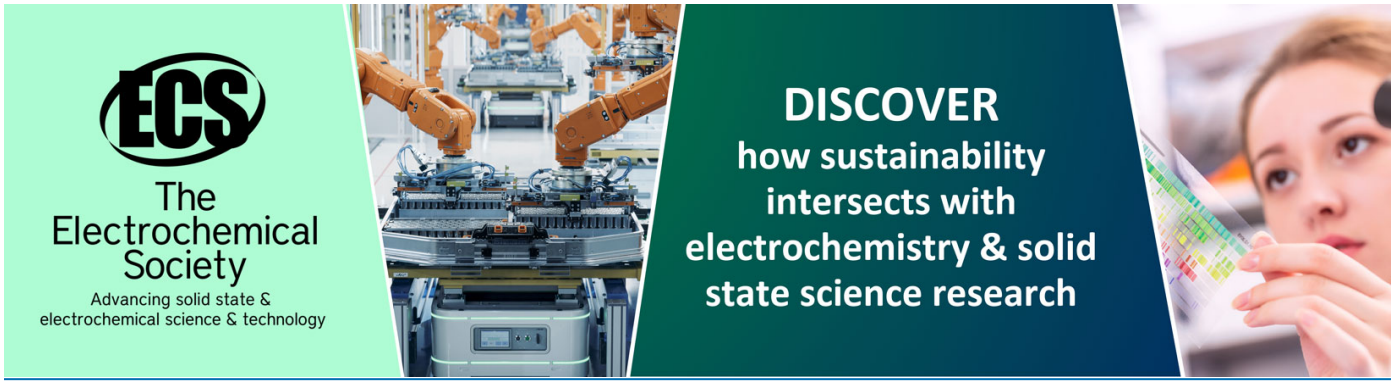
Python-based film review data acquisition and visualization design



Python-based film review data acquisition and visualization design

Yong Yang, Ying Xin Liu*, Yu Xi Zhang and Na Zhang

Shenyang University, 110044, China

*Corresponding author's Email: 327373832@qq.com

Abstract. With the rapid development of the Internet and artificial intelligence era, how to obtain effective information is particularly important in the complex network world. In order to obtain a real movie viewing experience and provide some reference for other users, this article is based on the Python language, taking the popular movie "Shuimen bridge of Changjin Lake" as an example on Douban website, and using web crawler technology to comment and rate users, etc. Relevant data is crawled, the crawled data is the object of analysis, and data visualization processing is carried out to more intuitively show the true feelings of moviegoers. The results show that the film meets the expectations of the public and is worthy of recommendation.

1. Introduction

The domestic film market is developing rapidly. China has become the second largest film market in the world, and the gap with the North American film market is also shrinking. The domestic film industry is booming [1]. Nowadays, audiences pay more attention to choosing movies that meet their own preferences when watching movies, rather than choosing movies that simply attract people's attention and use traffic to obviously earn box office. As the largest movie sharing and commenting community in China, Douban Movies can provide the latest movie introductions and movie reviews, and recommend movies to users based on their personal preferences. At the same time, users can also choose whether to watch the movie according to the movie's rating and the feedback from the reviews. It is very important to provide the correct orientation for the viewers [2].

Today, the World Wide Web has become an effective carrier of a large amount of information, and how to extract and utilize effective information data has also become a challenge [3]. Using the web crawler technology, we can easily obtain the data of the target web page through the legal channels allowed by the website, so as to analyze and process the obtained valid data and dig out the value behind the data. Based on Python language, this paper uses web crawler technology to crawl the review data of the movie "Shuimen bridge of Changjin Lake", and uses visualization tools to more intuitively show the theme color of the movie and the reality of the movie.

2. Web crawler

A web crawler is a program or script that can automatically crawl web page information according to specified rules. Simply put, a web crawler is a program that simulates the behavior of a user sending a request to a website [4]. Web crawlers are classified into four categories according to their functions: general web crawlers, focused web crawlers, incremental web crawlers, and deep web crawlers.

