Journal of Sociology and Ethnology (2021) 3: 56-60
Clausius Scientific Press, Canada

DOI: 10.23977/jsoce.2021.030109
ISSN 2616-2318

Design and Application of Public Opinion Analysis System Based on Python

Xinyue Wang, Nuo Qun, Liang Yan*, Weiheng Liang, Ran Wang, Weicheng Gu

Tibet University, Lhasa Tibet, 850000

*Corresponding author

Keywords: Post Bar, Network public opinion, Public opinion monitoring, Public opinion analysis

Abstract: The Internet has penetrated into every aspect of our lives, but freedom of speech does not mean that we can be outspoken on the Internet. Tibet is the border area of our country, so speech security is particularly important. This paper mainly takes the post bar of Tibet University as an example to complete the public opinion analysis system, and uses Python crawler to crawl all kinds of posts of the post bar of Tibet University in real time, set keywords, quickly and effectively screen out all kinds of sensitive information, so as to better maintain the speech security of the post bar of Tibet University.

1. Introduction

With the advent of the era of artificial intelligence, the rapid development of the network has brought great convenience to the society, but because of its openness, the network has become a place outside the law for some people, and pornographic, network violence and other remarks are full of various platforms. Strengthening the supervision of the network and actively promoting the construction of network rule of law is in progress, and the monitoring of network public opinion is also the top priority.

This paper takes the post bar of Tibet University as an example to carry out the analysis and research of network public opinion. It mainly focuses on the information published in the post bar of Tibet University, including text information and picture information. It uses the web crawler to crawl in real time, extract the post bar information, master the hot current events and emotional tendencies in the post bar, and then understand whether there are some negative information in the post bar of Tibet University, so as to quickly make some corresponding measures.

2. Design of Analysis System

The public opinion analysis system based on Python is decomposed, as shown in Figure 1. The system is mainly composed of two parts, namely public opinion monitoring and public opinion

analysis. Among them, public opinion monitoring is the most basic part, which is the foundation of the system. It includes crawler real-time crawling data and data storage; Public opinion analysis is the core part of the system. This part needs to edit (obtain) keywords according to user needs, and then make a comparative analysis between keywords and local data. Cosine similarity is used to calculate the similarity between keywords and data, and the data with high similarity with keywords is extracted to generate reports by the system.

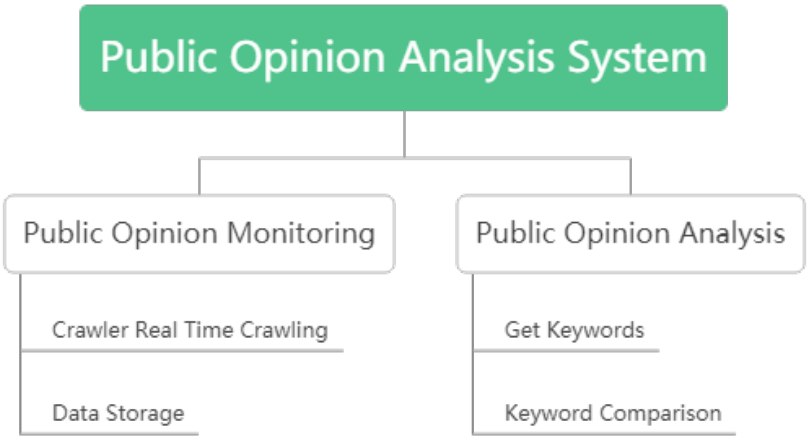

Figure 1: Breakdown of Public Opinion Analysis System

2.1 Public Opinion Monitoring

2.1.1 Real Time Crawling Post Bar Information

The most primitive Python crawler starts by getting the URL of the original website from the URL of one or more start sites. In the process of capturing the required websites, they always extract the latest URLs from the running pages, and put these URLs in the queue until the program meets some conditions of the system. According to a certain algorithm to filter and delete the pages with low correlation, only keep the links of the useful pages, put them in the queue, and repeat the above processing until there is a specific stop condition, the crawler will stop crawling. Generally speaking, a crawler is to simulate a browser to request a site and store the data returned by the site locally for extraction [1].

