Public Opinion Analysis of TikTok Hot Search Based on NLP

Zhangliao Xu
School of Information Engineering
Kunming University
Kunming, China
email: 1693241822@qq.com

Hongling He
Hongling HeSchool of Information Engineering
Kunming University
Kunming, China
email: 653353835@qq.com

Xisong Miao*
School of Information Engineering
Kunming University
Kunming, China
email: miaoxisong_kmu@163.com

Boyang Wang
School of Information Engineering
Kunming University
Kunming, China
email: 782931627@qq.com

2023 5th International Conference on Applied Machine Learning (ICAML) | 979-8-3503-4141-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAML60083.2023.00060

***Abstract*—If the outburst of network public opinion is not controlled in time, it may even allow netizens to map the emotions vented online to the real society, causing economic and other losses to the real society. This paper aims to establish a public opinion analysis system based on NLP for Tiktok hot search, which can quickly grasp the online public opinion of Tiktok hot search, reflect the attitude of various regions in China towards the hot spot through data visualization, grasp the trend of public opinion, avoid the escalation of public opinion and prevent public opinion crisis.The NLP natural language processing algorithm is analyzed, and the implementation process of Douyin crawler and the sentiment analysis of comments are described in detail. In this system, the review information of Douyin hot search is obtained by Scrapy framework, data preprocessing is carried out, then the Jieba Chinese word segmentation technology is adopted to put the review data into the SnowNLP model for emotion analysis, and the sentiment value of the review is obtained, and the data is visualized, and then the Flask visualization screen is built.**

Keywords-Online public opinion; Scrapy framework; NLP natural language processing; Emotion analysis; Data mining.

I. Introduction

A. *Research Background*

According to the statistical report on China's network development released by China Internet Network Information Center (CNNIC), as of June 2022, the number of netizens in China was 1.051 billion, 19.19 million more than that in December 2021, and the Internet penetration rate reached 74.4%. An increase of 1.4 percentage points from December 2021[1]. As shown in Fig. 1 below:

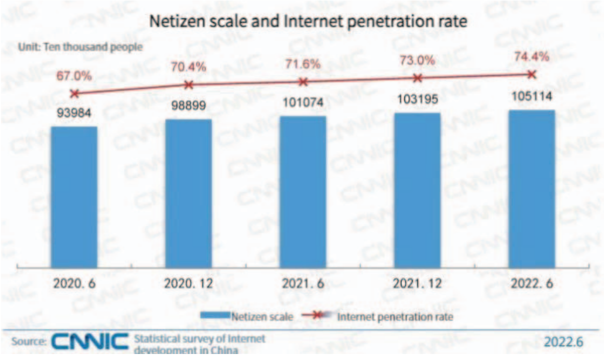


Figure 1. Our country netizen scale and internet popularity rate

Douyin, as the short video software with the most users, is a hotbed of online public opinion. If the outburst of network public opinion is not controlled in time, it may even allow netizens to map the emotions vented online to the real society, causing economic and other losses to the real society. China's Internet penetration rate has reached 74.4%. Under such a large netizen base, the speed of network public opinion will be immeasurable. The fermenting speed of online public opinion will be unimaginable. Many netizens will be led by some criminals to follow the trend of comments, deviating from the correct values and world view, so it is very important to analyze the public opinion of Douyin hot search comments.

B. *Research Purpose and Significance*

This paper proposes to use the web crawler method to crawl the hot search data of Douyin in real time, monitor the hot search of Douyin, and use the Natural Language Processing algorithm (NLP) for text processing and emotion analysis of hot search information and comments. The positive, neutral and negative attitudes of comments in different regions were evaluated, and the status of the trending searches in different provinces was displayed on the map of China through data visualization.

979-8-3503-4141-6/23/$31.00 ©2023 IEEE
DOI 10.1109/ICAML60083.2023.00060

NLP based Tiktok hot search public opinion analysis can quickly grasp the online public opinion of Tiktok hot search, reflect the attitude of different regions of China towards the hot spot through data visualization, grasp the trend of public opinion, avoid public opinion escalation, and prevent public opinion crisis.

II. Related Technologies and Concepts

A. *Web Crawler Technology*

Crawler technology refers to the technology of automatically obtaining website information, which can realize the mass acquisition of network information and quickly obtain all the information you can see on the web page. It makes a request to http by using URL, obtains the content of http response, analyzes the http response, extracts the content, and screens, cleans and saves the data. As shown in Fig. 2 below:

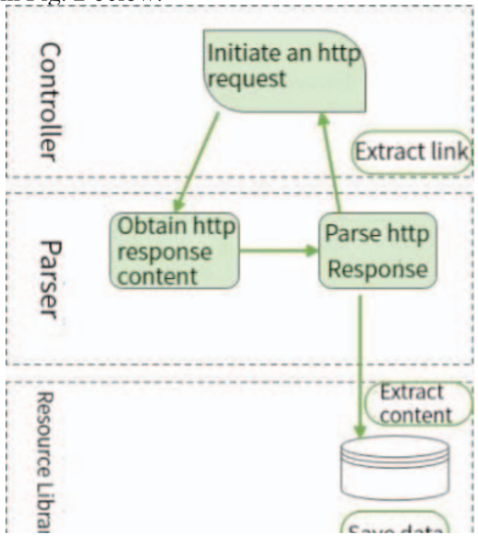


Figure 2. Flow chart of crawler technology

1) *Scrapy framework*

Scrapy is a crawler framework that applies to Python and is powerful enough to quickly crawl screen and Web data. First, the Spider extracts data and urls and sends them to the Scrapy Engine, which then transfers the URL received to the Scheduler scheduler. The Scheduler scheduler stores requests from the Scrapy Engine and sorts them and sends them back to the Scrapy Engine, After the Downloader downloads custom data and sets proxy, it returns to Scrapy Engine. Scrapy Engine receives the result and returns it to crawler Spider, which then parses and processes Html data. The parsed data of crawler Spider will be sent back to Scrapy Engine. Finally, Scrapy Engine receives the field data, stores it through the Item Pipeline, and sends the request to the Scheduler[2]. See Fig. 3 below.

