Analysis of JD Commodity Evaluation Word Cloud Based on Web Crawler

[1]School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China
[2]Xiangya Nursing School, Central South University, Changsha 410013, China

Wu Shiqi[1]
Email: 1305056857@qq.com

Zhao Xing-yu[2]
Email:3568694632@qq.com

Qiu Fenglin[1]
Email:1158136147@qq.com

Tang Zhi-hang[1]
Email: zhtang@hnie.edu.cn

--ABSTRACT--

This project is the design of word cloud analysis program based on web crawler. Taking Jingdong Mall as the platform, it crawls all comment information of designated products, conducts data cleaning and data analysis on the information obtained from the review and crawler, and generates word cloud map.At the same time, the visual analysis of review data can clearly show the advantages and disadvantages of customer-centered evaluations and commodities, and provide an important reference for consumers to choose commodities and businesses to improve decision-making and optimize services. This project is developed by using Python3 language, using PyChart as the IDE, using the requests library, JSON library, World Cloud library and PyMongo library, using Navicat to connect MongoDB, using PyQT5 library to achieve visual interface, and JavaScript+HTML5+ CSS3 +MySQL+ word cloud + Boozing and Bagging algorithm for data analysis and algorithm optimization. In addition to providing consumers with cost-effective, highly evaluated and highly rated goods, it also provides the sellers with more specific data to improve their own defects.

Keywords**: Jingdong crawler; Natural Language Processing; Data mining; Visualization**

Date of Submission: Mar 04, 2021 Date of Acceptance: Mar 29, 2021

I. INTRODUCTION

With the continuous update of Internet technology, there are many online e-commerce platforms, among which Taobao and Jingdong are the most well-known ones in China. With the rapid growth of electric business platform, information resources and the dynamics of the commodity information resources, and people in electric business platform, after purchase of products to the quality of the product, price, appearance, use the experience of giving their views and evaluation, filled with all sorts of reviews on e-commerce platform data, comment has already formed the massive amounts of data and information. However, in these massive comments data and information, there are different comments, some are consistent and some are contradictory in the comments, which will bring great inconvenience to users to obtain the real information.

Therefore, how to collect and extract resource information of e-commerce platform through technical means, obtain useful information to provide perfect real-time monitoring and fine operation strategy has become a problem worthy of attention at present. Aimed at allowing users to quickly and accurately understand the required products, to improve the shopping experience of consumers.

II. RESEARCH STATUS OF DATA MINING AND WORD CLOUD

2.1 Research status of data mining

In recent decades, people's ability to use information technology to produce and collect data has been greatly improved. In the era of countless data accumulation and information explosion, the biggest problem for people is how to process data and make use of data. "Demand is the mother of invention", and data mining technology arises at the historic moment.

What is Data Mining? It is the process of searching for information hidden in a large amount of data through algorithms. Data mining, also known as KDD(Knowledge Discover in Database), first appeared in the 1980s.

Data mining has spread from abroad, where it has been used in many industries, from retail to telecommunications to finance to business management and so on. With the deepening of data mining research, data mining technology and methods continue to mature and perfect, its application field is more and more extensive. The research focus of foreign countries gradually shifted from discovery method to system application to large-scale integrated system development, and paid attention to the integration of multiple discovery strategies and technologies. In China, although data mining started late, it has been developed and practiced for more than ten years. With the deepening of the application of big data, data mining has mined the value behind data for more and more industries such as communication, finance, retail, transportation, public safety, smart city and industry[1].

2.2 Research status of word cloud analysis

The term "word cloud" was coined by Rich Gordon, an associate professor of journalism and director of the new media program at Northwestern University. Gordon is a former editor, reporter and former director of the Miami Herald's new media section. He has been concerned about the latest forms of content distribution on the web - those that are available only on the Internet and that newspapers, radio, television and other media have no way of catching up. Often, the latest, most networked forms of transmission are also the best. Word cloud, also known as word cloud, is a visual display of the "key words" that appear more frequently in the text. Word cloud image filters out a large number of low-frequency and low-quality text information, so that visitors can appreciate the theme of the text as long as they scan the text[2].

