Journal of Computers Vol. 28, No. 6, 2017, pp. 301-313
doi:10.3966/199115992017122806027

Research on Keyword Extraction and Sentiment Orientation Analysis of Educational Texts

Lin Zhang[1,2], Xiao-Ping Li[2], Fan-Bo Zhang[3] and Bo Hu[2]

[1] Beijing University of Civil Engineering and Architecture, Beijing, China

[2] Beijing Institute of Technology, Beijing, China

[3] Dalian Maritime University, Dalian, China
zhanglin@bucea.edu.cn

Received 24 July 2017; Revised 19 September 2017; Accepted 19 October 2017

Abstract. Data acquisition, text keyword extraction and text sentiment orientation analysis are important technologies which can be applied in text content analysis. In this paper, a general crawler based on Scrapy framework is designed. It can be applied to many kinds of web sites and improved the efficiency and the versatility. A mixed method based on Textrank algorithm and TF-IDF algorithm is proposed. It can be applied to mobile client to extract the keywords. For the text tendency analysis, a new method based on SnowNLP is proposed. The experiment shows that crawling and extracting results and the tendency judge are more accurate for the long educational text.

Keywords: educational text, text analysis, text keyword extraction, text sentiment orientation analysis

1 A Summary of Text Analysis

Text analysis [1] refers to the representation of a text and the selection of its characteristics. Text analysis is one of the most fundamental problems in information retrieval. The text information can be expressed by extracting and quantifying the text feature items. Then they are analyzed by mathematical modeling and information processing so that they can be processed much easier. With the rapid development of Internet and artificial intelligence, text analysis also tends to be more intelligent, semantic and digitized. It will play a more important role on content resource management in the future.

1.1 Text Keyword Extraction

Text keywords extraction means extracting the most relevant information from the text. The appropriate keywords may significantly improve the retrieval efficiency of the article and help users to quickly determine whether they are their demand. It can improve the access efficiency and hit rate to the text. There are two kinds of keywords extraction methods. One is called keywords allocation. There must be a keyword lexicon in the system. When a text is analyzing, some related words will be assigned to it as its keywords. The other method is words extracted. Some words relevant to the subject can be extracted automatically from the text as the keywords. At present, most areas-unrelated keyword extraction algorithms are based on the second method.

The keywords extraction is the basis of text analysis and processing. The minimum unit expressing the text content is keyword, so it is important for natural language processing, such as text clustering [2], text classification, information retrieval [3]. It also is the foundation of text data mining [4]. After years of development, there are a lot of methods which can realize keywords extraction.

Keyword extraction based on statistics. This method is based on statistical model which is used to solve the problems about language recognition. It can judge if one sentence is correct by computing the probability of the sentence. This method can be applied to the implementation of keyword extraction.

First, build a distribution for the vocabulary information and vocabulary characteristics. Then compute the vocabulary frequency and location. At last, extract the keywords according to some constraints or calculation. Influenced by statistical method, some high-frequency words which are unrelated to the topic are easy to be extracted so that the accuracy will be influenced. But for its simplification, it is used by many researches.

Keyword extraction based on machine learning. This method is also commonly used in extracting keywords. Its main idea is to use the text classification. After setting a keyword classification model by training a large number of samples, some words will be extracted and determined if they are keywords. The most commonly used method is based on SVM (Support Vector Machine) [5]. If the training samples are enough, the extraction result will be more accuracy.

Keyword extraction based on semantics. This method is used to extract keyword by semantic analysis. Words semantics refers to the semantic similarity between words. It also is called the words distance and it uses the synonyms dictionary to calculate the distance between two words. This method can get semantic information by means of semantic knowledge and exploit the potential text content. So the quality of text information extraction is much higher than other methods.

1.2 Text Tendency Analysis

Text sentiment orientation analysis [6] mainly refers to explore the emotion of the text whether it is positive or negative. It belongs to the category of computer language processing which involves in machine learning, data mining and information retrieval technology. For some complex factors, different people have different opinion to the same text. There are four kinds of analysis methods.

Words sentiment orientation analysis. It can get the conclusion by analyzing some sentient words such as nouns, verbs, adjectives, etc. The words are commendatory or derogatory is the basic judgement standard. There are many methods can be applied to the words sentiment orientation analysis such as library-based, machine learning and artificial tagging.

Sentence sentiment orientation analysis. Sentence sentiment orientation analysis is based on words and is much more complicated than words. Because it depends on the context, it is important to analyze the relationship between words.

Text sentiment orientation analysis. It analyzes the tendency of text based on the average tendency of words or phrases. But it is not accurate enough. Another method adopts sample training method by labeling the large number of text manually to create learning classifier and complete the tendency analysis. Due to the samples focus on different areas and the effect of the classifier is different, the model is not universal enough. We cannot analyze the article as a whole research object because there may be some different viewpoints and we cannot get the correct conclusion. Therefore at present the method is not universally applicable.

2 The Implementation of Web Crawler

2.1 The Overview of Web Crawler

Web crawler [7] is also known as network robot or web spider. It is a kind of script or program which can automatically grabs the web information according to certain rules. If we consider the Internet as a directed graph, the content of the web as a node, link between web pages as edge, the web crawler will be a program which can traverse the graph. It can download the web information when it is in the process of traversing. The crawler is generally based on graph traversal algorithm. It starts working from a URL and looking for all URLs of the page, then adds them to the URL list and continues to traverse the URL list until the crawler is closed. Due to the characteristics of web pages, in the process of execution of the crawler, it will also involves some technical means, such as the structure of the links, data analysis, dynamic data acquisition and crawler rules setting. Combining all these techniques will constitute a complete crawler system.

