Team 23 Ying Wang, Jin Zhou, Yiru Li

NEWS TONE

A sentiment analysis implementation

INTRODUCTION

NEWS TONE is a sentiment analysis program which identifies polarity of particular input text, from the historical records of website 'News.com.au' saved by the website 'Wayback Machine' and demonstrates the result in visual charts. Moreover, users could inspect the change of polarity of the website 'News.com.au' with time by clicking the buttons.

This poster introduce WHAT objectives we have achieved, and HOW we have achieved these

OBJECTIVE

1. We aim to create a sentiment analyzing program which identifies the polarity of certain text material from 'news.com.au' and delivers the outcome precisely and clearly.

2. Extract the subjective information from the input material and identifies the polarity automatically, whether it is positive or negative.

3. The result of the sentiment analysis will be compared with human's identifications for the same information to assess the program.

ACHIEVEMENT

1. Obtaining data from recorded website 'News.com.au' and save these information in the designated template for further investigation purpose.

2. Transform data in a tidy format for algorithm scoring polarity based on multiple corresponding lexicons. A statictical sentiment analysis then practiced.

3. Visualize the result of sentiment analysis and present the change of trend of polarity (positive or Negative) for website 'New.com.au'.

USED SOFTWARE & TOOL

Java Servlet
Back end logic
Request handling

R Language
Statictical process
Sentiment analysis
Outcome Visualization

Python
Web crawler
Data obtaining
Data formatting

JavaScript
Webpage construct
Front end interaction
Data transfer

CSS
HTML
Style
Sheet

PROGRAM ARCHTECHTURE

Our sentiment analysis program involves three key components—the data obtaining which uses a web crawler to gather content sample on the Internet, the statistical algorithm analyzes the polarity of the sample text and the server application that visualizes the outcome and addresses the result to terminals on webpages. Each component is an individual system which executes independently. Our program is a combination of the three vital parts and optimizes their performance.

Figure 1. Flow chart of web crawler processing logic

STEP 1 web crawler gathering web pages and download them in temporary storage.
STEP 2 the program extracts the URL patterns in the HTML files and list the URL in the schedule. Meanwhile, the corresponding 'tags' and 'classes' will be inspected.
STEP 3 the crawler then exports the data in the designated format.
STEP 4 the crawler to continuing to inspect the web URLs in the scheduling pool.

Sentiment analysis is a form of vocabulary analysis that uses lexicons to measure the latent emotional charge of a body of the text. Numerous lexicons exist, scoring text on dimensions from positivity to negativity.

POLARITY

Negative   Positive

−1   +1

Figure 2. Polarity scoring

STEP 1 Tokenises input text into words, and format as being a table with one-token-per-row. A token is a meaningful unit of text, such as a word.
STEP 2 Score the polarity according to corresponding lexicons which are famous for sentiment analysis.
STEP 3 Count frequency of each word appears in the input text.
STEP 4 Output statistical result of polarity of each word.
STEP 5 Generate visualized outcome and repond to the server.

Figure 3. A visualized result of top ranked positive & negative words by statistics

