Financial mood

--Text mining final project report--

106022103 劉弘祥 106070020 何羿樺
106023021 王念筑 106070012 周久筠

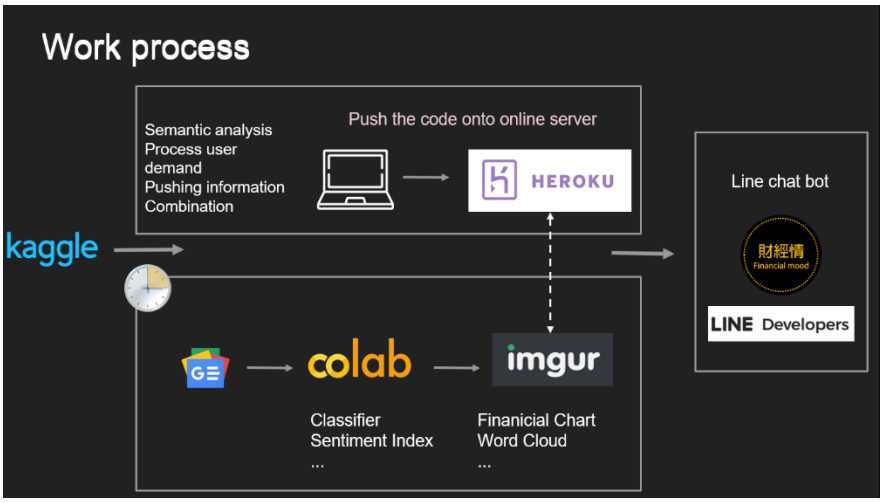
Motivations

There are two members of us that studied in quantitative finance which have a complete concept of finance. And our motivation is due to our interest in financial market and the difficulties we have met on searching financial information as students that are not free in daytime. Thus, we want to make a chat bot that combines many functions which enable the users to find financial information easily.

Objectives

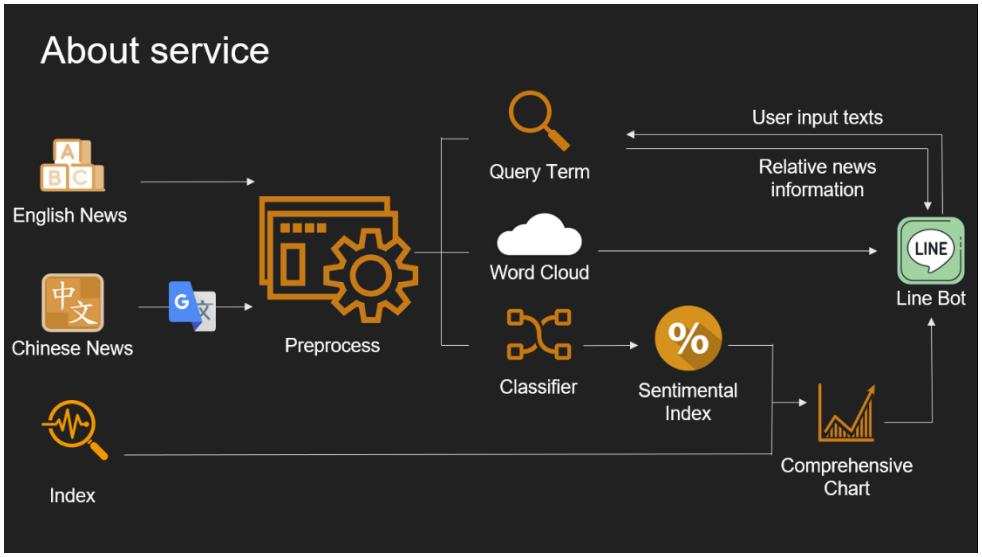
Our target users are those who concern the financial news and information without sufficient time to pay attention to the dynamic information from financial market. The objective of our project is to provide an chat bot that can quickly get integrated financial information with our sentimental index as an overview of the market and the in-depth information of each news.

System framework



The system of our framework is explained properly in the above chart. First, we have a complete data with labels from Kaggle. These texts are the headline of financial news and they are already labeled as “positive”, “neutral” and “negative”.

Then, we use these data to train our classifier model on the colab, and we are now using tfidf_vector to transform our text data to the vectors. Moreover, we are also trying to use doc2vec to obtain better performance and results.



Our service is based on the application of line-bot, and the back end of the system will use work scheduler to collect and input daily news, stock price and index, then finally gives users a readable output. At the same time, we’ll upload the picture of the preprocessed result to imgur, which can both reduce the system’s executing time and the user’s waiting time. Besides, about the server of the line-bot, we’ll push our code to HEROKU,which is a cloud application platform, then we can run our line-bot without local server.