2.2 The Design of Crawler Frame

Scrapy crawler frame. Scrapy is a high-level web scraping framework based on the rapid development of Python which is used to grab web site information and extract structured data from the page. Scrapy can be used in data mining, information processing and historical archives. Users only need to design the corresponding module based on rules, it will realize to fetch various web pages. Although Scrapy is designed to realize the web scraping, but at the same time it can also get the data by calling the API.

The principle of Scrapy is shown as in Fig. 1. The framework of the crawler program obtains the initial URL first and Scrapy Engine will pass it to the scheduler module. The module will assemble the URL into legal URL request, then pass it to the Downloader module to download. Downloader modules pass it to Spider to parse when it gets the web page information returned by the server. The Spider will parse out two kinds of results from the page, respectively is the link to be further grab and the data to be processed. The link will be transmitted the Scheduler module by the Engine to generate new request and repeat the above actions. The data will be transmitted to Pipiline module to complete some operations, such as data storage, data analysis, data filtering and so on. In the process of data transfer, we can design various Middlewares to realize the necessary processing for the pass data.

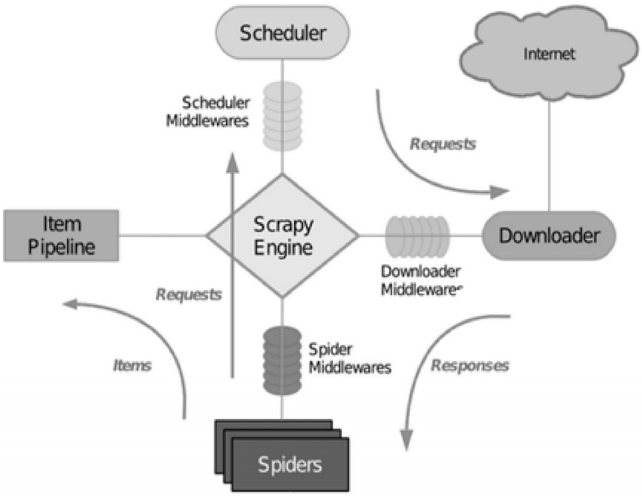

Fig. 1. The principle diagram of Scrapy

General Scrapy crawler process procedure. In order to apply to a variety of types site data fetching, a universal fetching process of General crawler based on Scrapy is designed as Fig. 2, it can try to avoid huge program changes caused by the different site structure.

The general Scrapy crawler process procedure is described as follow:

Step 1: The program starts from the entry URL.

Step 2: Get the current page of the specified URL and process them. Because most of URLs are expressed as relative path and the crawler only can deal with the absolute URL, so the procedure will construct them into executable paths first. Then they will be saved into the URL list in order to avoid repeating the crawl operations.

Step 3: Determine whether the current URL is the final data page. If it is, the procedure will continue, otherwise it will return to Step 2.

Step 4: Determine whether the current page needs to obtain the AJAX or JS dynamic data. If it needs, the procedure will execute the Step 5, otherwise the Step 6 will be executed.

Step 5: For the request of AJAX or JS dynamic data, the procedure will parse the page code by calling Ghost as Scrapy downloader middleware.

Step 6: Match the download page data and store the successful match data into database.

The design and implementation of General Scrapy crawler. The module of General Scrapy crawler is designed as Fig. 3.

