	34.90%
印度股市	26.53%
埃及股市	24.64%
菲律賓	21.47%
越南股市	19.21%
丹麥股市	19.00%

全球股市指數

○ 紅漲綠跌 ○ 綠漲紅跌 設定

2014/10/23 20:40:40

► 股市指數 ► 股市基金 ► 全球股市 ► 國際股市

亞洲股市行情 (Asian Markets)

股市	指數	漲跌	比例	今年	當地
紐西蘭	5292.83	13.13	0.25%	11.73%	10/23
澳洲股市	5369.90	-3.40	-0.06%	0.22%	16:37
日經225	15138.96	-56.81	-0.37%	-7.07%	10/23
東證一部	1232.34	-4.07	-0.33%	-5.37%	10/23
東證二部	3978.08	-2.47	-0.06%	11.44%	15:00
JASDAQ	99.57	0.27	0.27%	-3.35%	15:00
韓國股市	1931.65	-5.32	-0.27%	-3.96%	18:02
台灣加權	8731.07	-17.76	-0.20%	1.39%	10/23
台灣店頭	128.14	-0.68	-0.53%	-1.10%	10/23
上海綜合	2302.42	-24.14	-1.04%	8.81%	10/23
上海A股	2410.47	-25.28	-1.04%	8.85%	10/23
上海B股	256.23	-2.44	-0.94%	1.03%	15:29
深圳A股	1353.85	-20.92	-1.52%	22.65%	15:00
深圳B股	955.58	-9.39	-0.97%	10.08%	15:00
滬深300	2395.94	-22.71	-0.94%	2.83%	10/23
深證成指	7969.16	-94.04	-1.17%	-1.88%	10/23
香港恆生	23333.18	-70.79	-0.30%	0.11%	10/23
香港國企	10438.86	11.97	0.11%	-3.49%	10/23

歐洲股市行情 (European Markets)

股市	指數	漲跌	比例	今年	當地
俄羅斯	1036.22	-11.25	-1.07%	-28.15%	16:37
英國股市	6392.96	-6.77	-0.11%	-5.28%	13:21
法國股市	4126.66	21.57	0.53%	-3.94%	14:20
德國股市	8972.48	32.34	0.36%	-6.07%	14:1
土耳其	79139.00	1058.22	1.36%	16.72%	15:0
匈牙利	17298.24	-195.11	-1.12%	-6.82%	10/2
奧地利	2129.86	-1.64	-0.08%	-16.36%	14:2
波蘭股市	53233.98	-67.54	-0.13%	3.80%	14:2
捷克股市	945.09	3.16	0.34%	-4.44%	14:2
瑞典股市	1354.27	9.93	0.74%	1.60%	14:2
芬蘭股市	7482.26	69.81	0.94%	1.98%	15:2
挪威股市	522.66	-2.18	-0.42%	3.79%	14:2
希臘股市	977.12	-16.61	-1.67%	-15.96%	15:2
義大利	20318.40	-35.28	-0.17%	0.57%	14:2
比利時	3059.06	-2.69	-0.09%	4.63%	14:2
盧森堡	1464.00	3.88	0.27%	1.04%	14:0
荷蘭股市	395.06	-0.95	-0.24%	-1.68%	14:2
冰島股市	878.96	30.64	3.61%	1.62%	13:2

美洲股市行情 (American Markets)

股市	指數	漲跌	比例	今年	當地
道瓊工業	16461.32	-153.49	-0.92%	-0.70%	10/22
NASDAQ	4382.85	-36.63	-0.83%	4.94%	17:16
S&P 500	1927.11	-14.17	-0.73%	4.26%	10/22

2014 股市歷史資料
2013 股市歷史資料
2012 股市歷史資料
2011 股市歷史資料
2010 股市歷史資料
2009 股市歷史資料
2008 股市歷史資料
2007 股市歷史資料

● ○

1月						
一	二	三	四	五	六	日
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

2月						
一	二	三	四	五	六	日
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

3月						
一	二	三	四	五	六	日
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

4月						
一	二	三	四	五	六	日
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

5月						
一	二	三	四	五	六	日
			1	2	3	4

6月						
一	二	三	四	五	六	日
2	3	4	5	6	7	8

7月						
一	二	三	四	五	六	日
	1	2	3	4	5	6

8月						
一	二	三	四	五	六	日
				1	2	3

International Stock News

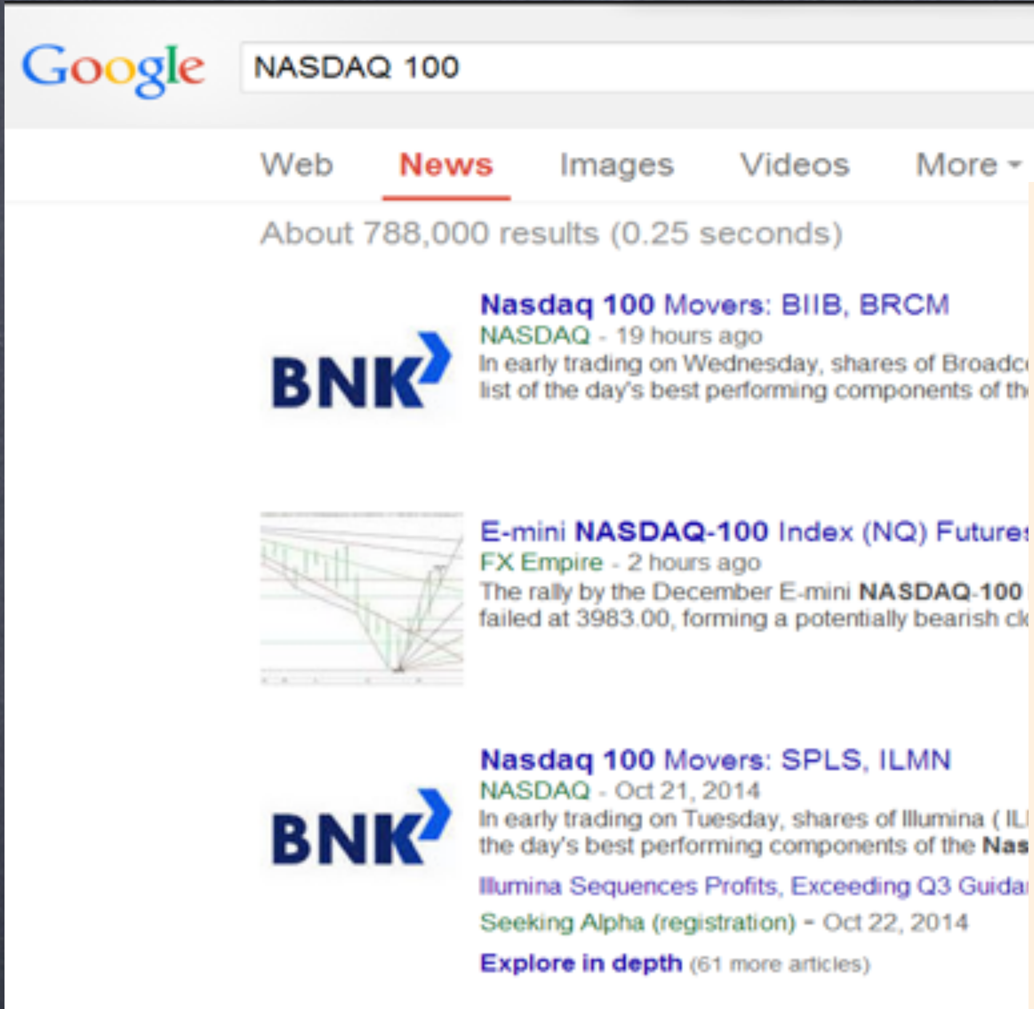
新聞內容

- 資料收集的內容為標題與摘要
- 擷取世界八大影響力較大的股市新聞，藉由這個影響力去預測有哪些股市受到連動的影響
- 資料設定以英文為主以符合國際情勢與即時性

新聞資料的處理

- 八大新聞標題與摘要以一天為單位儲存於文字檔
- 只擷取標題與摘要目的是因為摘要會顯示出對於主題所描述的關鍵字
  ex. Return, rise , down
- 擷取的資料要精準，盡量不必要的字詞可以優先忽略，以視為雜訊
- 在一開始就忽略雜訊，有助於作文章判斷正面與負面的準確度，進而提升預測的準確度

Stock News Is Look Like...



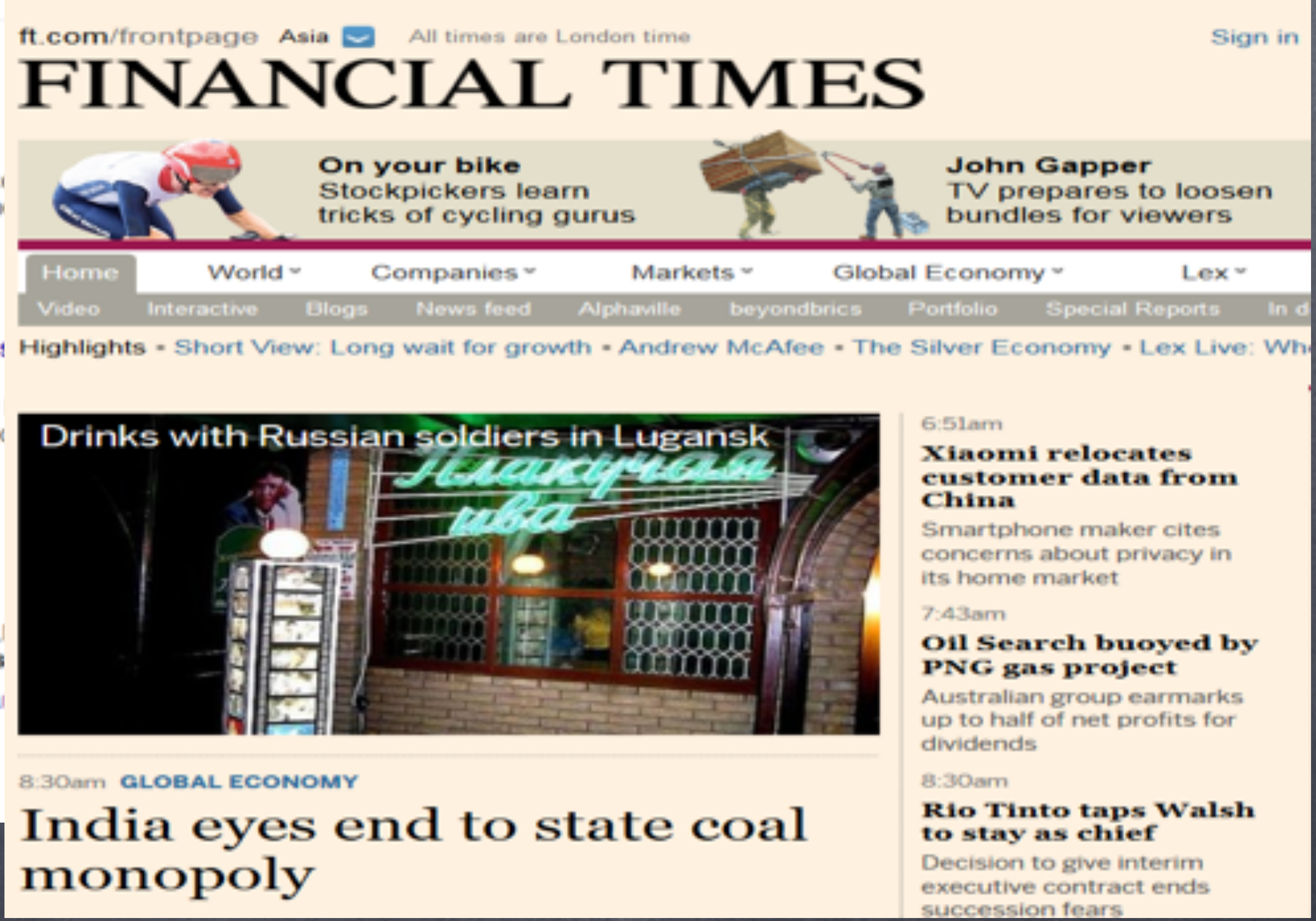
Google NASDAQ 100

About 788,000 results (0.25 seconds)

Nasdaq 100 Movers: BIIB, BRCM
NASDAQ - 19 hours ago
In early trading on Wednesday, shares of Broadc
list of the day's best performing components of th

E-mini NASDAQ-100 Index (NQ) Futures
FX Empire - 2 hours ago
The rally by the December E-mini **NASDAQ-100**
failed at 3983.00, forming a potentially bearish cl

Nasdaq 100 Movers: SPLS, ILMN
NASDAQ - Oct 21, 2014
In early trading on Tuesday, shares of Illumina (IL
the day's best performing components of the Nas
Illumina Sequences Profits, Exceeding Q3 Guida
Seeking Alpha (registration) - Oct 22, 2014
Explore in depth (61 more articles)



ft.com/frontpage Asia All times are London time Sign in

FINANCIAL TIMES

On your bike
Stockpickers learn tricks of cycling gurus

John Gapper
TV prepares to loosen bundles for viewers

Highlights - Short View: Long wait for growth - Andrew McAfee - The Silver Economy - Lex Live: Wh

Drinks with Russian soldiers in Lugansk

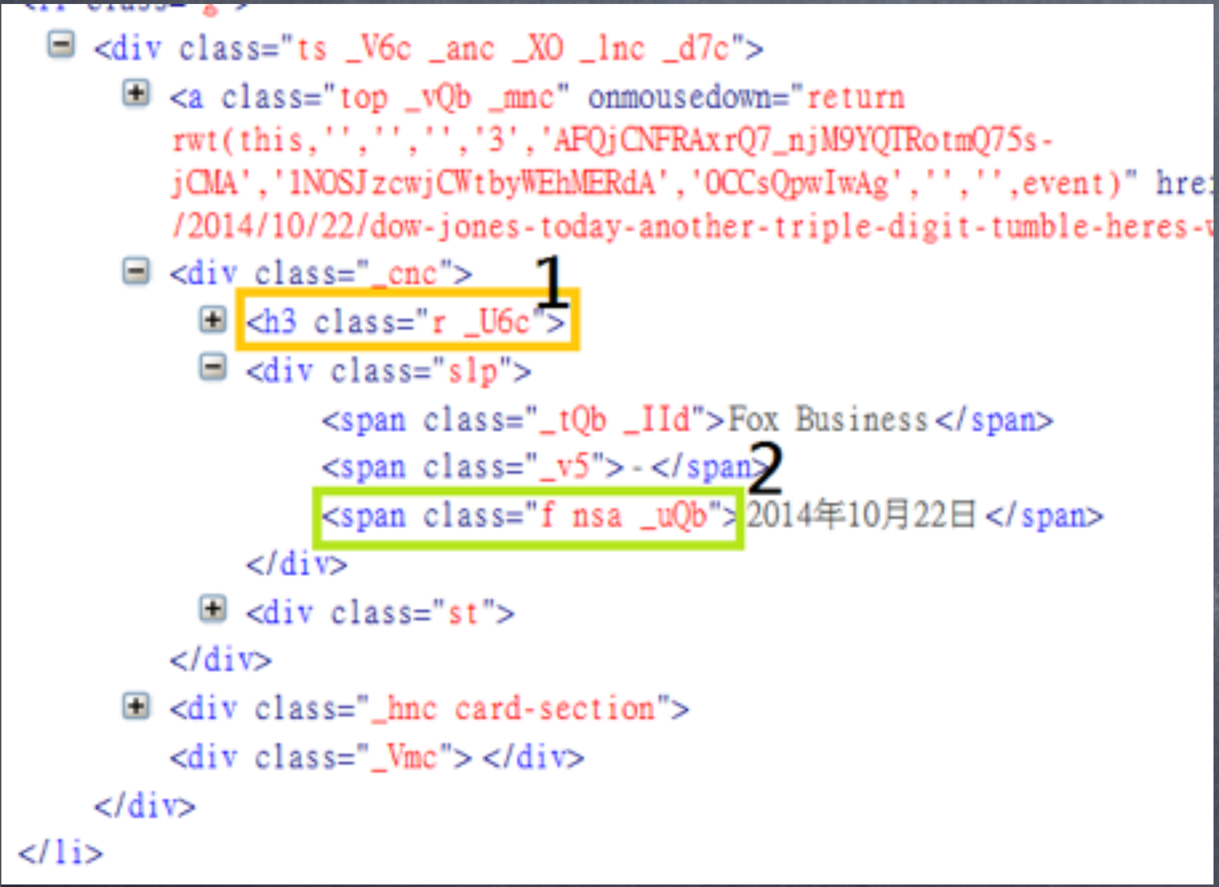
8:30am GLOBAL ECONOMY
India eyes end to state coal monopoly

6:51am
Xiaomi relocates customer data from China
Smartphone maker cites concerns about privacy in its home market

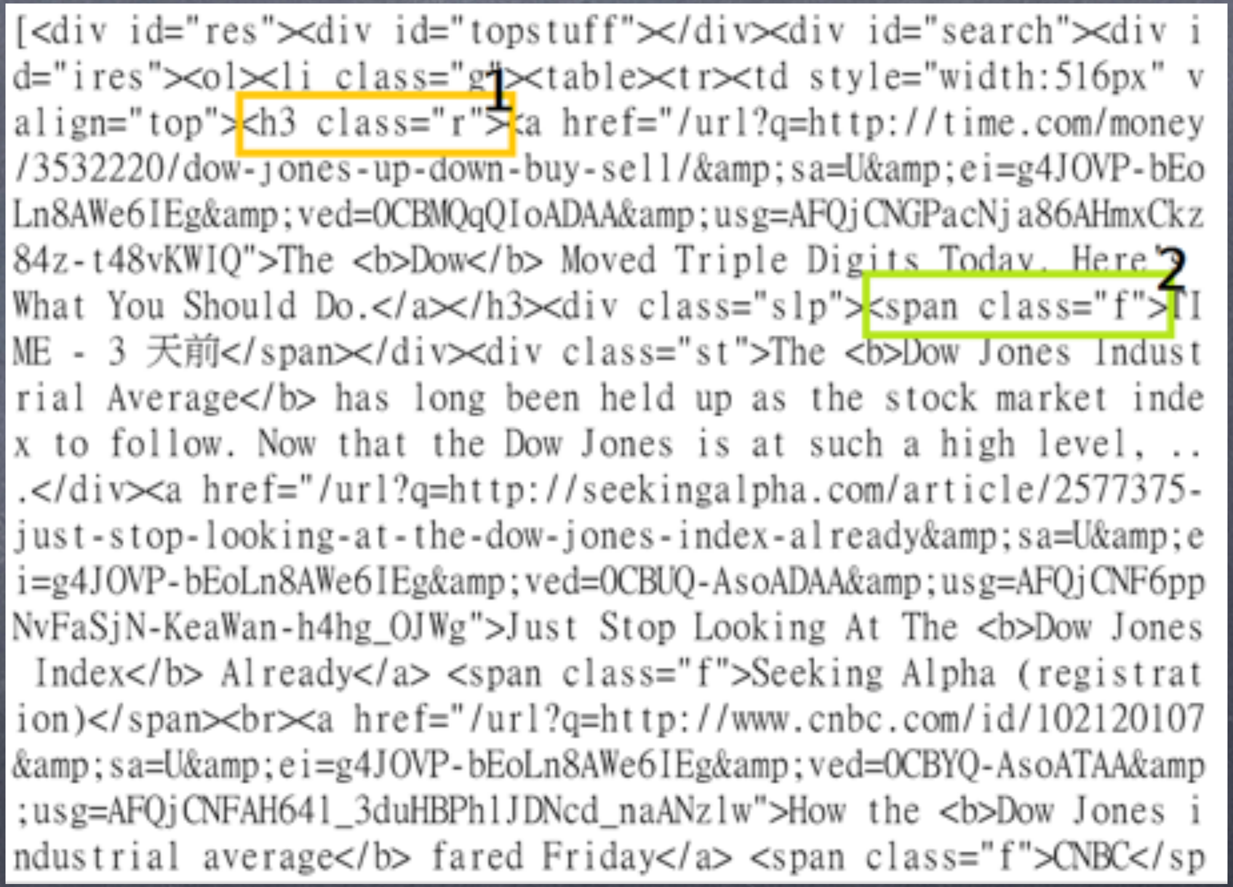
7:43am
Oil Search buoyed by PNG gas project
Australian group earmarks up to half of net profits for dividends

8:30am
Rio Tinto taps Walsh to stay as chief
Decision to give interim executive contract ends succession fears