(1) Universal web crawler

Universal crawlers are also called full-web crawlers. The crawling objects are expanded from some seed URLs to the entire Web. Universal crawlers are one of the important components of search engines, such as (Baidu, Google, etc.). It is mainly to download webpages on the Internet to the local area to form a mirror backup of Internet content.

(2) Focus on web crawlers

Focused web crawler, also known as "topic web crawler", is a kind of web crawler oriented to specific needs. The difference between it and the general crawler is: when the target website is crawled, the focused crawler will perform content processing on the content. Filtering and processing to ensure that only data information relevant to the needs is captured. This article is based on the crawler technology to crawl data on Douban Movie Network for specific needs, such as user name, time, rating, short comment and other basic information.

(3) Incremental web crawler

Incremental web crawler refers to incrementally updating the downloaded web pages in the downloaded website, or only crawling newly generated crawlers of changed types, which can ensure that the crawled web pages are as new as possible.

(4) Deep web crawler

Web pages can be divided into surface pages and deep pages according to some methods. The surface web page is the web page that traditional search engines can request, and it is composed of static types of web pages that can be reached by hyperlinks. Deep web pages are obtained through static links, hidden behind the search form, and only the keywords provided by the user can get relevant data. For example, after registering, users get more "deep" pages.

3. Related technology

(1) Requests is a native network request-based module in Python. It is powerful, simple, convenient, and efficient. It can simulate browser requests and obtain web page data.

(2) Xpath is the main method for web page data analysis. After obtaining the target web page data through requests, Xpath is used to obtain the data specified by itself. Compared with bs4, it provides a very concise path expression.

(3) Wordcloud is an excellent third-party library for displaying word clouds. Word clouds take words as the basic unit. When the word frequency is higher, the displayed font will be larger, and the text can be displayed more intuitively.

(4) Jieba, which can separate a Chinese text into Chinese word sequences. The Jieba library supports 3 word segmentation modes: exact mode, full mode, and search engine mode. This paper mainly uses jieba.lcut() precise word segmentation mode to analyze the processed text data.

(5) Matplotlib is one of the commonly used visualization tools in Python. It can easily create 2D charts and some basic 3D charts. You can define x and y axes according to the data set, and draw graphics (line charts, histograms, histograms) , density plots, scatter plots, etc.), can solve most needs.

3.1. Algorithm design

This article takes the Douban movie "Shuimen bridge of Changjin Lake" as an example, and designs the algorithm based on the Python language. The main process can be divided into three parts: Data acquisition, Data processing, Data visualization, the detailed design process is shown in Figure 1.

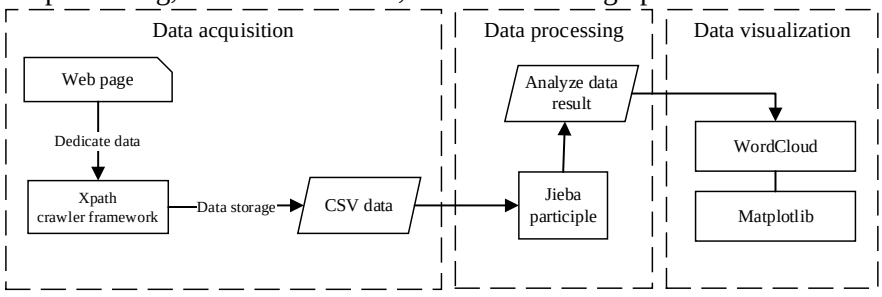


Figure 1. Design flow chart.

