CS practice project report template

Project Name: Analysis on movie comment by using python web spider and text mining

Your Name: Julian Ye

# Problem definition

## 1.1 What is the project about? (What is it? Why you want to do it? Possible applications)

The project aims to find a quick and effective way to analyze the comments on web. From the analysis results one can easily find the favorites of people. These results can not only be guidance for the new product, but also a good reference for the new customer. The websites for reviewing movies, like douban.com, are good sample, because there are always thousands of people give their comments after watching the movie. Therefore, this kind of technique is much important for both the commercial activity and the social science research as well.

## 1.2 What are the success criteria? (What are your goals that will guide you through design and implementation and can be evaluated)

1. The program can fetch the desired web pages in large amount continuously.

2. The program can make analysis on the content of the pages, and save the data in file, eg. in excel format.

3. The program can illustrate the analyses results in plots or figures.

# Design overview

How to solve the problem in part 1? You need to design your software system, design about the input, process, and output parts.

Please list the Identifier tables if needed.

INPUT: web site, including a series of web pages

Nezha was the most successful movie this summer. The comment pages on this movie in the Doban were selected to test the program.

PROCESS

Describe your key algorithms in flowchart or pseudocode.

Key algorithm 1: web spider

Using packages: urllib

Request a web page

Using the BeautifulSoup to analyze the web page

Obtain the interested element, like comments, time, and rating, etc

Key algorithm2: Saving the results data in excel format

Using packages: pandas

Combining the data from a serious web pages

Convering the data into a data structure

Using Pandas to write the data structure into a excel file, and making further analyses in Excel

Key algorithm 3: analyses of the words frequency

Using packages: Jieba, wordcloud

Separating the comments into words

Making statistical analyses of the word frequecy

Making statistical analyses of the word frequency and building wordcloud

Key algorithm 4: data visualization

Using packages: matplotlib

Showing variation of rating with time

Showing the wordcloud

OUTPUT:

1. Excel file of the data
2. Plot of the data

# Implementation

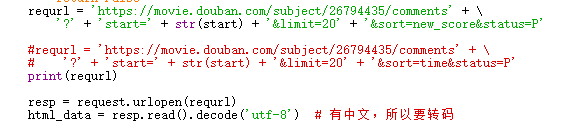
List the key techniques used in your software, like arrays, searching, sorting, dictionaries, python modules (third-party, like matplotlib, numpy, scipy, sklearn etc.)

Tech #1:

Why you use tech 1 in your project?

The package urllib provides way to fetch a web page as text.

How do you use it? Put your code screenshots here.



Tech #2:

Why you use tech 2 in your project?

The package pandas provides strong functions to operate excel file.

With pandas, the data can be written in excel format.

How do you use it? Put your code screenshots here.



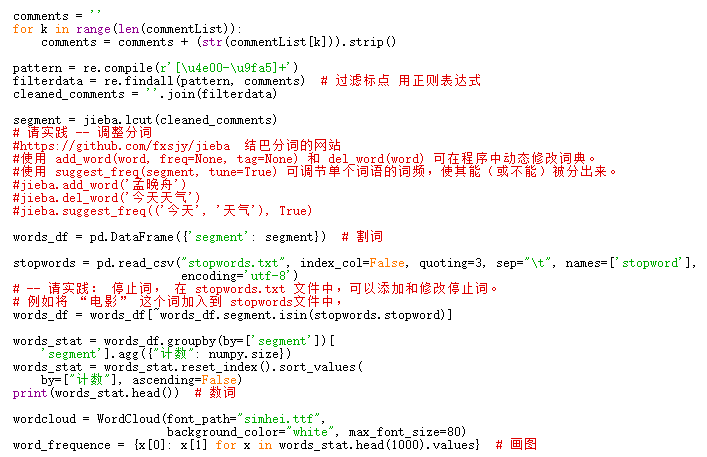
Tech #3:

Why you use tech 3 in your project?

The package Jieba helps to separate the text into words excluding the words with no specific meaning.

The package wordcloud can bulid a word cloud according to frequency of the word in a text.

How do you use it? Put your code screenshots here.