Word cloud analysis can be implemented in Python, and keyword content can be more clearly displayed through word cloud images formed after visual analysis of vocabulary capture. In the age of information explosion played a very good simplification effect.

III. AN OVERVIEW OF NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is currently one of the main application fields of deep learning. The overall goal of natural language processing is to enable computers to understand human language and work according to human language, so as to achieve effective communication between humans and computers. NLP plays an important role in the field of computer science and artificial intelligence. Natural language understanding and natural language generation are the two main technical areas that constitute NLP. The direction of natural language comprehension is to help machines better understand human language, including basic lexical, syntactic, and emotional level of high-level understanding[3]. Natural language generation direction, the main goal is to help machines to generate human understandable language, such as text generation, automatic summarization, etc.NLP technology is based on big data, knowledge mapping, machine learning, linguistics and other technologies and resources, and can form a specific application system of machine translation, in-depth question and answer, dialogue system, and then serve a variety of practical businesses and products. This paper will analyze the review data of Jingdong's designated commodities from the following three directions[4].

(1) Analysis of emotional tendency

The objective of the Sentiment Tendency Analysis (SAPA) task is to automatically determine the Sentiment Polarity category of the Chinese text with subjective description in the selected product reviews and to obtain the corresponding confidence level. Affective polarity is divided into positive, negative and neutral. Baidu sentiment orientation analysis is based on deep learning method, and its multi-granularity complete analysis tasks include sentence level, entity level and text level three. It can help merchants understand users' consumption habits, analyze hot topics and monitor public opinions in crisis, and provide powerful decision support for merchants[5].

(2) Recognition of conversation emotions

In view of the daily dialogue text crawled from the comment attribute of JD.com, the goal of dialogue emotion recognition is to automatically detect the implied emotional features, help merchants more comprehensively grasp the experience of designated goods and monitor the service quality of users of JD.com. It can automatically identify the user's emotions behind the dialogue according to the scene of the dialogue. The emotions can be divided into three kinds: positive emotions, neutral emotions and negative emotions, and positive emotions include love, happiness and gratitude. Negative emotions include complaint, anger, disgust, fear and sadness. Based on the recognition of users' negative emotions by the computer, targeted reference reply is given in combination with the context to help businesses to soothe customers' negative emotions in the first time. In addition to the automatic reply system, it can also trigger the intervention of human customer service to solve the problem[6].

(3) Text review

By convenient natural language processing technology, the better judgment of jingdong specified goods comments in a block of text content is in line with the specification, the network articles to complete automation, intelligent user comment information detection, so as to save the human cost, time, content audit to provide better services for business products, once found that detection of inclusion in the text, pornography, promotion, abusive, illegal, politics, irrigation, such as spam, can achieve automatic text review and real-time filtering, provide more reliable content security, guarantee the jingdong merchants have good user experience[7].

IV. COLLECTION AND PROCESSING OF JINGDONG COMMODITY REVIEW

4.1 Crawling process of commodity evaluation data

The overall design of JingDong comment collection and analysis system includes six modules: user interaction module, anti-reptile module, page grab module, page processing module, data storage module and data analysis module. The main function module diagram of the system is shown in figure 1.

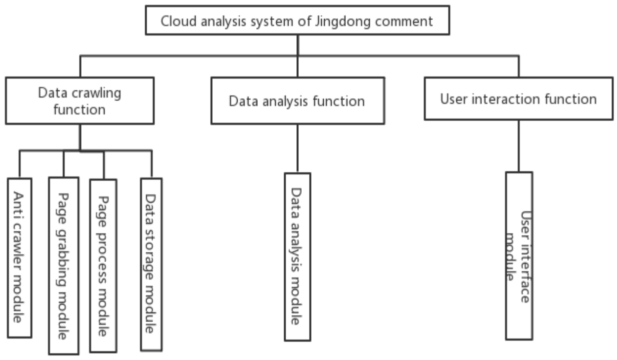

Figure 1 Main Function Module Diagram

4.1.1 Page grab module

Page grab is the initial operation of the entire crawler. Before crawling, Open your browser, Get the target page source code and post/get request, Then use the developer tool to detect the source code. A simulated browser login requires User-Agent, access For UA camouflage. Under the ul label class=" gl-warp clearfix", Different computers in computer search pages have different numbers, The numbers are stored under the data-sku properties of the li label, Then you can + by https://item.jd.com/+' number'. html' detailed page access. Through positioning, Scroll the number on the search computer detail page, You can crawl all the pages under this page.