For more detail explanation of our service, on the left side of the picture, three data resources are “English news”, “Chinese news” and “Index” which shows the data we collected from Internet. We translate the Chinese news to English and do the preprocessing which is the same as English news. The preprocess contains POS tagging, tokenization, and lemmatization. There are 2 modes of our service, Taiwan and US version. So the Chinese news and English news would create Query term, word cloud and sentimental index respectively.

The sentimental index would combine with the index we downloaded from Yahoo finance to present. The chart will look just like the last slide, four chart into one figure as the comprehensive index chart.

The word cloud and the comprehensive chart will push to the Line bot automatically everyday. The most difficult part is to analyze the texts that users typed and return with the proper recommendation of news title and the link immediately, since there are only a cloud application platform Heroku for us to operate.

Detailed steps

1. Classifier

   First, we used different model (tfidf2vec and doc2vec) to convert the processed text into vectors.

   Then, we used decision tree, random forest, XGboost, DNN to train the model.

   As the result, the model with highest accuracy is DNN with doc2vec but it’s slow. The model that takes the least time is decision tree with tfidf2vec.

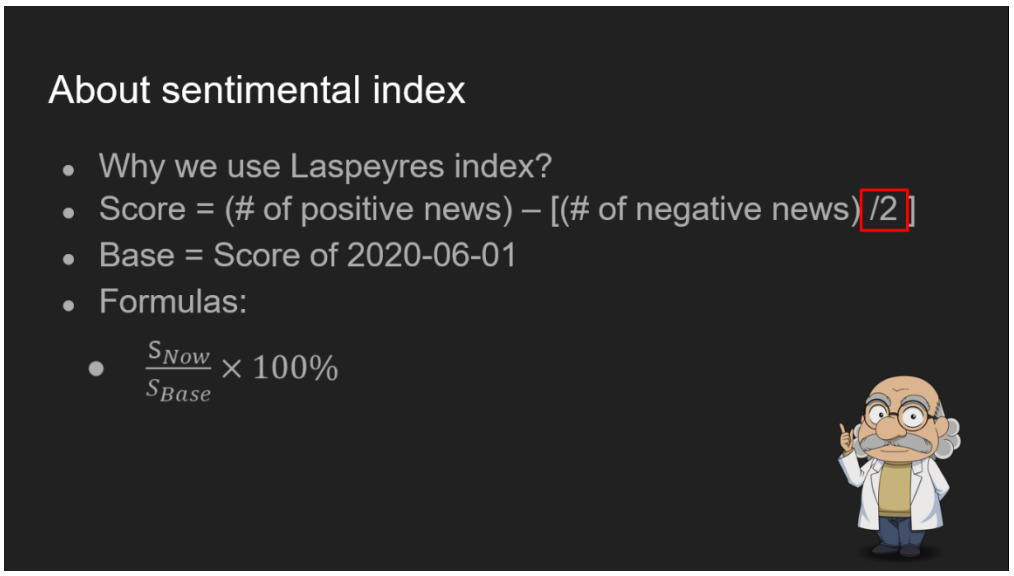
2. Data collection

   There are three main data source in our system: training data, news data and index data. The original data we use to train the classifier is already labeled from Kaggle. Then, we crawled down the News links and titles from Google news with beautifulsoup everyday. And the daily data of financial index from the real market is downloaded from the historical data of Yahoo finance.

3. Sentimental index

Why we choose Laspeyres index?

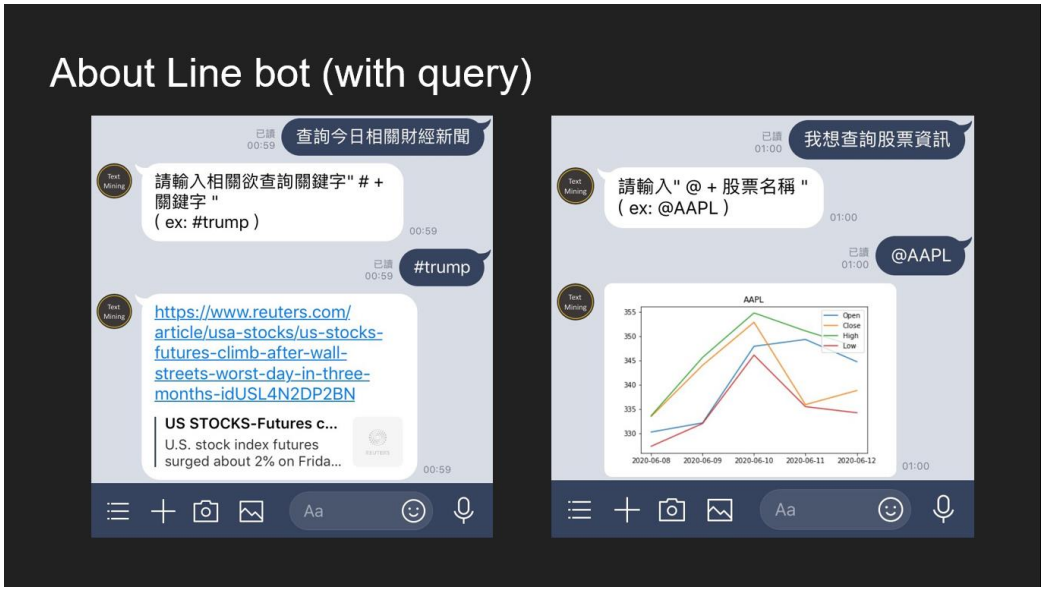
Laspeyres index is a very useful tool to evaluate the change of economical index, such as price index. Therefore, we tried to compute our sentimental index with this method in order to have a proper calculation of sentiment of popular news titles on Google news. First, we set the base date as June first for example, on June first, there are 70 news titles in total and 14 classified as positive, 6 negative, and the rest 50 as neutral.