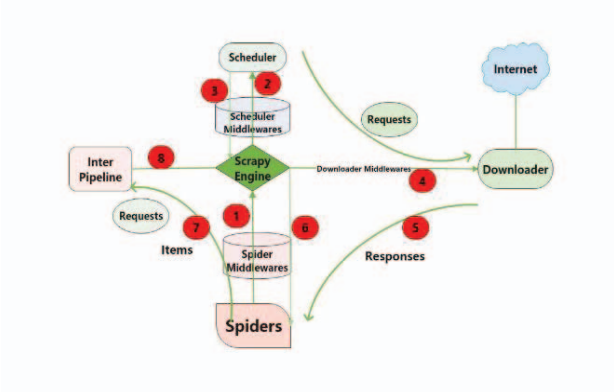


Figure 3. Operation flowchart of Scrapy framework

TABLE I. Component Explanation Table of Scrapy Framework

Name	Explain
Scrapy Engine	Engine, data transmission
Scheduler	Scheduler, which holds the request queue
Downloader	Downloader, download request information
Spider	Crawler, extract data and URL
Item Pipeline	Pipes, storing and transmitting data
Downloader Middlewares	Download the middleware and set the IP proxy
Spider Middlewares Spider	Middleware, custom HTTP requests

The implementation of Scrapy crawler framework is divided into five basic steps. Firstly, Scrapy startproject + projectname (projectname) is used to create the project, and then fields are clarified to specify the target fields to be crawled, and items.py file is written to make crawler. Write spider.py file for web crawling and store data after crawling. Design Item Pipeline to store the crawled data and write pipeline. py. Finally, run scrapy crawl search-s JOBDIR=crawls/search to start the entire Scrapy project[3]. As shown in Table I.

2) *Anti-reptile technology*

Due to the widespread use of web crawler technology, some criminals will use crawler technology to steal data on some websites, resulting in the loss of benefits of some companies. In addition, due to the rapid development of crawler technology, the number of requests that can be reached every second can be hundreds of visits to the website, and such high frequency visits to the website will lead to the direct collapse of the server. Users cannot access the site, limiting the use of IP leads to crawlers can not continue, users can not obtain the desired data[4].

In order to prevent the invasion of crawler software, some websites will add a lot of identity identification methods to analyze the identity information provided by the User to visit the website, user-agent and Cookie information, when the same IP information repeatedly visits the website, it will be considered by the website that the crawler program restricts access.

Tiktok is also equipped with an anti-crawler program, which will be terminated when it recognizes the same IP address and visits the website frequently. Therefore, we need to disguise the comments data of Tiktok hot search through the program Settings, so as to avoid restrictions on us after being identified by Tiktok's anti-crawler program. The following methods can be used for camouflage:

a)Use different computers and accounts to climb Douyin at the same time, and limit the speed and number of climbing for each account.

b)Reduce the crawling speed, and set a Sleep time for the program every time it climbs for a period of time, so that the program can continue data acquisition after a period of time.

c)Modify the User-Agent and Cookie information in Headers to disguise the identity through the IP proxy pool.

This paper adopts the third way to obtain the data of Douyin hot search comments.

B. Data Preprocessing Technology

Data preprocessing technology refers to cleaning the unwanted part of data after data acquisition, and converting the desired part into the desired data format.

The crawler program will obtain the HTML information of the webpage, and the hot search comment data we want is in the webpage information. We need to select the HTML information through Xpath technology, index the data under the tag div we want, and clean some data we do not need after obtaining the data.

Only after data is preprocessed can we get the data we want, and only after such data is brought into the model and algorithm verification can we get the expected results we want.

C. Chinese Word Segmentation Technology

Word segmentation technology originated earlier in foreign countries, and English word segmentation technology was the first to develop. The composition of English is simple, the grammar and logic are simple, and the word is taken as the unit to form sentences. However, the difficulty of Chinese word segmentation is much greater than that of English word segmentation, and the composition logic of Chinese and English is also very different. Therefore, Chinese word segmentation technology is different from English word segmentation technology.

Jieba word segmentation technology is a commonly used Chinese word segmentation technology, and there is a third-party library of Jieba word segmentation in Python. Easy to use, is the developer of Chinese word segmentation technology choice.

Jieba word segmentation combines the string matching algorithm and statistical model word segmentation idea. In fact, the string matching algorithm matches the string with the keyword in a dictionary, and the jieba word is matched with the dictionary. If the word is found in the machine dictionary, the jieba word segmentation is performed. The idea of word segmentation in statistical model is to consider the frequency of words that can be formed between each word. If it is found that two words in a sentence can form a word in a large number of frequency, it is considered that these two words form a word.

Jieba segmentation also stores the database of common jieba words, such as "zhi", "he", "ni", "mo", "we" and "so", which have no great practical meaning. By removing these words from the sentence by jieba segmentation method, not only can the algorithm be accelerated, but also the keyword density can be improved to make the segmentation more accurate. The following Fig. 4 The whole process of Jieba segmentation.

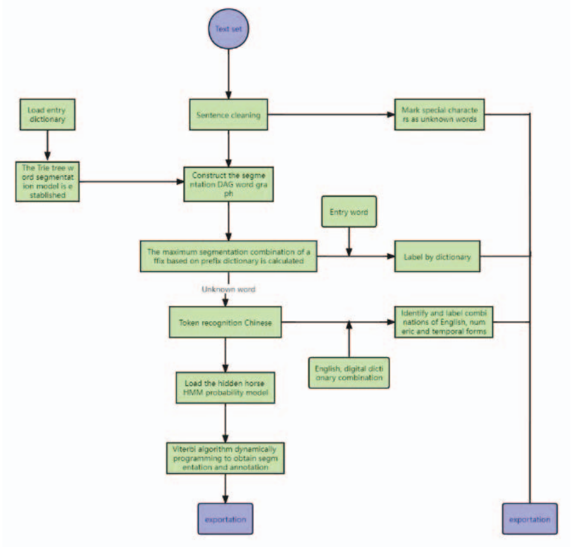

Text set
Load entry dictionary
Sentence cleaning
Mark special characte rs as unknown words
The Trie tree w ord segmentat ion model is e stablished
Construct the segme ntation DAG word gra ph
Entry word
The maximum segmentation combination of a ffix based on prefix dictionary is calculated
Label by dictionary
Unknown word
Token recognition Chinese
Identify and label combi nations of English, num eric and temporal forms
English, digital dicti onary combination
Load the hidden horse HMM probability model
Viterbi algorithm dynamically programming to obtain segm entation and annotation
exportation
exportation

Figure 4. Flow chart of Jieba word segmentation

III. NLP NATURAL LANGUAGE PROCESSING ALGORITHM

NLP (Natural Language Processing) is an important research direction in the field of artificial intelligence. It mainly processes and analyzes human natural language. NLP algorithm is mainly used in machine translation, Chinese OCR, text classification, speech recognition, sentiment analysis and so on.