4.1.2 Page Processing Module

For JingDong page processing, regular expressions and xpath are used to process. Regular expressions are the most commonly used, web source code is not much, and the required data is not newline, you can write regular expressions directly (. To match the information you need. If the page source code is increased, and the structure is complex and the data is distributed in multiple lines, it is necessary to write a very complex expression to extract information by regular expression, and then use xpath expression to accurately locate and crawl useful information according to the label of the page source code. Beautiful Soup library can use it to easily extract data from web pages [8]. Can be used for navigation, search, modification, analysis tree and other functions, and good handling of non-standard HTML tags, and then generate analysis tree.

4.1.3 Data Storage Module

MongoDB has the advantages of high performance, easy to deploy, easy to use, simple and fast to store data, which is very suitable for the requirements of this project. Install MongoDB, then open its service function and write code to connect to the database in the connect_mongo.py. The essence of the data storage module is to store all the crawling data in an orderly and accurate MongoDB database to realize the accurate connection with the Navicat. Navicat database management tools, select query tools, use commands to directly delete duplicate data in the database, write functions and methods to operate the MongoDB database in the connect_mongo.py file, use the Navicat to export the data to the local computer in the form of txt text format, CSV format or Excel data table for easy sharing[9].

4.2 Pre-processing of Jingdong Commodity Review Data

Data preprocessing is an important step in data analysis, data crawling and machine learning. The JingDong commodity comment data in crawling is often noisy and incomplete redundant data, which must be preprocessed. The main purpose is to improve the efficiency of data mining and obtain efficient and useful mining results. The evaluation preprocessing work is divided into data cleaning and data simplicity.

4.3 Does Page crawler breaches any security issues?

In the case of only scientific research, page crawler is not illegal.

Web crawler can not be used in the following three aspects: 1. Provide crawler related services for illegal organizations; 2. Capture and sell personal privacy data; 3. Profit from non copyright business data.

V. Analysis of JingDong reptile commodity review data

5.1. Data visualization analysis

Using Python Scikit-Learn library to read data, And then we use the groupby() method to group the data, And then use the count() function for statistics, Here are examples of women's sportswear, Such as visualizing the attributes of its evaluation, statement pd.groupby(' comment text will be written here'). count() to achieve type count, Then write code to draw comment text type statistics pie chart, Before writing to the values value of the pie chart, Finally, the comment text type statistics pie chart is shown in figure 2, As you can see in the diagram, the highest percentage of good reviews in the comments text attributes is 98.04, The second is the ratio of 1.56%, Finally, the poor rating rate is 0.40; As with comment text attribute statistics, Then analyze the comment response properties, As shown in Figure 3, As you can see in the diagram, there are comments and up to 69.40 words less than ten, Second, no comment ,27.59%, Then there is a comment reply and the number of words greater than 10 is 3.01.