Google news

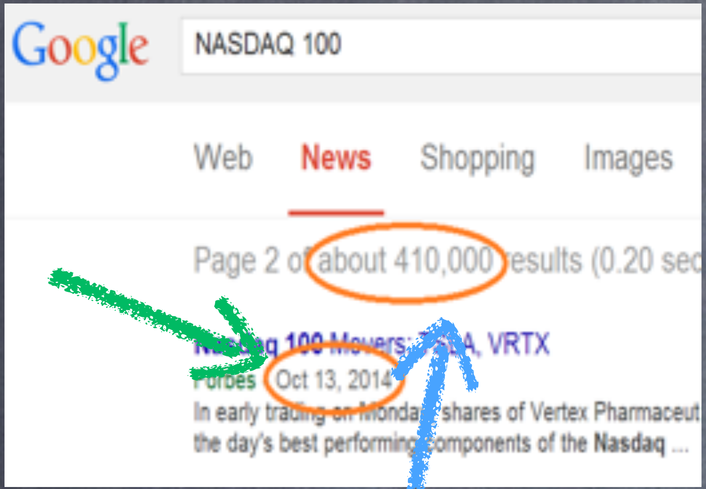


```
<div class="ts _V6c _anc _XO _lnc _d7c">
  <a class="top _vQb _mnc" onmousedown="return
  rwt(this,'','','','3','AFQjCNFRAxrQ7_njM9YQTRotmQ75s-
  jCMA','1NOSJzcwjCWtbyWEhMERdA','0CCsQpwIwAg','','',event)" hre
  /2014/10/22/dow-jones-today-another-triple-digit-tumble-heres-
  <div class="_cnc">
    <h3 class="r _U6c">
    <div class="slp">
        <span class="_tQb _IId">Fox Business</span>
        <span class="_v5">-</span>
        <span class="f nsa _uQb">2014年10月22日</span>
    </div>
    <div class="st">
  </div>
  <div class="_hnc card-section">
  <div class="_Vmc"></div>
</div>
</li>
```



```
[<div id="res"><div id="topstuff"></div><div id="search"><div i
d="ires"><ol><li class="g"><table><tr><td style="width:516px" v
align="top"><h3 class="r"><a href="/url?q=http://time.com/money
/3532220/dow-jones-up-down-buy-sell/&amp;sa=U&amp;ei=g4JOVP-bEo
Ln8AWe6IEg&amp;ved=0CBMQqQIoADAA&amp;usg=AFQjCNGPacNja86AHmxCkz
84z-t48vKWIQ">The <b>Dow</b> Moved Triple Digits Today. Here's
What You Should Do.</a></h3><div class="slp"><span class="f">TI
ME - 3 天前</span></div><div class="st">The <b>Dow Jones Indust
rial Average</b> has long been held up as the stock market inde
x to follow. Now that the Dow Jones is at such a high level, ..
.</div><a href="/url?q=http://seekingalpha.com/article/2577375-
just-stop-looking-at-the-dow-jones-index-already&amp;sa=U&amp;e
i=g4JOVP-bEoLn8AWe6IEg&amp;ved=0CBUQ-AsoADAA&amp;usg=AFQjCNF6pp
NvFaSjN-KeaWan-h4hg_OJWg">Just Stop Looking At The <b>Dow Jones
 Index</b> Already</a> <span class="f">Seeking Alpha (registrat
ion)</span><br><a href="/url?q=http://www.cnbc.com/id/102120107
&amp;sa=U&amp;ei=g4JOVP-bEoLn8AWe6IEg&amp;ved=0CBYQ-AsoATAA&amp
;usg=AFQjCNFAH641_3duHBPh1JDNcd_naANz1w">How the <b>Dow Jones i
ndustrial average</b> fared Friday</a> <span class="f">CNBC</sp
```

處理方法



Google NASDAQ 100

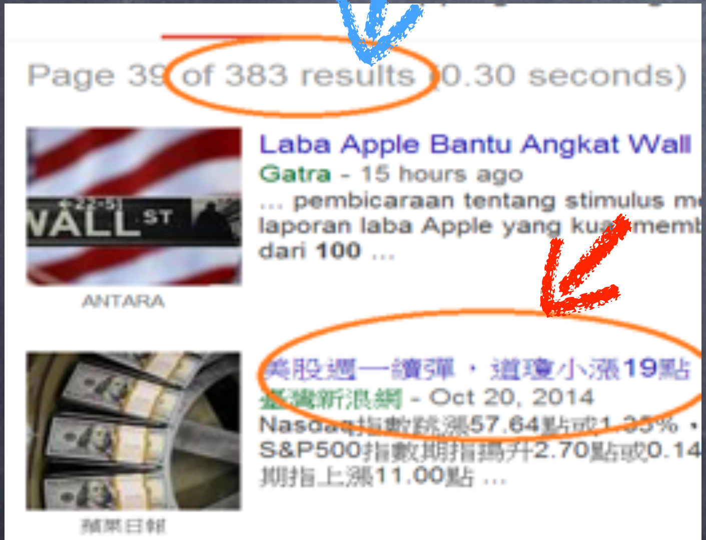
Web News Shopping Images

Page 2 of about 410,000 results (0.20 sec

Nasdaq 100 Movers: TSLA, VRTX

Forbes - Oct 13, 2014

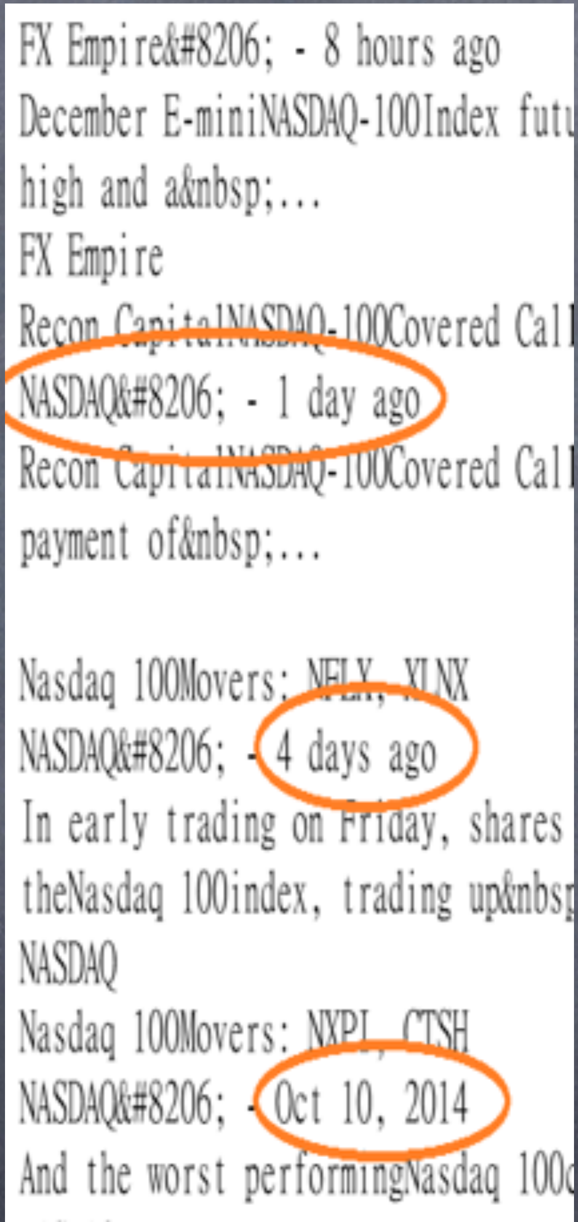
In early trading on Monday, shares of Vertex Pharmaceut
the day's best performing components of the Nasdaq ...



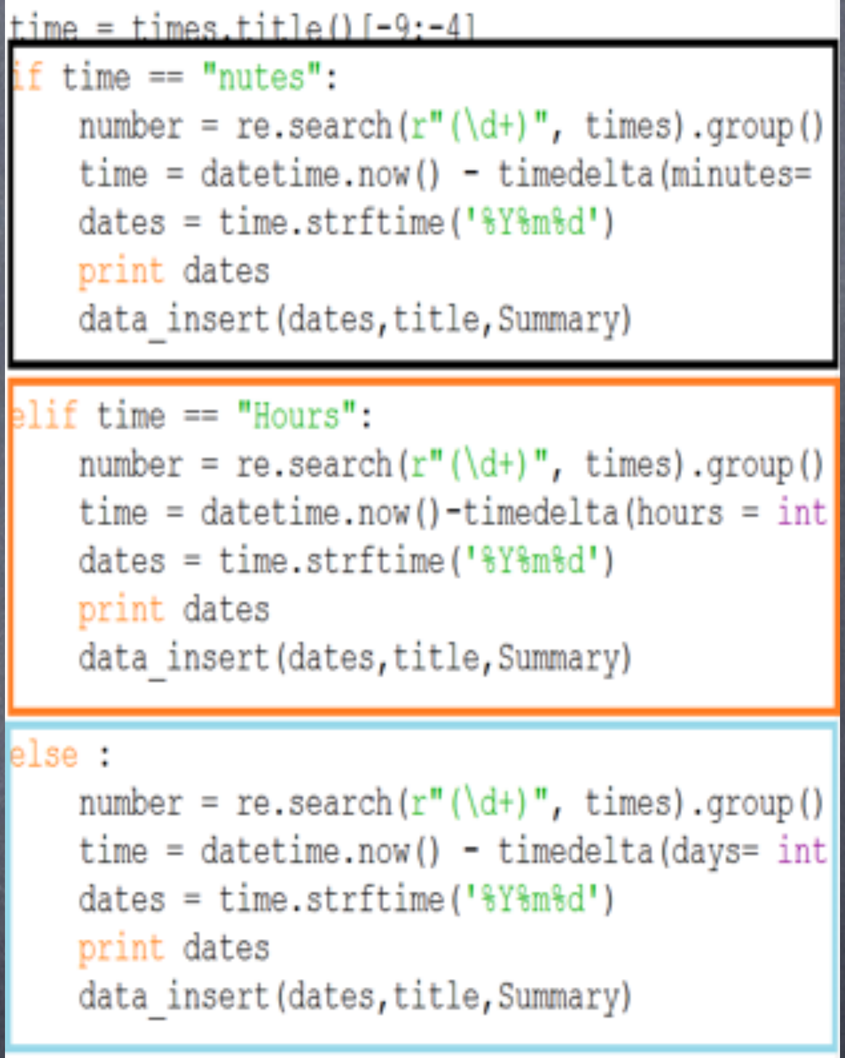
Page 39 of 383 results (0.30 seconds)

Laba Apple Bantu Angkat Wall
Gatra - 15 hours ago
... pembicaraan tentang stimulus me
laporan laba Apple yang kuat memb
dari 100 ...

美股週一續彈，道瓊小漲19點
臺灣新浪網 - Oct 20, 2014
Nasdaq指數跳漲57.64點或1.35%，
S&P500指數期指揚升2.70點或0.14
期指上漲11.00點 ...



```
FX Empire&#8206; - 8 hours ago
December E-miniNASDAQ-100Index futu
high and a&nbsp;...
FX Empire
Recon CapitalNASDAQ-100Covered Call
NASDAQ&#8206; - 1 day ago
Recon CapitalNASDAQ-100Covered Call
payment of&nbsp;...

Nasdaq 100Movers: NFLX, XLNX
NASDAQ&#8206; - 4 days ago
In early trading on Friday, shares
theNasdaq 100index, trading up&nbsp
NASDAQ
Nasdaq 100Movers: NXPI, CTSH
NASDAQ&#8206; - Oct 10, 2014
And the worst performingNasdaq 100c
```

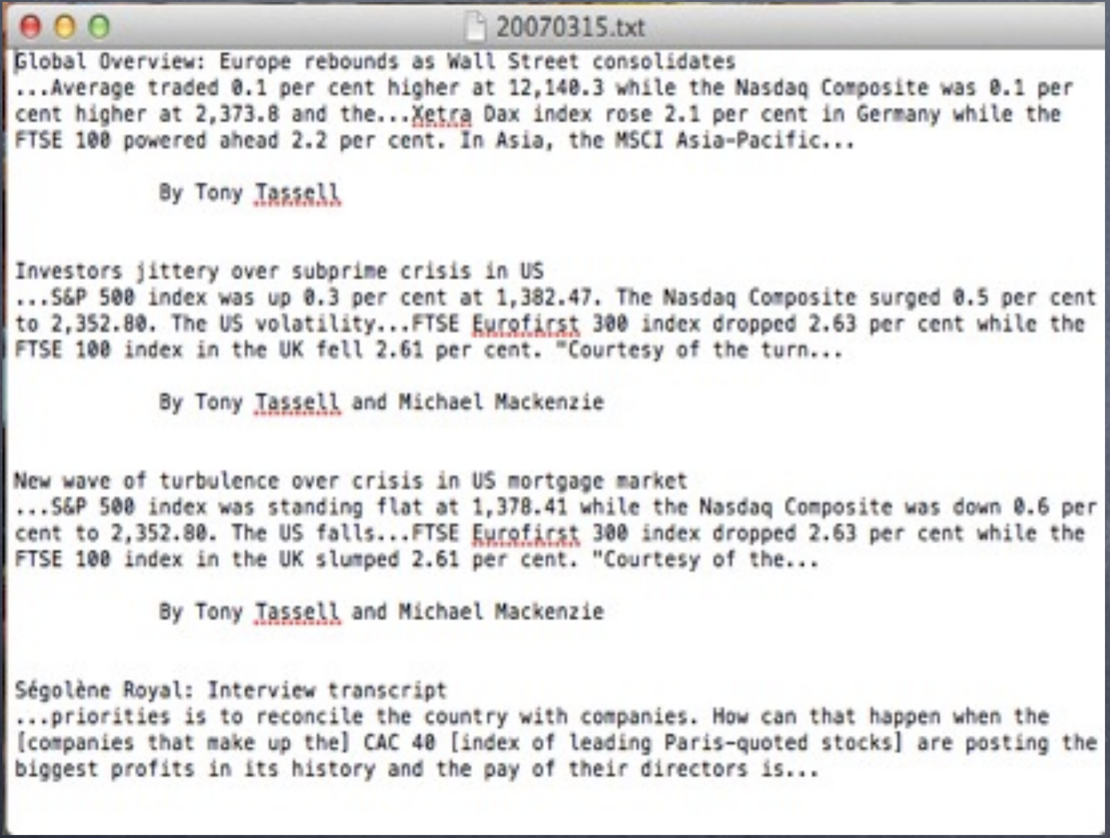


```
time = times.title()[-9:-4]
if time == "nutes":
    number = re.search(r"(\d+)", times).group()
    time = datetime.now() - timedelta(minutes=
    dates = time.strftime('%Y%m%d')
    print dates
    data_insert(dates,title,Summary)
elif time == "Hours":
    number = re.search(r"(\d+)", times).group()
    time = datetime.now()-timedelta(hours = int
    dates = time.strftime('%Y%m%d')
    print dates
    data_insert(dates,title,Summary)
else :
    number = re.search(r"(\d+)", times).group()
    time = datetime.now() - timedelta(days= int
    dates = time.strftime('%Y%m%d')
    print dates
    data_insert(dates,title,Summary)
```

Google news
FINANCIAL TIMES

StockQ.org

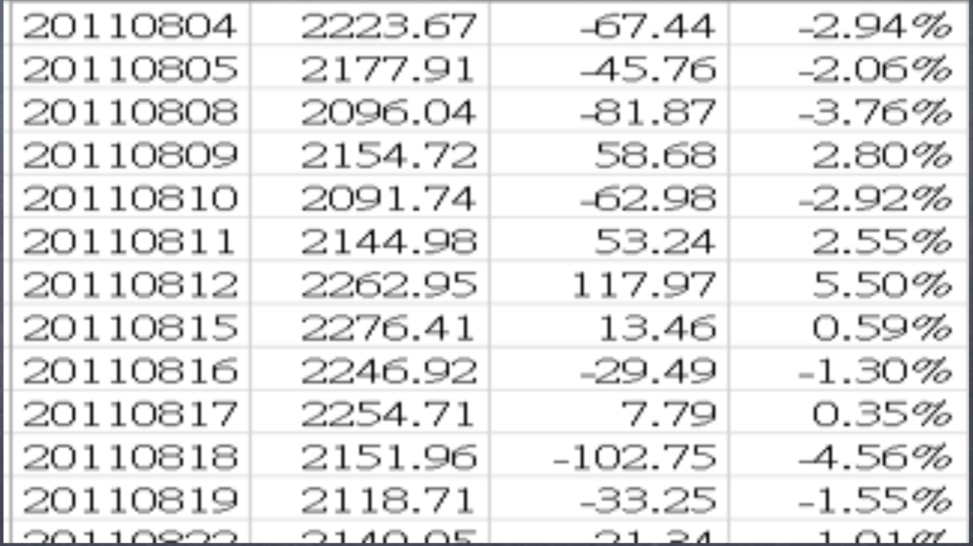
國際主要股市新聞



20070315.txt

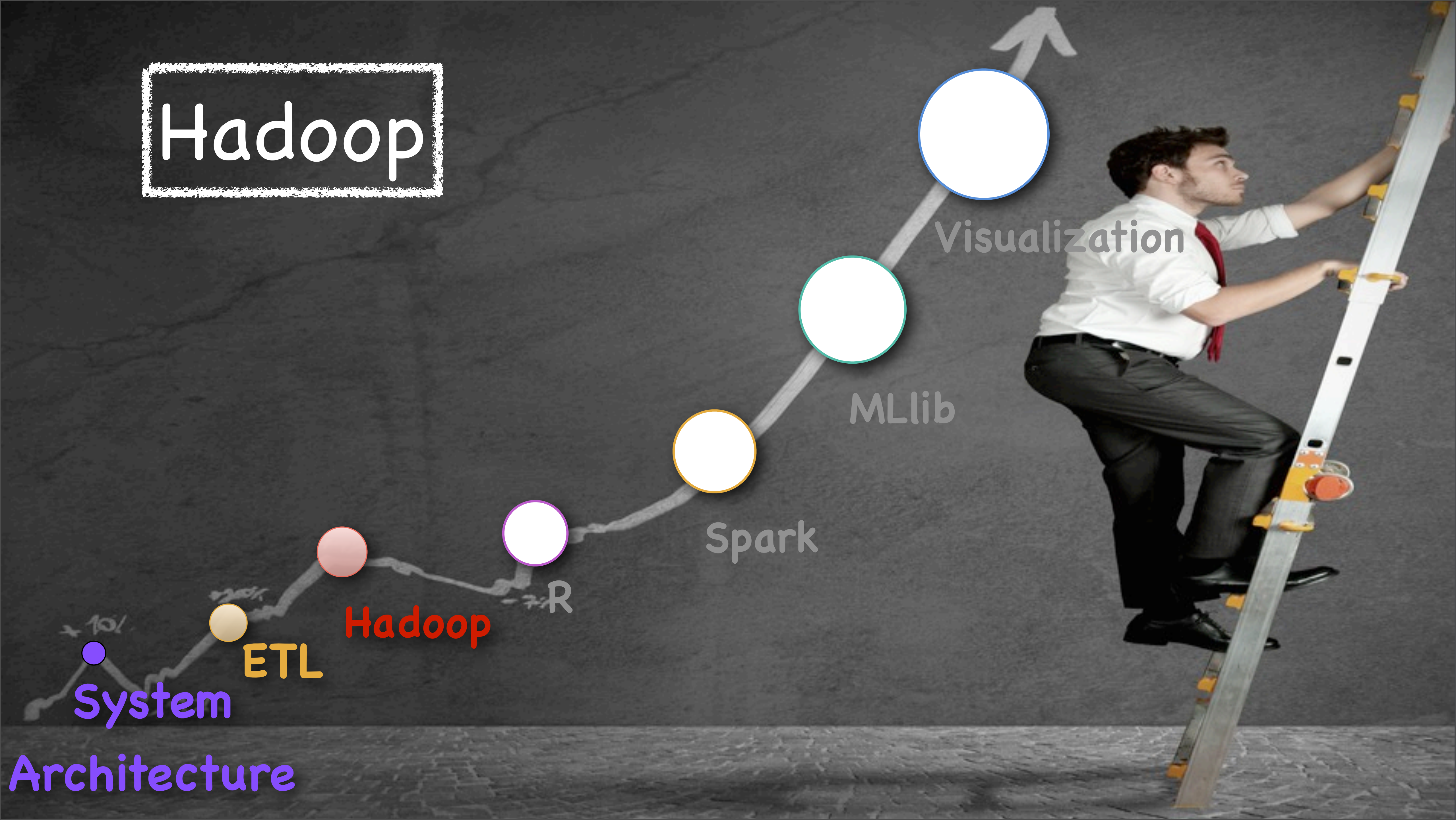
```
Global Overview: Europe rebounds as Wall Street consolidates
...Average traded 0.1 per cent higher at 12,140.3 while the Nasdaq Composite was 0.1 per
cent higher at 2,373.8 and the...Xetra Dax index rose 2.1 per cent in Germany while the
FTSE 100 powered ahead 2.2 per cent. In Asia, the MSCI Asia-Pacific...

         By Tony Tassell


Investors jittery over subprime crisis in US
...S&P 500 index was up 0.3 per cent at 1,382.47. The Nasdaq Composite surged 0.5 per cent
to 2,352.80. The US volatility...FTSE Eurofirst 300 index dropped 2.63 per cent while the
FTSE 100 index in the UK fell 2.61 per cent. "Courtesy of the turn...

         By Tony Tassell and Michael Mackenzie


New wave of turbulence over crisis in US mortgage market
...S&P 500 index was standing flat at 1,378.41 while the Nasdaq Composite was down 0.6 per
cent to 2,352.80. The US falls...FTSE Eurofirst 300 index dropped 2.63 per cent while the
FTSE 100 index in the UK slumped 2.61 per cent. "Courtesy of the...

         By Tony Tassell and Michael Mackenzie


Ségolène Royal: Interview transcript
...priorities is to reconcile the country with companies. How can that happen when the
[companies that make up the] CAC 40 [index of leading Paris-quoted stocks] are posting the
biggest profits in its history and the pay of their directors is...
```

國際股市指數



日期	指數	漲跌	漲跌比例
20110804	2223.67	-67.44	-2.94%
20110805	2177.91	-45.76	-2.06%
20110808	2096.04	-81.87	-3.76%
20110809	2154.72	58.68	2.80%
20110810	2091.74	-62.98	-2.92%
20110811	2144.98	53.24	2.55%
20110812	2262.95	117.97	5.50%
20110815	2276.41	13.46	0.59%
20110816	2246.92	-29.49	-1.30%
20110817	2254.71	7.79	0.35%
20110818	2151.96	-102.75	-4.56%
20110819	2118.71	-33.25	-1.55%
20110822	2140.05	21.34	1.01%