D. TF-IDF Algorithm

Tf-idf algorithm, where TF stands for word frequency, IDF stands for inverse document frequency, it is a weighting technology. The algorithm takes the frequency of a word in a sentence as an evaluation index to judge the difficulty of the important planning of the word. If the word appears more frequently in this corpus, but not significantly in other corpus, it is considered that the word is very important in this corpus. Of course, words with high frequency in this corpus and other corpus, such as "of", "we" and "so", will also appear in this corpus. This kind of words will think that the importance of these words is not high in this corpus. TF-IDF algorithm has the advantages of simple and easy to understand, fast running speed and strong filtering ability.

The word frequency TF formula in the TF-IDf algorithm is as follows:

$$TF = n(A) \tag{1}$$

Where A represents A certain word, n() represents the number of times, and n(A) represents the number of times of a certain word in this corpus.

Since the length of discourse is different, TF can be standardized, and the formula is as follows:

$$TF = \frac{n(A)}{n(B\ )} \tag{2}$$

Where n(B) represents the total number of words in the corpus.

IDF inverse document frequency, also known as inverse document frequency, is the reciprocal of the document frequency. The formula is as follows:

$$IDF = \log\left(\frac{n(A)}{n(B\ )+1}\right) \tag{3}$$

Where, n(A) is the total number of corpus documents, and n(B) is the number of documents containing the word.

The formula of TF-IDF is as follows:

$$TF - IDF = TF \cdot IDF \tag{4}$$

According to the formula of tF-IDf, it can be seen that when the frequency of a word in a certain corpus is high, that is, when TF is large, and the number of documents containing the word is small, that is, when the frequency of IDF inverse document is high, the larger tf-IDf is and the higher the importance of the word is.

E. *Text Sentiment Analysis Model*

Many users will comment on each hot search in Douyin, and the online public opinion will develop along with the comments of these users. Some users will leave some negative comments, which will mislead some users to follow the trend and make negative comments. Therefore, in order to grasp the online public opinion of Douyin hot search, we need to grasp the emotional information of users' comments. Sentiment analysis of user comments is needed.

1) Overview of text sentiment analysis

Sentiment Analysis is an important research direction in NLP algorithms and an important technology for emotion recognition. General text emotion analysis mainly classifies a paragraph of text or a sentence or a word, and determines whether the emotion color belongs to positive or negative. Of course, there are also some emotion analysis algorithms to classify the analysis results, including happiness, sadness, anger, surprise and so on. This paper mainly studies the public opinion orientation of hot search comments on Douyin, so the results of emotion analysis are divided into positive public opinion and negative public opinion.

Text sentiment analysis first requires word stopping, custom dictionary, and word segmentation of the original text. These operations can be completed by Jieba word segmentation technology, and the pre-processed text needs to be divided into training set and test set. Training set, in the training set to extract corpus information, TF-IDF algorithm can be used to extract corpus information, and then emotion extraction, classification, retrieval[5]. The flow chart of text sentiment analysis is shown in Fig. 5 below:

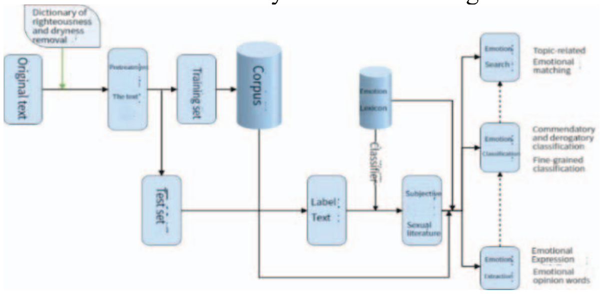


Figure 5. Flow chart of text sentiment analysis

2) SnowNLP emotion analysis

SnowNLP is a third-party library in Python that has its origins in TextBlob. The Python third-party libraries SnowNLP, Jieba, BosonNLP are all for Chinese language processing. The main functions of SnowNLP include Chinese word segmentation, partof speech tagging, sentiment analysis, text classification, pinyin conversion, text keywords extraction, abstract extraction, text similarity calculation, etc. The SnowNLP library comes with a set of positive and negative emotional training, with trained models ready for immediate use.

The principle of SnowNLP's emotional prediction of Douyin hotsearch comments is as follows:

Douyin hot search comments are divided into positive comments A1 and negative comments A2. Each comment has a total of n words: $\boldsymbol{B_1}, \boldsymbol{B_2}, \cdots$**and** $\boldsymbol{B_n}$. The naive Bayes formula is as follows:

$$P(A_1|B_1,\cdots,B_n) = \frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n)} \tag{5}$$

Where the probability that a comment composed of $\boldsymbol{B_1}, \boldsymbol{B_2}, \cdots, \boldsymbol{B_n}$ words is positive comment A1 is equal to the probability that the comment has $\boldsymbol{B_1}, \boldsymbol{B_2}, \cdots, \boldsymbol{B_n}$ words if the comment is positive comment A1 times the probability that the comment is positive comment A1, divided by the probability that the comment has $\boldsymbol{B_1}, \boldsymbol{B_2}, \cdots, \boldsymbol{B_n}$ words.

