Red Envelope Analysis of Applications in WeChat Group

Chen Liang Student ID: 46275313

Background

WeChat is the most popular chatting application in China, covering nearly 1 billion people. Therefore, in China, whether it is takeaway, taxi calling or video application, the operator would post red envelopes through WeChat to do business promotion. Red envelopes are shared through WeChat chatting group, and both the person who sends it and the one who receives it could get extra rewards or discount for services such as takeaway or calling taxi provided by the app.

Objective

Through analysing a dataset of electronic red envelopes being shared in a Wechat chatting group composed of the undergraduates in a certain university in Beijing, to understand the overall use of sorts of types of apps and the characteristics of their use at various times. At the same time, study the frequency of the advertising words used in the shared red envelopes to infer the advertising words that are more attractive to customers.

Method

The dataset contains 19675 observations from August 2015 to August 2017. The main variables which were used to analyse were the time of red envelope being shared (variable 'time'), advertising words and URLs. The advertising words and URLs were kept in 'content' variable as string. All the analysis process was based on python, and the packages used are Tkinter, datetime, Jieba, OS, SciPy, numpy, wordcloud, Matplotlib and collections.

1. Data Processing

Firstly, Since the advertising words are all Chinese, the `string_code_identify(b: bytes)` function is defined to identify the Chinese code in the document to avoid gibberish or loss of information.

For the red envelopes in the data set, it was necessary to identify the domain names contained in the URLs to determine which app the red envelope belonged to. Therefore, the `get_domain_name` function was defined to identify the domain name from the red envelope URL, and then a list of all domain names was formed through the `get_domain_list` function. A histogram was drawn through draw_bar(domain_main) to show the top five domain names.

Due to the fact that there were some domain names appeared rarely, there was a lack of contribution of those observations to the statistical process. Therefore, in order to ensure the reference of the data and reduce the workload of type recognition, the observations which were included in the main domain name appearing less than 30 times was removed when conducting `process_data(data, domain_main)` function, and the data set then contained 18907 observations.

2. Data Analysis

To make the data more consistent, class Red () defined a class Red, representing the red envelope observation. The attributes of Red were, the time when red envelope was shared (obtained by using `get_time(datetimestr)` function), the advertising words on the red envelope (got through `get_words(sentence)` function), the type of the application and the domain name of the red envelope. The function was mainly based on `jieba`, which was a Chinese word segmentation module based on

Python, which could accurately segment Chinese documents based on large-scale corpus.

In order to find out which advertising words were most popular for different types of apps, the advertising words appearing most frequently should be figured out. The wordfrequency (word_list) function based on ‘collections’ package was used to calculate the number of appearance of various advertising words for each category of app. Next, use draw_wordcloud (freq, n) function to make word cloud to visualize the result.

In order to study the number of red envelopes belonging to different types of apps being shared in different time periods during a day, the timeseries_dic (data) function was used to select time and apptype of Red for statistics, and then all the appearance time of each type was got. After that, use timeseries_draw (timetype) to draw a time series plot for each type so as to check what time periods of red envelopes for these types were shared the most frequently. In the process of drawing, in order to facilitate statistics, the time was rounded as the hour that had the least time difference. When transformation, todate (time, formattime) was defined to convert the string into the time format for comparing.

3. User Interaction

The purpose of the data analysis process was to provide app companies with the time in which the red envelopes were shared most frequently and the most popular advertising words in those red envelopes. Therefore, an interactive window was designed to offer users those information, which is achieved by defining class recword () function.

Result

Data Processing

From the histogram (Appendix a), it could be seen that the highest count of domain names was ‘h.ele.m’, with a total number of nearly eight thousand, followed by ‘ pay.xiaojukeji.com’.

Data Analysis

The word cloud for each type of app was shown in appendix (since the unreasonable results obtained for some kinds data because their number of observations were low, only four types were kept at last). The following example was a word cloud for Takeaway type app.