- We set the score as the number of positive news minus the number of negative news divided by 2. Because of the “fallacy of gambler’s”, people tend to believe in optimistic information. So we divide the number of negative news by 2 to be closer to the real interaction of the market.
- The formula of our sentimental index is “the score of today divided by the score of the base date.

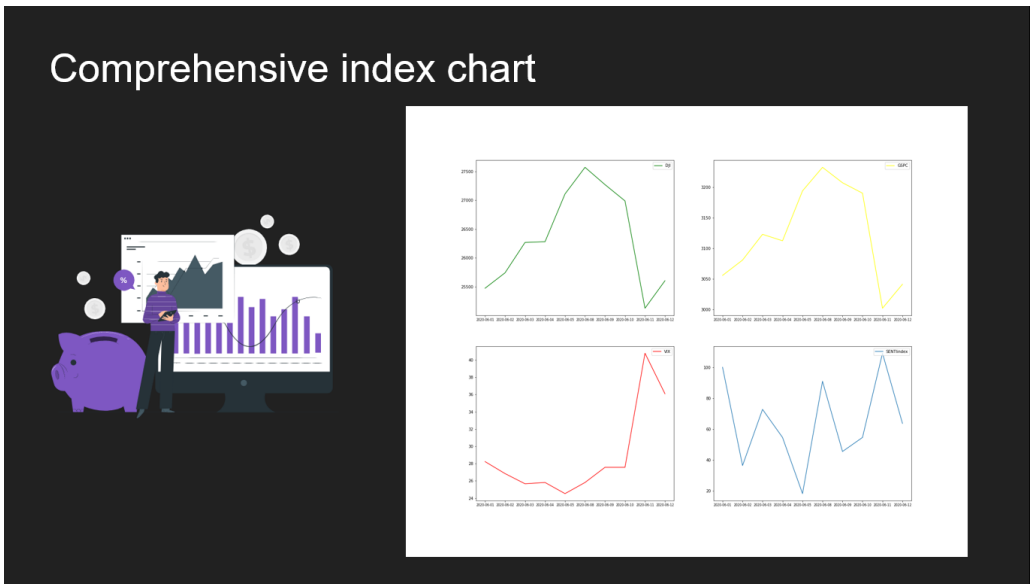
4. Line bot

- Register a line chatbot account and get the channel ID, secret key and access token of the chatbot.
- Register a Heroku account and establish the app.
- Use webhook to link the chatbot and online server (Heroku).
- Write the code to deal with the communication with user and link the service (query, predict, word cloud…) to chatbot.
- Push the code to Heroku and we can run the chat bot now.



Results and evaluation

To verified our result quantitatively, we compare our sentimental with an index called “VIX”, which is a popular measure of the stock market’s expectation of volatility based on S&P 500 index options. As you can see, our index is moving with the same direction as VIX, and the margin of our sentimental index is a lot more significant. But it is quite accurate overall.



We have asked for our family and friends to give us some advice for the chat bot. And we get some suggestions of the improvement of our line bot:

- Shorten the time of updating the index and news.
- Improve the GUI more user friendly.
- As the importance of numbers to the financial market, we should add the number onto the chart.

Reflection

It is easy to plan something ideally but hard to practice, so we have met a lot of difficulties in doing this project. Thus, it took us lots of time to figure out the problems and then try to solve it, and we also felt frustrated after trying many methods but still let the problems unsolved. During the upset situation, we understand the importance of the cooperation, passionate and persistent team atmosphere; therefore, by helping each other, and working hard every week, we finally achieved our goal and felt satisfied with the output of the project. Through the process of the project, we learned a lot of practical experience and realize how hard to practice the ideas. Moreover, every knowledge taught in the class is useful and interesting, and we do gain much from this class.

Appendix (Q&A)

We answered the question from classmates in the online Excel file and we provide the link of the file as below: link.