Hadoop



System Architecture

ETL

Hadoop

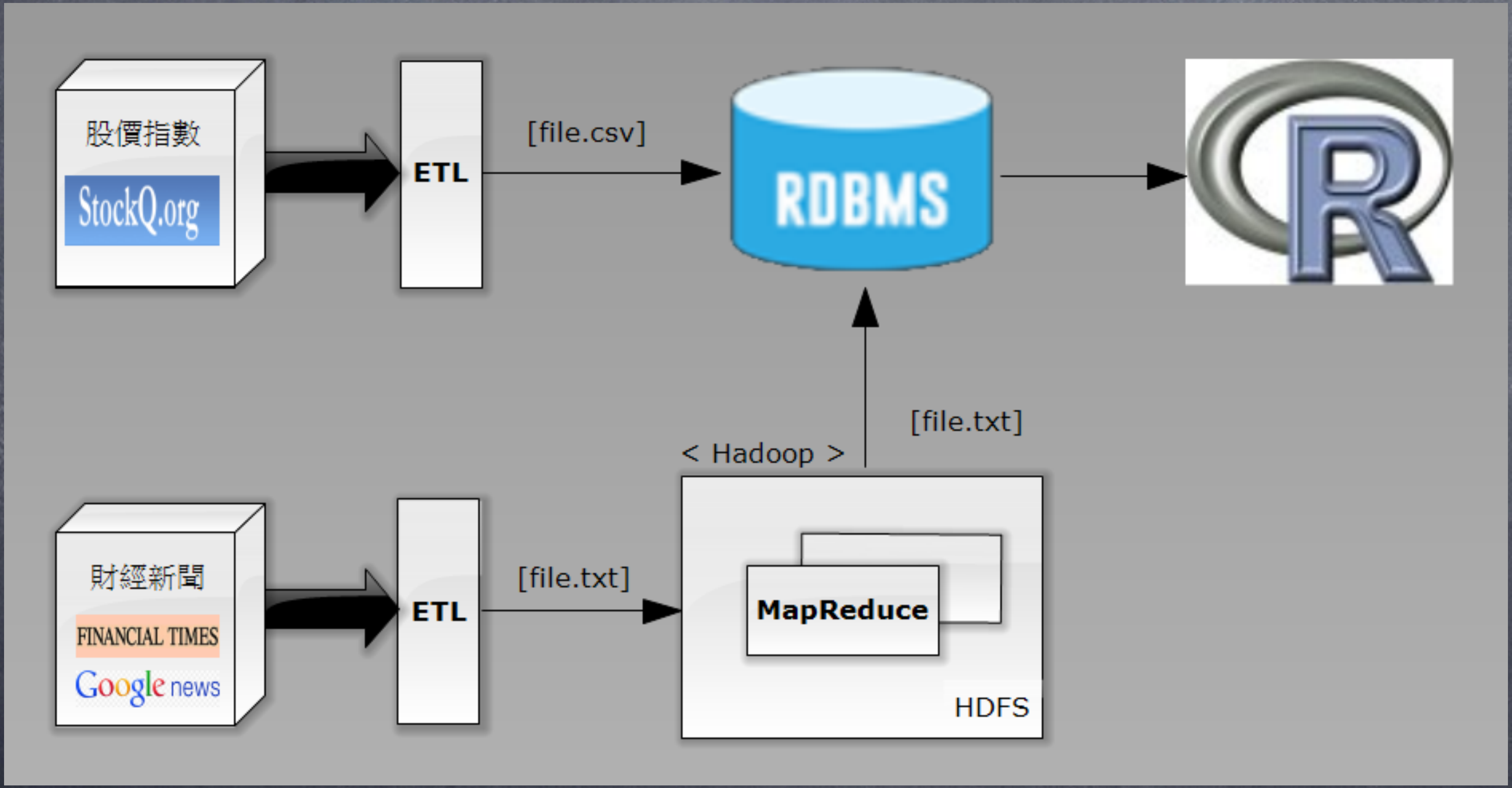
R

Spark

MLlib

Visualization

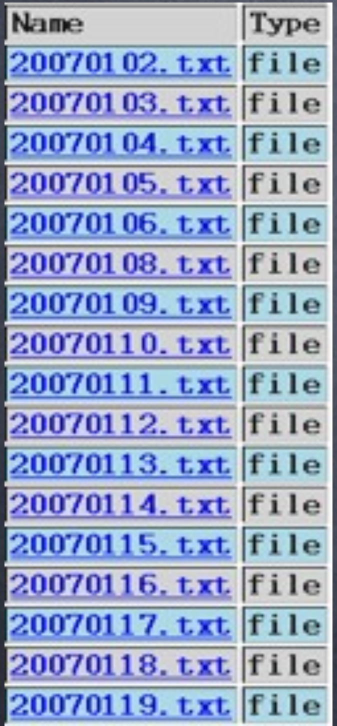
Hadoop應用流程


股價指數
StockQ.org
ETL
[file.csv]
RDBMS
R
[file.txt]
< Hadoop >
財經新聞
FINANCIAL TIMES
Google news
ETL
[file.txt]
MapReduce
HDFS

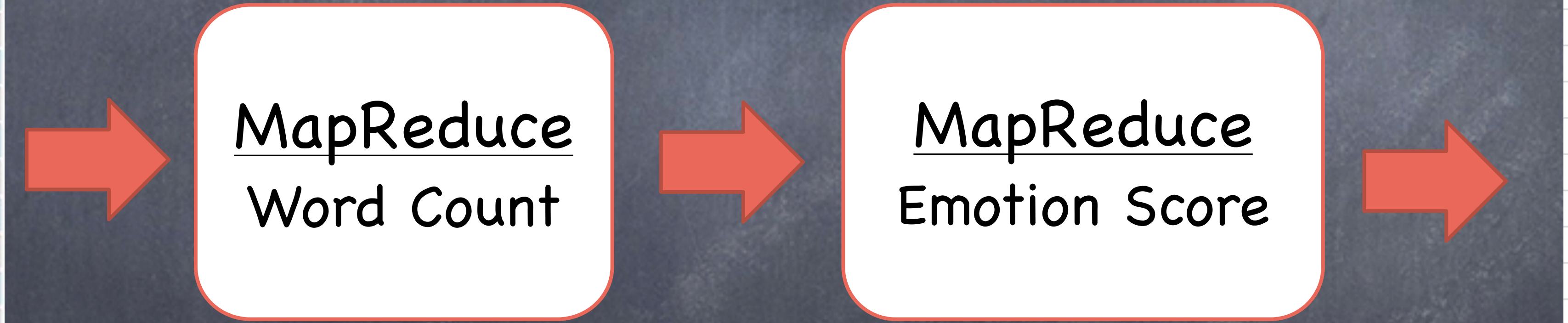
Transform Data - MapReduce

1.針對每天的新聞利用MapReduce計算單詞次數

2.取得每天新聞的單詞計數利用MapReduce與情緒字典做比對，產生每日的情緒分數

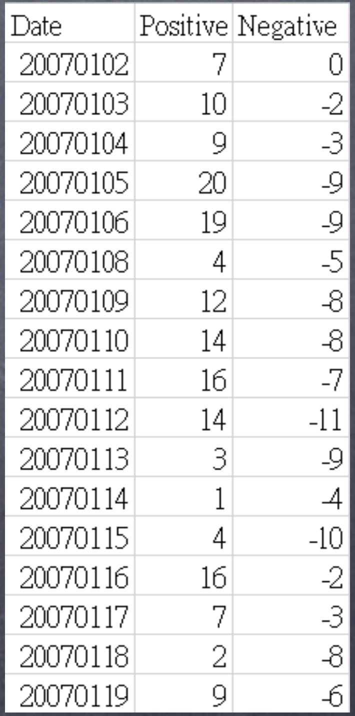


Name	Type
20070102.txt	file
20070103.txt	file
20070104.txt	file
20070105.txt	file
20070106.txt	file
20070108.txt	file
20070109.txt	file
20070110.txt	file
20070111.txt	file
20070112.txt	file
20070113.txt	file
20070114.txt	file
20070115.txt	file
20070116.txt	file
20070117.txt	file
20070118.txt	file
20070119.txt	file



MapReduce
Word Count

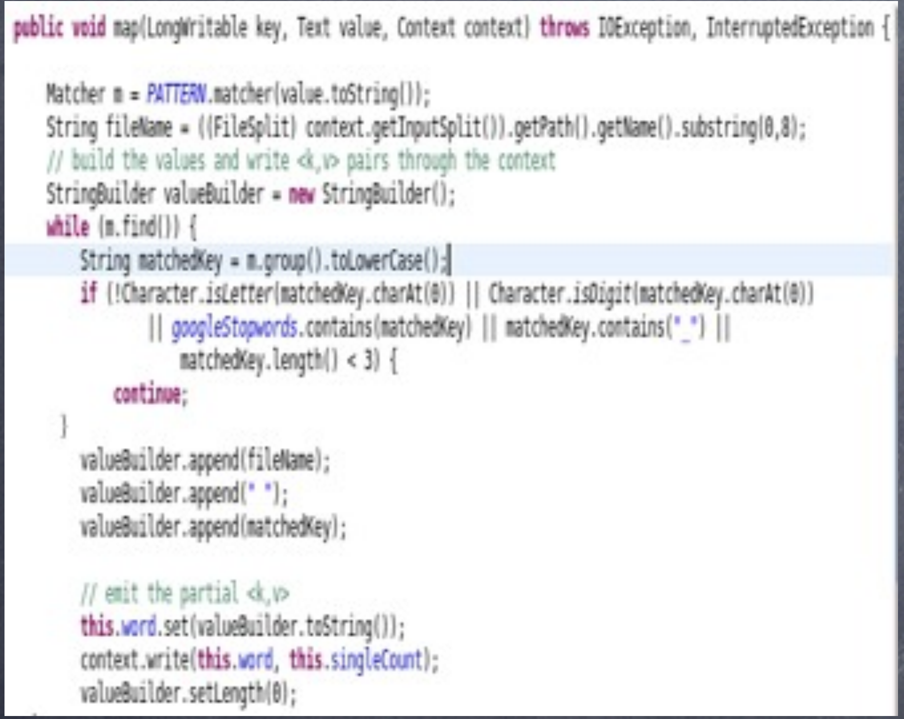
MapReduce
Emotion Score



Date	Positive	Negative
20070102	7	0
20070103	10	-2
20070104	9	-3
20070105	20	-9
20070106	19	-9
20070108	4	-5
20070109	12	-8
20070110	14	-8
20070111	16	-7
20070112	14	-11
20070113	3	-9
20070114	1	-4
20070115	4	-10
20070116	16	-2
20070117	7	-3
20070118	2	-8
20070119	9	-6

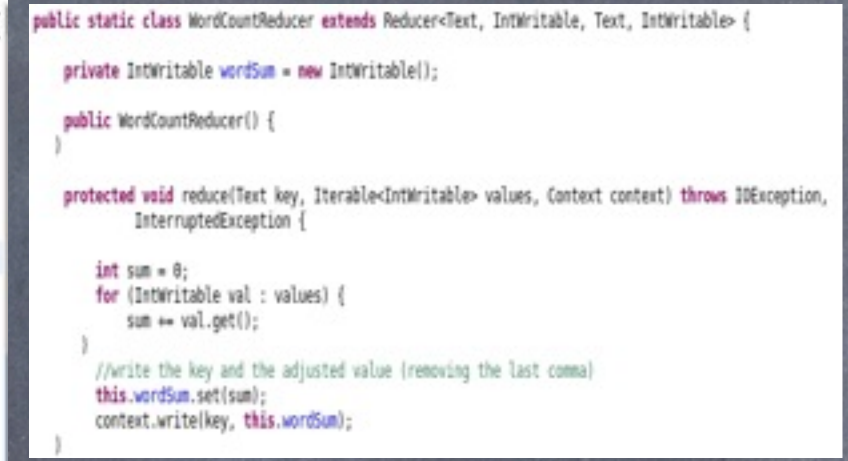
MapReduce – Word Count

Map:

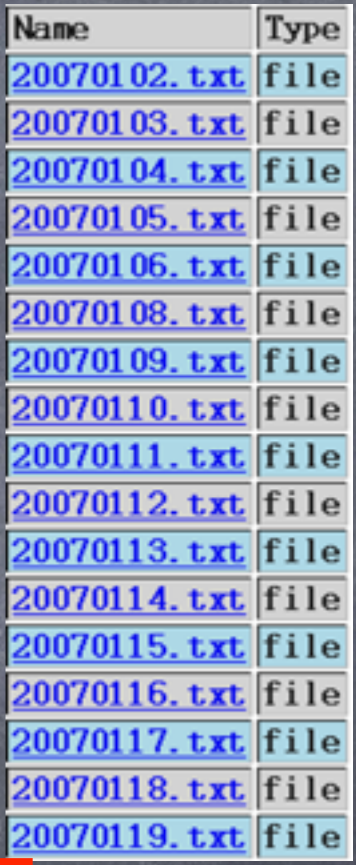


```
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

    Matcher m = PATTERN.matcher(value.toString());
    String fileName = ((FileSplit) context.getInputSplit()).getPath().getName().substring(0,8);
    // build the values and write <k,v> pairs through the context
    StringBuilder valueBuilder = new StringBuilder();
    while (m.find()) {
        String matchedKey = m.group().toLowerCase();
        if (!Character.isLetter(matchedKey.charAt(0)) || Character.isDigit(matchedKey.charAt(0))
                || googleStopwords.contains(matchedKey) || matchedKey.contains("_") ||
                  matchedKey.length() < 3) {
            continue;
      }
        valueBuilder.append(fileName);
        valueBuilder.append(" ");
        valueBuilder.append(matchedKey);

        // emit the partial <k,v>
        this.word.set(valueBuilder.toString());
        context.write(this.word, this.singleCount);
        valueBuilder.setLength(0);
```

Reduce:

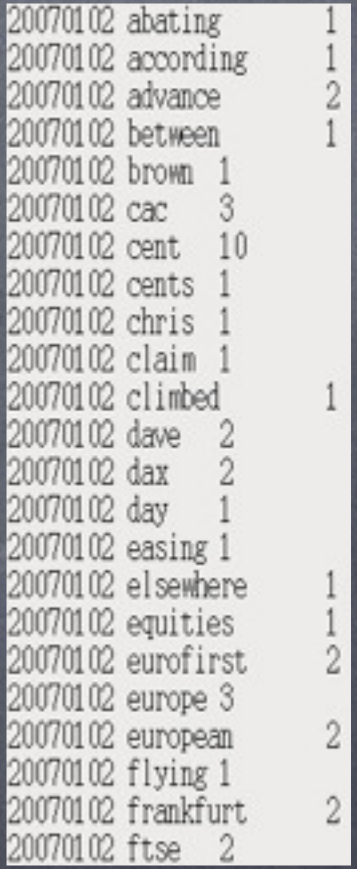


```
public static class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable wordSum = new IntWritable();

    public WordCountReducer() {
   }

    protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException,
            InterruptedException {

        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
       }
        //write the key and the adjusted value (removing the last comma)
        this.wordSum.set(sum);
        context.write(key, this.wordSum);
   }
```

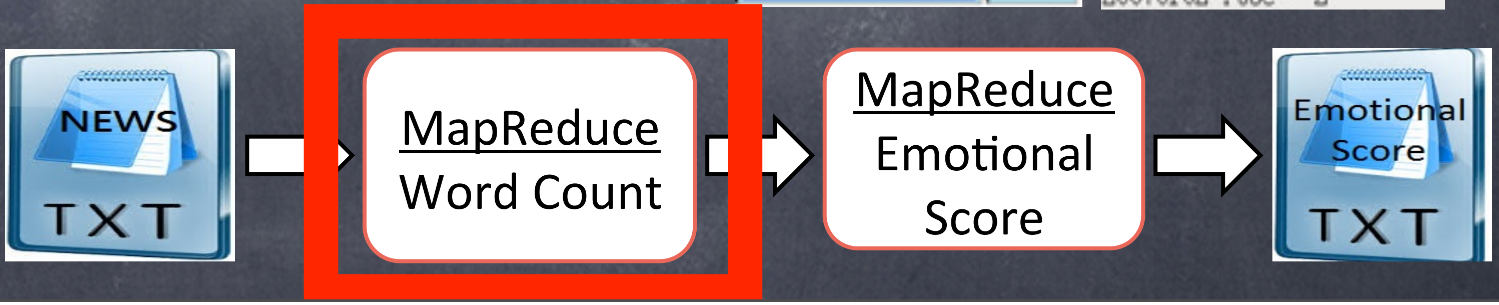


Name	Type
20070102.txt	file
20070103.txt	file
20070104.txt	file
20070105.txt	file
20070106.txt	file
20070108.txt	file
20070109.txt	file
20070110.txt	file
20070111.txt	file
20070112.txt	file
20070113.txt	file
20070114.txt	file
20070115.txt	file
20070116.txt	file
20070117.txt	file
20070118.txt	file
20070119.txt	file

Date Word Count



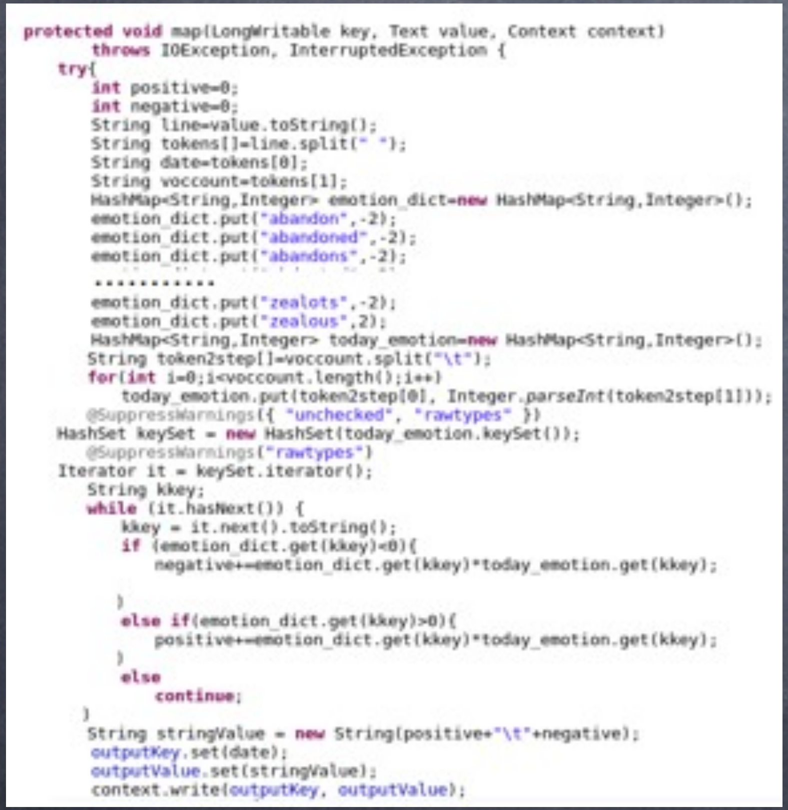
```
20070102 abating        1
20070102 according      1
20070102 advance        2
20070102 between        1
20070102 brown  1
20070102 cac    3
20070102 cent   10
20070102 cents  1
20070102 chris  1
20070102 claim  1
20070102 climbed        1
20070102 dave   2
20070102 dax    2
20070102 day    1
20070102 easing 1
20070102 elsewhere      1
20070102 equities       1
20070102 eurofirst      2
20070102 europe 3
20070102 european       2
20070102 flying 1
20070102 frankfurt      2
20070102 ftse   2
```



MapReduce - Emotional Score

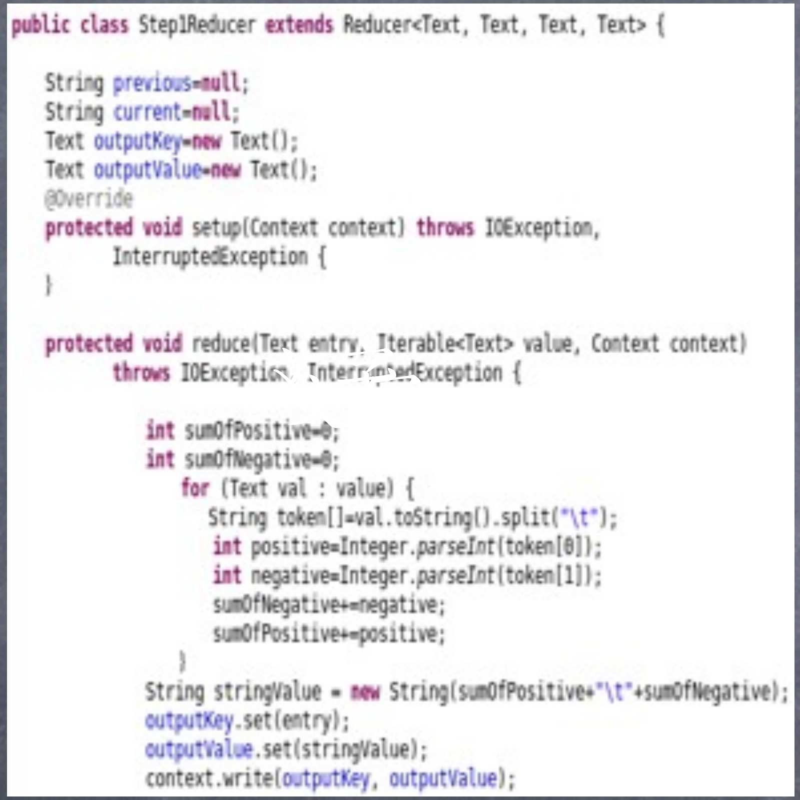
Map:

走訪情緒字典



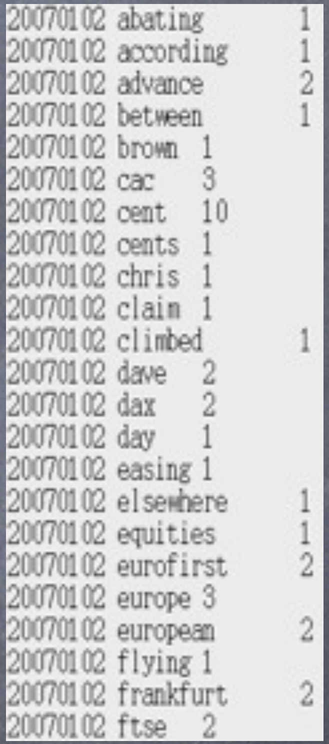
```
protected void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException {
    try{
        int positive=0;
        int negative=0;
        String line=value.toString();
        String tokens[]=line.split(" ");
        String date=tokens[0];
        String voccount=tokens[1];
        HashMap<String,Integer> emotion_dict=new HashMap<String,Integer>();
        emotion_dict.put("abandon",-2);
        emotion_dict.put("abandoned",-2);
        emotion_dict.put("abandons",-2);
        ............
        emotion_dict.put("zealots",-2);
        emotion_dict.put("zealous",2);
        HashMap<String,Integer> today_emotion=new HashMap<String,Integer>();
        String token2step[]=voccount.split("\t");
        for(int i=0;i<voccount.length();i++)
            today_emotion.put(token2step[0], Integer.parseInt(token2step[1]));
        @SuppressWarnings({ "unchecked", "rawtypes" })
    HashSet keySet = new HashSet(today_emotion.keySet());
        @SuppressWarnings("rawtypes")
    Iterator it = keySet.iterator();
        String kkey;
        while (it.hasNext()) {
            kkey = it.next().toString();
            if (emotion_dict.get(kkey)<0){
                negative+=emotion_dict.get(kkey)*today_emotion.get(kkey);

            }
            else if(emotion_dict.get(kkey)>0){
                positive+=emotion_dict.get(kkey)*today_emotion.get(kkey);
            }
            else
                continue;
        }
        String stringValue = new String(positive+"\t"+negative);
        outputKey.set(date);
        outputValue.set(stringValue);
        context.write(outputKey, outputValue);
```

Reduce:

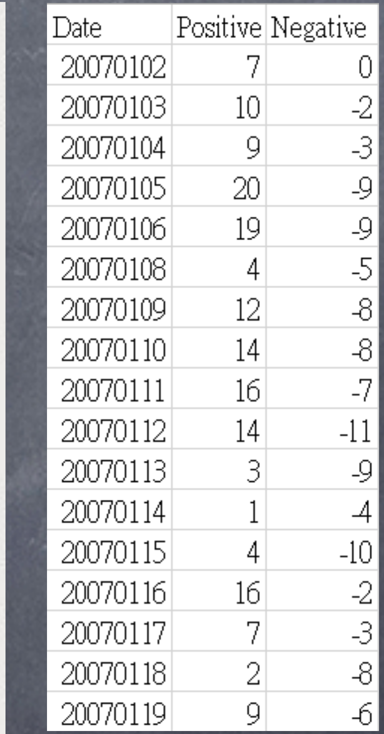


```
public class Step1Reducer extends Reducer<Text, Text, Text, Text> {

    String previous=null;
    String current=null;
    Text outputKey=new Text();
    Text outputValue=new Text();
    @Override
    protected void setup(Context context) throws IOException,
            InterruptedException {
    }

    protected void reduce(Text entry, Iterable<Text> value, Context context)
            throws IOException, InterruptedException {

            int sumOfPositive=0;
            int sumOfNegative=0;
                for (Text val : value) {
                    String token[]=val.toString().split("\t");
                    int positive=Integer.parseInt(token[0]);
                    int negative=Integer.parseInt(token[1]);
                    sumOfNegative+=negative;
                    sumOfPositive+=positive;
                }
            String stringValue = new String(sumOfPositive+"\t"+sumOfNegative);
            outputKey.set(entry);
            outputValue.set(stringValue);
            context.write(outputKey, outputValue);
```

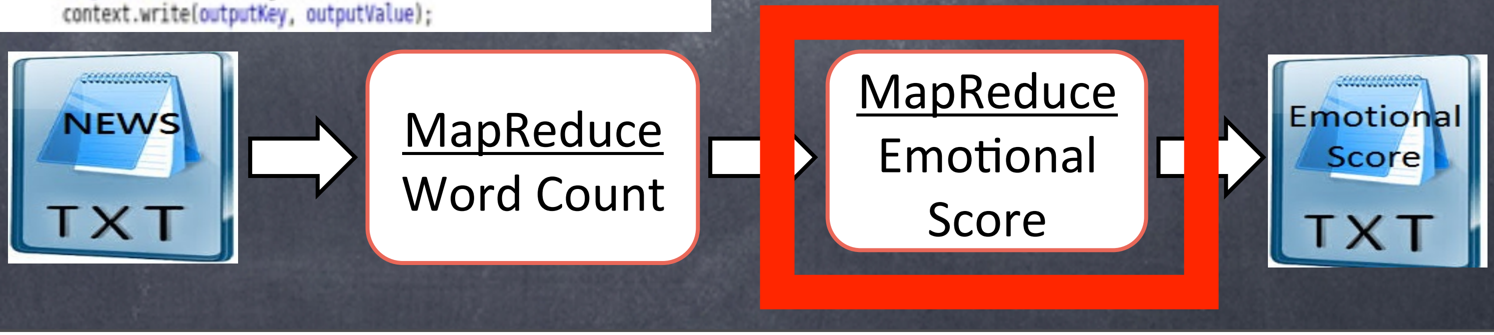
Date Word Count



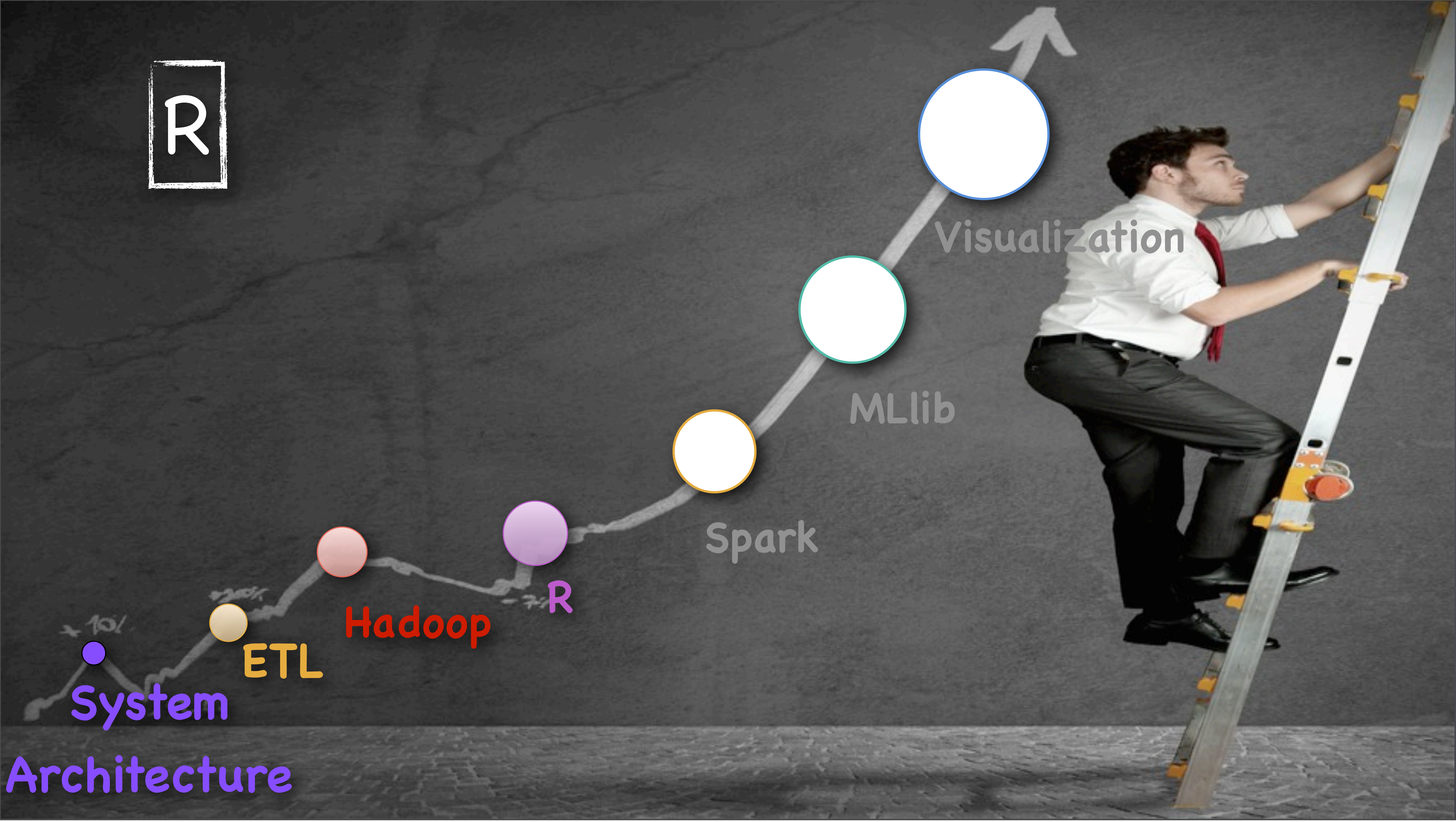
```
20070102 abating        1
20070102 according      1
20070102 advance        2
20070102 between        1
20070102 brown   1
20070102 cac     3
20070102 cent    10
20070102 cents   1
20070102 chris   1
20070102 clain   1
20070102 climbed        1
20070102 dave    2
20070102 dax     2
20070102 day     1
20070102 easing 1
20070102 elsewhere      1
20070102 equities       1
20070102 eurofirst      2
20070102 europe 3
20070102 european       2
20070102 flying 1
20070102 frankfurt      2
20070102 ftse    2
```



Date	Positive	Negative
20070102	7	0
20070103	10	-2
20070104	9	-3
20070105	20	-9
20070106	19	-9
20070108	4	-5
20070109	12	-8
20070110	14	-8
20070111	16	-7
20070112	14	-11
20070113	3	-9
20070114	1	-4
20070115	4	-10
20070116	16	-2
20070117	7	-3
20070118	2	-8
20070119	9	-6



R


+10%
+20%
-7%
System Architecture
ETL
Hadoop
R
Spark
MLlib
Visualization

RHmm & Market Trend

模型：

Hidden Markov Model (HMM，隱馬可夫模型)

一連串事件接續發生的機率，簡單的說，「隱馬可夫模型」提供了一套數學的理論以及工具，讓我們可以利用「看得到的」連續現象去探究、預測另一個「看不到的」連續現象。

演算法：

Viterbi algorithm (維特比演算法)

一種動態規劃演算法。它用於尋找最有可能產生觀測事件序列的-維特比路徑-隱含狀態序列,特別是在馬爾可夫信息源上下文和隱馬爾可夫模型中。被用於尋找觀察結果最有可能解釋相關的動態規劃算法。

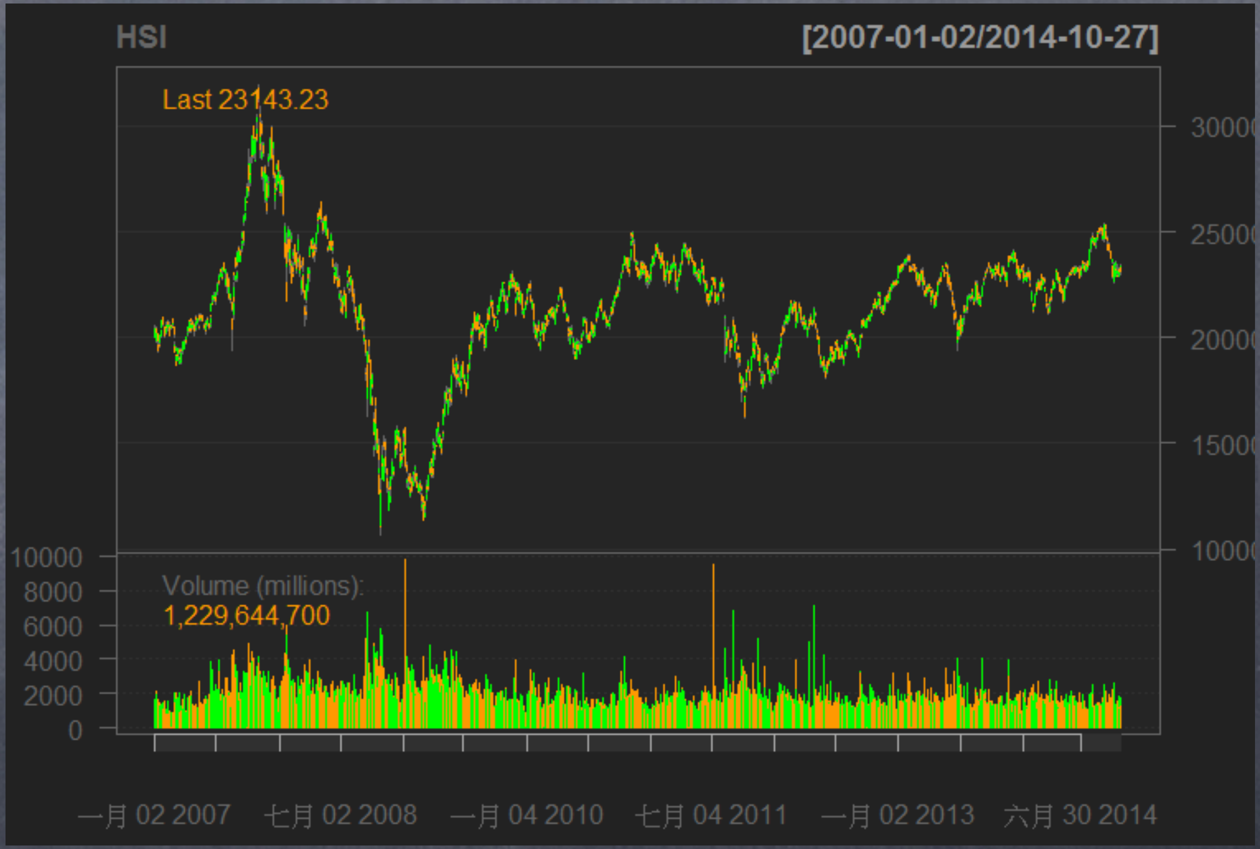
Data Analysis-R

- 利用歷史的指數資料以及新聞情緒分數建置隱馬可夫模型(Hidden Markov Model，HMM)

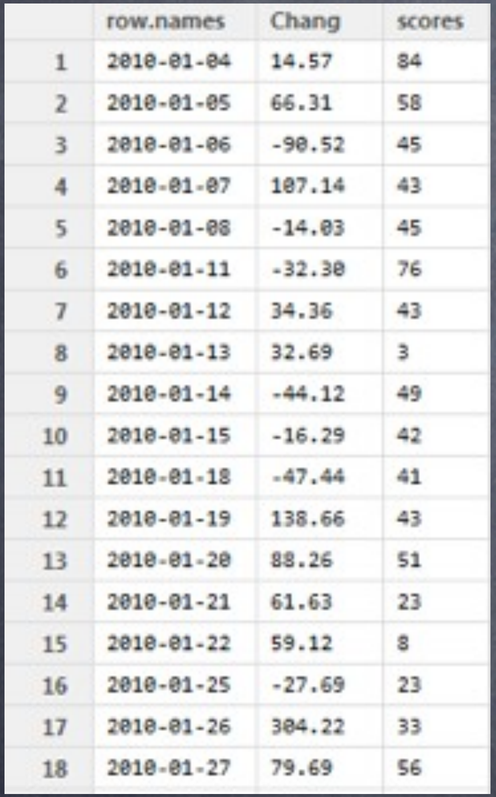
- 運用維特比演算法(Viterbi algorithm)套用已建置模型來預測股價走勢

- 將結果以時間序列圖形做視覺化呈現

香港恆生指數



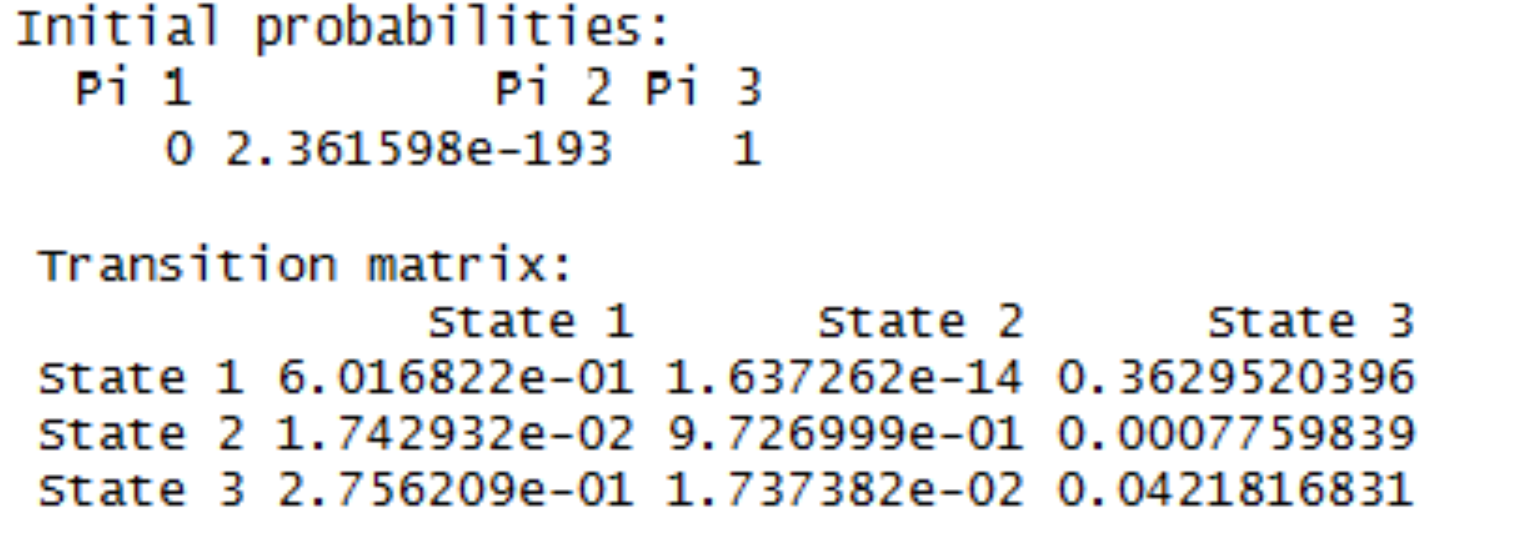
R - Package(RHmm)



	row.names	Chang	scores
1	2010-01-04	14.57	84
2	2010-01-05	66.31	58
3	2010-01-06	-90.52	45
4	2010-01-07	107.14	43
5	2010-01-08	-14.03	45
6	2010-01-11	-32.30	76
7	2010-01-12	34.36	43
8	2010-01-13	32.69	3
9	2010-01-14	-44.12	49
10	2010-01-15	-16.29	42
11	2010-01-18	-47.44	41
12	2010-01-19	138.66	43
13	2010-01-20	88.26	51
14	2010-01-21	61.63	23
15	2010-01-22	59.12	8
16	2010-01-25	-27.69	23
17	2010-01-26	304.22	33
18	2010-01-27	79.69	56

建立HMM模型:

```
HSI_hm_model <- HMMFit(obs =HSI_Train, nStates = 3)
```



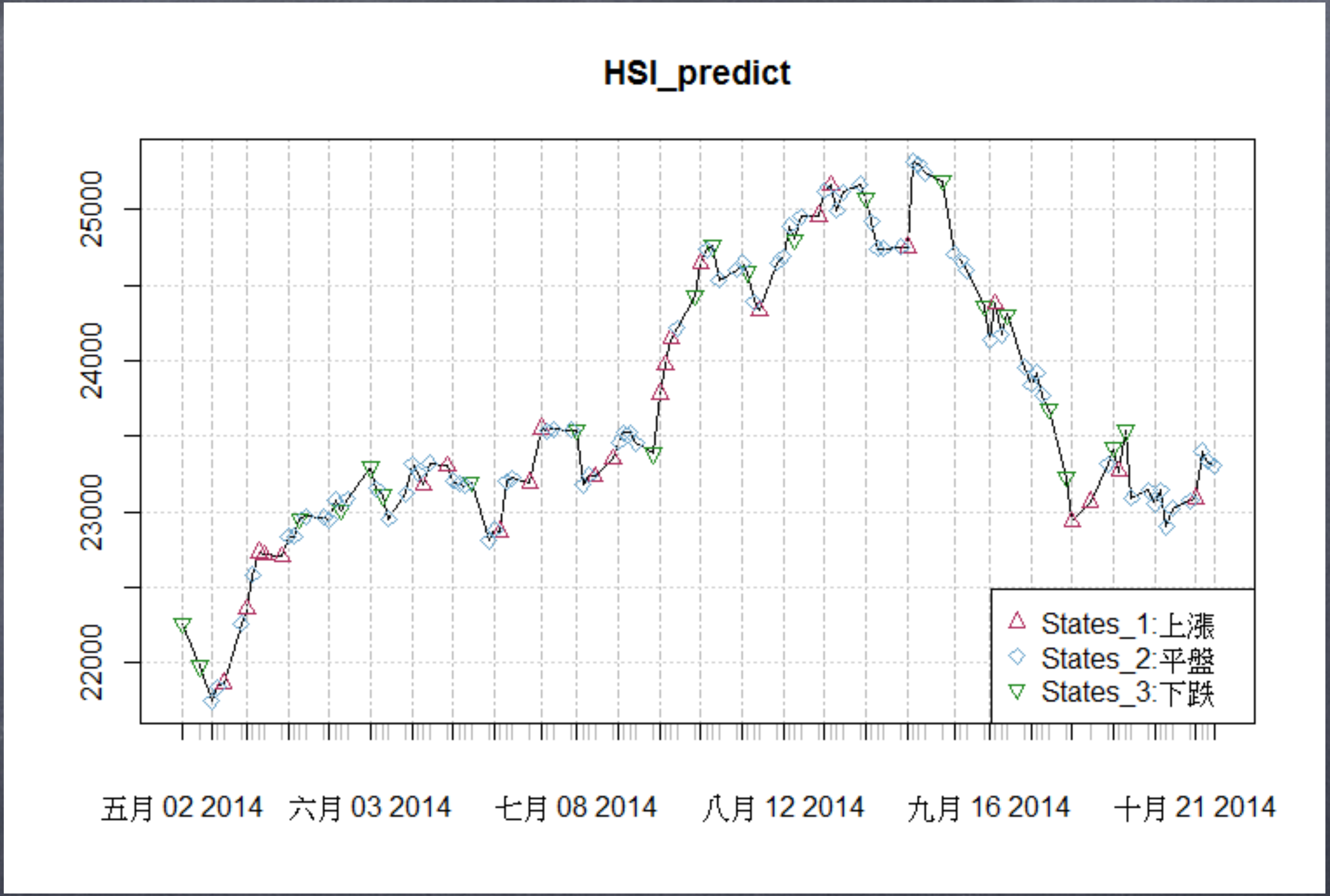
```
Initial probabilities:
  Pi 1           Pi 2 Pi 3
     0 2.361598e-193    1

 Transition matrix:
             State 1      State 2      State 3
State 1 6.016822e-01 1.637262e-14 0.3629520396
State 2 1.742932e-02 9.726999e-01 0.0007759839
State 3 2.756209e-01 1.737382e-02 0.0421816831
```

