Mathematical Computation
Volume 7, 2018, PP. 1-5

The Application of R Software in Text Analysis

Weihao Shi [1], Zhezhi Jin [1†]

1.College of Science, Yanbian University, Yanji 133002, China

[†]Email: jinzhezhi@sina.com

Abstract

With the development of Web 2.0, more and more people choose to use the Internet to express their opinions. All this opinions together into a new form text which contains a lot of valuable emotional information, this is why how to deal with these texts and analysis the emotional information is significant for us. We get three main tasks of sentiment analysis, including sentiment extraction, sentiment classification, sentiment application and summarization. In this paper, based on the R software, we introduced the steps of sentiment analysis in detail. Finally, we collect the movie reviews from the Internet, and use R software to do sentiment analysis in order to judge the emotional tendency of the text.

Keywords: Text Sentiment Analysis; R Software; Information Extraction; Emotion Recognition

1 INTRODUCTION

With the development of Internet technology, people has changed the traditional way of life, the wide use of computer, making more and more documents appear digitally, which can quickly spread over the Internet. It can be said that in the modern information technology life, electronic documents have become one of the main carriers of information transmission. Today, electronic documents are still growing rapidly, and the main sources of these electronic documents are Web pages, Email, Microblogs, Product reviews, etc. However, facing the batch of document data in the network, how to process and extract the information we really need is a hot topic in the computer field. The text emotion analysis arises in this context.

Text emotion analysis [1] refers to the process of processing, analyzing and applying the emotive text, which is the frontier research field in natural language processing. Text emotion analysis has high application value, which is often combined with online social media. It can be used in public opinion supervision and event prediction in public domain.,and can be used to collect commodity evaluation in business area. A complete process of Chinese text sentiment analysis should include text collection, text preprocessing(text participle, cleaning, classification, feature extraction), choosing classifier and emotional induction step, etc. In R, we can use rJava, rCurl, XML, slam, word-cloud and other R package to conduct the preliminary text processing. In this paper, firstly, we give an overview of the text emotion analysis, and give the definition of the text emotion analysis and the related processing steps. Then, we apply the statistical software to the processing steps of the text emotion analysis by R, and we carry out the simple text emotion analysis for the online movie reviews what we have crawled from the website.

2 FUNDAMENTAL

2.1 The Extraction of Emotional Information

The emotional analysis based on machine learning mainly consists of three steps [2]: machine learning classification model training stage, training model testing stage and text emotion classification stage. After preprocessing the text, we also need to extract the valuable information which relates to emotional inclination, it mainly includes: evaluation object, viewpoint holder, evaluation words, polarity intensity, etc. The valuable information in the text is called the feature item, and there are two main methods: emotional extraction and emotional dictionary, which can extract feature items.

If the network text is not processed, all the features are treated as feature vectors, the difficulty of calculation is

higher, and there is a large number of noise information, to avoid such situation, we should be handled by the method of dimension reduction, reserve the main information; remove the characteristics of the small influence. Feature extraction refers to the selection of distinguishing features from the original characteristics based on certain criteria. Common methods include document frequency (DF), word frequency (WF), and information gain (IG), mutual information (PMI) and CHI method.

The emotional dictionary is a collection of phrases and sentences that contain emotional colors, emotional words constitute an important part of affective lexicon [3]. By judging the emotional color of the words in the text, we can infer the emotional inclination of the text. Therefore, good emotional dictionary is the focus of text emotion analysis. There are three ways to build an emotional dictionary: artificial selection, emotional dictionary based method and corpus based method.

2.2 The Emotion Dictionary

The corpus can be constructed by manual tagging [4] method, but this method has the disadvantages of limited corpus size, uneven distribution of emotional words in corpus, large workload and prone to human error.In the analysis of text affective tendency based on semantic method, the core step is the construction of emotional dictionary, because the perfection of the emotional dictionary determines whether the results of the text emotion classification are accurate. In the steps of constructing the emotional dictionary, someone has expanded on a widely recognized emotional dictionary (such as HowNet, WordNet, etc). Some people construct specific dictionaries by artificial means for the characteristics of the comments text. Zhang Chenggong[5] sorted out the comprehensive dictionary including the basic emotion dictionary and domain dictionary, the network word dictionary and the modifier dictionary. Would like to point out in particular, constructing the specific dictionary by the artificial method can add information what you need, this emotional dictionary is more targeted, which makes text sentiment analysis is more efficient, but at the same time it also larger quota.