4. Algorithm implementation

4.1. Requests data acquisition

Use Google Chrome browser to open Douban Movies website, search for "Water Gate Bridge of Changjin Lake", enter the target website https://movie.douban.com/subject/35613853/, and click all comments. Right-click the webpage and click Check to enter the developer mode to facilitate the analysis of the page. Click Network to view Fetch/XHR, refresh the current page, and view the parameter information in Headers. The request method is GET.

```
import requests
headers = {"User-Agent":"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.60 Safari/537.36"}
for page in range(0,20):
url = f'https://movie.douban.com/subject/35613853/comments?start={page*20}&limit=20&status=P&sort=new_score'
    resp = requests.get(url=url,headers=headers,cookies=cookies)
    resp.encoding = "utf-8"
```

4.2. Anti-climbing solution

Nowadays, many websites have set up anti-crawling methods for web crawlers, and the first principle of building web crawlers is: "So all information is "forged" [5].

(1) Set the request header file Headers. The request header file is similar to the visitor's ID card when visiting the website. If it is not set, it will be blocked by the server as a robot and cannot be accessed. Obtain User-Agent through Google Chrome browser developer tools to simulate user login for disguise.

(2) Setting Cookies. In Douban.com, if you do not set cookies to disguise, you can only crawl a small amount of web page data and you will be banned from visiting. To obtain cookies, you need to register and log in to the Douban.com account. Each account user has its own unique cookies. Cookies are also an important way to identify users.

(3) Set up delayed access. When using a program to crawl web page data, it is often easy to be monitored by the website due to the excessive access speed, and access is prohibited. To simulate a real user, you can use sleep() to set the crawling speed of the crawler.

4.3. Xpath parsing web page

On the movie review page, according to the text collected from the data, it is parsed and captured through Xpath. The key code is implemented as follows:

```
from lxml import etree
et = etree.HTML(resp.text)
divs = et.xpath("//div[@class='mod-bd']/div")
for div in divs:
    name=div.xpath(""./div[2]/h3/span[2]/a/text()")
    comment=div.xpath("./div[2]/p/span[@class='short']/text()")
    score=div.xpath("./div[2]/h3/span[2]/span[@class='allstar50 rating']/@title")
    time=div.xpath("./div[2]/h3/span[2]/span[@class='comment-time']/text()")
```

4.4. Data storage

Operate through the built-in CSV file in Python, and write the acquired data into the CSV file to realize data preservation. The part of the crawled data stored in the CSV file is shown in Table 1, where the header part includes user nickname: "Name", user comment: "Comment", rating: "Score". However, due to the large amount of table data, only the first few words of the user comments are displayed.

```
f = open("Movie_review_data.csv"，mode='w'，encoding="utf-8")
    f.write(f"{name}\n")
    f.write(f"{comment}\n")
    f.write(f"{score}\n")
f.close()
```

Table 1. CSV section data table.

Name	Comment	Score
fog	Movies worth recommending	Highly recommend
Li's six-year-old baby kiss	The plot is very good	Highly recommend
Singular passerby armor	generation history	recommend
remembrance	The protagonist is hot	good
ANNa	in this era	good
Zhang 3'er	two hours of torment	very poor
over six feet	It is said that the upper	good

4.5. *Jieba word segmentationgood*

Import Movie_review_data.csv file, use the jieba.lcut() function to segment the comment phrases in the document, and for some important keywords, you can use the jieba.add_word() function to add new words to the word segmentation dictionary to prevent participle.

```
import jieba
text = open("../../Movie_review_data.csv","rb").read()
jieba.add_word("watergate bridge")
jieba.add_word("Changjin Lake")
wordlist = jieba.cut(text,cut_all=True)
wl = " ".join(wordlist)
```

4.6. Data visualization

4.6.1. Word cloud

Use wordcloud and imageio to combine graphic pictures to draw word frequency statistics word cloud map, and draw word cloud map with heart-shaped photos as background graphics to show viewers' evaluation of the movie. The word cloud drawing result is shown in Figure 2.

```
wc = WordCloud(background_color = "white",
    mask = imread('heart_background.jpg')，
    max_words = 200, stopwords = ["this"],
    font_path = "C:\Windows\Fonts\simkai.ttf",
    min_font_size = 20,max_font_size = 60,random_state = 10)
myword = wc.generate(wl)
wc.to_file('heart_chart.jpg')
plt.imshow(myword)
plt.axis("off")
plt.show()
```