利用模型，套用Viterbi演算法:

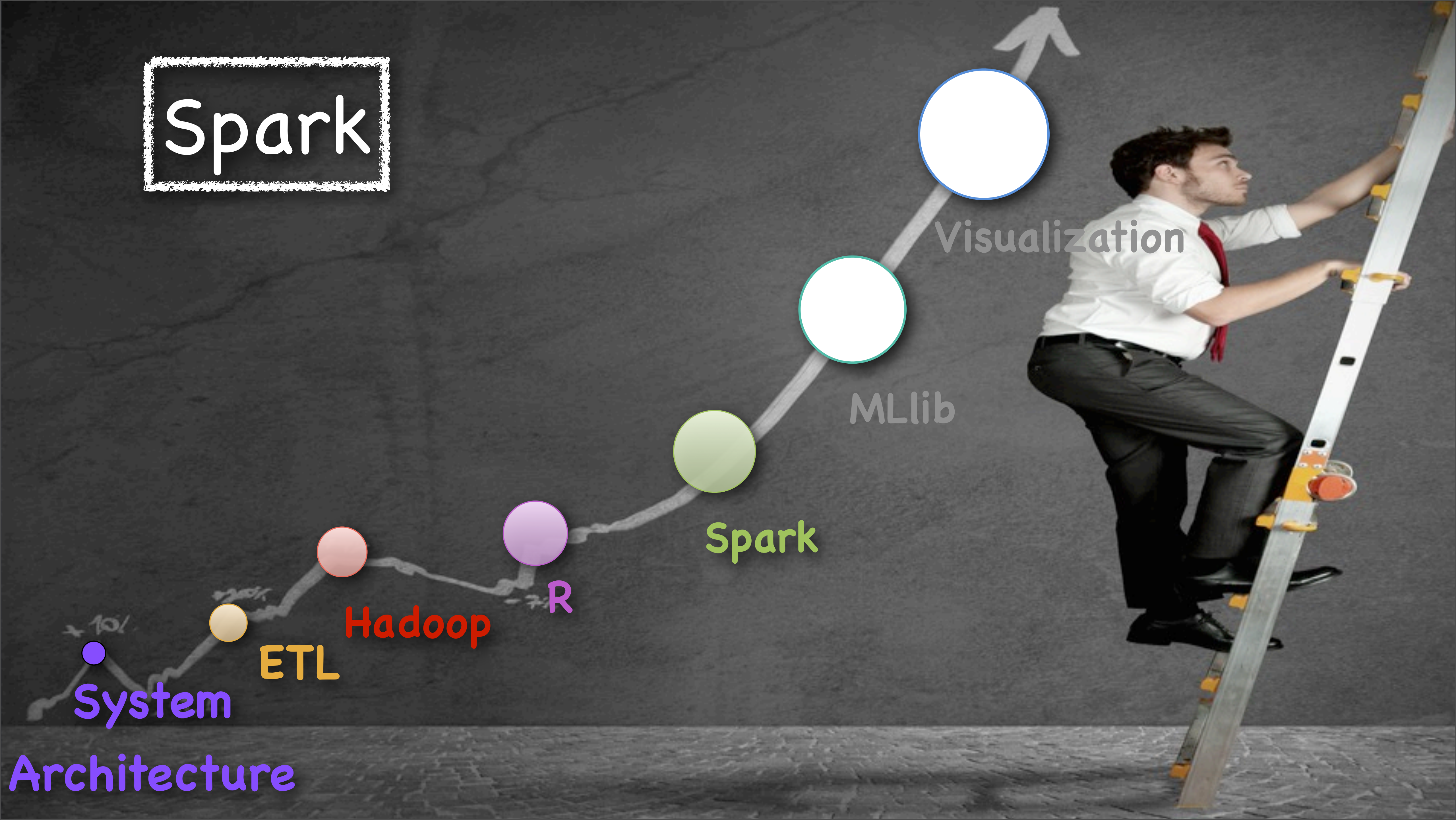
```
HSI_VitPath <- viterbi(HSI_hm_model, HSI_Predict)
```

R - Package(xts)

	實際狀態	預測狀態	準確率
上漲	30	17	56.77%
平盤	70	36	51.43%
下跌	20	12	60%

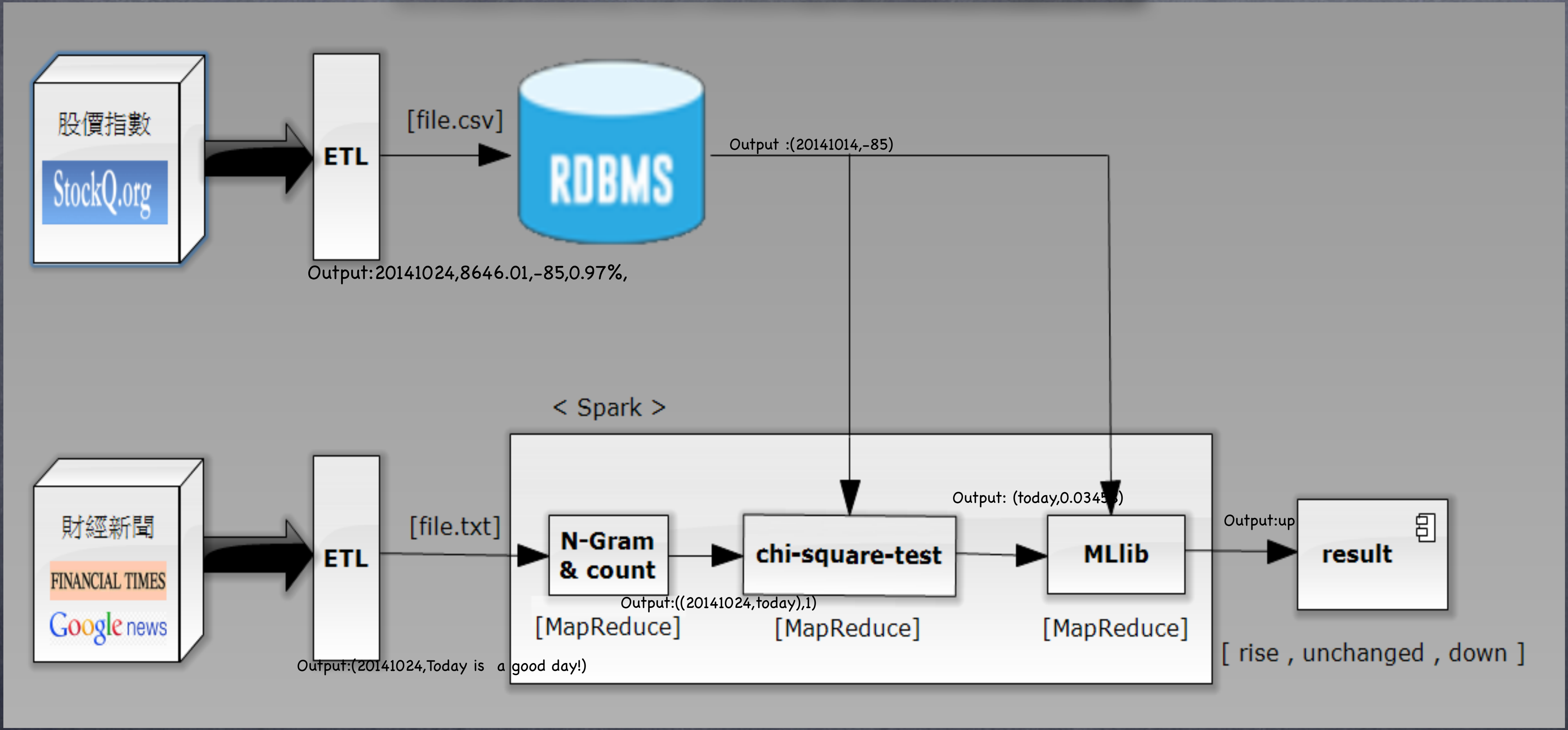


Spark



System Architecture
ETL
Hadoop
R
Spark
MLlib
Visualization

Spark Flow Chart



股價指數

StockQ.org

ETL

[file.csv]

RDBMS

Output:20141024,8646.01,-85,0.97%,

Output :(20141014,-85)

< Spark >

財經新聞

FINANCIAL TIMES

Google news

ETL

[file.txt]

N-Gram & count

[MapReduce]

Output:((20141024,today),1)

chi-square-test

[MapReduce]

Output: (today,0.03453)

MLlib

[MapReduce]

Output:up

result

[rise , unchanged , down]

Output:(20141024,Today is a good day!)

N-Gram

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus.

Example: Today is a good day

1.Today

2.Today is

3.is

4.is a

5.a

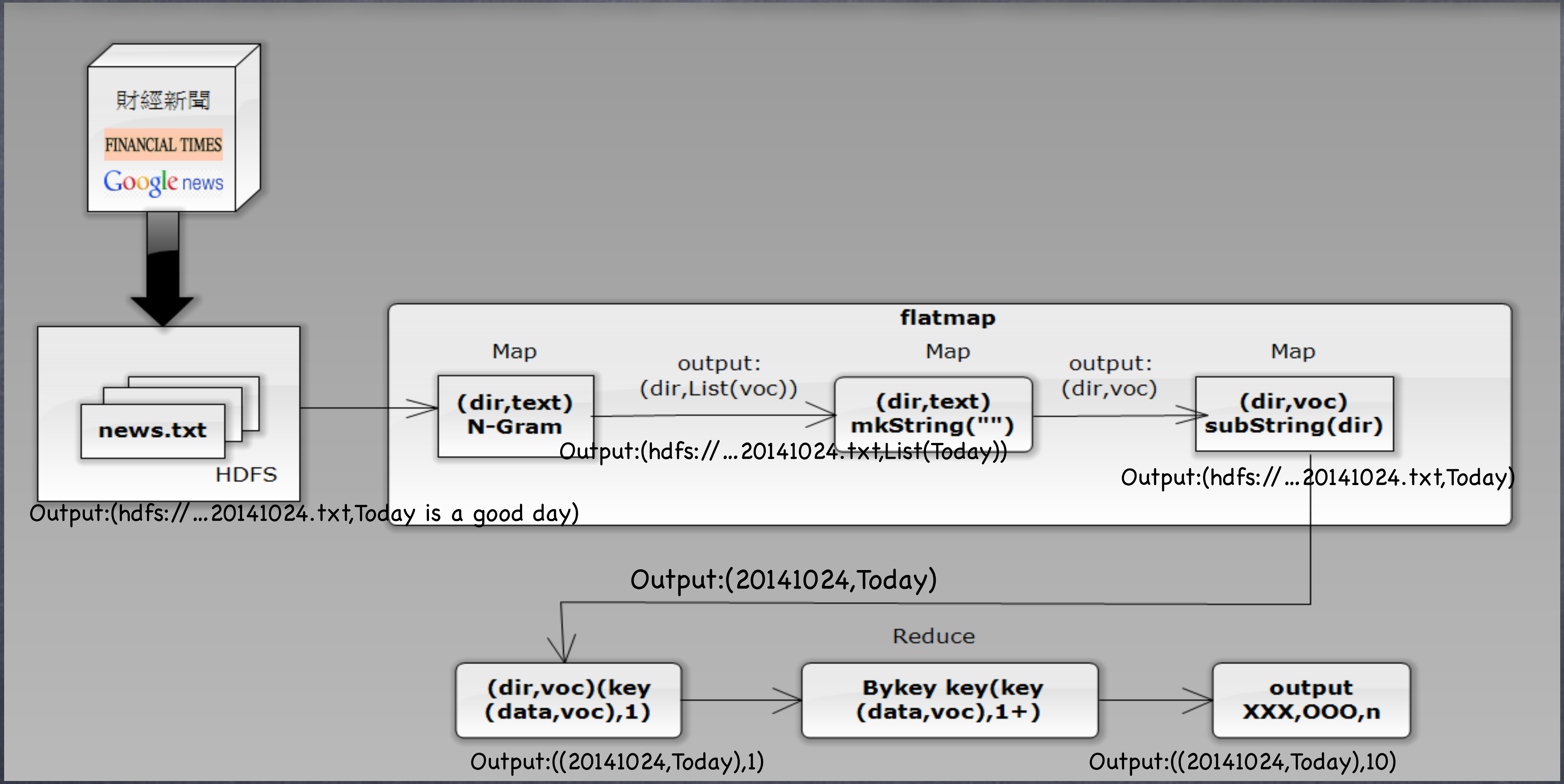
6.a good

7.good

8.good day

9.day

N-Gram Flow Chart



財經新聞

FINANCIAL TIMES

Google news

news.txt

HDFS

Output:(hdfs://…20141024.txt,Today is a good day)

flatmap

Map

(dir,text)
N-Gram

output:
(dir,List(voc))

Map

(dir,text)
mkString("")

Output:(hdfs://…20141024.txt,List(Today))

output:
(dir,voc)

Map

(dir,voc)
subString(dir)

Output:(hdfs://…20141024.txt,Today)

Output:(20141024,Today)

Reduce

(dir,voc)(key
(data,voc),1)

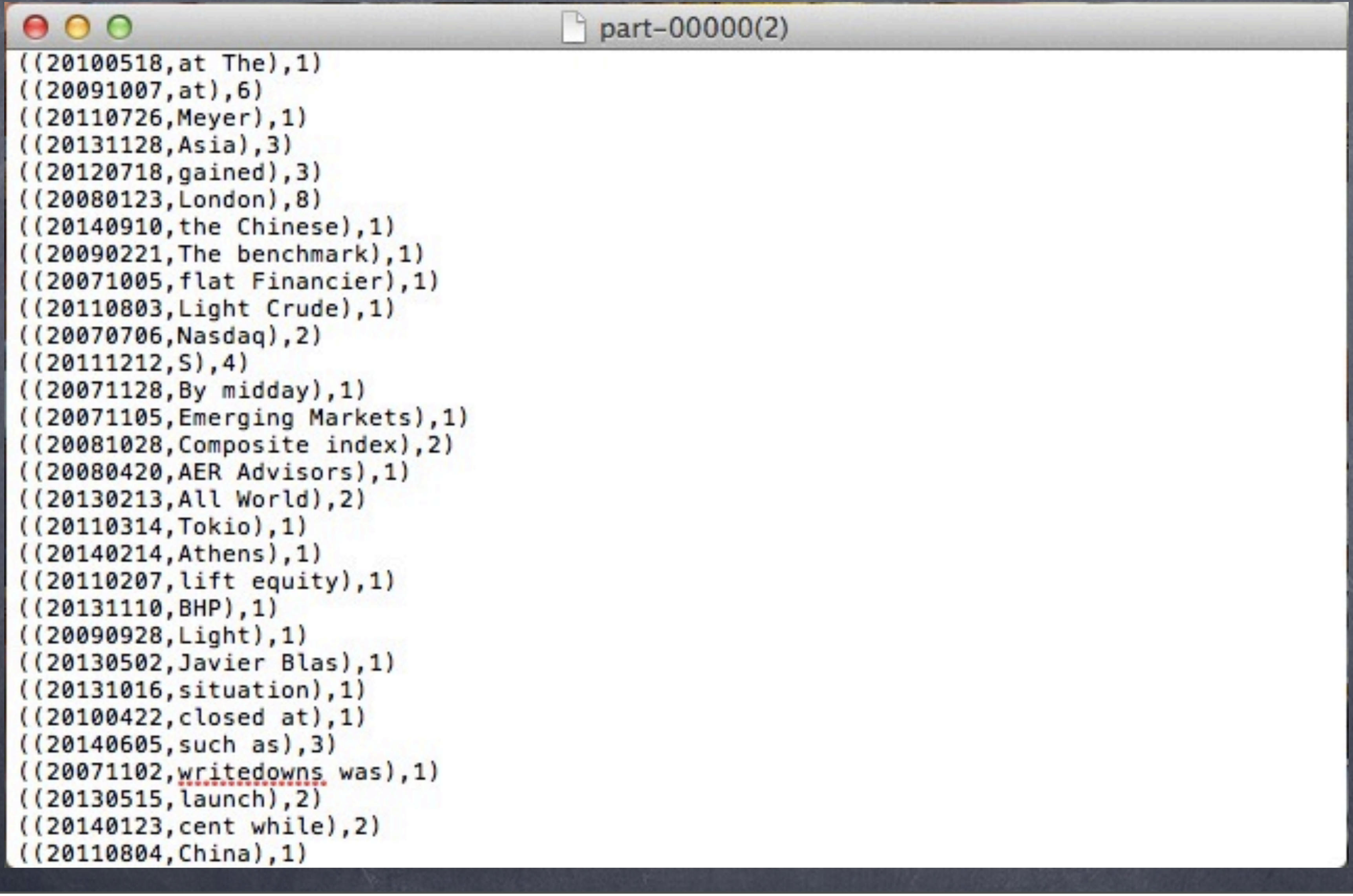
Output:((20141024,Today),1)

Bykey key(key
(data,voc),1+)

output
XXX,OOO,n

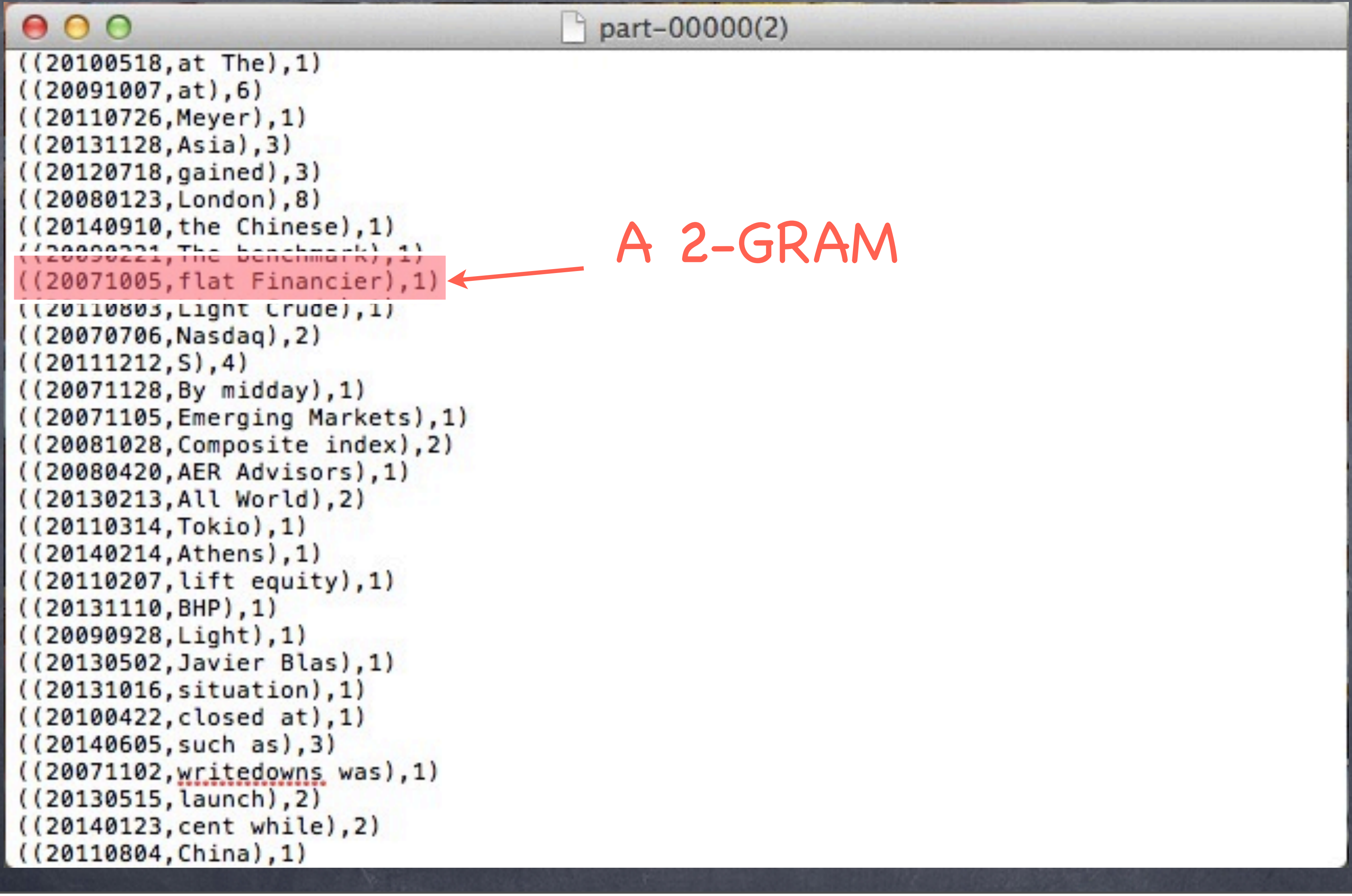
Output:((20141024,Today),10)

N-Gram Result



```
((20100518,at The),1)
((20091007,at),6)
((20110726,Meyer),1)
((20131128,Asia),3)
((20120718,gained),3)
((20080123,London),8)
((20140910,the Chinese),1)
((20090221,The benchmark),1)
((20071005,flat Financier),1)
((20110803,Light Crude),1)
((20070706,Nasdaq),2)
((20111212,S),4)
((20071128,By midday),1)
((20071105,Emerging Markets),1)
((20081028,Composite index),2)
((20080420,AER Advisors),1)
((20130213,All World),2)
((20110314,Tokio),1)
((20140214,Athens),1)
((20110207,lift equity),1)
((20131110,BHP),1)
((20090928,Light),1)
((20130502,Javier Blas),1)
((20131016,situation),1)
((20100422,closed at),1)
((20140605,such as),3)
((20071102,writedowns was),1)
((20130515,launch),2)
((20140123,cent while),2)
((20110804,China),1)
```