2.3 The Calculation of Emotional Intensity

The previous steps have been able to distinguish between the positive and negative tendencies of the participle, but the emotional inclination of the article cannot be simply read by the positive and negative tendencies of the participle, and the exact emotional intensity information is needed. On the basis, we can give the strength information by scoring. as follows[6]**:**

- For each document, we calculate S=P/(P+ N), where P is the sum of all the positive frequencies, and N is the sum of the negative term frequencies.
- If S> t, we classify is as positive, the threshold is usually 50%. Otherwise, we classify it as a negative document.
- we calculate the bias index, FI is the global proportion of all S. FI= (number of active documents)/(number of negative documents).

3 THE EMOTIONAL ANALYSIS OF MOVIE TEXT REVIEWS WITH R

This section we made text sentiment analysis for movie reviews which is collected online by R. Firstly, we use web crawler movie reviews on the network, and then we make text preprocessing, extracte of emotional information, and calculate the emotional polarity operations.

3.1 The Preparation of Review Corpus

First of all, we need to get the information that we need from the comments on weibo, and get the actionable text. We use the octopus collector. Through octopus collector [7], we climbed to the user's ID, gender, release date, comment content, forwarding amount and other information. For the sake of the next processing, we leave only the text of the comment and release time, as shown in figure 1.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
url	用户id(us	用户昵称(:	用户资料(user_profil	发布时间(:	微博内容(:	发自(sour	转发数	评论数	点赞数	视频地	原微博(or	转发(repo	评论(comm	关键字(ke	爬取时间
http://m.	1.83E+09	信仰战Yul	{"sex":"男","intro"	1.49E+09	#看过《权	豆瓣								《长城》评	1.49E+(
http://m.	2.01E+09	XX明要长p	{"sex":"","intro":"	1.49E+09	#看过《长	豆瓣								《长城》评	1.49E+(
http://m.	5.15E+09	路人甲没乙	{"sex":"","intro":"	1.49E+09	《长城》的	iPhone 7	1		1					《长城》评	1.49E+(
http://m.	3.35E+09	燃燒的火焰	{"sex":"","intro":"	1.49E+09	#降临#</a	微博 weibo.com		1	1					《长城》评	1.49E+(
http://m.	3.09E+09	山山山山山	{"sex":"","intro":"	1.49E+09	看了张艺谋	三星 GALA	1	2	4					《长城》评	1.49E+(
http://m.	5.18E+09	YZG-CD	{"sex":"","intro":"	1.49E+09	#每日一喵	HUAWEI Mate S								《长城》	1.49E+(
http://m.	5.59E+09	甜美系鹿小	{"sex":"","intro":"	1.49E+09	The Great	iPhone 5s								《长城》	1.49E+(
http://m.	6.03E+09	邵邵的梦幻	{"sex":"","intro":"	1.49E+09	越来越爱王	iPhone 6s								《长城》	1.49E+(

FIG. 1 THE FILM REVIEW INFORMATION

3.2 The Pretreatment of Corpus

The Chinese word segmentation is the first step in the analysis of this paper. The accuracy of the word segmentation will affect the accuracy of subsequent text analysis. R has a good performance in terms of participles. The following steps in this paper we use the "jiebaR", which is a Chinese word for R language. The stuttering word is an efficient R language Chinese word pack, which can be used for free, which enables R software to deal with Chinese text more efficiently. Next, we used the review data in "Great Wall" as the analysis object to carry out the Chinese word segmentation, and the result of the word segmentation is shown in figure 2.

```
>  two<-read.csv("C:/users/Administrator/Desktop/yajie.csv")
> head(two)
                                                  字段1
1                                        美国电影高逼格
2 不吹不黑 一些心里话 看过的 没看过的 水军 喷子 ...
3                                  电影，还是靠质量说话
4                              长城这部电影不配叫做长城。
5                          为什么批评景甜的帖子都被删了！
6                                  昨天被儿子忽悠看了长城
> txt<-segmentCN(as.character(two$字段1))
> head(txt)
[[1]]
[1] "美国" "电影" "高"   "逼"   "格"

[[2]]
 [1] "不"     "吹"     "不"     "黑"     "一些"   "心里话" "看"

[[3]]
[1] "电影" "还"   "是"   "靠"   "质量" "说话"

[[4]]
[1] "长城" "这部" "电影" "不配" "叫做" "长城"

[[5]]
 [1] "为什么" "批评"   "景"     "甜"     "的"     "帖"     "子"

[[6]]
[1] "昨天" "被"   "儿子" "忽悠" "看"   "了"   "长城"
```

FIG.2 THE RESULTS OF WORD SEGMENTATION

3.3 Extract Emotional Information

In the above text preprocessing process, we split the text simply. However, after the initial participle, some stop-words are added to the list, and these words are not useful for our text emotional identification. We need to further delete the word stop. In order to distinguish the emotional inclination of the vocabulary, we need to read the positive and negative word library. The positive and negative word libraries used in this paper are all the prepared emotional word libraries provided online. In this way, we get the text after we have deleted the words we use, and the emotional lexicon used to identify emotional tendencies.