Meanwhile, the full probability formula is as follows:

$$P(B) = P(B|A)\cdot P(A) + P\left(B\middle|A^{'}\right)\cdot P\left(A^{'}\right) \tag{6}$$

Formula 3-7 can be obtained according to the full probability formula:

$$P(B_1,\cdots,B_n) = P(B_1,\cdots,B_n|A_1)\cdot P(A_1) + P(B_1,\cdots,B_n|A_2)\cdot P(A_2) \tag{7}$$

That is, the comment has the word $B_1,\cdots,B_n$ probability = if the comment is positive comment A_1, then the comment has the word $B_1,\cdots,B_n$ probability times the comment is positive comment A_1, plus if the comment is negative comment A_2, then the comment has the word $B_1,\cdots,B_n$ probability. The product of the probability that the comment is a positive comment A_1. Formula 3-5 can then be converted into Formula 3-8, as follows:

$$P(A_1|B_1,\cdots,B_n)=\frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)+P(B_1,\cdots,B_n|A_2)\cdot P(A_2)} \tag{8}$$

Equation (8) is the basic formula used in SnowNLP emotion prediction. The denominator and numerator can be divided by a numerator at the same time, which can be simplified into (9) as follows:

$$\begin{aligned} &P(A_1|B_1,\cdots,B_n) \\ &=\frac{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)+P(B_1,\cdots,B_n|A_2)\cdot P(A_2)} \\ &=\frac{1}{1+\frac{P(B_1,\cdots,B_n|A_2)\cdot P(A_2)}{P(B_1,\cdots,B_n|A_2)\cdot P(A_1)}} \\ &=\frac{1}{1+\exp\left\{\lg\left[\frac{P(B_1,\cdots,B_n|A_2)\cdot P(A_2)}{P(B_1,\cdots,B_n|A_1)\cdot P(A_1)}\right]\right\}} \\ &=\frac{1}{1+\exp\{\lg[P(B_1,\cdots,B_n|A_2)\cdot P(A_2)]-\lg[P(B_1,\cdots,B_n|A_1)\cdot P(A_1)]\}} \end{aligned} \tag{9}$$

We have a 1 in the denominator and we can rewrite it as 3-10

$$1 = \exp\{\lg[P(B_1,\cdots,B_n|A_1)\cdot P(A_1)]-\lg[P(B_1,\cdots,B_n|A_1)\cdot P(A_1)]\} \tag{10}$$

The steps of SnowNLP sentiment analysis are shown in Fig. 6 below:

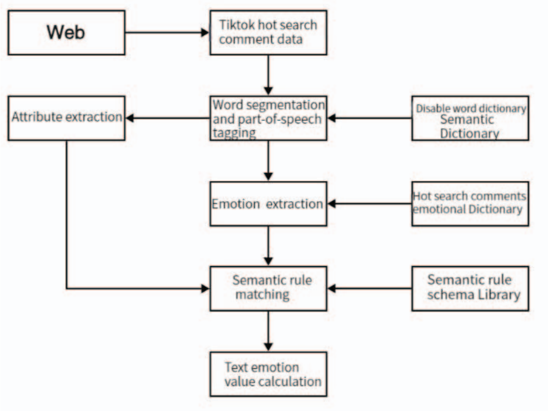

Web
Tiktok hot search comment data
Attribute extraction
Word segmentation and part-of-speech tagging
Disable word dictionary Semantic Dictionary
Emotion extraction
Hot search comments emotional Dictionary
Semantic rule matching
Semantic rule schema Library
Text emotion value calculation

Figure 6. SnowNLP emotion analysis flow chart

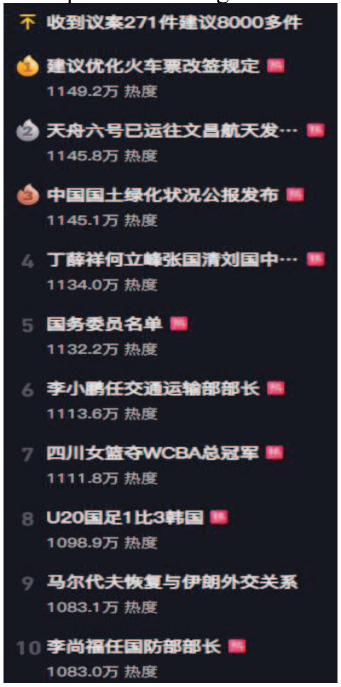
Firstly, the Jieba word segmentation technology is used to conduct word segmentation and partof speech tagging for Douyin hot search comments, then the emotion words obtained from the text data after word segmentation and partof speech tagging are extracted, and the semantic rules are matched, the emotion matrix is established, and the emotion score is calculated, which will be standardized to the interval (0,1). The closer a standardized sentiment score is to zero, the more negative the sentiment is. Conversely, the closer the sentiment score is to 1, the more positive the sentiment is. This allows a threshold to be set programmatically to filter desired emotional attitudes. This method is more objective and faster than the original artificial emotion analysis[6].

IV. Implementation of Network Public Opinion System

A. Data Acquisition

1) Climb the Hot Search list

Firstly, enter the hot search list of Douyin, capture the interface of Douyin hot search list by Fiddler, and then analyze the url composition of the interface and the encryption mode of parameters. Fig. 7 below.



收到议案271件建议8000多件

1 建议优化火车票改签规定
1149.2万 热度

2 天舟六号已运往文昌航天发…
1145.8万 热度

3 中国国土绿化状况公报发布
1145.1万 热度

4 丁薛祥何立峰张国清刘国中…
1134.0万 热度

5 国务委员名单
1132.2万 热度

6 李小鹏任交通运输部部长
1113.6万 热度

7 四川女篮夺WCBA总冠军
1111.8万 热度

8 U20国足1比3韩国
1098.9万 热度

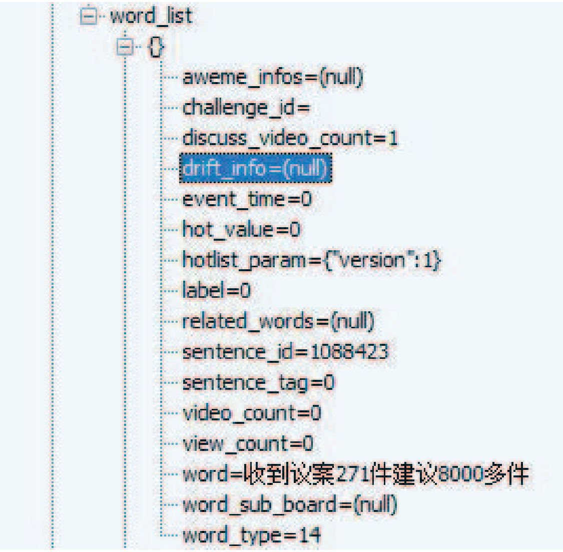
9 马尔代夫恢复与伊朗外交关系
1083.1万 热度

10 李尚福任国防部部长
1083.0万 热度

Figure 7. Hot Search list display

Note:The Chinese characters in Figure 7 are the title names of the TikTok hot search list.