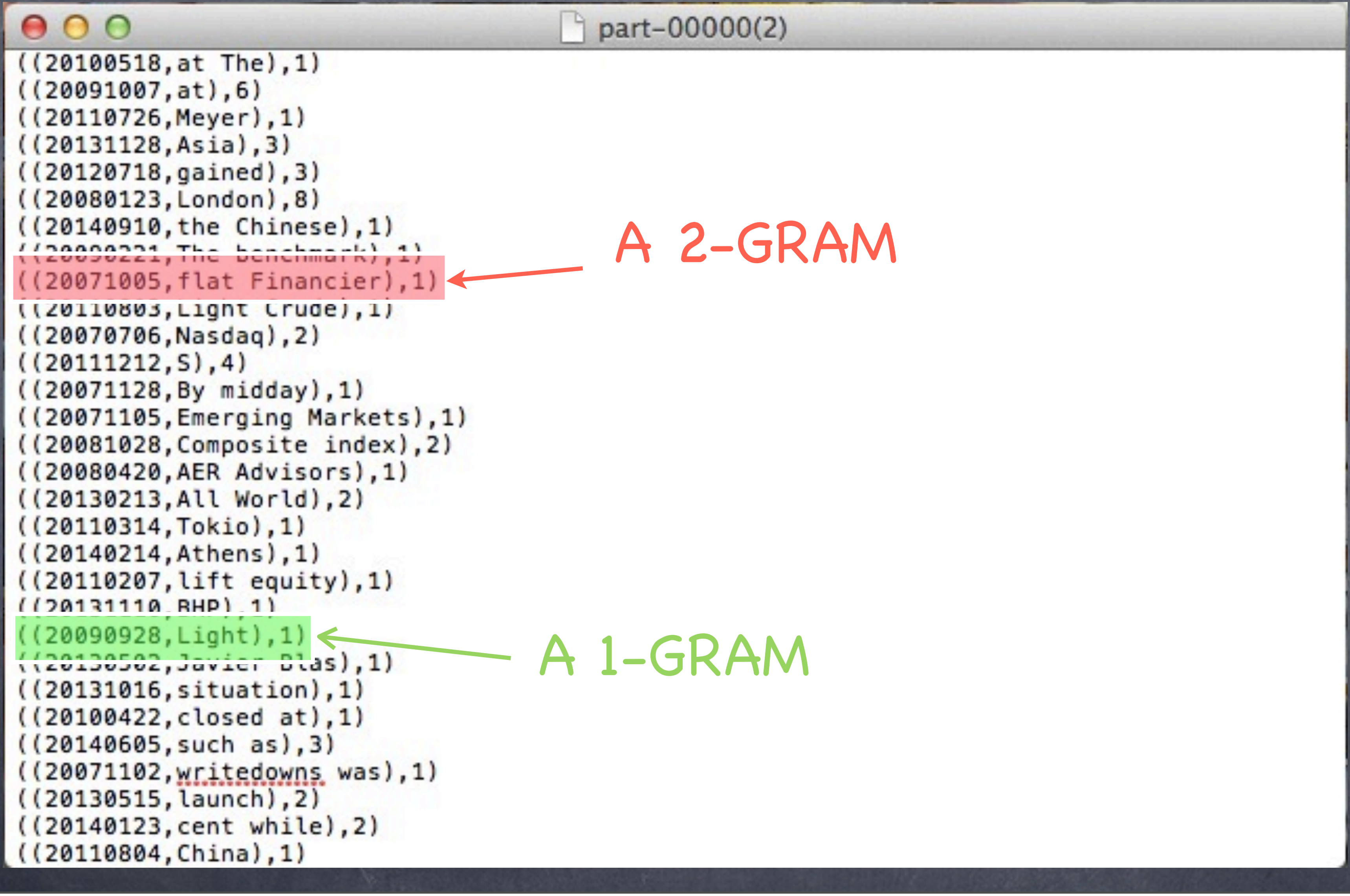
N-Gram Result



```
((20100518,at The),1)
((20091007,at),6)
((20110726,Meyer),1)
((20131128,Asia),3)
((20120718,gained),3)
((20080123,London),8)
((20140910,the Chinese),1)
((20090221,The benchmark),1)
((20071005,flat Financier),1)
((20110803,Light Crude),1)
((20070706,Nasdaq),2)
((20111212,S),4)
((20071128,By midday),1)
((20071105,Emerging Markets),1)
((20081028,Composite index),2)
((20080420,AER Advisors),1)
((20130213,All World),2)
((20110314,Tokio),1)
((20140214,Athens),1)
((20110207,lift equity),1)
((20131110,BHP),1)
((20090928,Light),1)
((20130502,Javier Blas),1)
((20131016,situation),1)
((20100422,closed at),1)
((20140605,such as),3)
((20071102,writedowns was),1)
((20130515,launch),2)
((20140123,cent while),2)
((20110804,China),1)
```

A 2-GRAM

N-Gram Result



```
((20100518,at The),1)
((20091007,at),6)
((20110726,Meyer),1)
((20131128,Asia),3)
((20120718,gained),3)
((20080123,London),8)
((20140910,the Chinese),1)
((20090221,The benchmark),1)
((20071005,flat Financier),1)
((20110803,Light Crude),1)
((20070706,Nasdaq),2)
((20111212,S),4)
((20071128,By midday),1)
((20071105,Emerging Markets),1)
((20081028,Composite index),2)
((20080420,AER Advisors),1)
((20130213,All World),2)
((20110314,Tokio),1)
((20140214,Athens),1)
((20110207,lift equity),1)
((20131110,BHP),1)
((20090928,Light),1)
((20130502,Javier Blas),1)
((20131016,situation),1)
((20100422,closed at),1)
((20140605,such as),3)
((20071102,writedowns was),1)
((20130515,launch),2)
((20140123,cent while),2)
((20110804,China),1)
```

A 2-GRAM

A 1-GRAM

Chi-Squared Test For Independence

A chi-squared test, also referred to as χ^2 test, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.

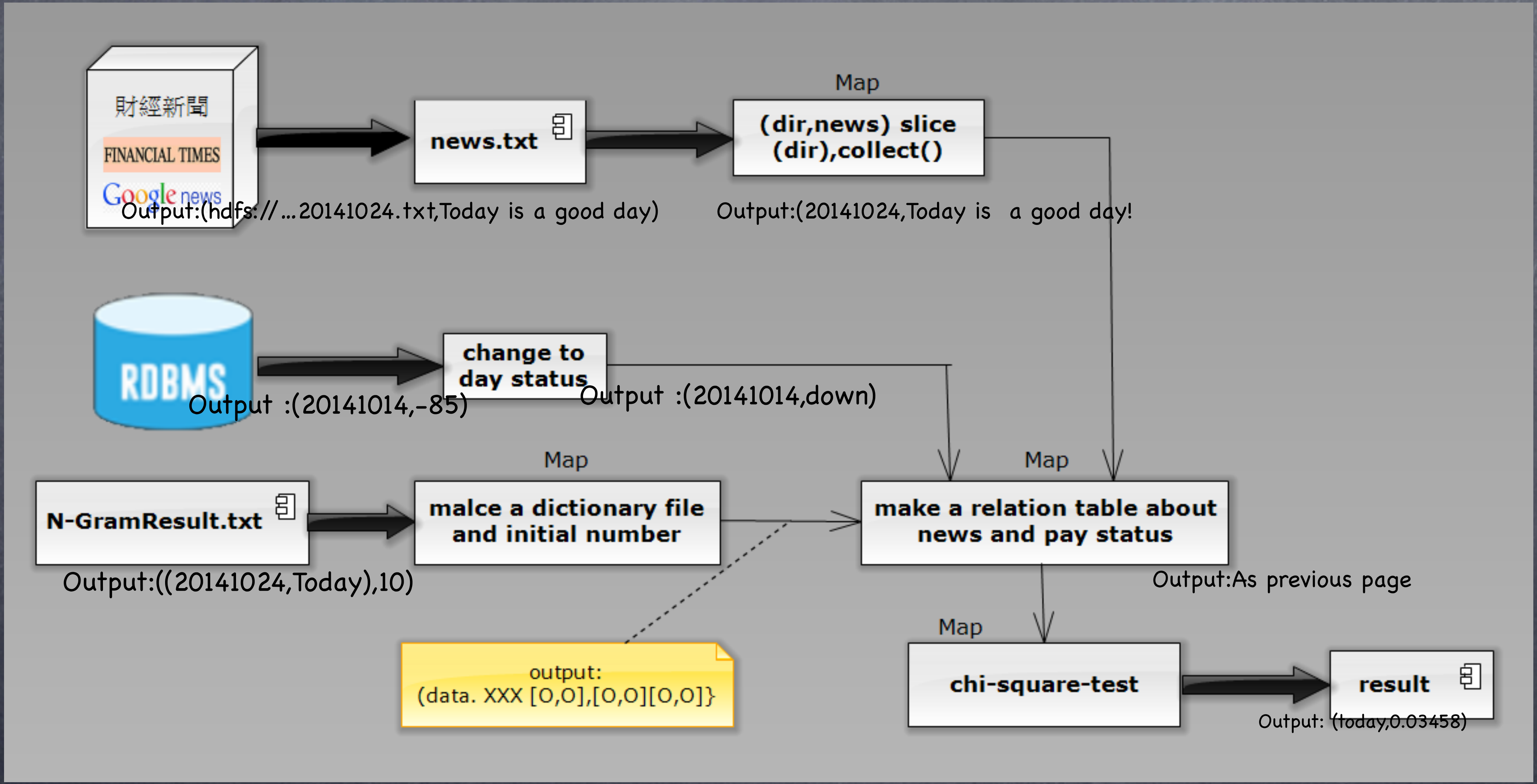
	Yes	No
Up	10	15
Unchange	13	17
Down	14	11

Hypothesis Test:

H0:The word “Today” in first day is independence with it’s next day‘s index status.

H1:The word “Today” in first day is dependence with it’s next day‘s index status.

Chi-squared Test On Spark Flow Chart



財經新聞

FINANCIAL TIMES

Google news

news.txt

Output:(hdfs://...20141024.txt,Today is a good day)

Map

(dir,news) slice
(dir),collect()

Output:(20141024,Today is a good day!

RDBMS

Output :(20141014,-85)

change to
day status

Output :(20141014,down)

Map

N-GramResult.txt

Output:((20141024,Today),10)

malce a dictionary file
and initial number

output:
(data. XXX [0,0],[0,0][0,0]}

Map

make a relation table about
news and pay status

Output:As previous page

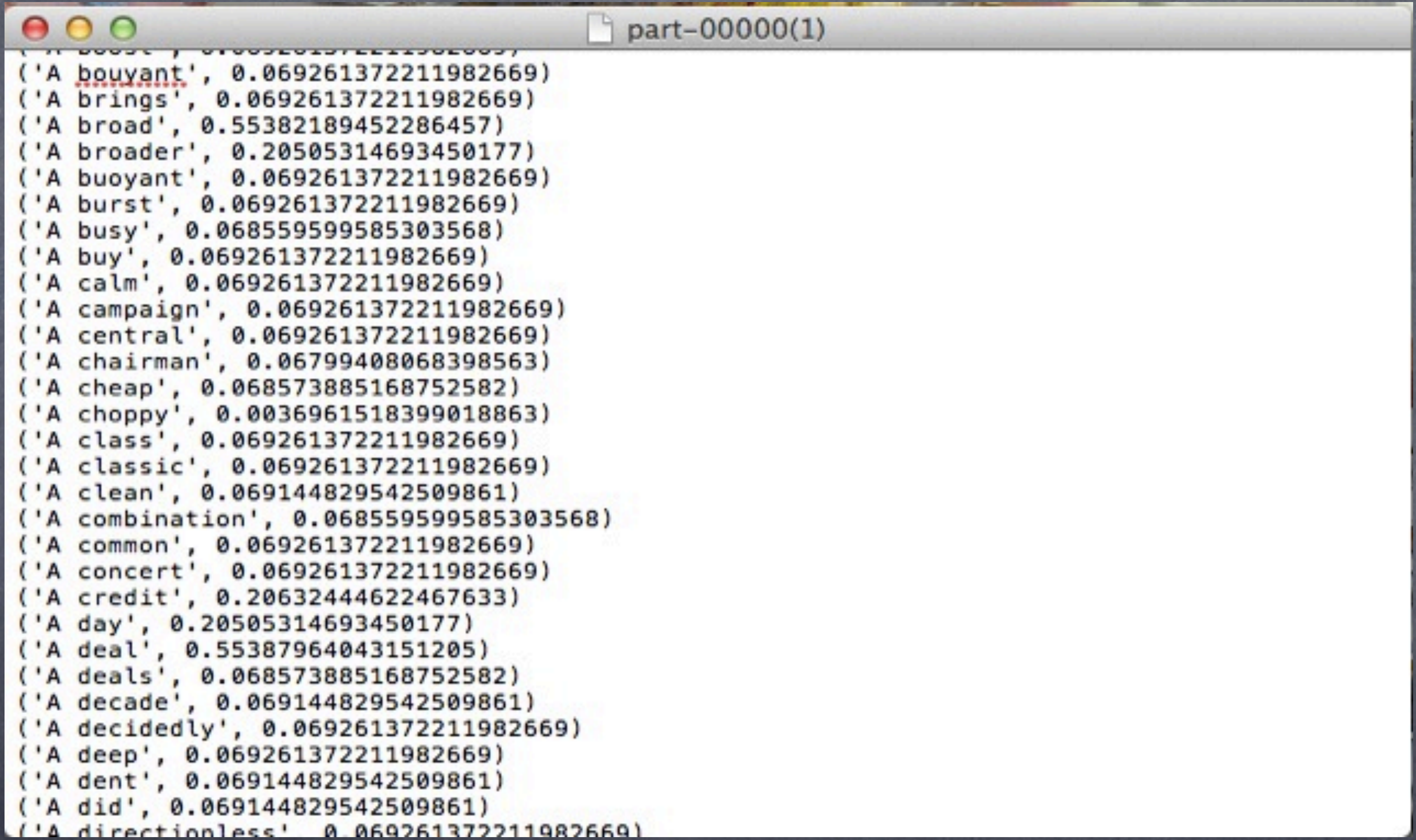
Map

chi-square-test

result

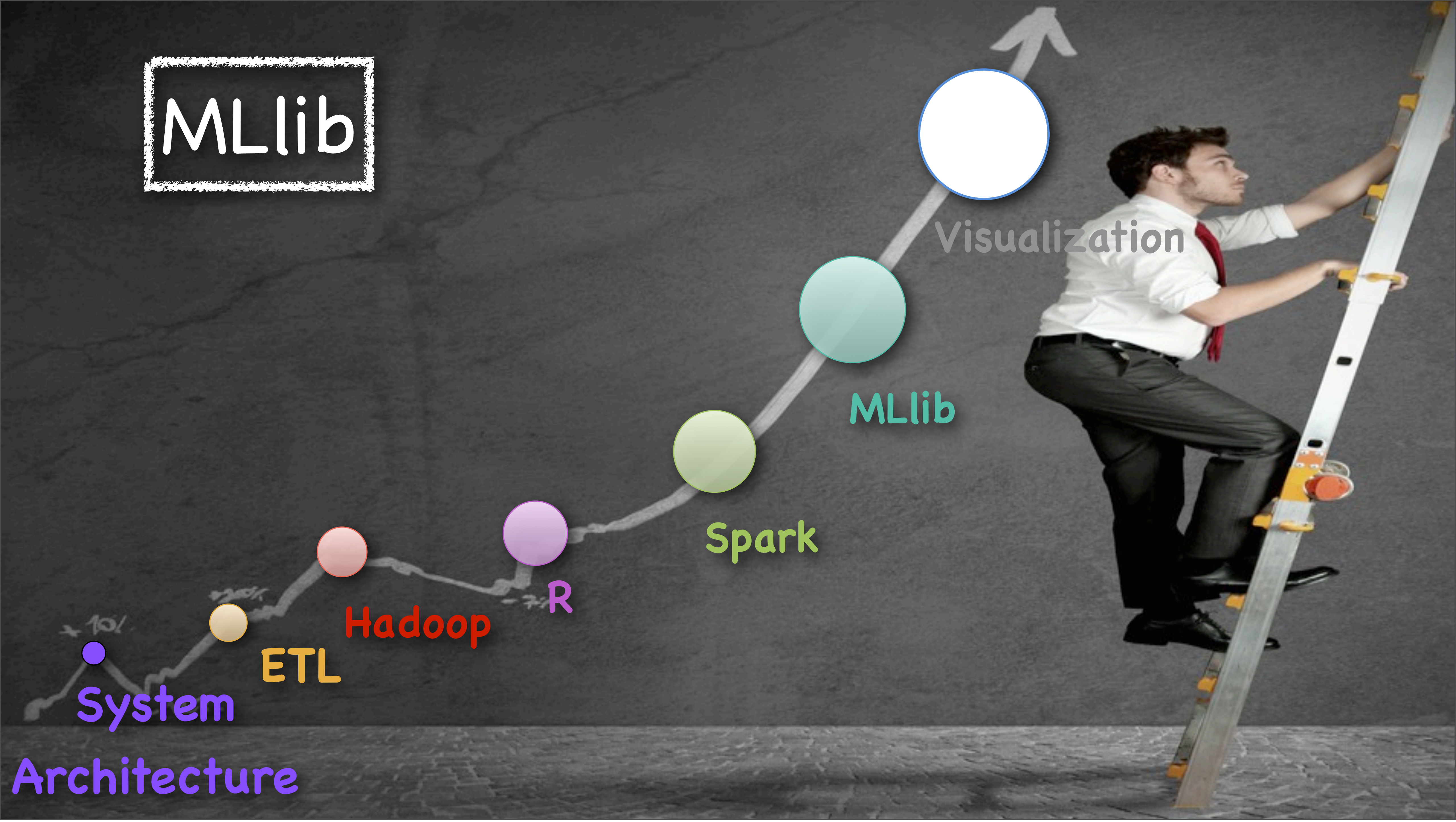
Output: (today,0.03458)

Chi-Squared Test Result



```
('A bouyant', 0.069261372211982669)
('A brings', 0.069261372211982669)
('A broad', 0.55382189452286457)
('A broader', 0.20505314693450177)
('A buoyant', 0.069261372211982669)
('A burst', 0.069261372211982669)
('A busy', 0.068559599585303568)
('A buy', 0.069261372211982669)
('A calm', 0.069261372211982669)
('A campaign', 0.069261372211982669)
('A central', 0.069261372211982669)
('A chairman', 0.067994080068398563)
('A cheap', 0.068573885168752582)
('A choppy', 0.0036961518399018863)
('A class', 0.069261372211982669)
('A classic', 0.069261372211982669)
('A clean', 0.069144829542509861)
('A combination', 0.068559599585303568)
('A common', 0.069261372211982669)
('A concert', 0.069261372211982669)
('A credit', 0.20632444622467633)
('A day', 0.20505314693450177)
('A deal', 0.55387964043151205)
('A deals', 0.068573885168752582)
('A decade', 0.069144829542509861)
('A decidedly', 0.069261372211982669)
('A deep', 0.069261372211982669)
('A dent', 0.069144829542509861)
('A did', 0.069144829542509861)
('A directionless', 0.069261372211982669)
```

MLlib



System Architecture

ETL

Hadoop

R

Spark

MLlib

Visualization

Sparse Matrix In Naïve Bayes Model

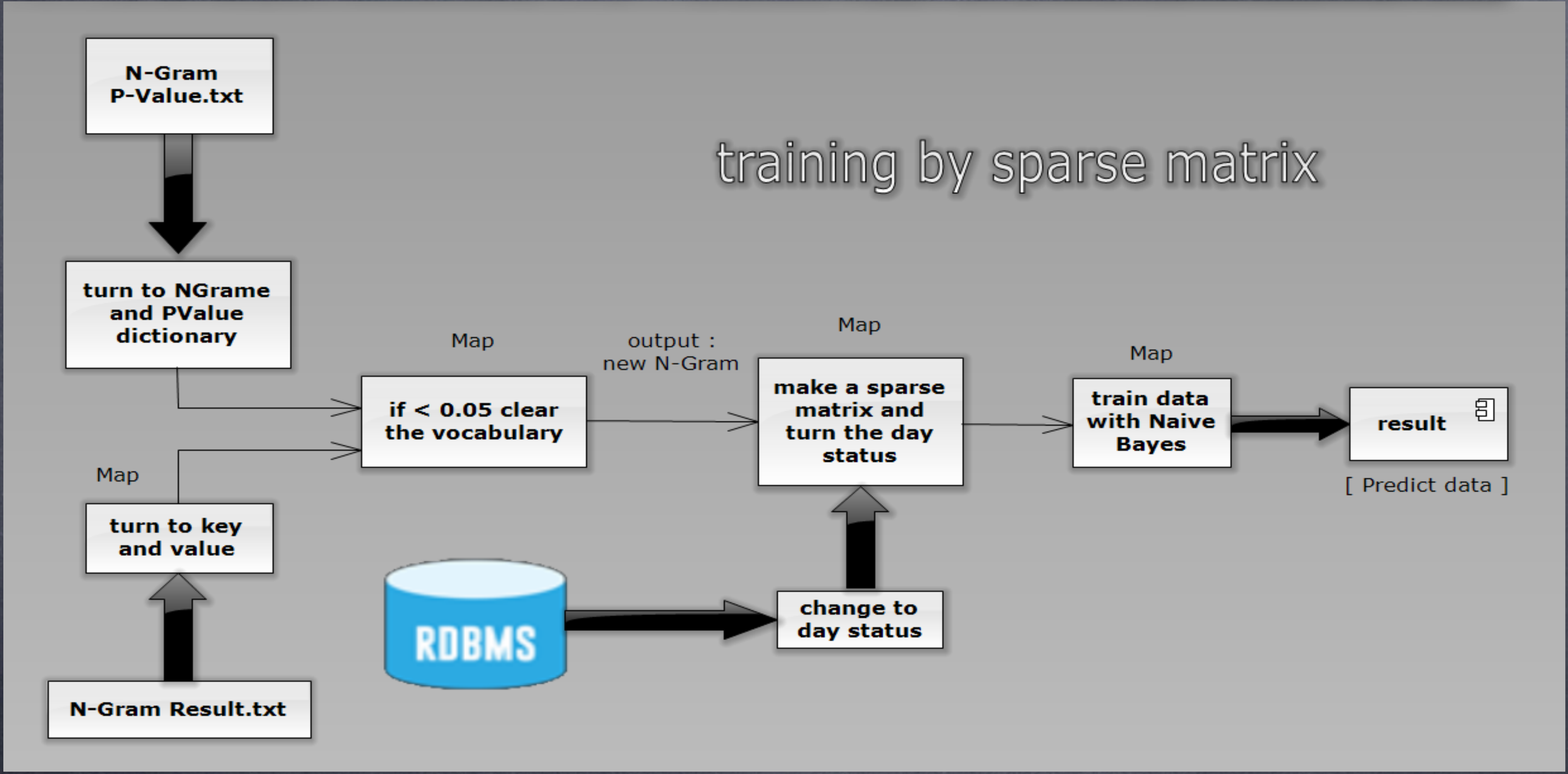
In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(status|Fi)=P(status)P(Fi|status)/P(Fi)$$