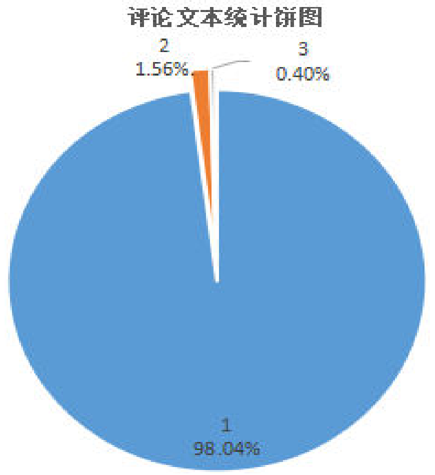

Figure 2 Review text statistics pie chart

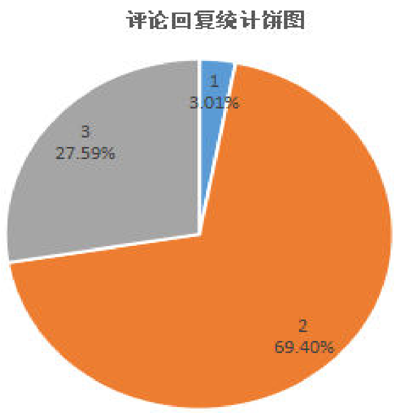

Figure 3 Review Response Statistics Pie

As shown in Figure 4, What you can see in the Pie Chart is that the number of days released is 45.16%, Because the double eleven during the period led to the JingDong women's sportswear economy, The second is the number of days from six months to nine months, Then the number of days released is 16.23% within three months, and the least is 2.55% between six months and three months. Finally, the comment likes the attribute statistics to carry on the visualization operation, A statistical chart of comments and likes is generated as shown in figure 5, As you can see in the diagram, the number of comment likes is zero, The second is that the number of reviews is one to ten, Then, the number of comments is 9.68% in 10 to 50,2.34% in 50 to 100, and 1.32% in the least.

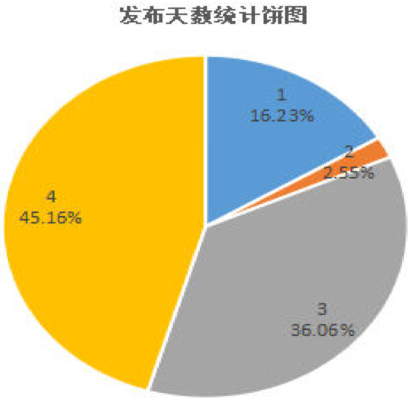

Figure 4 Pie Chart of Days of

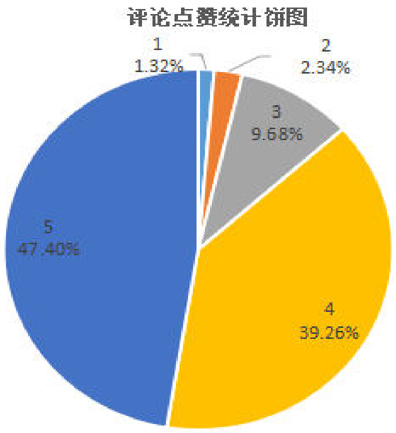

Figure 5.Comments and likes statistics

According to figure 6, we can see that the Chinese version of the review is the most normal, accounting for 98.64%; the second is the suspected abnormal text, which may be the audit dimension of violent terrorist violations, text pornography, political sensitivity, malicious promotion, vulgar abuse, low quality irrigation failed, accounting for 0.91%, the need for manual review of the text. Finally, the text is abnormal, that is, the audit dimension has not passed, the proportion is 0.45.

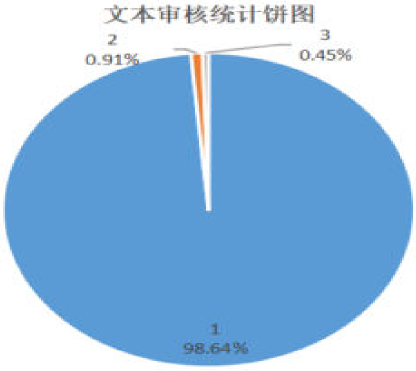

Figure 6 Review Text Review Statistics Pie

Because people's emotional expression is rich and colorful, emotional tendency analysis and dialogue emotion recognition will show the emotional trend of comments more clearly through two ways. First, the normalization process is used to express whether the comments are related to the degree of emotional preference. As shown in figure 7, of the 26000 comment data after preprocessing, the comment data are mainly focused on the positive emotional polarity, and the confidence is 0.8.At the same time, we can also see that some of the comment data are scattered in the negative and neutral parts, and it can be concluded that there is no linear correlation between the comment and the emotional analysis. As shown in Figure 8, The most positive emotional polarity is 94.68 and the negative emotional polarity is 3.19. From both diagrams, we can conclude that 26000 comments and emotional analysis are mostly positive.