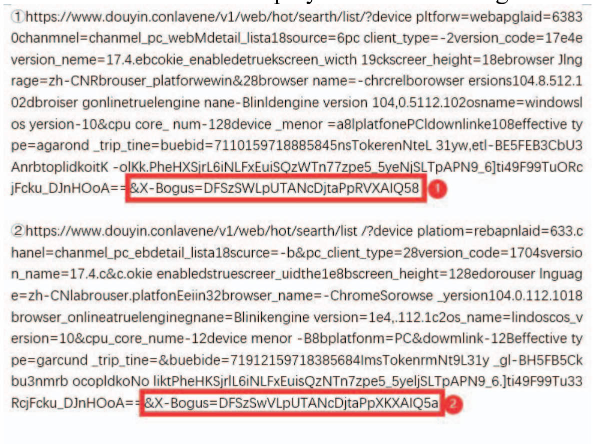
The interface return result is shown in Fig. 8 below.



```
word_list
  {}
    aweme_infos=(null)
    challenge_id=
    discuss_video_count=1
    drift_info=(null)
    event_time=0
    hot_value=0
    hotlist_param={"version":1}
    label=0
    related_words=(null)
    sentence_id=1088423
    sentence_tag=0
    video_count=0
    view_count=0
    word=收到议案271件建议8000多件
    word_sub_board=(null)
    word_type=14
```

Figure 8. Display of interface return results

The two url interface displays are shown in Fig. 9 below.

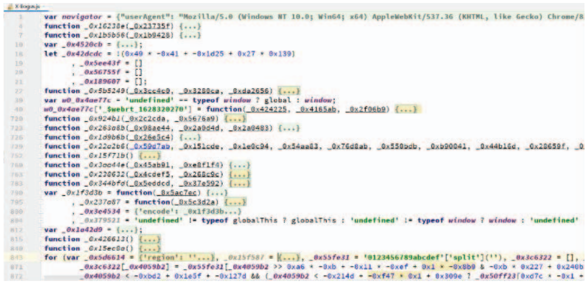


```
①https://www.douyin.conlavene/v1/web/hot/searth/list/?device pltforw=webapglaid=6383
0chanmnel=chanmel_pc_webMdetail_lista18source=6pc client_type=-2version_code=17e4e
version_neme=17.4.ebcokie_enabledetruekscreen_wicth 19ckscreer_height=18ebrowser Jlng
rage=zh-CNRbrouser_platforwewin&28browser name=-chrcrelborowser ersions104.8.512.1
02dbroiser gonlinetruelengine nane-Blinldengine version 104,0.5112.102osname=windowsl
os yersion-10&cpu core_ num-128device _menor =a8lplatfonePCldownlinke108effective ty
pe=agarond _trip_tine=buebid=7110159718885845nsTokerenNteL 31yw,etl-BE5FEB3CbU3
AnrbtoplidkoitK -olKk.PheHXSjrL6iNLFxEuiSQzWTn77zpe5_5yeNjSLTpAPN9_6]ti49F99TuORc
jFcku_DJnHOoA==&X-Bogus=DFSzSWLpUTANcDjtaPpRVXAIQ58 ❶

②https://www.douyin.conlavene/v1/web/hot/searth/list /?device platiom=rebapnlaid=633.c
hanel=chanmel_pc_ebdetail_lista18scurce=-b&pc_client_type=28version_code=1704versio
n_name=17.4.c&c.okie enabledstruescreer_uidthe1e8bscreen_height=128edorouser lnguag
e=zh-CNlabrouser.platfonEeiin32browser_name=-ChromeSorowse _yersion104.0.112.1018
browser_onlineatruelenginegnane=Blinikengine version=1e4,.112.1c2os_name=lindoscos_v
ersion=10&cpu_core_nume-12device menor -B8bplatfonm=PC&dowmlink-12Beffective ty
pe=garcund _trip_tine=&buebide=71912159718385684lmsTokenrmNt9L31y _gl-BH5FB5Ck
bu3nmrb ocopldkoNo liktPheHKSjrlL6iNLFxEuisQzNTn7zpe5_5yeljSLTpAPN9_6.]ti49F99Tu33
RcjFcku_DJnHOoA==&X-Bogus=DFSzSwVLpUTANcDjtaPpXKXAIQ5a ❷
```

Figure 9. Two url interface displays

Comparing the url interface twice, we can see that the X-Bogus parameter in the url changes each time, so we need to analyze how the X-Bogus parameter was generated before we can solve the anti-climb problem. Through analysis, it is found that the X-Bogus parameter is passed through the JS file. Before accessing the interface, the string before the X-Bogus parameter inside the interface is passed into the JS function. Then after generating the X-Bogus parameter, the interface can be correctly accessed with this parameter Fig. 10 below.

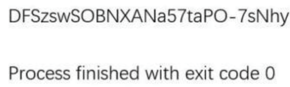


X-Bogus.js

```
var navigator = {"userAgent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/8
function _0x16238e(_0x23735f) {...}
function _0x1b5b56(_0x1b9428) {...}
var _0x4520cb = {...};
let _0x42dcdc = !(0x49 * -0x41 + -0x1d25 + 0x27 * 0x139)
      , _0x5ee43f = []
      , _0x56755f = []
      , _0x189607 = [];
function _0x5b5249(_0x3cc4c0, _0x3280ca, _0xda2656) {...}
var w0_0x4ae77c = 'undefined' == typeof window ? global : window;
w0_0x4ae77c['_$webrt_1628320270'] = function(_0x424225, _0x4165ab, _0x2f06b9) {...}
function _0x924b1(_0x2c2cda, _0x5676a9) {...}
function _0x263a8b(_0x98ae44, _0x2a9d4d, _0x2a0483) {...}
function _0x1d9b6b(_0x26e5c4) {...}
function _0x22a2b6(_0x59d7ab, _0x151cde, _0x1e0c94, _0x54aa83, _0x76d8ab, _0x558bdb, _0xb90041, _0x44b16d, _0x28659f, _0
function _0x15f71b() {...}
function _0x3aa44e(_0x45ab91, _0xe8f1f4) {...}
function _0x238632(_0x4cdef5, _0x268c9c) {...}
function _0x344bfd(_0x5eddcd, _0x37e592) {...}
var _0x1f3d3b = function(_0x5ac7ec) {...}
      , _0x237a87 = function(_0x5c3d2a) {...}
      , _0x3e4534 = {'encode': _0x1f3d3b...}
      , _0x379521 = 'undefined' != typeof globalThis ? globalThis : 'undefined' != typeof window ? window : 'undefined'
var _0x1a42d9 = {...};
function _0x426613() {...}
function _0x15ec8a() {...}
for (var _0x5d6614 = {'region': ''...}, _0x15f587 = {...}, _0x55fe31 = '0123456789abcdef'['split'](''), _0x3c6322 = [],
        _0x3c6322[_0x4059b2] = _0x55fe31[_0x4059b2 >> 0xa6 * -0xb + -0x11 * -0xef + 0x1 * -0x8b9 & -0xb * 0x227 + 0x240b
        _0x4059b2 < -0xbd2 + 0x1e5f + -0x127d && (_0x4059b2 < -0x214d - -0xf47 * 0x1 + 0x309e ? _0x50ff23[0xd7c * -0x1 +
```