1.Where $\{i=1,2,3....\infty\}$ is the word that has appeared

2.Fi means count's of the word

Naïve Bayes Model In MLlib Flow Chart

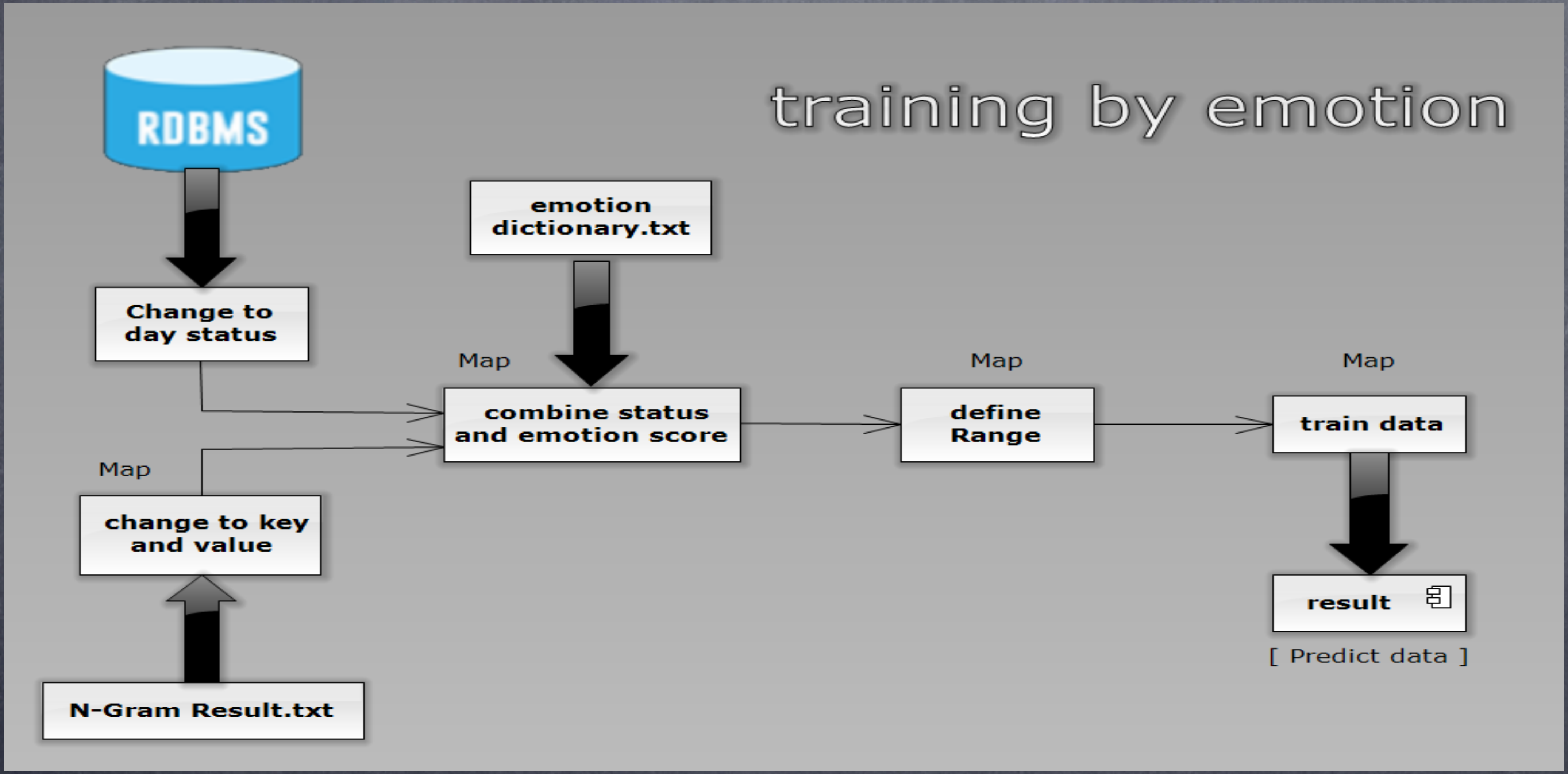

N-Gram P-Value.txt
turn to NGrame and PValue dictionary
training by sparse matrix
Map
if < 0.05 clear the vocabulary
output : new N-Gram
Map
make a sparse matrix and turn the day status
Map
train data with Naive Bayes
result
[Predict data]
Map
turn to key and value
N-Gram Result.txt
RDBMS
change to day status

Emotion Score In Naïve Bayes Model

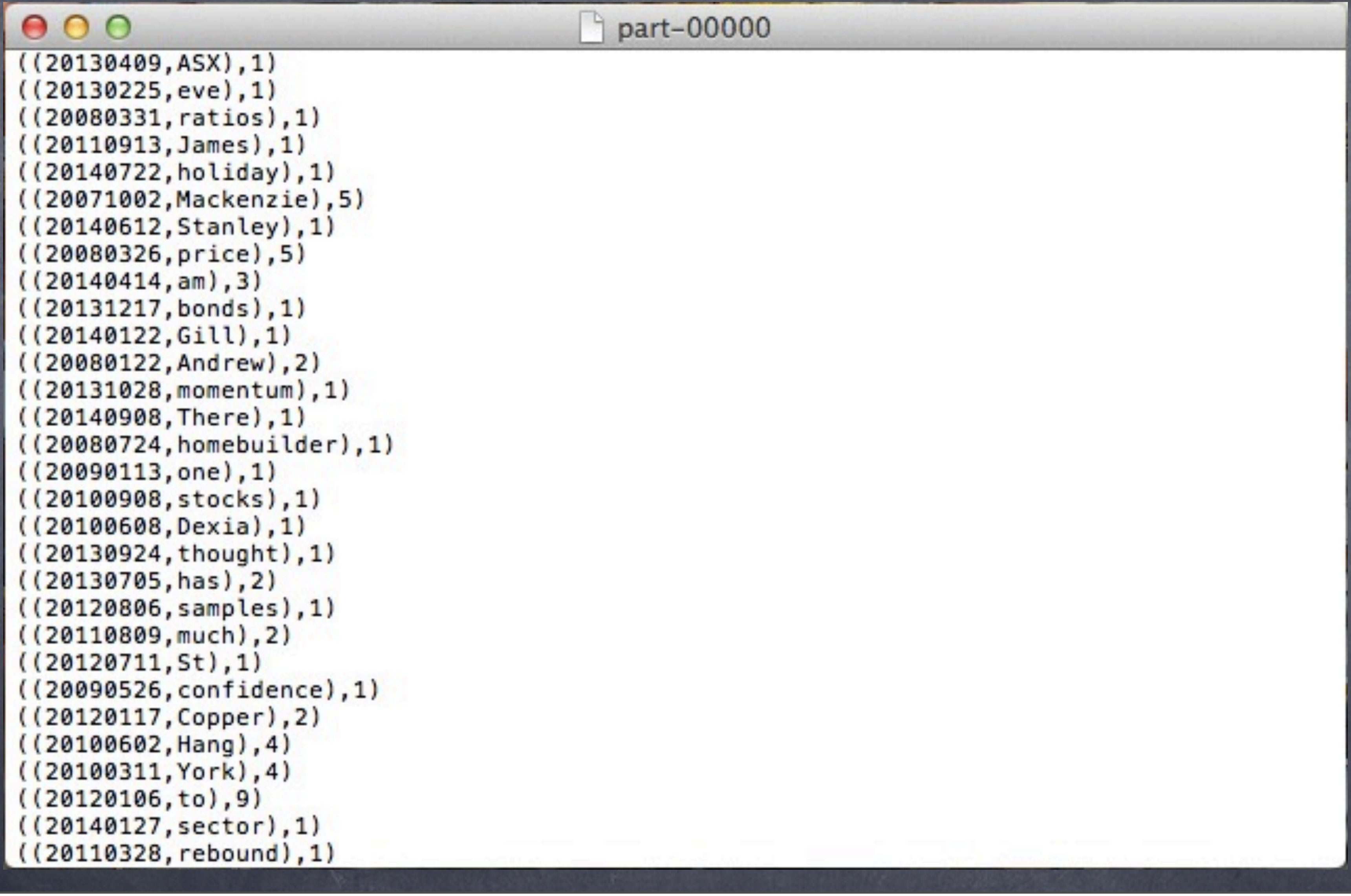
In the previous page we saw that the memory in D-Ram can not handle the sparse Matrix.So we reduce the dimension by emotion dictionary into 2-D vector and put it into Naive Bayes Model to training the data again.....

$$P(\text{status}|[\text{Positive Score, Negative Score}])$$
$$=P(\text{status})P([\text{Positive Score, Negative Score}]|\text{status})$$
$$/P([\text{Positive Score, Negative Score}])$$

MLlib-Naïve Bayes Model Flow Chart

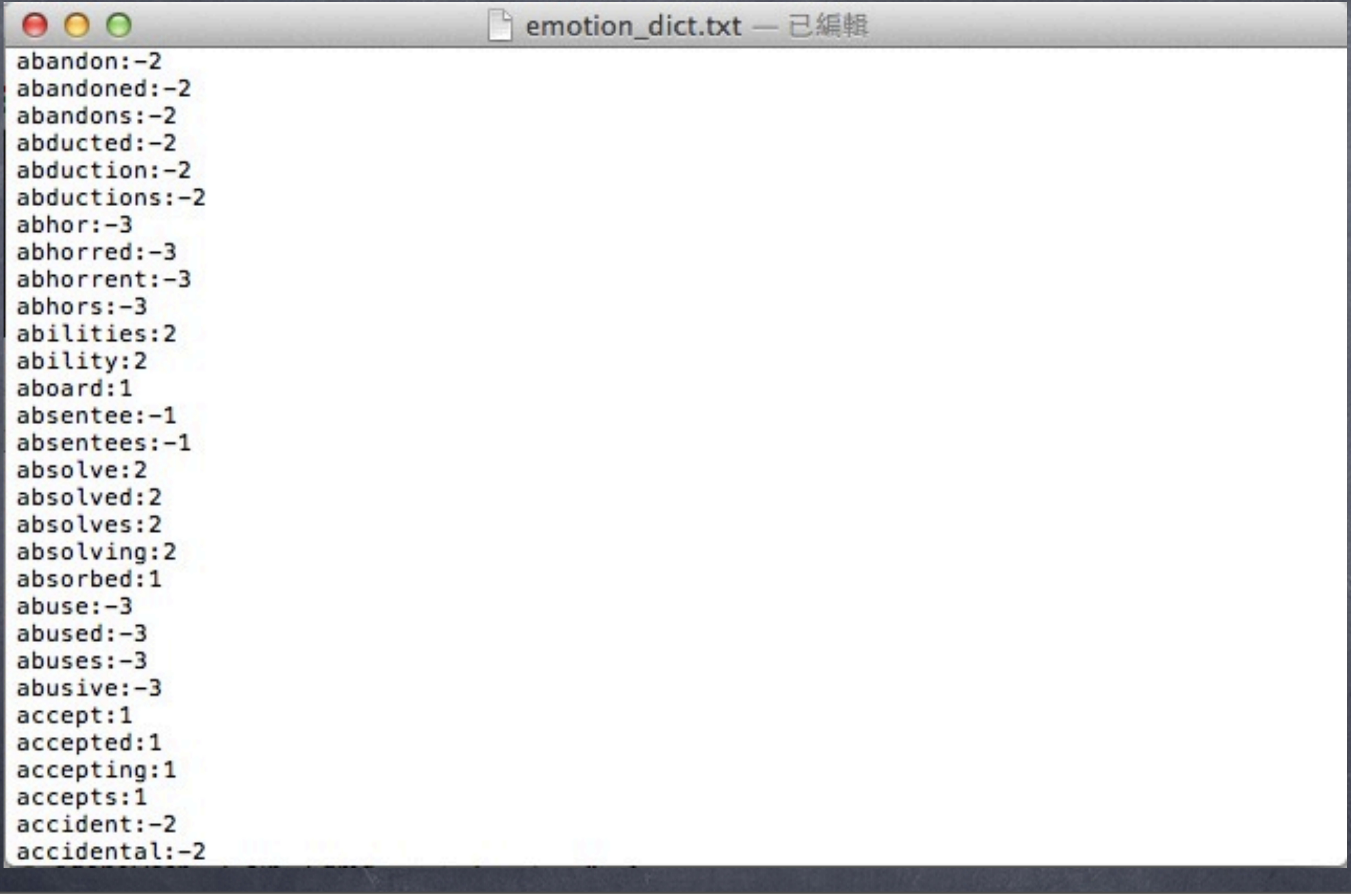


Make Another N-Gram...



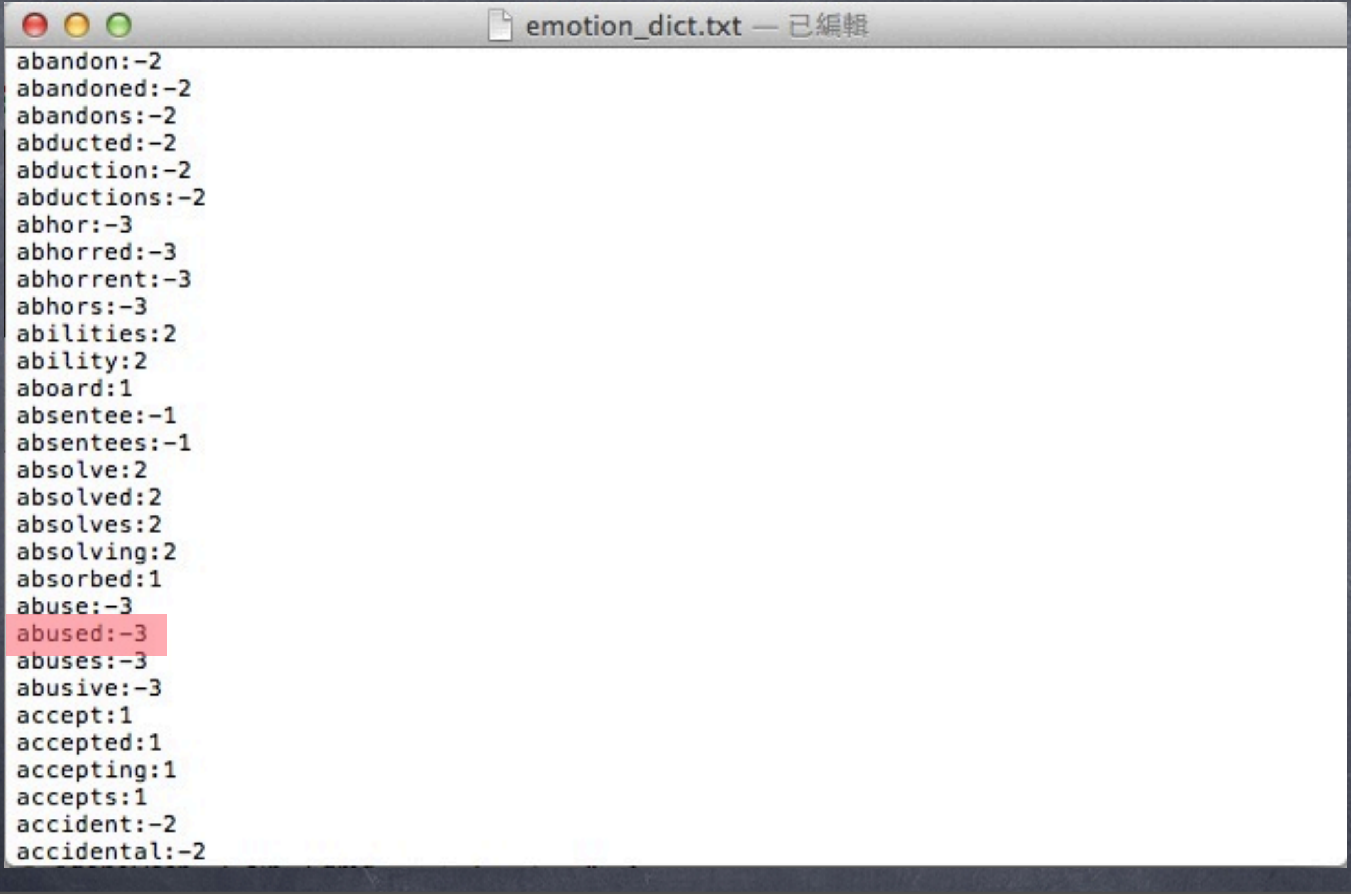
```
((20130409,ASX),1)
((20130225,eve),1)
((20080331,ratios),1)
((20110913,James),1)
((20140722,holiday),1)
((20071002,Mackenzie),5)
((20140612,Stanley),1)
((20080326,price),5)
((20140414,am),3)
((20131217,bonds),1)
((20140122,Gill),1)
((20080122,Andrew),2)
((20131028,momentum),1)
((20140908,There),1)
((20080724,homebuilder),1)
((20090113,one),1)
((20100908,stocks),1)
((20100608,Dexia),1)
((20130924,thought),1)
((20130705,has),2)
((20120806,samples),1)
((20110809,much),2)
((20120711,St),1)
((20090526,confidence),1)
((20120117,Copper),2)
((20100602,Hang),4)
((20100311,York),4)
((20120106,to),9)
((20140127,sector),1)
((20110328,rebound),1)
```

Emotion Dictionary



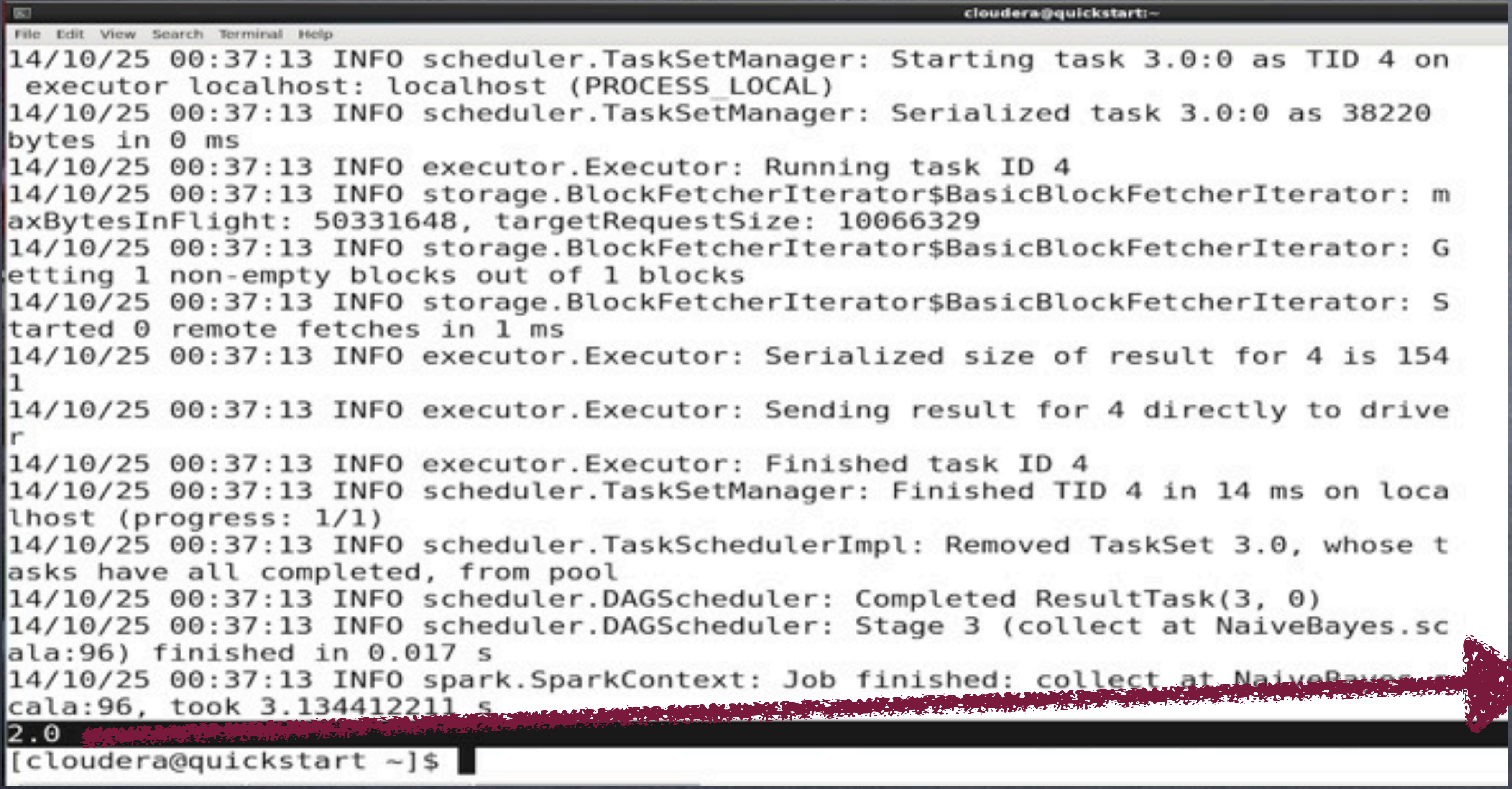
```
abandon:-2
abandoned:-2
abandons:-2
abducted:-2
abduction:-2
abductions:-2
abhor:-3
abhorred:-3
abhorrent:-3
abhors:-3
abilities:2
ability:2
aboard:1
absentee:-1
absentees:-1
absolve:2
absolved:2
absolves:2
absolving:2
absorbed:1
abuse:-3
abused:-3
abuses:-3
abusive:-3
accept:1
accepted:1
accepting:1
accepts:1
accident:-2
accidental:-2
```