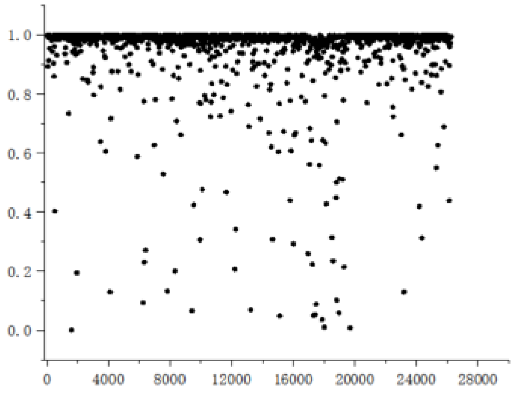

Figure 7 Comment on sentiment analysis scatter plot

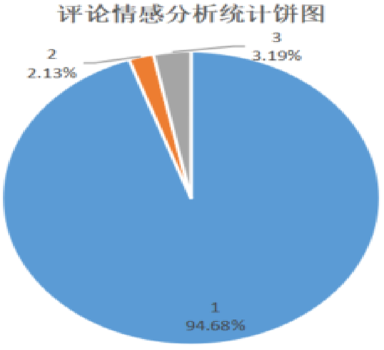

Figure 8 Comment on sentiment analysis statistics pie chart

As shown in Figure 9, Conversational emotion recognition scatter plots focus on 0.5 to 1, Below 0.5 are only very few scattered points, But the closer to 1, the higher the density, The proportion of positive emotions is the highest. Through the comments dialogue mood pie Figure 10 know, The highest proportion of positive emotions was 75.29, Positive emotions include affection, pleasure and gratitude; The second place is the neutral ratio of 23.12; The least negative emotions, The ratio is 1.60, Among them, negative emotions include complaint, anger, disgust, fear and sadness. Both graphs show that the dialogue emotion recognition of 28000 comments is mostly positive.

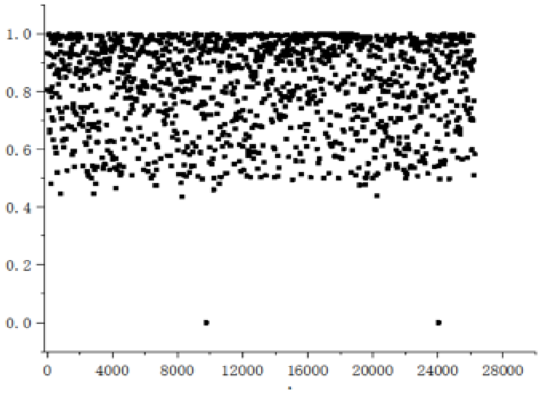

Figure 9 Comment Dialogue Emotion Recognition Scatter Chart

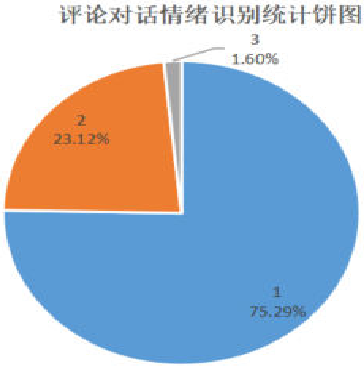

Figure 10 Comment Dialogue Emotion Recognition Statistics

5.2 Cloud Analysis of Words

After analyzing the seven attributes in the previous section, the Python program is used to analyze the word cloud map. By forming the key word cloud layer, the word cloud map highlights the keywords with high frequency to screen a large number of useless text information. Quickly understand the main meaning of the text. In the analysis of word cloud map, the text in crawling data is processed by word segmentation, and the key words are obtained according to the frequency of text appearance. The larger the font in the word cloud map, the higher the frequency of the text appears[10].

This paper directly calls the Python word cloud extension library when making the word cloud diagram, which is a powerful word cloud display third party library. The first need to enter in the cmd such as "pip install word cloud" installation, then import the word cloud module, and finally generate the word cloud diagram. For example, the font, color and shape of the word cloud can be set, the word cloud can be used as the object in the word cloud extension library, and the word cloud map can be drawn according to the frequency and other parameter information of the keyword in the text[11]。

The flow chart of word cloud map generation module is shown in figure 11. First, the screening conditions and contents are established, the result data satisfying the conditions are screened out, the frequency of each keyword is counted, and the related parameters of word cloud map are set.