Scrapy crawler framework is an application framework for crawling website data and collecting structure data. It can be used in many fields, including data collection, information processing or historical data preservation. Through this framework, we can capture the content or image of a given website [2].

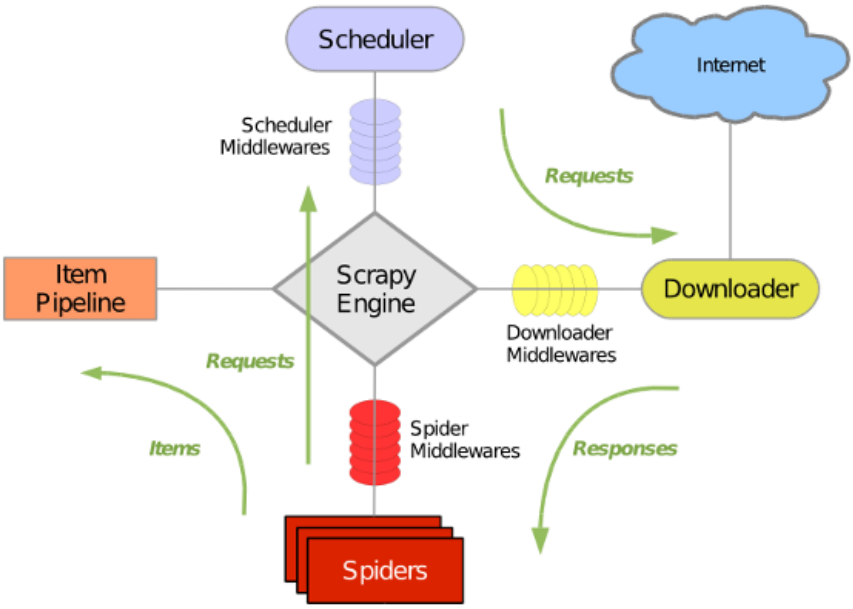

Figure 2: Scrapy Framework

In Python crawler real-time crawling, it mainly includes the following four steps.

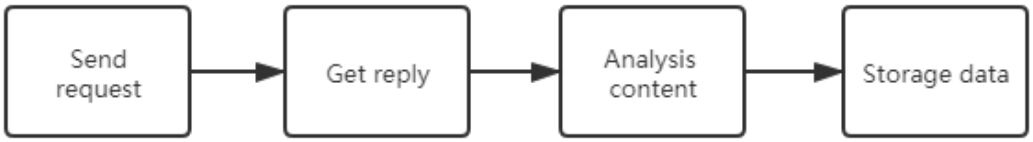

Figure 3: Crawler Real Time Crawling Flow Chart

a. Send request: use HTTP library to send the request to the target page, that is, send a request, which includes request header, request body, etc.

b. Get reply: if the requested content exists on the target server, the server will return the requested content. Requests include: HTML, JSON string, image, video, etc.

c. Analysis content: for users, this is to find the information they need. For Python crawlers, regular expressions or other libraries are used to extract target information.

d. Storage data: the read-in data can be stored locally in different forms such as text, audio and video.

2.1.2 Data Storage

In terms of database, this system uses MySQL database, which is created before Python crawls the post bar. Create an xzdx database in mysql, use pymysql library to operate mysql, connect to the created xzdx database, and create a table in this database, named Tibet, to store the crawled post bar data. There are eight fields in this table: id，title，author，author_id，content，reply_time，floor，content_pho.