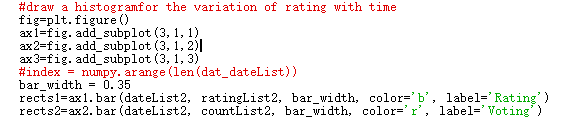


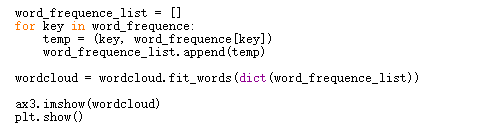
Tech #4:

Why you use tech 4 in your project?

The package matplotlib can visualize the data into plot or figure.

How do you use it? Put your code screenshots here.



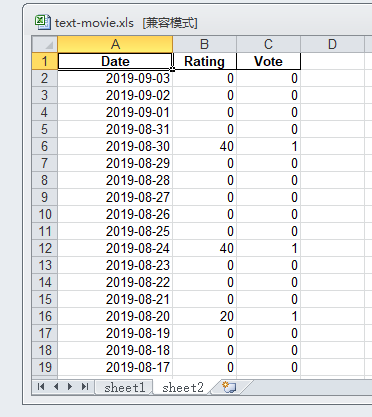
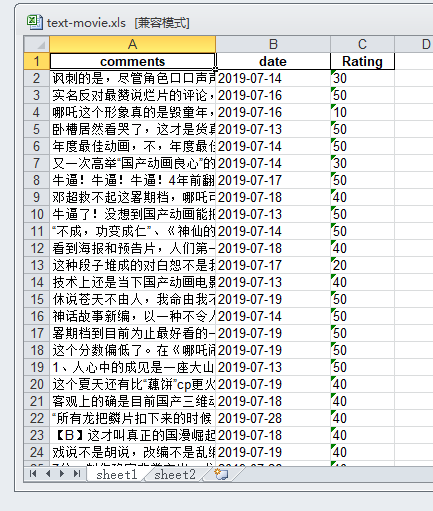


# Results

Show the outcomes/results of your projects

Twenty pages including 200 comments were fetched by the program and analyzed

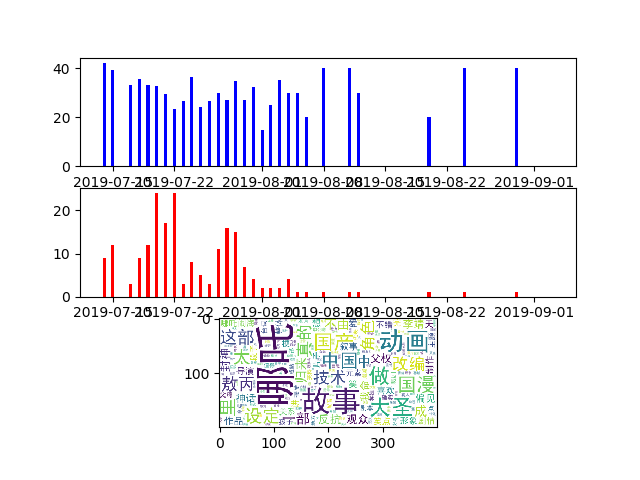
1. An excel file with two worksheets was created. Sheet1 contains the comments, dates and rating of the movie. Sheet2 contains the accumulate voting number and the average rating. The highest rating score is 50.



1. Visual illustration of the data

It can be seen that most comments were issued in July and August when the movie was on the screen. The rating was high at beginning in the first week, and decreased a little.

The word cloud show the favorite subject from this movie. It is interesting that another movie, the monkey king was often mentioned.



# Evaluation and conclusion

Evaluate the project based on your success criteria in Part 1.

All the criterial have been reached. The project was successful.

Some further analyses can be made.

The program should include a login procedure, so that it can download more pages at one time.

What is the conclusion of the project?

The web page can be analyzed by using web spider and text mining. The results can be a guidance for both the company and customer. The analyses results show that people like this movie because of the good character, good tale, good technique, and good combination of Chinese elements.

# References

List all the materials you referred to.

1. Douban, <https://movie.douban.com/subject/26794435/comments?start=0&limit=20&sort=new_score&status=P>
2. python爬虫实战一：分析豆瓣中最新电影的影评, <https://segmentfault.com/a/1190000010473819>
3. 用pandas读取excel并画图展示, <https://blog.csdn.net/WPP1989/article/details/82220209>
4. Python 绘图，我只用 Matplotlib（三）—— 柱状图, <https://zhuanlan.zhihu.com/p/47679593>
5. python 多图绘制, <https://blog.csdn.net/qq_41455420/article/details/79729284>