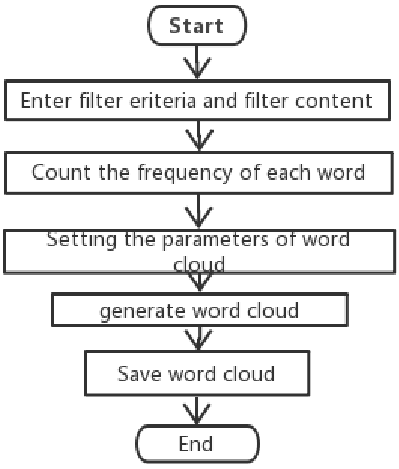
Start
Enter filter eriteria and filter content
Count the frequency of each word
Setting the parameters of word cloud
generate word cloud
Save word cloud
End

Figure 11: Flowchart of High Frequency Keyword Module

These key words show the main characteristics of this kind of clothing. Users can quickly understand the main meaning of the comment text, generally grasp the consumer evaluation preference of this kind of clothing goods, and can also optimize the title and detailed page Integration learning was first discovered by Hansen and Salamon, which mainly solves the same problem by description of sports women's clothing goods to help merchants ensure the quality of this kind of goods, Develop better marketing strategies to provide the basis [12].

5.3 Analysis of positive rate

For further verification of the popularity of this kind of goods, we first randomly select a part of the number of comments in women's sportswear, then analyze and generate a visual web view through the pyecharts library. The final results are shown in figure 12. The diagram Shows the percentage of praise, evaluation and evaluation.

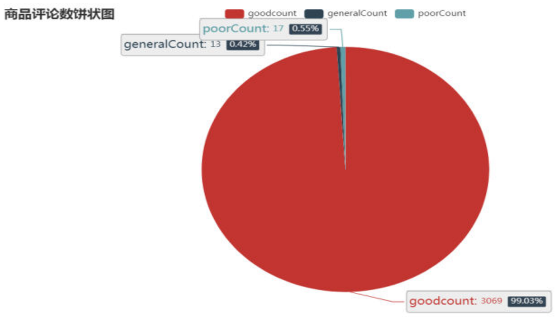

Figure 12 Selected commodity review rates

VI. MACHINE LEARNING ALGORITHMS FOR ONLINE REVIEWS

After the specified commodity information is crawled through the network crawler, the data is also quantified and analyzed. Then, the random forest algorithm, decision tree algorithm, Bagging and Boosting algorithm are used to compare and optimize the accuracy of the prediction model for evaluating the usefulness of information, and the optimal machine learning algorithm is selected and recommended to the e-commerce platform[13].

6.1 Random Forest Theory

6.1.1 Integrated learning

establishing several model combinations together. Multiple JD women's sportswear evaluation classifiers are generated to independently learn and obtain predictions respectively. These predictions are fused into single predictions, and the final result is better than the prediction made by any single classification. Random forest is a subclass of ensemble learning, which depends on the voting selection of decision tree to determine the final classification result[14].

There are two main ways to integrate algorithms, which are Bagging and Boosting algorithm, whose full

name is Bootstrap Aggregation algorithm, is a simple and effective integration algorithm. Firstly, the task of sample

extraction can be carried out independently through the original sample data set. M training samples are randomly selected from the training samples, and T training sets can be obtained after the extraction of T round. And then through T training sets to build T models; Finally, the results of the training model are selected. The selection method can be based on the difference of classification results. The voting method is the best choice for classification problems, and the average method is the best choice for classification problems. Boosting algorithm is a set of "weak" form of combination, belongs to the integrated learning algorithm of a cluster, first of all use for training on each sample set an initial value, training on training set 1 "weak learning 1", then according to the result of "weak learning 1" forecast timely adjustment and get the training set 2, each iteration will get a weak learning; Finally, the results of each iteration are combined and the above process is repeated to predict the weight sum of each basic learner[15].Bagging algorithm adopts random sample selection with replacements. Each sample data has the same proportion in each training data set and the weight of each sample is the same. Therefore, each prediction function can be generated in parallel. Boosting algorithm, with fixed training samples, adjusts the weight of samples according to the error rate and changes the weight of each sample. The larger the error rate, the larger the weight of the samples will be. With the progress of iterative training, each prediction function will be generated by iteration in order[16].