2.2 Public Opinion Analysis

2.2.1 Get Keywords

In this system, the generation of keywords is manually set according to the monitoring requirements, using "and" (&), "or" (|), "not" (!), "and" (|) Brackets () and other logical symbols realize the connection between keywords. When monitoring whether there is sensitive information in the post bar of Tibet University, The keywords can be set to ((Tibet|…|Gaize County | Coqin county) [3] & (Dalai | Buddhism | Tibetan | Han | religion | contact information)). The first part includes the Tibet Autonomous Region itself and its 72 districts and counties, which belong to the place determiner; the latter part is the sensitive event vocabulary, which belongs to the specific event determiner. No matter the location qualifier or event qualifier, they can be modified, added or deleted according to the user's needs.

2.2.2 Data Keyword Comparison

Before obtaining keywords, the system will also process the data stored in the local database for Chinese word segmentation. This system uses Jieba Chinese word segmentation method for word segmentation. The biggest advantage of this method is that it has a thesaurus named dict.txt, which contains the number of entries and part of speech. It is the result of Python developers' training based on the corpus resources of people's daily [4].

After the completion of Chinese word segmentation, the system compares the similarity between the segmentation results and keywords. The most commonly used similarity measurement methods are minimum edit distance, Euclidean distance, cosine distance, jacquard similarity, etc. [5]. The measurement method used in this system is cosine distance method. Cosine distance method is derived from the cosine value of the angle between two vectors in space in mathematics, and the formula is as follows:

$$\cos(\alpha) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

Cosine similarity is to virtualize two individuals into two vectors in the space, and measure the difference between two individuals by comparing the cosine value of the angle between two vectors in the vector space. Let 1 be the same and 0 be different, then the similarity value is between 0 and 1, and the similarity range of all things should be between 0 and 1. If the angle between the cosines of the vectors fitted by two individuals is closer to 1, then the similarity between the two individuals is higher. The basic idea is: if the word sentence similarity of two sentences is higher, the angle between their transformed vectors should be smaller, and the cosine angle should be closer to 1, then the cosine similarity is higher, so we can calculate the similarity of two sentences from the word frequency [6]. When the system searches the sensitive information of the post bar, it can set a certain threshold according to the user's needs. Only when the similarity exceeds a certain threshold, it will be displayed to the post bar administrator for filtering.

3. Conclusion

The design of public opinion analysis system based on Python is to facilitate the daily management of the post bar of Tibet University. The system is divided into two modules: public opinion monitoring and public opinion analysis. Due to my limited professional ability, too large framework of public opinion analysis module and too much workload, it needs team cooperation to complete, so only the public opinion monitoring module has been implemented, Combined with the existing public opinion analysis system on the market, this system needs to conduct in-depth research on the following issues in the later stage. For example: How to use page display instead of Excel after crawling post bar? What is the bridge from database to page? In addition to the directional crawling information of the post bar of Tibet University, can the later research crawl the whole network? For example, wechat, microblog, forum, blog, news, website, client, how to realize these? How to export after crawling information? What is the format of the exported file? Can the system generate keywords automatically? In the follow-up research work, we will combine the current popular artificial intelligence technology to make the data better serve the ideological and political education of colleges and universities.

References

[1] Li Wenhua. Design and implementation analysis of Python based web crawler system [J]. Neijiang science and technology, 2021, 42 (02): 58-59 + 26

[2] Zou Wei, Li Tingyuan. Crawling domain website files based on scrapy crawler framework [J]. Modern information technology, 2020, 4 (21): 6-9

[3] The Yearbook is selected from Tibet Statistical Yearbook (2013). Edited by Zhao Rui, China Local Chronicles Yearbook, 2014215, Yearbook

[4] Zeng Xiaoqin. Implementation of Chinese stuttering segmentation technology based on Python [J]. Information and computer (theoretical Edition), 2019, 31 (18): 38-39 + 42

[5] Zhao Zhijing, Jiang Di. Research on language classification based on editing distance [J]. Language research, 2020, 40 (02): 43-50

[6] Wu Sen, Gao Xiaonan, he Huixia. Topic discovery algorithm based on bidirectional improved cosine similarity [J]. Operations research and management, 2021, 30 (02): 75-83