Figure 10. JS file display

The generated X-Bogus parameter is shown in Fig. 11 below.

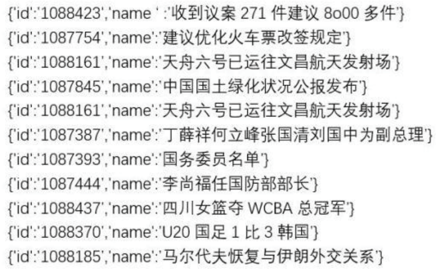


```
DFSzswSOBNXANa57taPO-7sNhyMV

Process finished with exit code 0
```

Figure 11. X-Bogus generated parameter display

Then through this method, access the hot search list interface, climb out the top 10 list names and list video ID. Fig. 12.



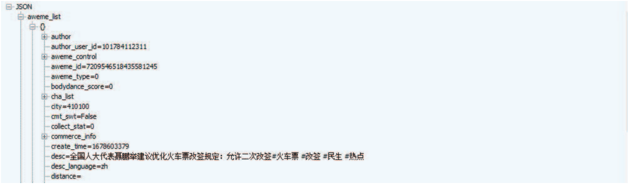
```
{'id':'1088423','name ' :'收到议案 271 件建议 8o00 多件'}
{'id':'1087754','name':'建议优化火车票改签规定'}
{'id':'1088161','name':'天舟六号已运往文昌航天发射场'}
{'id':'1087845','name':'中国国土绿化状况公报发布'}
{'id':'1088161','name':'天舟六号已运往文昌航天发射场'}
{'id':'1087387','name':'丁薛祥何立峰张国清刘国中为副总理'}
{'id':'1087393','name':'国务委员名单'}
{'id':'1087444','name':'李尚福任国防部部长'}
{'id':'1088437','name':'四川女篮夺 WCBA 总冠军'}
{'id':'1088370','name':'U20 国足 1 比 3 韩国'}
{'id':'1088185','name':'马尔代夫恢复与伊朗外交关系'}
```

Figure 12. Display of crawl results

Note:The Chinese in Figure 12 is crawling out the top 10 list names of the hot search list Call video ID with list.

2) Crawl the comments and IP addresses of list videos

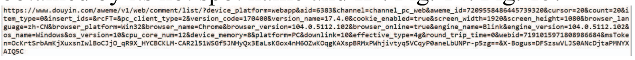
Firstly, the interface url of the video comment on the list is analyzed. Through the structure of the url, we find that the parameter aweme_id in the url represents the aweme_id of the video under the hot search list, but it is not the ID of the hot search list obtained in the first step. Therefore, interface analysis is needed on how aweme_id is obtained. The analysis shows that when clicking the hot hand list, a new interface will be accessed, and the data returned by the interface contains the aweme_id parameter. Fig. 13.



```
JSON
  aweme_list
    {}
      author
      author_user_id=101784112311
      aweme_control
      aweme_id=7209546518435581245
      aweme_type=0
      bodydance_score=0
      cha_list
      city=410100
      cmt_swt=False
      collect_stat=0
      commerce_info
      create_time=1678603379
      desc=全国人大代表聂鹏举建议优化火车票改签规定：允许二次改签#火车票 #改签 #民生 #热点
      desc_language=zh
      distance=
```

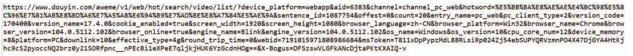
Figure 13. Display of interface results

After analyzing the url of this interface, it is found that there are two parameters, hotword and sentence_id, which correspond to the list name and ID obtained in the first step. aweme_id in the url to obtain each hot video comment is obtained by its hot spot name and ID. Fig. 14. Fig. 15.



```
https://www.douyin.com/aweme/v1/web/comment/list/?device_platform=webapp&aid=6383&channel=channel_pc_web&aweme_id=7209558486445739320&cursor=20&count=20&i
tem_type=0&insert_ids=&rcFT=&pc_client_type=2&version_code=170400&version_name=17.4.0&cookie_enabled=true&screen_width=1920&screen_height=1080&browser_lan
guage=zh-CN&browser_platform=Win32&browser_name=Chrome&browser_version=104.0.5112.102&browser_online=true&engine_name=Blink&engine_version=104.0.5112.102&
os_name=Windows&os_version=10&cpu_core_num=12&device_memory=8&platform=PC&downlink=10&effective_type=4g&round_trip_time=0&webid=7191015971808986684&msToke
n=OcKrtSrbAmKjXuxsnIwlBoCJjO_qR9X_HYCBCKLM-CAR2151WSGf5JNHyQx3EaLsKGox4nH6OZwKOqgKAXspBRMxPWhjivtyq5VCqyP0aneLbUNPr-p5zg==&X-Bogus=DFSzswVLJS0ANcDjtaPMNYX
AIQ5C
```