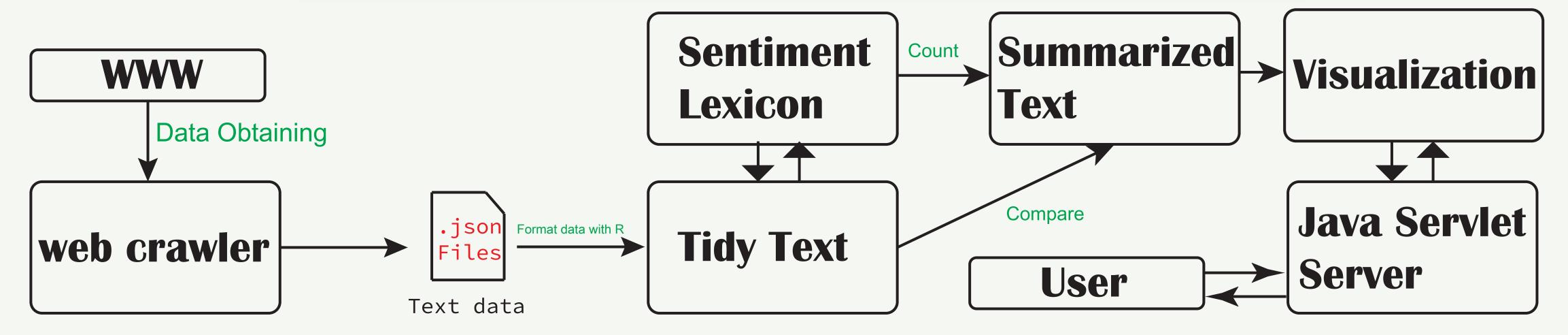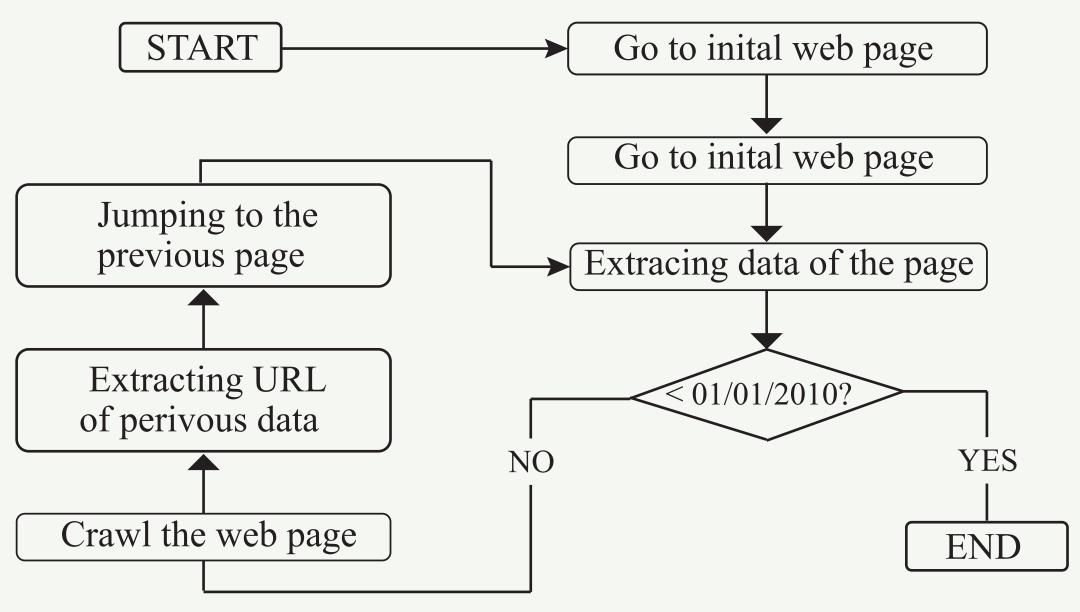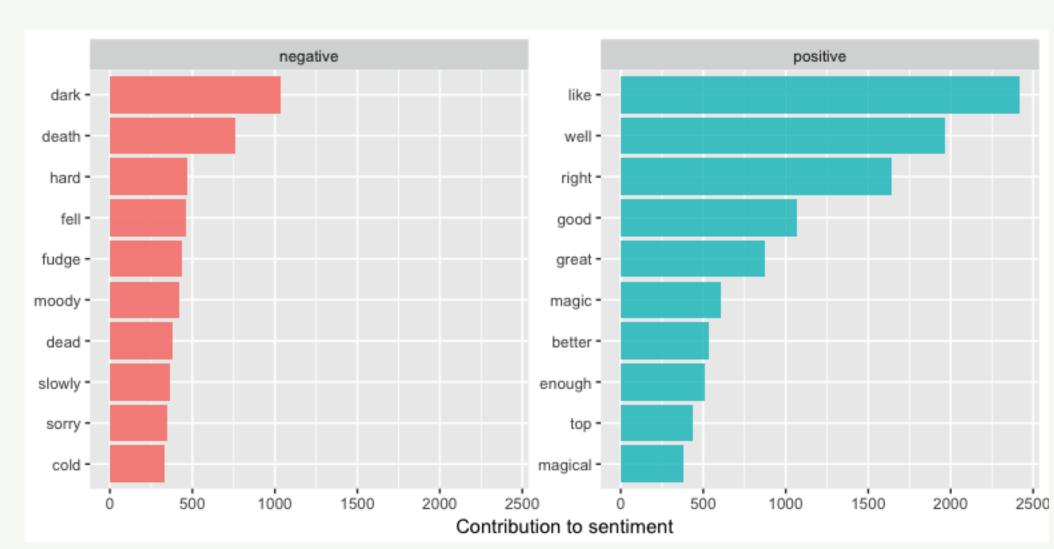## RESULT

The achievement of News tone is delivered through webpage. Scan the QR code to open the webpage, which illustrates how our program delivers the result of sentiment analysis.

EXTENSION

FIRST Of course, any purely automated technique has certain limitations that must be taken into account when applying it to a content analysis project. Our sentiment mining program cannot avoid obtaining advertising data from the web pages. These unexpected ads data will bring a negative impact to our analysis to some extent which cannot be measured at this stage. But considering the advertising data only occupied less than 2% on average of all data collection. We eliminate the influence as a correction.

SECOND Computers have no background knowledge outside of their sentiment lexicons and thus must evaluate a text based purely on its linguistic choice. Thus, lexicon construction determines the correctness of sentiment analysis to a large extent. How to make accurate lexicon is a good topic to investigate.

CONCLUSION

Sentiment analysis is a highly specialized form of content analysis that examines the polarity of particular content to measure even complex characteristics like emotion.

The lexicon we applied in the project is constructed by machine learning technique that has been developed in the recent years. But unlike human, automated sentiment analyzing does not consider whether a concept itself is positive or negative, only how it is portrayed linguistically in the given text.

REFERENCE

[1] Leetaru, K., 2012. Data mining methods for the content analyst: An introduction to the computational analysis of content. Routledge.
[2] Silge, J. and Robinson, D., 2017. Text mining with R: A tidy approach. "O'Reilly Media, Inc.".