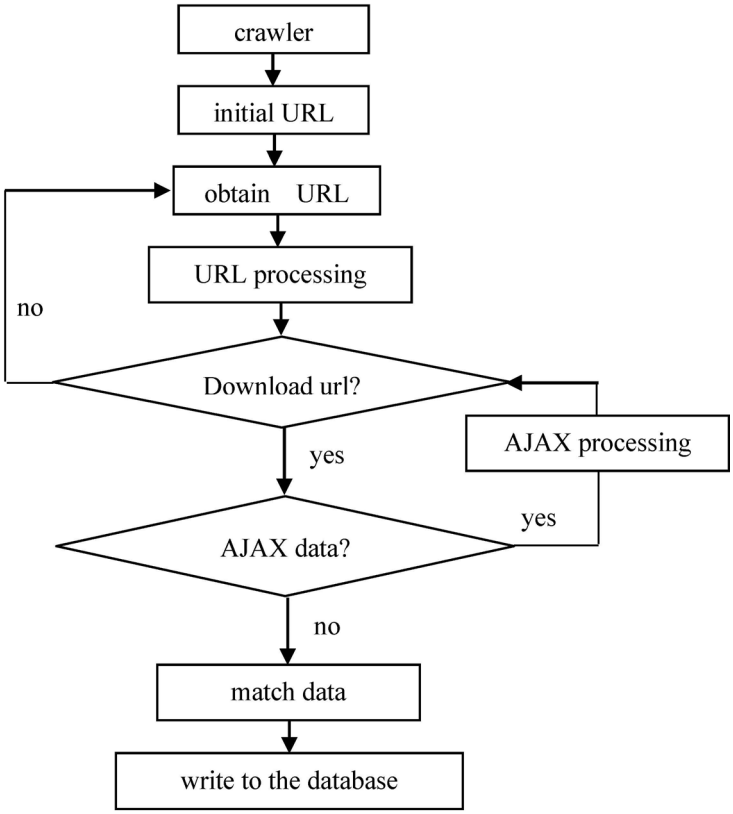

Fig. 2. General Scrapy crawler process procedure

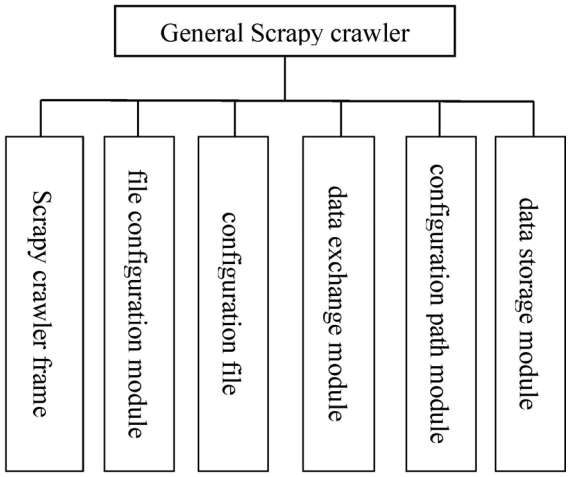

Fig. 3. The module of General Scrapy crawler

To realize General Scrapy crawler, the major problem we must solve is the acquisition of dynamic interactive data which mainly are requested by AJAX. AJAX is an asynchronous interactive network application which can improve the user experience in the Web2.0 era. More and more web sites began to adopt this technology. Just because of the rapid development of this technology, the design of General Scrapy crawler is becoming more difficult.

Currently there are two solutions for this problem. One is to construct the browser through the simulation so that crawlers automatically achieve dynamic data interaction process. Another way is to do specific analysis for specific sites. Look for data request source from the JS code so that the program automatically request data from the data source.

The first method is a general method, which is very suitable for the implementation of the General

Scrapy crawler. But because of the need to simulate the browser to parse the data, the crawl complexity will greatly increase. So the program time and space load will increase and affect crawling efficiency seriously. The second method is very high efficiency, but it is only for specific crawling, so its versatility is very poor.

This paper combines the advantages and disadvantages of the two approaches so that the crawler program can improve the efficiency and the versatility as much as possible. For the second method, the parser was extended and defines the 'json' for the field of type in the json configuration file. When we parse in this way, we can directly read the field name. Of course, this will increase the difficulty of the file configuration and it will take more time to read the site page source code and find the data request address.

In order to simulate the browser to parse AJAX as the first method, this paper adopts the Ghost library provided by Python so as to achieve dynamic interactive data acquisition. Ghost is a python web client based Webkit and it can easily use the code to achieve the function of the browser simulation. Using the open ( ) method, we can easily open a web page and return two parameters, one saves all the page HTML information, and the other loads all the resources (CSS, javascripts, images, etc.). This makes it easy to parse page AJAX data. Ghost provides other methods are as follows:

(1) Ghost.evaluate( )

This method can get DOM elements in JS way and execute JS program.

(2) Ghost.set_field_value( )

This method simulates to assign values to form.

(3) Ghost.fill( )

This method simulates to realize input operation of the text data box.

(4) Ghost.call( )

This method simulates to submit the form or call the function.

(5) Ghost.click( )

This method simulates to complete the browser click event.

Using the above method, we can simply simulate the operation of the browser to achieve the basic data manipulation, so that the crawler skip some of the validation page and targeted selection of pages. Of course, the biggest flaw of simulation browser program is the increase of space and time load, so the program is inefficient. The simulation browser will download all page resources when it loads the page, which greatly increases the crawling corresponding time. Ghost provides several loading methods to achieve page loading without the need for all script scripts are executed:

(1) Ghost.wait_for_alert( )：

Wait for finishing alert() of js.

(2) Ghost.wait_for_page_loaded( ):

Wait for finishing loading for a new page.

(3) Ghost.wait_for_selector(selector):

Wait for finishing loading for the element that matches the selector.

(4) Ghost.wait_for_text (text):

Wait for finishing loading matching text.

We can greatly enhance the efficiency of crawling by combining with these methods to end the unnecessary script analysis for different sites. Combine Ghost with Scrapy by downloading middleware and define class of GhostDownloader as a middleware, we can analyze the dynamic data. But if each page is analyzed, the crawler efficiency is still very low. So we add the judgement of URL and only analyze some of web pages in order to improve efficiency.

During the process of crawling, access link and capture data directly will greatly improve the efficiency of the implementation of the program. We define three classes in extractor.py:

(1) BaseExtractor: It completes the most basic data extraction, including direct static data analysis and the most primitive Xpath data analysis.

(2) ItemExtractor: It is derived from class BaseExtractor. ItemExtractor achieves the analytic extraction for extraction data items.

(3) FLinksExtractor: It is also inherited from class BaseExtractor. FLinksExtractor achieves the analytic extraction for the URL link.

The analytical extraction methods of ItemExtractor and FLinksExtractor are all based on the analysis way which combined regular expression and Xpath.

The basic methods of these classes include a variety of page types of data analysis. Which method will be called should be determined by the content of type and value in the configuration file.

The implement of crawler is shown as Fig. 4.