Emotion Dictionary

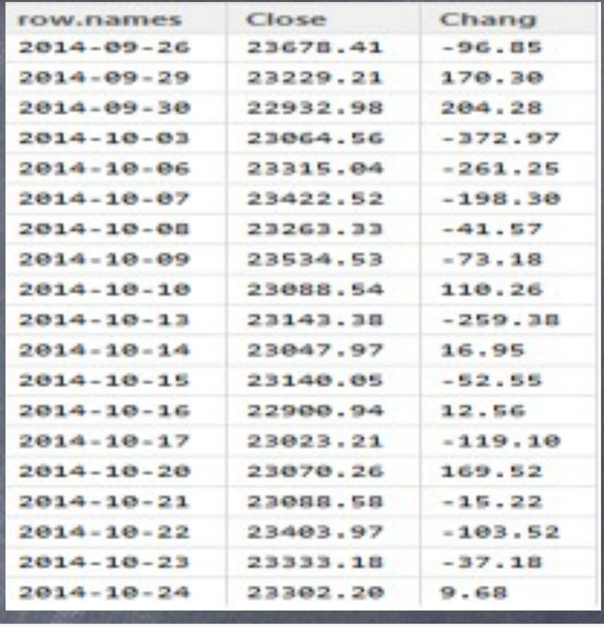


```
abandon:-2
abandoned:-2
abandons:-2
abducted:-2
abduction:-2
abductions:-2
abhor:-3
abhorred:-3
abhorrent:-3
abhors:-3
abilities:2
ability:2
aboard:1
absentee:-1
absentees:-1
absolve:2
absolved:2
absolves:2
absolving:2
absorbed:1
abuse:-3
abused:-3
abuses:-3
abusive:-3
accept:1
accepted:1
accepting:1
accepts:1
accident:-2
accidental:-2
```

Emotion Score training Result



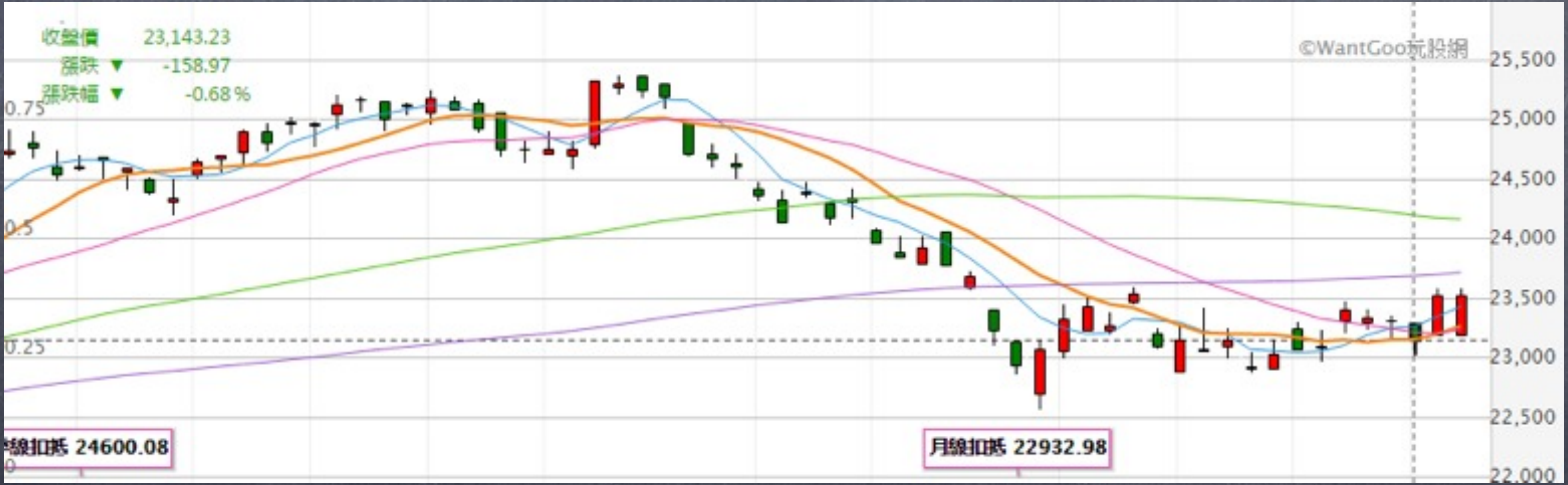
```
14/10/25 00:37:13 INFO scheduler.TaskSetManager: Starting task 3.0:0 as TID 4 on
 executor localhost: localhost (PROCESS_LOCAL)
14/10/25 00:37:13 INFO scheduler.TaskSetManager: Serialized task 3.0:0 as 38220
bytes in 0 ms
14/10/25 00:37:13 INFO executor.Executor: Running task ID 4
14/10/25 00:37:13 INFO storage.BlockFetcherIterator$BasicBlockFetcherIterator: m
axBytesInFlight: 50331648, targetRequestSize: 10066329
14/10/25 00:37:13 INFO storage.BlockFetcherIterator$BasicBlockFetcherIterator: G
etting 1 non-empty blocks out of 1 blocks
14/10/25 00:37:13 INFO storage.BlockFetcherIterator$BasicBlockFetcherIterator: S
tarted 0 remote fetches in 1 ms
14/10/25 00:37:13 INFO executor.Executor: Serialized size of result for 4 is 154
1
14/10/25 00:37:13 INFO executor.Executor: Sending result for 4 directly to drive
r
14/10/25 00:37:13 INFO executor.Executor: Finished task ID 4
14/10/25 00:37:13 INFO scheduler.TaskSetManager: Finished TID 4 in 14 ms on loca
lhost (progress: 1/1)
14/10/25 00:37:13 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose t
asks have all completed, from pool
14/10/25 00:37:13 INFO scheduler.DAGScheduler: Completed ResultTask(3, 0)
14/10/25 00:37:13 INFO scheduler.DAGScheduler: Stage 3 (collect at NaiveBayes.sc
ala:96) finished in 0.017 s
14/10/25 00:37:13 INFO spark.SparkContext: Job finished: collect at NaiveBayes.s
cala:96, took 3.134412211 s
2.0
[cloudera@quickstart ~]$
```



row.names	Close	Chang
2014-09-26	23678.41	-96.85
2014-09-29	23229.21	170.30
2014-09-30	22932.98	204.28
2014-10-03	23064.56	-372.97
2014-10-06	23315.04	-261.25
2014-10-07	23422.52	-198.30
2014-10-08	23263.33	-41.57
2014-10-09	23534.53	-73.18
2014-10-10	23088.54	110.26
2014-10-13	23143.38	-259.38
2014-10-14	23047.97	16.95
2014-10-15	23140.05	-52.55
2014-10-16	22900.94	12.56
2014-10-17	23023.21	-119.10
2014-10-20	23070.26	169.52
2014-10-21	23088.58	-15.22
2014-10-22	23403.97	-103.52
2014-10-23	23333.18	-37.18
2014-10-24	23302.20	9.68

Training Result	Status
1.0	UP
2.0	UNCHANGE
3.0	DOWN

How's The Status Of HSI At 10/27?



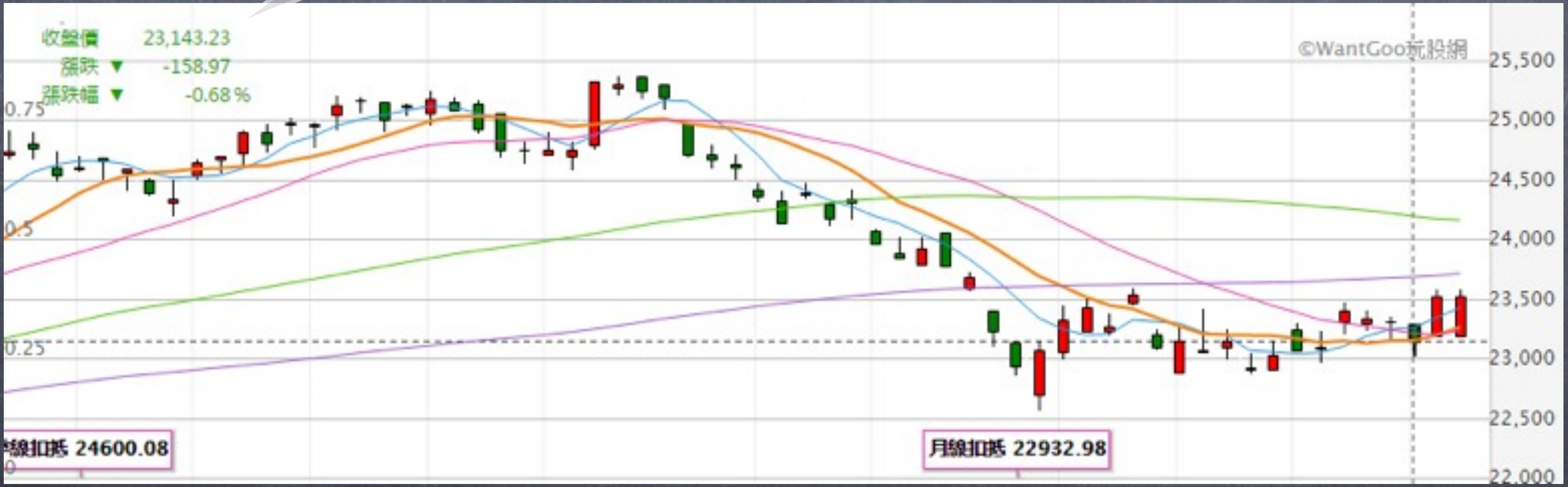
How's The Status Of HSI At 10/27?

-158.97

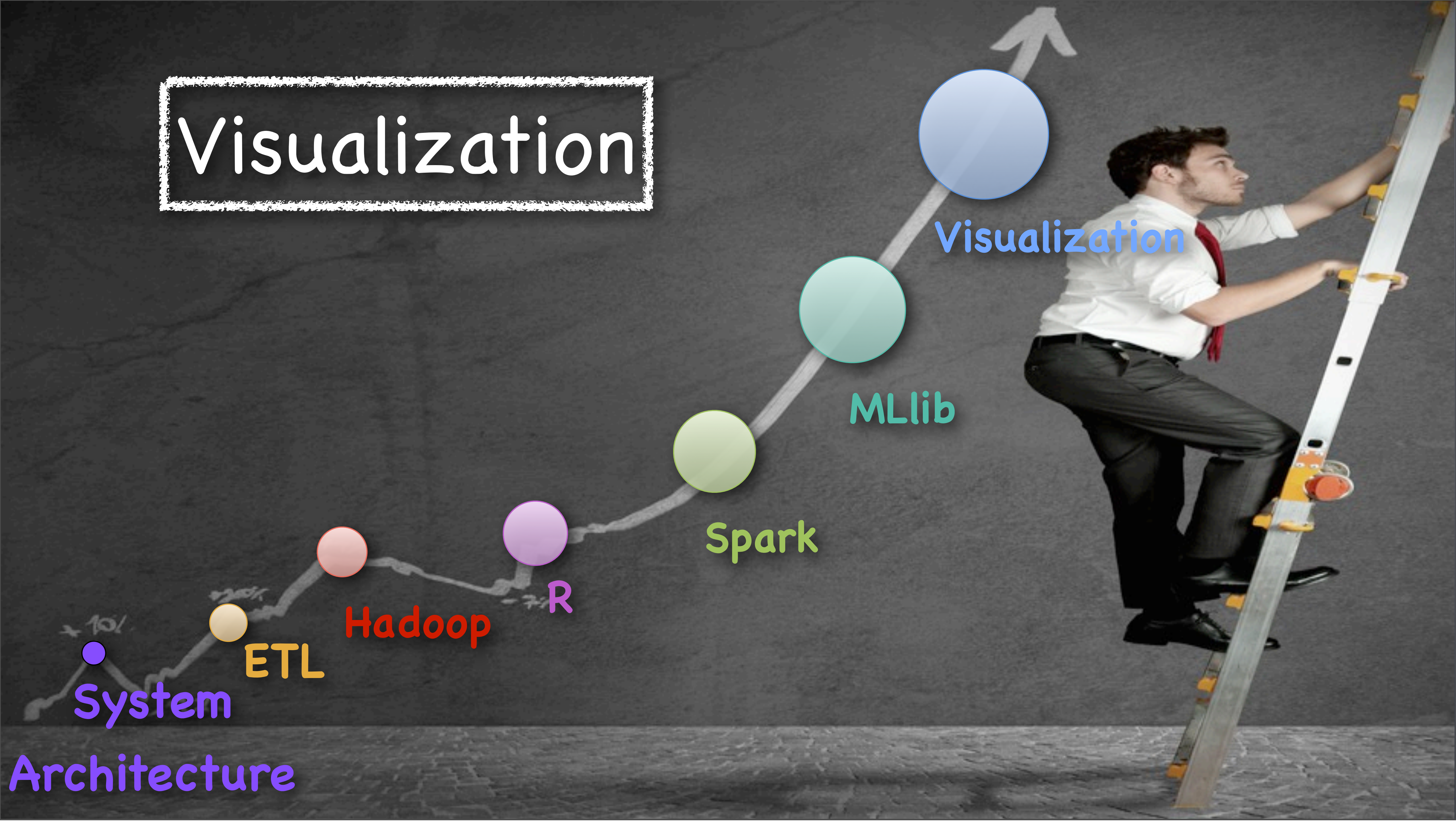
STANDARD ERROR:345.5

345.5 X 0.5 =172.75 > 158.97

Less Than Half Of STANDARD ERROR!!



Visualization



System Architecture

ETL

Hadoop

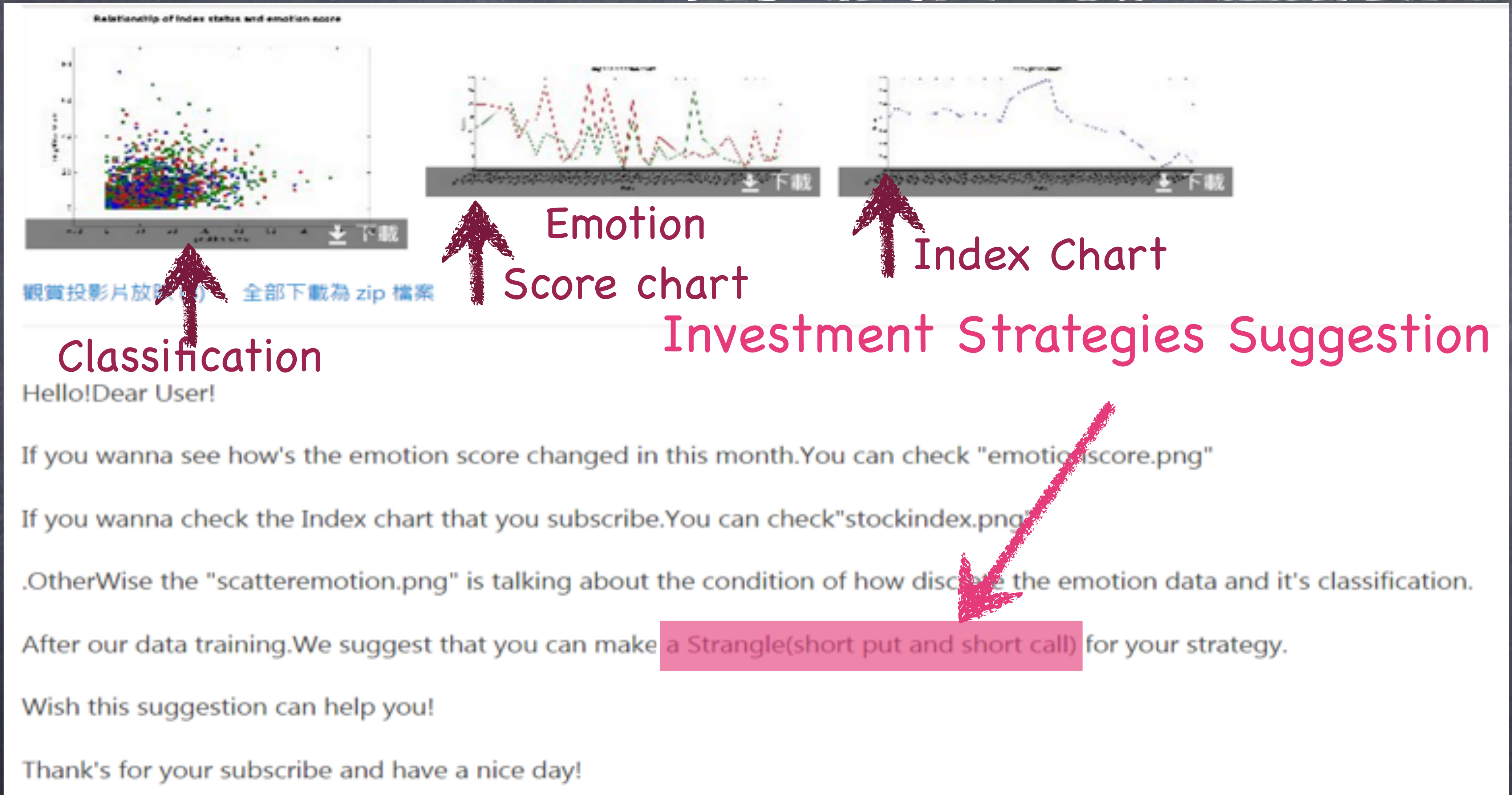
R

Spark

MLlib

Visualization

Sent To User



Classification

Emotion Score chart

Index Chart

Investment Strategies Suggestion

Hello!Dear User!

If you wanna see how's the emotion score changed in this month.You can check "emotionscore.png"

If you wanna check the Index chart that you subscribe.You can check"stockindex.png"

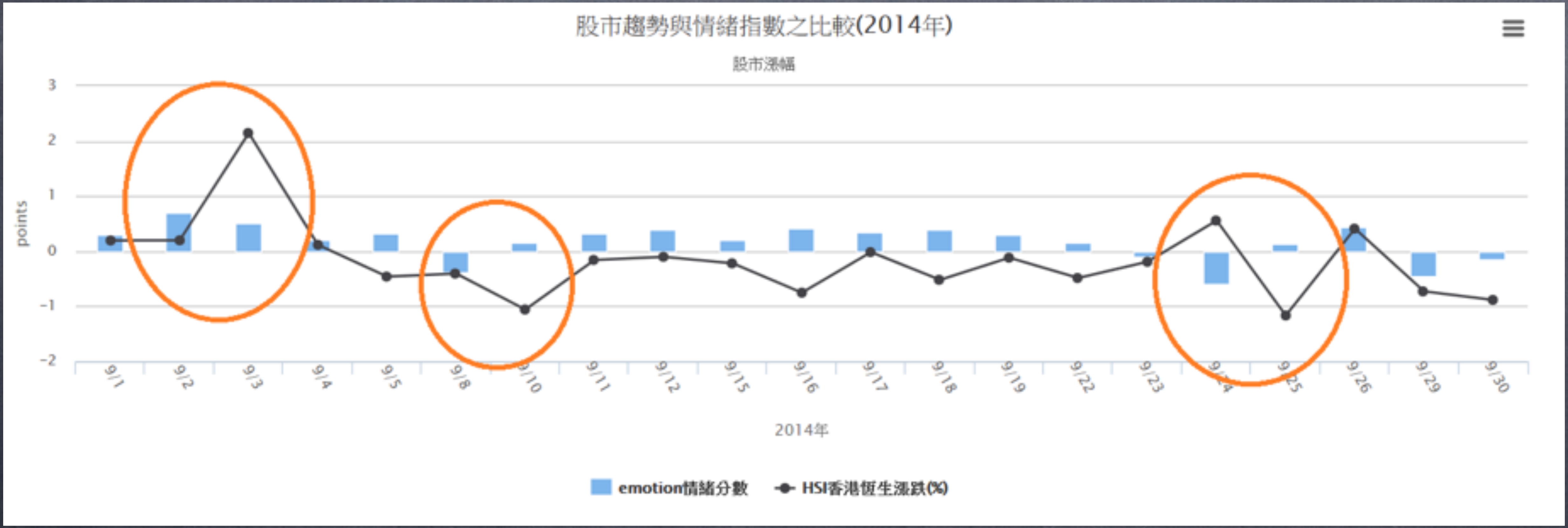
.OtherWise the "scatteremotion.png" is talking about the condition of how discrete the emotion data and it's classification.

After our data training.We suggest that you can make a Strangle(short put and short call) for your strategy.

Wish this suggestion can help you!

Thank's for your subscribe and have a nice day!

Web-HTML5


股市趨勢與情緒指數之比較(2014年)
股市漲幅
points
3
2
1
0
-1
-2
9/1
9/2
9/3
9/4
9/5
9/8
9/10
9/11
9/12
9/15
9/16
9/17
9/18
9/19
9/22
9/23
9/24
9/25
9/26
9/29
9/30
2014年
emotion情緒分數
HSI香港恆生漲跌(%)

Thanks For Your Listening

And Have A Nice Day!