6.1.2 Principle of Random Forest Algorithm

The earliest by Leo Breiman and Adele Cutler forest randomly, random forest algorithm is a kind of precise integration algorithm, through the integrated study of thought more decision tree will eventually reach the integration of an algorithm, has stronger ability to resist noise the, also can to deal with discrete data and continuous data Random forest algorithm is based on tree to training and forecasting samples According to the following algorithm to build every tree, every decision tree classification results, after statistical classification results, choose which has the most votes as a result of random forest classification[17].

VII. CONCLUSION

In this paper, natural language processing (NLP)and word cloud methods are used to construct JD commodity review text, establish topic analysis model, and objectively evaluate the quality, format and price of commodities. It is convenient to bring cost-effective products to users and provide data for merchants to improve their products. In the experiment, a lot of time needs to be saved in the process of crawling, filtering, importing Python projects and making themes, and the filtering of annotated text is also very important, which can be washed several times more. According to the consideration of the experimental situation, we can modify the stop word list for many times to clean up the data and achieve better results.The quantity of experimental data is too small and horizontal comparison is not achieved, which needs to be strengthened. More work will be done in this link in the future.

ACKNOWLEDGEMENTS

Project supported by Innovation and entrepreneurship training program for college students in Hunan Province in 2020" Analysis of Jingdong commodity evaluation word cloud based on Web Crawler" and Provincial Natural Science Foundation of Hunan（2018JJ2082）

References

[1] Xu Lei, Zhang Kewei.Analysis of Jingdong commodity reviews based on text mining [J]. Inner Mongolia Science, Technology and Economy, 2020,(3):41,43.

[2] Yan Ming, Zheng Changxing. Text segmentation and word cloud production in Python environment [J].Modern Computers, 2018,(34):86-89.

[3] Feng Feng Xingjie and Zeng Yunze. In-depth recommendation model based on scoring Matrix and review text. Journal of Computer Science, 2020, 43(5) : 884-900.

[4] Li Jun, Zhou Yuying, Tang Zhihang.Clothing Information Collection Based on Topic Web Crawler. Information Technology and Information Technology,

2018, (8):97-99.

[5] Zeng Xiaoqin, Yu Hong.Sentiment analysis of commodity review text based on Python [J]. Computer Knowledge and Technology, 2020, 16(8):181-183.

[6] Zhang Yan, Wu Yuquan.Design of Network Data Crawler Program Based on Python [J]. Computer Programming Skills and Maintenance, 2020,(4):26-27.

[7] Mona Nasr, Andrew karam, Mina Atef, Kirollos Boles, Kirollos Samir, Mario Raouf,Natural Language Processing: Text Categorization And Classifications[J].Int. J. Advanced Networking and Applications, 2020,12(02), 4542-4548.

[8]ZUO Wei, ZHANG Xi, DONG Hongjuan, et al.Review of Topic Web Crawler Research [J]. Software Guide, 2020, 19(2):278-281.

[9] Li Junhua. Research on Web Crawler Based on Python [J]. Modern Information Science and Technology, 2019, 3(20):26-27,30.

[10] Li Lin. Design and Implementation of Web Crawler System Based on Python [J].Information and Communications Technology, 2017,(9):26-27.

[11] Li Huiyun, He Zhenwei, Li Li, et al. Research on HTML5 Technology and Application Mode [J].Communications Science and Technology,2012,28(5):24-29.

[12] Sun Jianyan, Ma Yuxin, Wu Wenjie.Web Crawler System Based on Python [J]. Computer Knowledge and Technology, 2019,15(26):61-63.

[13] Bi Sen, Yang Yubing.Research on Web Crawler Technology Based on Python [J]. Digital Communications World, 2019, (12):107-108.

[14] Huo Bingliang. Analysis of Web Crawler Technology Based on Python [J]. Digital World, 2020,(4):73-74.

[15] Rahul Desai, Dr. B P Patil, Maximizing throughput using adaptive routing based on reinforcement learning[J].Int. J. Advanced Networking and Applications,2017, 09(02), 3391-3395

[16]Zhao G, Shu X. Analysis and application of SQL language in Navicat for MySQL platform[J]. Wireless Internet Technology, 2017, (19):74-75.

[17] Li Pei. Research on web crawler and anti-crawler technology based on Python [J].Computer and Digital Engineering, 2019, 47(6):1415-1420.