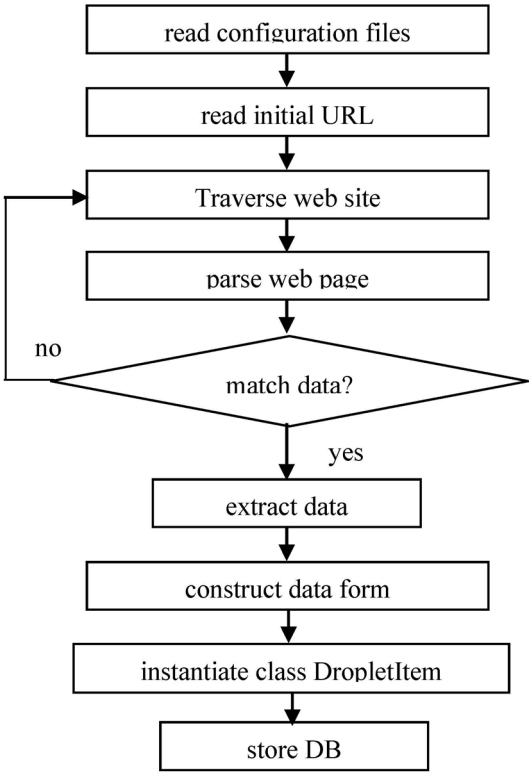

Fig. 4. The implement of crawler

The main program of droplet implements the crawler execution class DropletSpider (Spider), which inherits from the base class Spider of the Scrapy framework and defines the method of __init __ (self, config, ** kwargs) so as to read the legal crawler configuration file when the class is initialized.

The entire process begins with Start_requests (). It reads the entry URL and traverses web site according to the rules of the configuration files. The page will be passed to traversal () to be parsed when the traversal is going on. The main tasks are recognizing traversal rules of PAGE_LINK and LINKS, judging and finding the relative path address, and constructing URL of the current page request by the deal_URL () method.

When the crawler matches the data item, it calls make_item () to extract the data item and constructs them into the stored data format. Then the DropletItem class will be instantiated and the basic properties of the crawler will be added, which includes crawler name, web name, crawling date and site type, which will be stored into Mongodb database with the help of scrapy_mongodb. Finally, use the command "scetch crawl droplet-a config_file = config / tieba.json" to select the configuration file and start the crawler.

Mobile data capture based on mobile agent. The main purpose of the program of capturing client data based on the proxy server is to take advantage of the network proxy server and set the mobile client network as agent. After the client sends a request to the site, the data is sent to the client by proxy server. The proxy server captures the data packets during they are transmitting, and then filter and parse the data in order to obtain them. The main work flow is shown as Fig. 5.

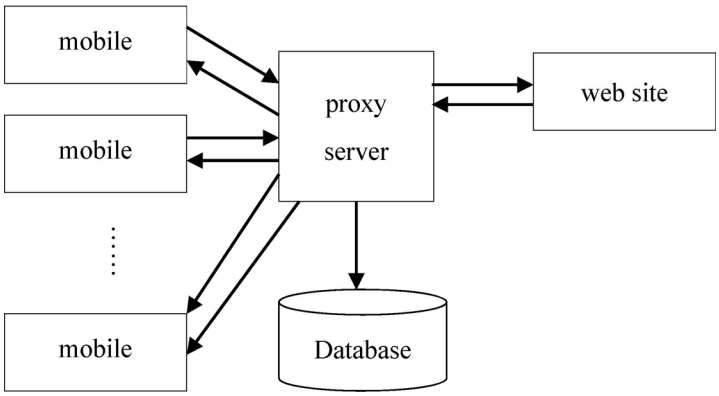

Fig. 5. The main work flow of proxy server

We can install an analog APP software to the mobile device and set the software to replace the manual operation of mobile client. The network request can be sent automatically by clicking on the public account article. The data are sent by proxy server which copies the data packets to local machine and the server program will parse and store them into the database.

During the data processing process, the proxy server is being in the monitoring state and monitoring the request sent to the site by the mobile device.

When the request URL conforms to the rules of public platform, the return data will be copied. Otherwise any other network requests are ignored. This method can filter out some unnecessary data. After obtaining the data packets, the data will be filtered with regular expression and store the analytical data into the database finally by pymongo. For the evaluation of content resources, the target data we want to obtain are the text of the WeChat, text reading quantity and Thumb up quantity.

In order to realize data capture based on the network proxy server, we must first establish a proxy server. In this project, we use python Mitmproxy as a proxy server. Mitmproxy is a framework for a middleman agent based on python. It can intercept and modify the http or https packets for the analysis of packet interaction applications. While for the mobile communication analysis, libmproxy library provided by Mitmproxy is a very good packet interception tool.The Libmproxy library contains many files. The most critical one is the file flow.py which includes class Request and class Response.

The Request methods include:

Get_guery( ):get the requested url parameters and save them into the dictionary.

Set_query(odict): set the request url parameter.

Set_url( ):get the request url.

Set_url( ):set the domain of the url.

Get_cookies( ):get the request cookie.

The Response methods include:

Headers( ):return the dictionary of the header.

Code( ): return the status of the packet.

Httpversion( ): return the http version.

