News tone

Pitch draft

Ying wang: a1733805

Yiru Li: a1736798

Jin Zhou：a1758331

1. What do we do?

   Our project is an application of sentiment analysis named ‘News tone’. We are going to develop a program that can grab the index web page of a news site on a day to analyze the sentiment of it. By analyzing the historical data of the site for a given period, it will display the sentiment outcomes on a dashboard for the convenience of users’ utilization. The purpose of our project is to provide a sentiment analysis algorithm for news to distinguish whether a news report is positive or negative.

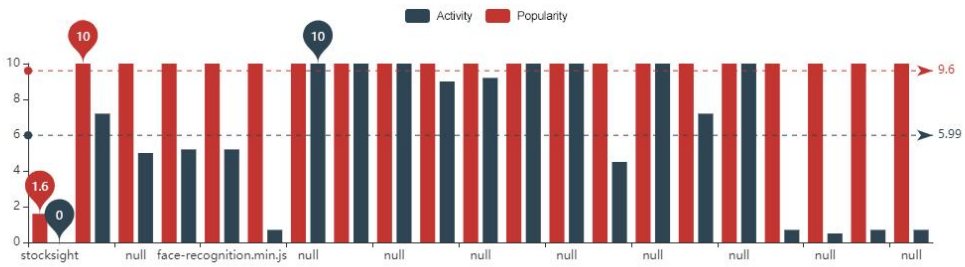
2. Why do we do this?

   People can easily identify each other's emotions such as happiness, sadness, horror and anger during an interaction base on body languages, mood, and expression. However, it is difficult for us to analyze the emotional tendencies of all people in society due to a large number of samples. Understanding the overall sentiment of society is a vital technique which has a variety of advantages in many aspects. For instance, it could assist the financial sector to adopt appropriate strategies to deal with market panics that may occur in the future in advance. This may help relevant stakeholders to avoid a huge loss of wealth. Moreover, analyzing and mining news material information through sentiment analysis to obtain public opinion's sentiment on some hot issues so that it can provide a scientific basis for the strategic decision direction of the government and enterprises.

   As a sentiment analysis program, it is necessary to analyze the emotional tendencies of news reports and news reviews at the same time, in order to correctly analyze the public's attitudes to a hot issue such as a certain policy and corporate dynamic.

3. Example of sentiment analysis application

Stocksight - a prediction/analysis platform of the stock market based on sentiment analysis of twitter and titles of news.



Stock sight is a crowd-sourced stock analysis open-source software that uses Elasticsearch to store Twitter and news headlines data for stocks. Stock sight analyzes the emotions of what the author writes and does sentiment analysis on the text to determine how the author "feels" about a stock. Stock sight makes an aggregated analysis of all collected data from all sources.

Each user running stock sight has a unique fingerprint: specific stocks they are following, news sites and twitter users they follow to find information for those stocks. This creates a unique sentiment analysis for each user, based on what data sources they are getting stocksight to search. Users can have the same stocks, but their data sources could vary significantly creating different sentiment analysis for the same stock. stocksight website will allow each user to see other sentiment analysis results from other stock sight user app results and a combined aggregated view of all.

4. How would the problem be solved?

Through sentiment analysis, we divide news headlines on news sites into positive and negative types based on machine learning. Due to time constraint, our program is structured in 6 steps. We will introduce the process first and then explain how it can be applicable to our project.

Step 1：Extract news content for training: extracting the features of content by crawling content from websites through web crawlers.

We will extract raw data from website – news.com.au. Article titles on the front page of the website will be extracted as the raw data through web crawlers.

Step 2：Feature extractions for training news: different types of news have different features. Sentiment analysis of news will be identified by dictionaries.

The raw data will be filtered that structural words in the article titles should be deleted. Once this progress is done, we can obtain the training data set.

Step 3：Analyze the eigenvalues of training set: in the process of sentiment tendency recognition, the eigenvalue is input into the model, and the model obtains a news sentimental tendency value. It can judge the tendency of the input.

We are going to conduct sentiment analysis on the training data which is a set of words representing a web page. Then sentiment analysis tools (e.g. sentiment analysis dictionary/words database) will be applied to identify whether the page is positive or negative.

Step 4: Training based on news features and machine learning algorithms and establishing/improve News sentiment analysis model.

The results we obtained in step 3 will be input into a machine learning algorithm model for sentiment analysis model training to obtain a news report sentiment model.

Step 5: Classify all the training set.

We are going to label all pages from a period (could be 10 years) with their corresponding results of sentiment analysis (positive or negative).

Step 6：Extract news content to be identified.

Step 7：Extracting features of news to be identified.

Step 6 and 7 have the same operation with step 1 to 2.

Step 8：Determine whether the news is positive or negative based on the features of the news to be identified and the news sentiment model.

At the end of the project, we can set the results into a graph so that it is intuitive to visualize analysis results on a dashboard or a panel. These results shall include at least what types of news show up most often on which days.

5. Project allocation

Role	Name
Developers	Ying Wang, Yiru Li
Testers	Yiru Li
Documentation	Jin Zhou

In our team, each member has his/her responsibility for this project. Based on the table above, Ying and Yiru are responsible for developing and testing the program and Jin is responsible for the documentation. The risks or issues, which may occur during the process will be recorded in the document as

challenges and at the end of the project, we can learn from these experiences and improve efficiency for further developing. Necessary modifications may apply to adjust or improve the final performance of the program, in words, programming algorithm procedures may vary from the plan which described in section four.

At present, we are going to focus on the title on the front page. We may consider the comparison between the title and contents if the pervious task has been solved with enough time left.