Information Retrival Final Project

A Stock News Analysis System

Vic Tai & Li-Yu Li & Howard Lin & Huang

NTU CSIE

國立臺灣大學 National Taiwan University

Introduction

There's a great deal of people throughout the world eagerly pursuing to be a successful investor, especially on stock market, a relative intense and highly profitable market. Information is what matters most to the fluctuation of stock price. Those who can get the first-hand information and quickly digest them are usually those who can make a great fortune. In an era of information explosion and thousands corporations being traded on stock market, how can we efficiently gather necessary and organized information? Though there are already plenty of tools and websites providing these information, there's none that is user-friendly and interactive enough that can fulfill our needs. Therefore, we present a system that can summarize news of public traded companies on TWSE (Taiwan Stock Exchange).

System Description

Query Expansion

To let users find related companies efficiently, we expand user's query using pre-trained word embedding.

```
('建材', 0.7222533226013184)
('石材', 0.705674171447539)
('混凝土', 0.695707380771637)
('油漆', 0.653332890800476)
('鋼筋', 0.648490786552429 2)
```

Figure 1: 5 most similar words to 'cement'

```
('衣物', 0.6842703819274902)
('器皿', 0.6551379561424255)
('用品', 0.646360397338867 2)
('器具', 0.6347752809524536)
('生活用品', 0.6341350078582764)
```

Figure 2: 5 most similar words to 'utensil'

Company Selection

All enterprises' yield are categorized into 1655 categories, such as semiconductor, insurance, electronic products, etc. All companies are then represented by a 1655-dimention vector, the value of each dimension is the profit yield percentage. Companies are then ranked by the value, and those with nonzero score would be considered as related companies.

News summary

After getting the target company name, our system will fetch the related news from the database, analyze the content and finally extract the most important information in the news.
To get useful information in the news, we use textrank algorithm to summarize the news and output sentences ranked by their importance. In this project, we process the news by three steps. First, we cut the news into several sentences separated by period and further analyze the sentence to get the specific word in a sentence. Second, we compose a graph with the sentences mentioned above and compute the weight between each sentence by the following formula(Eq.1). Third, we use the pagerank(Eq.2) algorithm to summarize the news based on the undirected weighted graph.

Eq.1 Sentence Similarity:

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{1}$$

where S_i, S_j are two sentences in the news and w_k is a word appear in both sentences.

This formula(Eq.1) is used to compute similarity. Similarity between two sentence can be viewed as a process of "recommendation": a sentence that address certain concept in a text, gives the reader a "recommendation" to refer to other sentences in the text that address the same concepts, and therefore a link can be drawn between any two such sentences that share common content.

Eq.2 Pagerank:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} (\frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)) \tag{2}$$

where d is a damping factor(set to 0.85 in this project) that can be set between 0 and 1, which has the role of integrating into the model the probability of jumping from a given vertex to another random vertex in the graph.

Data

We use three kinds of public traded companies information.

1. Basic Information
   We get these information from https://tw.stock.yahoo.com. Detailed information of specific company can further be parsed from links such as https://tw.stock.yahoo.com/d/s/company_2330.html, while the *2330* indicates the stock symbol. The information includes its industry category and profit. For example, TSMC is in the semiconductor industry, and its profit is mainly from producing wafer.

公司資料			
基本資料		股東會及107年配股	
產業類別	半導體	現金股利	8.00元
成立時間	76/02/21	股票股利	-
上市(櫃)時間	83/09/05	盈餘配股	-
董事長	劉德音	公積配股	-
總經理	魏哲家	股東會日期	108/06/05
發言人	何麗梅		
股本(詳細說明)	2593.04億		
股務代理	中信託02-66365566		
公司電話	03-5636688		
營收比重	晶圓88.35%、其他11.65% (2018年)		
網　址	http://www.tsmc.com/		
工　廠	新竹、台南、大陸上海、南京、美國、新加坡		

獲利能力(108第1季)		最新四季每股盈餘		最近四年每股盈餘	
營業毛利率	41.31%	108第1季	2.37元	107年	13.54元
營業利益率	29.38%	107第4季	3.86元	106年	13.23元
稅前淨利率	31.18%	107第3季	3.44元	105年	12.89元
資產報酬率	2.90%	107第2季	2.79元	104年	11.82元
股東權益報酬率	3.59%	每股淨值:	67.21元		