We can set the proxy server as follow:

```
from libmproxy.proxy.server import ProxyServer
args = init_parser()
init_logging(args.log_dir)
server = ProxyServer(proxy.ProxyConfig(port=args.port))
HttpSniffer(server, args.data_dir).run()
```

The program can be run as follow:

```
python http_sniffer.py --port 12513 --log-dir D:\logs --data-dir D:\data
```

Ini_parser ( ) can analyze the command line, including obtaining listening port of the proxy server, log storage address and data storage address. Init_logging(args.log_dir) stores the basic information of monitor procedure provided by log storage of python. run( ) directly starts the entire proxy service program.

3 Text Keyword Extraction Algorithm

For the evaluation research of content resources, the study results should not only be a simple type of index data because they only reflect the visual value of content resources and cannot reflect the connotation of content resources. By analyzing text and extracting keywords using the keyword extraction technology, the main related contents of the texts can be obtained [8]. They can make people understand the topics covered in the content resources more conveniently and more intuitively. Using text orientation analysis technology, we can learn the social public opinion tendency of the content and the quantity of commendatory or derogatory content resources. It responses the society reaction of content resources more intuitively and it makes the evaluation more objective.

There are a lot of keyword extraction algorithms and they have different solution methods. The keyword extraction method based on statistics is the most common one. This topic will choose two basic keyword extraction algorithms (TF - IDF algorithm and textrank algorithm) for keyword extraction.

3.1 TF-IDF Keyword Extraction Algorithm

TF-IDF (Term Frequency-Inverse Document Frequency) is a common weighted technology of information retrieval and data mining. When we extract keywords in a text, we often compute the frequency of each word. The result shows that the high-frequency words often are not the important words, such as "a", "in" and "of" etc. In order to extract those real keywords, we define them as stop words. When we analyze the text, we must ignore these stop words first. There are still a lot of words which can't reflect the importance in the text. For example, in "the University's 50th Anniversary", there are three words "University", "50^{th}", "Anniversary" and their frequencies are the same. In fact, "Anniversary" should be the most important in the text ant it can express the main idea of the article. So there must be a reasonable coefficient to measure the importance of a word. If an infrequent word appears many times in the article, it may be the keyword we are extracting.

From a statistical standpoint, besides the word frequency, we also can assign a weight to each word according to the importance which is called "Inverse Document Frequency" (abbreviated as IDF). Its value is inversely proportional to its frequency in the text.

The TF-IDF value of a word is the product of the "word frequency (TF)" and "Inverse Document Frequency (IDF)". The greater the value of TF-IDF, the more important the word is for the article. So the top-ranked words will be the keywords in the result.

The algorithm is described as follows:

Calculate word frequency.

$$TF = \frac{\text{The number of a word in the article}}{\text{The total number of words in the article}} \tag{1}$$

Calculate IDF. There should be a corpus to simulate the language environment.

$$IDF = \log\left(\frac{\text{The number of documents in corpus}}{\text{Total number of documents containing the word} + 1}\right) \tag{2}$$

Calculate TF-IDF.

$$\text{TF-IDF} = \text{TF*IDF} \tag{3}$$

Obviously, TF-IDF value is proportional to TF and is inversely proportional to the usage frequency in the text. This algorithm outputs the keywords by computing TF-IDF values.

3.2 TextRank Keyword Extraction Algorithm

TextRank algorithm is based on PageRank and it is used to create the abstract and keywords of a text. PageRank is used to calculate the importance of the web page so that web pages can be sorted reasonably. In the algorithm, the Internet is regarded as a digraph and web page is regarded as a node. If a web page is linked by many other web pages, it proves that the web page is popular and it will rank a lot higher than others.

The algorithm is described as follow:

Assume vector B is the rank of n web pages. Matrix A is the number of links between web pages and a_{mn} stands for the link number from the No. m web page to the No. n web page.

$$B = (b_1, b_2, \ldots, b_N)^T \tag{4}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} & \cdots & a_{1M} \\ \cdots & & & & \cdots \\ a_{m1} & & a_{mn} & & a_{mM} \\ \cdots & & & & \cdots \\ a_{M1} & & a_{Mn} & \cdots & a_{MM} \end{bmatrix} \tag{5}$$

Clearly A is the known quantity and B is the unknown quantity. Do, hypothesis is the first iteration results. Suppose $B_1 = A \cdot B_{i-1}$ is the iterative value.

At first, we suppose the rank of each web page is $\frac{1}{N}$, that is:

$$B_0 = \left(\frac{1}{N}, \frac{1}{N}, \cdots, \cdots \frac{1}{N} \right) \tag{6}$$

After the matrix iterative operation for many times, we can compute the values of $B_1, B_2, B_3 \cdots$

3.3 Keyword Extraction Based on Jieba

Jieba is a word segmentation library based on python which provides precise mode, full mode and search engine mode. It scans the words based on the Trie tree structure, find the maximum probability path based on dynamic programming and segment words based on HMM.

Because the system cannot provide a perfect dictionary to identify the word of some names, jieba.load_userdict (file_name) can load user-defined dictionary. The dictionary is a TXT file and each line is divided into three parts: the word, word frequency and part of speech. Another TXT file jieba.analyse.set_stop_words (file_name) can load the stop word and there is one stop word in each line. We can remove a lot of topics-unrelated vocabulary by setting stop word.