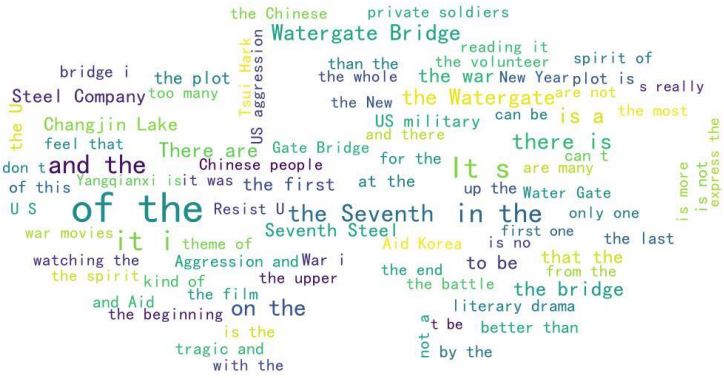


Figure 2. Short comment word cloud.

4.6.2. Matplotlib pie chart

Use pie() in Matplotlib to draw a pie chart to describe the proportional relationship between values. Extract the score column in the CSV file, and use the list comprehension formula to count the occurrences of non-null values in the score column. The proportion of the scoring level corresponding to the arc length of each sector in the pie chart. The parameter autopct="%1.2f%%" is passed in the pie() method to automatically calculate the percentage, the distance of each sector from the center of the explode, which is used to achieve image optimization and save the pie chart as a JPG file format. The pie chart drawing result is shown in Figure 3.

```
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.pie(nums,explode=explode,labels=labels,autopct="%1.2f%%")
plt.title("Rating scale distribution")
plt.savefig(r"C:\Users\Administrator\Pictures\Saved Pictures\01.jpg")
plt.show()
```

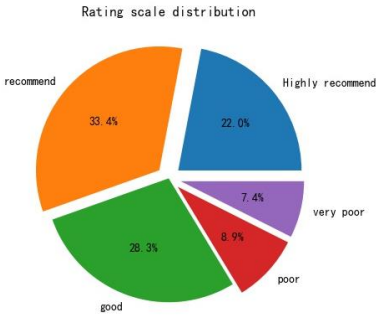


Figure 3. Rating scale pie chart.

5. Conclusion

Based on the Python language, this paper takes the Douban movie "Shuimen bridge of Changjin Lake" as the crawling target, realizes the crawling and analysis of the film review data, and displays it through data visualization, and excavates the key information behind the film review. Filmmakers and the film industry study audience preferences for reference. Through the implementation of this case crawling, the shortcomings of insufficient crawling breadth have been found. The next step will be based on the research of this paper, increase the crawling breadth, analyze the characteristics of the audience, and achieve a more comprehensive analysis.

Acknowledgments

This research was sponsored by the National Natural Science Foundation of China (No. 71601126) and the Natural Science Foundation of Liaoning Province of China (No. 20180550423).

References

[1] Yuxuan, X., Xiaodong, W. (2021) Research on sentiment analysis of movie reviews based on text mining [J]. *Journal of Mudanjiang Normal University (Natural Science Edition)*, **47**: 25-28.
[2] Yufei, G., Hongxia, M. (2020) Data collection and analysis of Douban film and television short reviews based on Python [J]. *Modern Information Technology*, **4**: 10-12+16.
[3] Jianhua, J., Jinsong, J. (2018) Design and implementation of a search engine based on crawler-focused [J]. *System Simulation Technology*, **14**: 221-226.
[4] Yang, Y. (2018) Design and implementation of web crawler based on Scrapy [J]. *Computer programming skills and maintenance*, **25**: 19-21+58.
[5] Pei, L. (2019) Research on Python-based web crawler and anti-crawler technology [J]. *Computer and Digital Engineering*, **47**: 1415-1420+1496.