Table 1 Popular Advertising Words for each type

OFO	low carbon, yellow bicycle, riding, awards, winter
Online shopping	receive, finance, discount ticket, skin care production, house hold
Taxi	pick up, once, minutes and seconds, pay the bill, one touch
Takeout	sale, rich food, million, highest, delicious

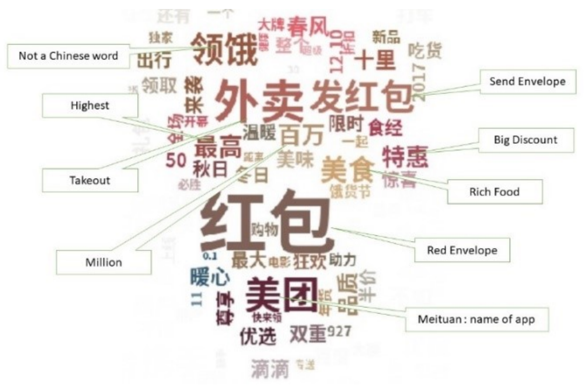


Figure 1 Word Cloud Map for Takeaway Type App

As can be seen from Figure 1, the most popular advertising words for takeaway apps were ‘red envelopes’, ‘takeaway’, ‘sending red envelopes’, ‘rich food’, ‘special benefits’, ‘meituan’, ‘millions’, etc. However, in these words, "red envelopes" was the necessary word in all the red envelopes, and "meituan" was the name of an app, so they should not belong to the popular words which be recommended to the users. All the popular words were shown in Table 1.

The time series for each type of app were put in Appendix and Figure 2 was that for takeaway type.

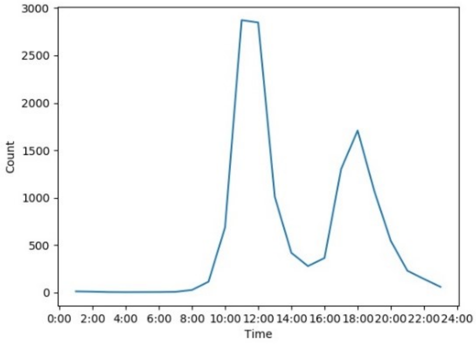


Figure 2 Time Series Map for Takeaway Type App

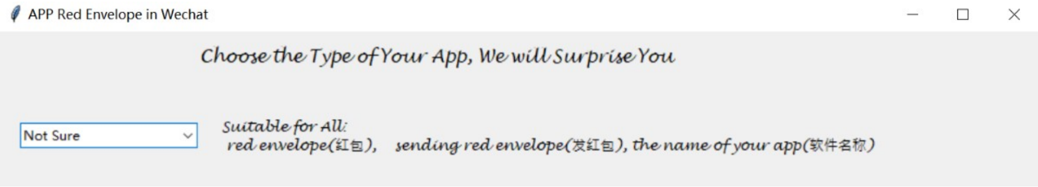
Table 2 Popular Time period for each type

OFO	13:00~14:00, 17:00~19:00
Online shopping	11:00~12:00
Taxi	14:00~16:00, 18:00~20:00
Takeout	11:00~12:00, 17:00~18:00

It could be seen from Figure 2, the number of red envelopes being shared of takeaway app reached the peak at 11:00am to 12:00am, and the period from 17:00 to 19:00 was the second highest peak for this type of apps.

User Interaction

Figure 3 shows the interaction window for the user, users could choose their own type of app, and then the window would recommend the most popular advertising words and the highest usage time during a day.