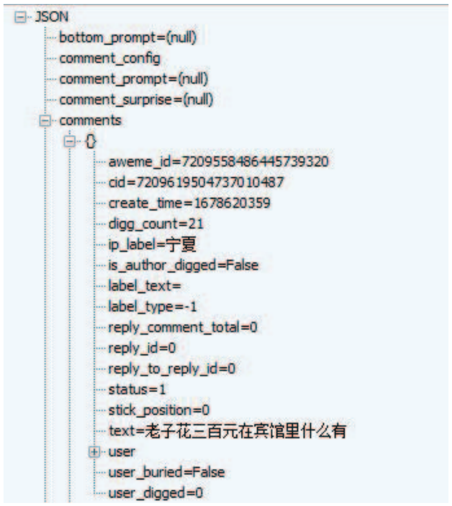
Figure 14. url display of comment interface



```
https://www.douyin.com/aweme/v1/web/hot/search/video/list/?device_platform=webapp&aid=6383&channel=channel_pc_web&hotword=%E5%BB%BA%E8%AE%AE%E4%BC%98%E5%8
C%96%E7%81%AB%E8%BD%A6%E7%A5%A8%E6%94%B9%E7%AD%BE%E8%A7%84%E5%AE%9A&sentence_id=1087754&offest=0&count=20&entry_name=pc_web&pc_client_type=2&version_code=
170400&version_name=17.4.0&cookie_enabled=true&screen_width=1920&screen_height=1080&browser_language=zh-CN&browser_platform=Win32&browser_name=Chrome&brow
ser_version=104.0.5112.102&browser_online=true&engine_name=Blink&engine_version=104.0.5112.102&os_name=Windows&os_version=10&cpu_core_num=12&device_memory
=8&platform=PC&downlink=10&effective_type=4g&round_trip_time=0&webid=7191015971808986684&msToken=T811xDpPypzMdL88RLsiRp024Zj54ebSUPYQRVzmnPO4X47DjGYA4HtKj
hc9cS2pyoccNQ2brz0y21SORfpnc__nPEcBi1eXPeE7qljkjHUK6YzGcdnHOg==&X-Bogus=DFSzswVLGFkANcDjtaPKtKXAIQ-v
```

Figure 15. Interface for obtaining aweme_id

Then, the aweme_id obtained can be used to access the evaluation interface, and the comments and IP location of the video under the hotspot can be obtained normally. Fig. 16 shows.



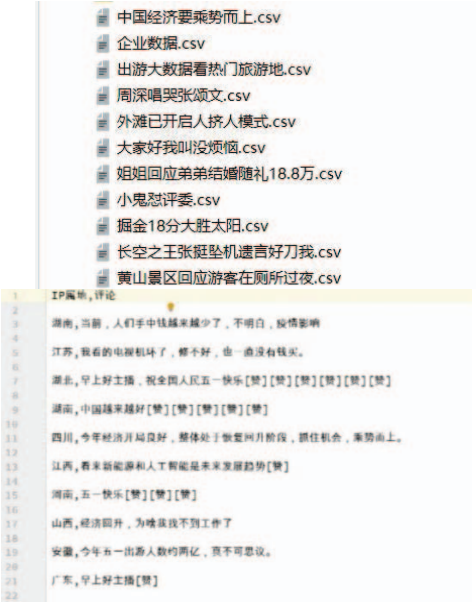
```
⊟ JSON
    bottom_prompt=(null)
    comment_config
    comment_prompt=(null)
    comment_surprise=(null)
  ⊟ comments
    ⊟ {}
        aweme_id=7209558486445739320
        cid=7209619504737010487
        create_time=1678620359
        digg_count=21
        ip_label=宁夏
        is_author_digged=False
        label_text=
        label_type=-1
        reply_comment_total=0
        reply_id=0
        reply_to_reply_id=0
        status=1
        stick_position=0
        text=老子花三百元在宾馆里什么有
      ⊞ user
        user_buried=False
        user_digged=0
```

Figure 16. Comments and IP homing display

Finally, this method successfully realizes the comments and IP location data crawling of hot search list videos.

3) Data storage

The comment data obtained in the second step is stored separately. Each hot search comment is stored separately, and the hot search name is used as the file name and saved as a CSV file. Fig. 17



中国经济要乘势而上.csv
企业数据.csv
出游大数据看热门旅游地.csv
周深唱哭张颂文.csv
外滩已开启人挤人模式.csv
大家好我叫没烦恼.csv
姐姐回应弟弟结婚随礼18.8万.csv
小鬼怼评委.csv
掘金18分大胜太阳.csv
长空之王张挺坠机遗言好刀我.csv
黄山景区回应游客在厕所过夜.csv

```
IP属地,评论

湖南,当前，人们手中钱越来越少了，不明白，疫情影响

江苏,我看的电视机坏了，修不好，也一直没有钱买。

湖北,早上好主播，祝全国人民五一快乐[赞][赞][赞][赞][赞][赞]

湖南,中国越来越好[赞][赞][赞][赞][赞]

四川,今年经济开局良好，整体处于恢复回升阶段，抓住机会，乘势而上。

江西,看来新能源和人工智能是未来发展趋势[赞]

河南,五一快乐[赞][赞][赞]

山西,经济回升，为啥我找不到工作了

安徽,今年五一出游人数约两亿，真不可思议。

广东,早上好主播[赞]
```

Figure 17. Display of storage results

Note:The obtained comment data in Figure 17 were stored separately, and the hot search name was used as the file name and saved as a CSV file.

B. Data Preprocessing

Data obtained by crawler needs data preprocessing operation, and the data obtained is the user IP location and comment of each hot search comment. The comments for each IP territory are counted by IP territory classification. Review data is cleaned, empty lines are filtered out, and Jieba is performed. Word frequency statistics of comments are performed by Jieba word segmentation technology, and the word frequency data of comments are finally obtained.

Write python code to read the csv file of each hot search comment and establish the Jieba function to process each hot search Jieba word segmentation.

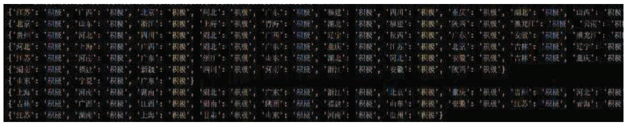
C. Comment on Sentiment Analysis

SnowNLP emotion analysis method was used to conduct emotion analysis on the comment information of Douyin hot search, and the emotion score was obtained. If the emotion score was 0-0.33, the comment was considered negative; if the emotion score was 0.33-0.66, the comment was considered neutral; if the emotion score was 0.66-1.00, the comment was considered positive. As shown in Table II.