除權資料		除息資料	
除權日期	-	除息日期	108/06/24
最後過戶日	-	最後過戶日	108/06/25
融券最後回補日	-	融券最後回補日	108/06/18
停止過戶期間	-	停止過戶期間	108/06/26-108/06/30
停止融資期間	-	停止融資期間	-
停止融券期間	-	停止融券期間	108/06/18-108/06/21

2. Stock Information
   We get the stock information such as weekly stock price from https://hk.finance.yahoo.com/.
3. Related News
   We get the related news from https://tw.stock.yahoo.com, aiming to find out the reasons of the fluctuation of stock price.

For **Basic Information** and **Related News**, we make use of *urllib*, a python module, to parse information we need from different websites. As for **Stock Information**, we make use of *Selenium*, an tool that can automatically browse websites, and download csv files containing stock price.

Results

Following is the the result of our project. In the following example, we searched "electric vehicle" and we fetched news about Tesla. As we can see in the Figure 3, there are fifteen sentences in this articles and 4 sentences are in the middle of the picture which mean they are more important than the others.

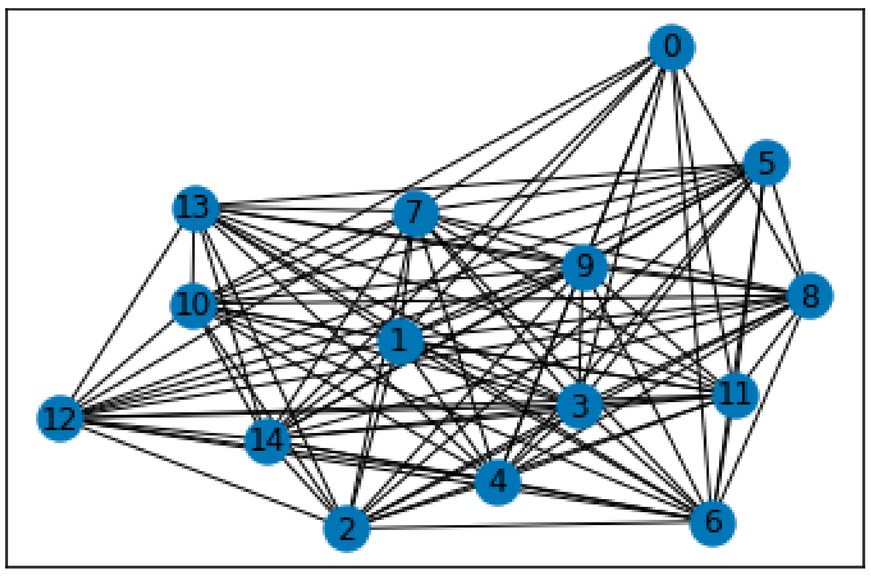

Figure 3: textrank graph

After composing the undirected weighted graph, we used pagerank algorithm to rank the sentences. Figure 4 show the top 3 sentences we extract from the news.

Top three :

Num 0: 馬斯克的高壓管理可從去年 3 月特斯拉董事會確定的績效方案窺見，該方案要求馬斯克 10 年內將特斯拉市值從 590 億美元提升至 6,500 億美元
Num 1: 6 月 19 日，任職兩年的特斯拉人力資源副總裁兼多樣化主管 Felicia Mayo 確認離職
Num 2: 針對特斯拉高層離職頻繁，投資者查諾斯（Jim Chanos）曾表示，當幾乎所有高層都在股價很高時離開，放棄了股權獎勵，這絕不是好兆頭

Figure 4: textrank result

Future Work

We plan to investigate deeper to find out the relationship between enterprises. Our experiments on finding the correlation between stock price movement and profit yield composition failed. We would try to extract keywords from news and even utilize text embedding models to find out latent relationship between companies. We believe our project would be developed into a useful tool for investors to quickly extract information they need, and even make ideal suggestions.

References

[1] Fangze Zhu Baotian Hu, Qingcai Chen. Lcsts: A large scale chinese short text summarization dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing:19671972, September 2015.

[2] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing:404–411, July 2004.