APP Red Envelope in Wechat

Choose the Type of Your App, We will Surprise You

Not Sure

Suitable for All:
red envelope(红包), sending red envelope(发红包), the name of your app(软件名称)

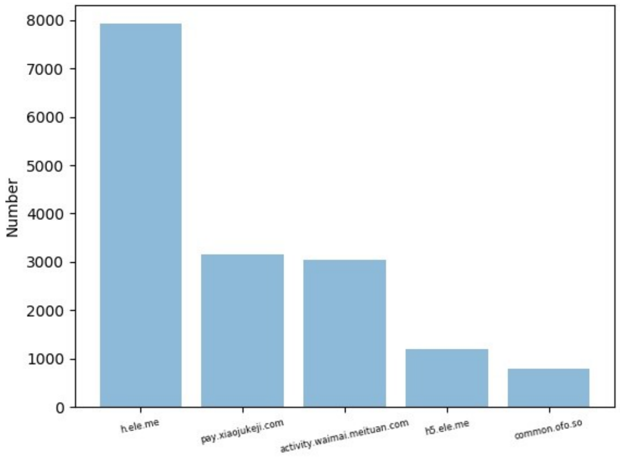
Figure 3 User Interaction

Conclusion and Discussion

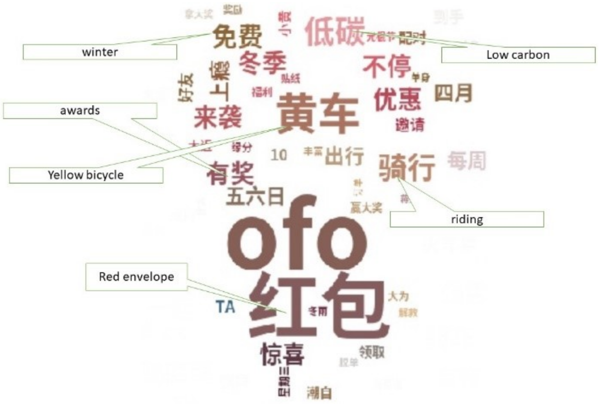
This analysis used Python to get the most popular ads words on red envelopes in the WeChat group for four types of apps, which were of OFO, Taxi, Online shopping and takeout, as well as the time periods when their red envelopes being shared the most frequently. The App Company could use this information to determine the time for releasing more red envelopes, and whether it is necessary to use more attractive advertising words or other measures to increase the number of red envelopes that are shared at other time periods.

However, the WeChat group being used consists of university students and located in Beijing, so it has some limitations of the results. In addition, because some apps provide two types of services, there is no clear division of their type. This case is not excluded in analysis, which will also affect the results.