TABLE II. SCORE TABLE OF COMMENT SENTIMENT ANALYSIS

Emotion score	Sentiment analysis
0-0.33	Passive
0.33-0.66	Neutral
0.66-1.00	Vigorous

The variables negative, neutral and positive are the attitude of each comment of the hot search, so as to judge the number of negative, positive and neutral comments in the hot search. Where data[" comment "][j] is a string format of a comment, a parameter in the interval (0,1). The smaller the return value, the greater the negative impact. In addition, if the negative users account for more than 60%, it is considered that the region has a negative attitude towards the hot search. If the positive users account for more than 60%, it is considered that the region has a positive attitude towards the hot search. The rest of the group considered the comment neutral. The preliminary statistics of attitudes in different regions of each hot search are shown in Fig. 18 below:



```
{'江苏': '积极', '广西': '积极', '北京': '积极', '河北': '积极', '广东': '积极', '福建': '积极', '四川': '积极', '重庆': '积极', '湖北': '积极', '山西': '积极
{'北京': '积极', '山东': '积极', '浙江': '积极', '上海': '积极', '青海': '积极', '湖北': '积极', '福建': '积极', '陕西': '积极', '黑龙江': '积极', '云南': '积
{'贵州': '积极', '河北': '积极', '四川': '积极', '湖北': '积极', '广西': '积极', '辽宁': '积极', '陕西': '积极', '广东': '积极', '安徽': '积极', '黑龙江': '积
{'河北': '积极', '上海': '积极', '广西': '积极', '湖北': '积极', '广东': '积极', '重庆': '积极', '江苏': '积极', '北京': '积极', '吉林': '积极', '辽宁': '积极
{'江苏': '积极', '河南': '积极', '广东': '积极', '浙江': '积极', '山东': '积极', '湖北': '积极', '河北': '积极', '安徽': '积极', '吉林': '积极', '重庆': '积极
{'湖南': '积极', '福建': '积极', '新疆': '积极', '四川': '积极', '河南': '积极', '浙江': '积极', '安徽': '积极', '陕西': '积极'}
{'山东': '积极', '宁夏': '积极', '广东': '积极'}
{'上海': '积极', '河南': '积极', '湖南': '积极', '湖北': '积极', '广东': '积极', '浙江': '积极', '北京': '积极', '重庆': '积极', '贵州': '积极', '河北': '积极
{'吉林': '积极', '广西': '积极', '江西': '积极', '湖南': '积极', '陕西': '积极', '福建': '积极', '山东': '积极', '安徽': '积极', '江苏': '积极', '青海': '积极
{'江苏': '积极', '湖南': '积极', '上海': '积极', '甘肃': '积极', '山东': '积极', '河南': '积极', '贵州': '积极'}
```

Figure 18. Attitudes towards hot searches in different regions.

Note:The Chinese characters in Figure 18 are preliminary statistics on the attitudes of different regions of each hot search, which are the name and attitude of the region.

D. *Data Visualization*

Through data processing, the number of comments in different regions of each hot search is analyzed, and the Flask visualization screen is built. The Douyin hot search list is built on the left side. You can click the name of the top search list on the left to jump to the graph of the distribution of the number of comments in the current trending search. Flask data visualization screen is shown in Fig. 19.

Figure 19. Distribution of the number of comments

Note: The Chinese word cloud map in Figure 19 shows the general situation of the comment data.

V. Conclusion

In this paper, public opinion analysis of Douyin hot search comments based on NLP has been basically completed. This paper proposes to use Scrapy crawler framework to crawl the hot search data of Douyin in real time, monitor the hot search of Douyin, carry out data cleaning, weight removal and void removal on the acquired data, then use the Chinese word segmentation technology of Jieba to segment the review data, and then use TF-IDF algorithm to vectorize the data after word segmentation. The NLP natural language processing algorithm is used to process the text and analyze the sentiment of the hot search information and comments. The positive, neutral and negative attitudes of comments in different regions were evaluated, and the status of the trending searches in different provinces was displayed on the map of China through data visualization. You can intuitively check the attitude of each region to Douyin hotsearch. Real-time control of Douyin hot search public opinion.

Due to my practical experience and lack of writing cycle, there are still many places in this paper can continue to improve and sublimate. Here are some examples[7].

- Add more online public opinion evaluation indicators.
- More Douyin data can be crawled for analysis. Not only the IP address of the original video and user comments, but also the comments and IP under the forwarded video can be added to the analysis.
- Add speech recognition or text extraction method to identify the language content or text in the video, analyze the content, and give greater weight to this part when judging the network public opinion.

Acknowledgment

This work was funded by Yunnan Province local undergraduate universities basic research joint special fund project.

Reference

[1] Y. B. Lai, Y. Chen, X. C. Hu and X. H, “Sentiment analysis of public health emergency microblogs based on prompt embedding,” Data analysis and knowledge discovery: p. 1, 2023. HTTP://kns.cnki.net/kcms/detail/10.1478.g2.20230206.1809.001.html

[2] Q. X. Song, K. Liu and Q. F. Xun, “Research on decision modeling of police-related network public opinion multi-stage intervention based on Big Data,” Mathematics in Practice and Understanding, vol. 53, no. 2, p. 13, 2023.

[3] Y. Li, Z. S. XU and X. X. Wang, “Sentiment analysis method based on online reviews and application,” Control and decision, 2022. The DOI: 10.13195 / j.k. zyjc. 2022.1788.

[4] W. Wei, X. S. Sun, L. F. Li and Y. X. Zhang, “Sentiment analysis method based on text discusses, “ Journal of digital technology and applications, 2022, 40 (12) : 1-3 + 22. DOI: 10.19695 / j.carol carroll nki cn12-1369.2022.12.01.

[5] J. W. Liu and S. Q. Zhang, “Evolution analysis of hot topics in Tiktok based on Sentiment Analysis,” Information Systems Engineering, no. 12, pp. 137-140, 2022.

[6] Z. X. Jia, “Design and implementation of short text Sentiment analysis system based on Deep learning,” Nanjing university of posts and telecommunications, 2022. doi:10.27251/d.cnki.gnjdc.2022.000420.

[7] S. Gao, Y. Huang, and W. J. Dai, “Speech words intelligent recognition method based on NLP,” Journal of information technology, vol. 46, no. 11, pp. 66-70, 2022, doi: 10.13274/j.cnki.hdzj.2022.11.012.