Jieba library provides TF-IDF algorithm and TextRank algorithm. TF- IDF keyword extraction technology is easier to implement and calculate much faster than TextRank. It is fit to deal a lot of data if setting enough stop words. On the contrary, TextRank is more accurate than TF- IDF. But the extracting time is longer. So we combine these two algorithms to analyze the source text in our research.

4 Text Tendency Analysis

Text sentiment orientation analysis means getting the subjective information by analyzing the attitude, the point of view or the emotion of the text. The data of content resource evaluation come from articles, community comments and so on. In order to obtain the direction of the social public opinion, we evaluate the number of positive report, negative report and neutral report. In this paper, we apply Naive Bayes classifier to judge the tendency of the text.

4.1 Naive Bayes Classifier

Naive Bayes classifier is based on Bayes' theorem and widely applied to text classification. There are two random events A and B, if A has happened, we can compute the probability of event B as follow:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \tag{7}$$

P(AB) is the probability of events A and B happened at the same time. P(B) is the probability of event

B.

Suppose we have a set of data, random sample C means the sample probability in class C, F_1 stands for the characteristics probability of test sample 1. Bayesian formula is as follows:

$$P(C \mid F_1) = \frac{P(CF_1)}{P(F_1)} = \frac{P(C) \cdot P(F_1 \mid C)}{P(F_1)} \tag{8}$$

Formula (8) means the sample probability in class C when feature F_1 occurs.

If the sample space is divided into N classes and each class is expressed as F_i, the Bayes' theorem can be extended to:

$$P(C \mid F_1, F_2, \cdots F_n) = \frac{P(C) \cdot P(F_1, F_2 \cdots F_n \mid C)}{P(F_1\ F_2 \cdots F_n)} = \frac{P(C) \cdot P(F_2, \cdots F_n \mid CF_1)}{P(F_1\ F_2 \cdots F_n)} = \cdots$$
$$= \frac{P(C) \cdot P(F_1 \mid C) \cdot P(F_2 \mid CF_1) \cdots P(F_n \mid CF_1F_2 \cdots F_{n-1})}{P(F_1\ F_2 \cdots F_n)} \tag{9}$$

It can be simplified as formula (10):

$$P(C) \cdot P(F_1 \mid C) \cdot P(F_2 \mid C) \cdots P(F_n \mid C) \tag{10}$$

When doing text tendency analysis, assume a text is consisted of D words. A is expressed as positive class and B expressed as negative class, then

$$P(A \mid D) = \frac{P(A) \cdot P(D \mid A)}{P(D)} \tag{11}$$

$$P(B \mid D) = \frac{P(B) \cdot P(D \mid B)}{P(D)} \tag{12}$$

The D words is consisted of $d_1, d_2 \ldots d_n$, so

$$P(D \mid A) = P(d_1, \mid d_2 \cdots d_n \mid A) = P(d_1 \mid A) \cdot P(d_2 \mid d_1, A) \cdot P(d_3 \mid d_2, d_1, A) \cdots \tag{13}$$

$$P(D \mid B) = P(d_1, \mid d_2 \cdots d_n \mid B) = P(d_1 \mid B) \cdot P(d_2 \mid d_1, B) \cdot P(d_3 \mid d_2, d_1, B) \cdots \tag{14}$$

Using conditional independence assumption, the emergence of each word is independent of each other, formulas (13) and (14) will be:

$$P(D \mid A) = P(d_1 \mid A) \cdot P(d_2 \mid A) \cdot P(d_3 \mid A) \cdots P(d_n \mid A) \tag{15}$$

$$P(D \mid B) = P(d_1 \mid B) \cdot P(d_2 \mid B) \cdot P(d_3 \mid B) \cdots P(d_n \mid B) \tag{16}$$

$P(d_1 \mid A)$ means the probability of d in the positive sample space. By setting a threshold, we can infer if d_1 is in this class.

4.2 SnowNLP

This paper analyzes the emotion tendency by SnowNLP. SnowNLP is a text analysis library based on python and all the codes is completed by Unicode. It tags the emotion of training set by NLP statistical method and judges the emotion tendency by analyzing the text.

SnowNLP provides basic Naive Bayes classifier and the basic positive and negative libraries. At first, train enough sample data, then extract appropriate classification model and classify the data according to the model, at last, compute the tendency probability. In this paper, the data come from lots of famous video websites. The source codes are:

```
from snownlp import seg
seg.train('positive.txt')
```

```
seg.save('seg.marshal2')
from snownlp import SnowNLP
s = SnowNLP(str1.decode('utf8'))
s.sentiments
```

The method above can get the tendency result which is a probability to the positive sample space and the value is between 0 and 1. If the value is greater than 0.6, we think the text is a positive article and if it belongs to 0.4-0.6, it should be a neutral article. But it is less than 0.4, it perhaps is a negative topic.