3.4 Calculate Emotional Polarity

To determine polarity text emotion, we need to calculate the text of the emotional score, we customize the emotional score function, the word positive emotion is assigned 1, the negative emotions is -1, we calculate the positive and negative score.

3.5 The Analysis of Experimental Results

From the above results, positive and negative sentiment comments were basically the same in comments on the movie "the Great Wall" collected by douban.com. Next, we randomly select 10 comments from the comments to make verification. Verification results are shown in Figure 3.

```
> txt.score<-cbind(two,score)
> txt.score<-transform(txt.score,emotion=ifelse(total>=0,'pos','neg'))
> set.seed(1)
> validation<-txt.score[sample(1:nrow(txt.score),size=10),]
> validation[,c(1,5)]
                                                字段1 emotion
14 人民日报好阴险！故意以拙劣的表现钓出一大群1星革...     neg
19  总有人喜欢把个人的审美趣味上升到生殖器的高度。...     pos
29  确实不错，把中国很多丢失的好东西找回来，也真实...     pos
44                          景甜也啥这么牛？？牛？？     pos
10                                      长城永留     pos
42 鹿晗陈学冬景甜王俊凯-最终票房会有15亿么？15号首...     pos
43  为什么有些人只看了预告片就在那里瞎给分，要给一...     neg
30                                  来来来，互相伤害     neg
28  当导演本身变成一个争议性话题，那他的作品还剩下...     pos
3                                电影，还是靠质量说话     pos
> 
```

FIG. 3 VERIFICATION RESULTS

Through the above verification, we found that the 10 "Great Wall" comments randomly selected were very consistent with those of our previous steps. Associating with watercress online again, User gives the comprehensive score of 5.5, we can think scores given by douban more reasonable, and the comments reflect emotions tend to be more consistent.

4 CONCLUSIONS

This paper mainly discusses the application of R in text analysis, in which the text analysis is concretely translated into the emotional analysis of text. The comments on the web contain a lot of emotional information that can be used to predict the direction of a thing or to judge the general public's attitude towards one thing, which is the public opinion supervision. As open source software, R has the advantages of wide application scope, free use and independent programming. We can use R software for a series of processing and analysis to maximize the use of network text resources. For example, Rcurl is used as a web crawler to crawl online text, use Rwordseg for word segmentation, and use wordcloud to draw word clouds. Therefore, the popularity of R software has promoted our progress in the Internet text processing technology. In this paper, we use the network movie review, we analyse the net friend comment on the movie "the Great Wall" emotion by R, form a hierarchy and expanded Chinese text analysis. It provides a basic framework for emotional analysis of network text, and further study of text analysis can be further developed within this framework.

REFERENCES

[1] ZHAO Yanyan, QIN Bing, LIU Ting. Sentiment analysis [J]. Journal of Software, 2010, 21(8): 1834-1848.

[2] Liu Xianyou. Research on sentiment analysis of Electronic Commerce review text [D]. Anhui: University of Science & Technology China, 2015.

[3] Shao Muo Weisi, Zhang Tong, Predictive text mining foundation, Xi'an, Xi'an Jiaotong University Press, 2012.

[4] WIEBE J,WILSON T,CARDIE C. Annotating expressions of opinions and emotions in language[J].Language Resources and Evaluation,2005,39(2/3):164－210.

[5] ZHANG Chenggong, LIU Peiyu, ZHU Zhenfang et al. A sentiment analysis method based on a polarity lexicon [J]. Journal of Shangdong University: Natural Science,2012,3: 47-50

[6] Zhou Yongmei, Yang Jianeng, Yang Aimin. Construction method of Chinese sentiment dictionary for text sentiment analysis [J]. Journal of Shandong University (Engineering Edition), 2013 (06): 27-33

[7] CUI Yujie, LIAO Kun. Automatic Extraction of Overlapped Web-based Metadata with Octopus Collectors [J]. Editiology, 2016(05): 485-488.

AUTHORS

[1]**Weihao Shi** (1996-), Master graduate, research direction: application statistics. E-mail: 809109795@qq.com）.

[2]**Zhezhi Jin** (1977-), lecturer, graduate tutor, research direction: information statistics and insurance actuarial. Now, he is a professor in the mathematics, Yanbian University.zhzhjin@ybu.edu.cn.