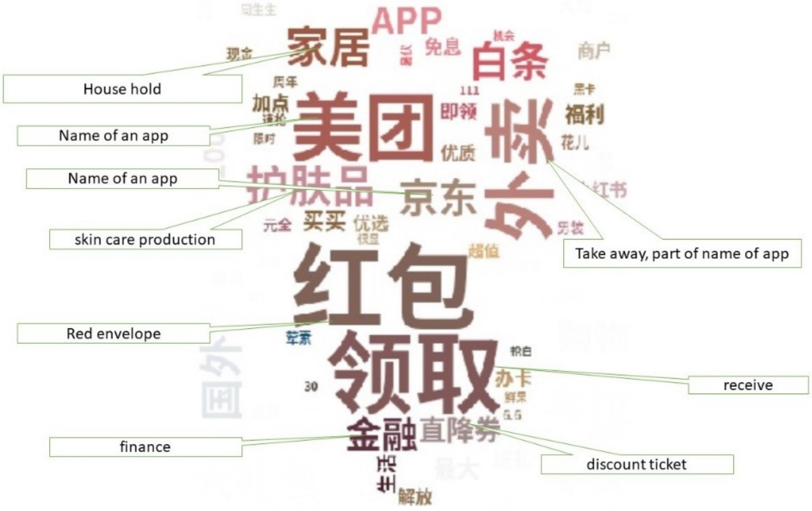
APPENDIX



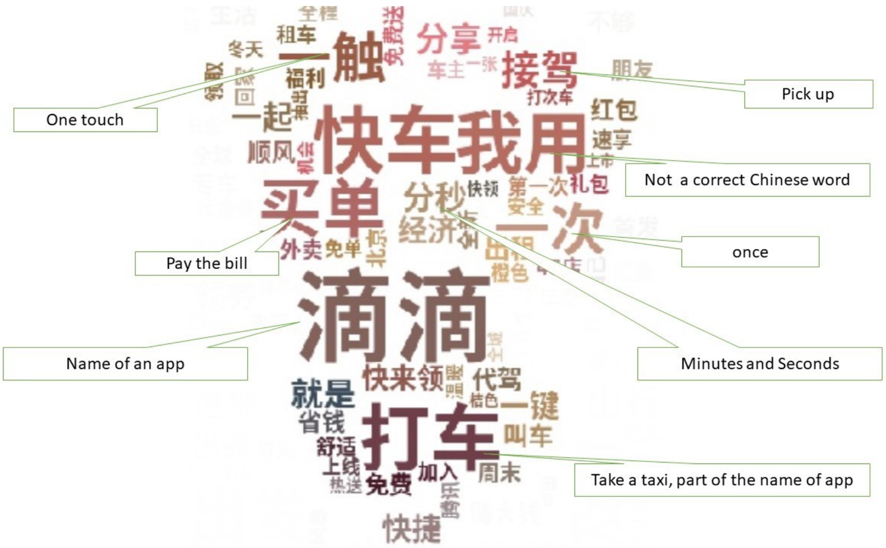
(a) Bar Chart of Domain Name



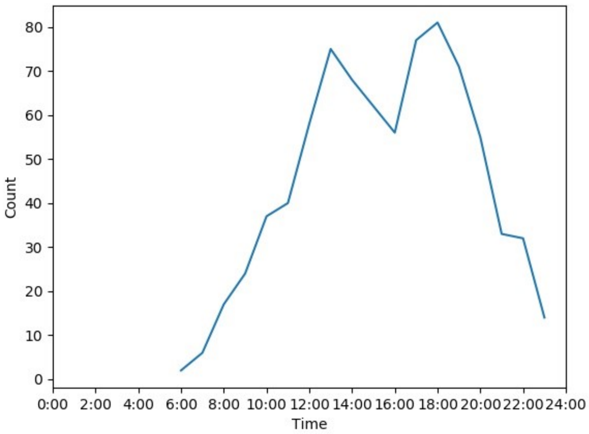
(b) Word Cloud for OFO Type



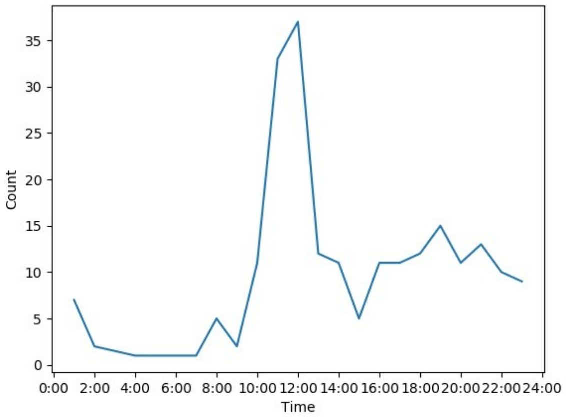
(c) Word Cloud for Online shopping



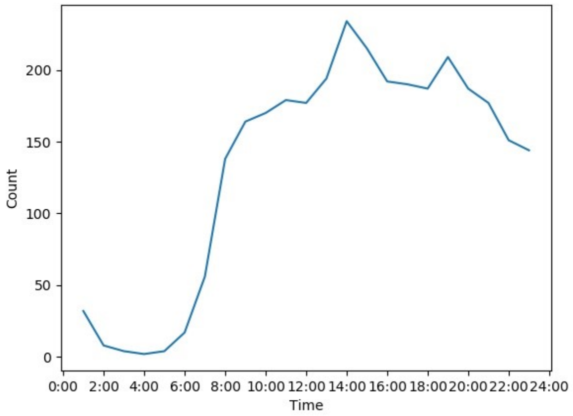
(d) Word Cloud for Taxi



(e) Time Series for OFO



(f) Time Series for Online Shopping



(g) Time Series for Taxi