5 The Result of Experiments

5.1 The Keyword Extraction

In the experiment, we choose an educational article as a text analysis data. Because the structure of the site is different, in the process of text data fetching, inevitably there are a number of HTML tags information. We use TextRank and TF-IDF to extract 20 keywords. The experiment text data and calculation time are shown in Fig. 6 and Table 1.

```
            html = '''<div class="left_zw" style="position:relative"><p>According to the
voice of China "epicenter wide news" report, released by the China university students' em
ployment entrepreneurship development report shows that there are seven pairs of 2015 coll
ege graduates in employment pleased with the result,</p><p>The the China business developm
ent report of university students' employment is a fudan university released yesterday by
the North-East, the report also shows that 72.28% of the class of 2015 were satisfied with
 the results of the employment of university graduates, the report is the ministry of educ
ation in the development of philosophy and social sciences project through to the national
 more than 20 fresh university graduateBirth and successive nearly 5000 entrepreneurial un
iversity students, which compiled the questionnaire analysis.</p><p>Report shows that 67.9
3% of the fresh college graduates, employment and professional is matching.Training and le
arning good college students' employment and professional matching degree is high.</p><p>,
 says 68.66% of graduates work in accordance with the occupational expectation has a facto
r.Higher occupational expectation inosculation of the first three factors are working loca
tion, interests, hobbies and development prospects.Salary and development prospect are gra
duates value most when choosing the unit, and working stability, work place and professing
 is graduates choose to consider the factors most employment units.<table border=0 cellspa
cing=0 cellpadding=0 align=left style="padding-right:10px"><tr><td><div id=adhzh name=hzh>
<script type="text/javascript" src="http://i8.chinanews.com/gg/yichuanmei/k.js"></script><
/div><!--[4,175,19] published at 2015-08-31 23:40:35 from #10 by zhengzhenhai--></td></tr>
</table></p><p>In higher vocational college graduates to work stability, solve the registe
red permanent residence and professing attention significantly higher than that of other t
ypes of colleges and universities,In addition, on the working stability, professing opinio
ns and family attention sort of specialized subject, undergraduate and graduate students.T
he attention on development prospect of sorting was followed by graduate students, undergr
aduates and students.</p><p>In terms of business report also showed that the 2015 session
of the national college graduates entrepreneurship rate was 2.86%,One of the specialized s
ubject graduates entrepreneurship rate is higher than the other degree students.<journalis
t liule><div id="function_code_page"></div>...
```

Fig. 6. Experiment text data

Table 1. Textrank and TF-IDF methods

Method	Time (s)	Keywords
TextRank	0.029000005722	Graduates, employment, entrepreneurship, reports, college students, colleges and universities, job, choice, professing, development, unit, fresh, factors, the nation, shows that above, specialty, China, professional, financial
TF-IDF	0.00600004196167	Graduates, 2015, obtain employment, colleges and universities, entrepreneurship, professional counterparts, specialized subject, report, this year, college students, work, development prospects, style tr td table id left show that satisfied

The article is about college students' employment situation. By extracting 20 keywords, the time efficiency of TextRank algorithm is almost 5 times of TF-IDF. But from the extraction result, we can see that both of the two algorithms can show the main idea of the text. The top keywords are similar to each other, but TextRank is better than TF-IDF at the latter keyword because TF-IDF can select some HTML as the keyword. In order to avoid extracting HTML tags, we can filter the text by matching tags with regular expression. Another method is making massive changes in the stop word dictionary. This paper adopts both methods to extract the keywords. The experiment text data and results are shown as Fig. 7 and Table 2.

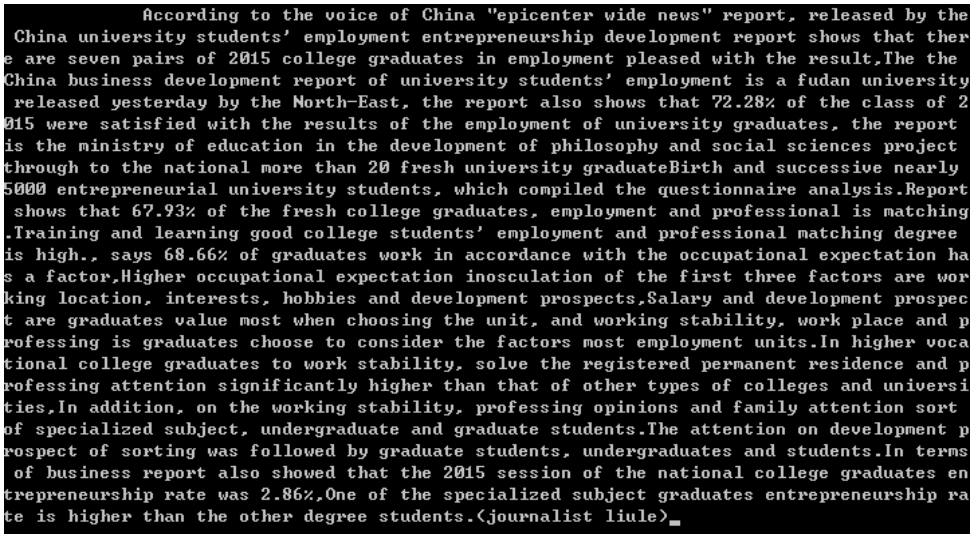
According to the voice of China "epicenter wide news" report, released by the China university students' employment entrepreneurship development report shows that there are seven pairs of 2015 college graduates in employment pleased with the result,The the China business development report of university students' employment is a fudan university released yesterday by the North-East, the report also shows that 72.28% of the class of 2015 were satisfied with the results of the employment of university graduates, the report is the ministry of education in the development of philosophy and social sciences project through to the national more than 20 fresh university graduateBirth and successive nearly 5000 entrepreneurial university students, which compiled the questionnaire analysis.Report shows that 67.93% of the fresh college graduates, employment and professional is matching.Training and learning good college students' employment and professional matching degree is high., says 68.66% of graduates work in accordance with the occupational expectation has a factor,Higher occupational expectation inosculation of the first three factors are working location, interests, hobbies and development prospects,Salary and development prospect are graduates value most when choosing the unit, and working stability, work place and professing is graduates choose to consider the factors most employment units.In higher vocational college graduates to work stability, solve the registered permanent residence and professing attention significantly higher than that of other types of colleges and universities,In addition, on the working stability, professing opinions and family attention sort of specialized subject, undergraduate and graduate students.The attention on development prospect of sorting was followed by graduate students, undergraduates and students.In terms of business report also showed that the 2015 session of the national college graduates entrepreneurship rate was 2.86%,One of the specialized subject graduates entrepreneurship rate is higher than the other degree students.(journalist liule)

Fig. 7. Experiment text data

Table 2. Improved algorithm result

Methods	Time (s)	Keywords
Textrank	0.0159999847412	Graduates, employment, entrepreneurship, reports, college students, colleges and universities, job, choice, professing, development, unit, fresh, factors, the nation, shows that above, specialty, China, professional, financial
TF-IDF	0.00399994850159	Graduates, employment, colleges and universities, 2015, entrepreneurship, professional counterparts, the follows: firstly, the report, the fresh, job, college students, development prospects, show that satisfied, undergraduate, factors, professional, sorting, most, attention

After data filtering, the result of the TF-IDF algorithm is optimized obviously and the processing time is much less than before. Textrank is not affected by filtering data. We take lots of texts as the testing samples in the experiment. The average time of TF-IDF is less a lot and the veracity is lower than Textrank because the total words number is different from each sample.

5.2 Text Sentiment Orientation Analysis

Text sentiment orientation analysis experiment is based on SnowNLP. Before the experiment, we selected 1000 5 star long comments as positive training sample from douban website. The training results are saved as the file of seg.marshal. Modify the point of data_path in the snownlp/seg/_init_.py file and start to analyze the text tendency. We select some 1-5 star long comments to analyze. Suppose the telegraphs are unrelated with each other, we segment the text to many paragraphs and compute the emotional tendency for them. At last, average all the values. The experiment result is shown as Table 3.

Table 3. Douban sentiment orientation analysis

The star	The average results
1 star	0.188779959979
2 star	0.239102686913
3 star	0.425684900104
4 star	0.752354456488
5 star	0.845219565462

The data shows that text sentiment orientation analysis results are reasonable. Because of the training sample data are 5 stars in the douban website, the average values of 1 star and 2 stars are very small and have no obvious difference between each other. But for the training data is not large enough so that 4 star and 5 star data is not large obviously. So more training samples should be selected for websites and we also can train them according to different classes so that the accurateness is improved greatly.

6 Conclusion

This paper determines the research target of the text analysis and realizes the keyword extraction of text content and text tendency analysis. The main work includes three parts:

6.1 Data Acquisition

A general Scrapy crawler based on Scrapy framework is designed. The crawler adopts configuration file to adapt many kinds of websites. It realizes dynamic data acquisition by using Ghost simulated browser and interface request. A Proxy server is developed to achieve the mobile client data acquisition.

6.2 Keyword Extraction

The TF-IDF and Textrank algorithms are analyzed and introduced, then realize the two algorithms using the python jieba text analysis library. By comparing and combining the two algorithms, the Textrank algorithm is applied to the mobile client and TF-IDF for web general spiders.

6.3 Text Tendency Analysis

The Naive Bayes classification algorithm based on SnowNLP is used to judge the long text by segmenting and averaging all the segments. The experiment shows that the results for the educational text analysis are more accurate and fast.

With the rapid development of Internet technology and the exponential growth of network information, the evaluation of net content resources will become the main means of resource evaluation in all fields of life. Especially in the field of educational technology, it is great significant for the evaluation of teaching content, selection of quality of teaching content and recommendation of courses according to the social hot spots.

References

[1] A.V. Arzhakov, D.S. Silnov, New approach to designing an educational automated test generation system based on text analysis, ARPN Journal of Engineering and Applied Sciences 11(5)(2006) 2993-2997.

[2] T. Wei, Y. Lu, H. Chang, Q. Zhou, X. Bao, A semantic approach for text clustering using WordNet and lexical chains, Expert Systems with Applications 42(4)(2015) 2264-2275.

[3] S.K. Dwivedi, C. Arya, Automatic text classification in information retrieval: a survey, in: Proc. International Conference on Information & Communication Technology for Competitive Strategies, 2016.

[4] T.M. Eldos, Arabic text data mining: a root-based hierarchical indexing model, International Journal of Modelling & Simulation 23(2015) 158-166.

[5] A. Kale, M.D. Ingle, SVM based feature extraction for novel class detection from streaming data, International Journal of Computer Applications 110(9)(2015) 1-3.

[6] J. Zhang, B. Wang, H. Tang, L. Tiancai, Unsupervised sentiment orientation analysis on micro-blog based on Biterm topic model, Computer Engineering 2015(7)(2015) 219-223.

[7] R. Kumar, A. Jain, C. Agrawal, Survey of web crawling algorithms, Advances in Vision Computing: An International Journal 2014(9)(2014) 1-8.

[8] M. Rezaei, N. Gali, P. Franti, ClRank: a method for keyword extraction from web pages using clustering and distribution of nouns, in: Proc. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on IEEE, 2016.