

The Analysis and Prediction of New Avengers

This project is a data analysis project including initial data scraping from Official APIs, BeautifulSoup Package and JSON data format, data cleaning through Python Script and Excel, data analysis through statistics & Entropy method and final process of data visualization. We used five data sources to analyze and figure out which character will be the next generation's Avengers.

1 Installation

1.1 Python Environment

Download the python version on <https://www.python.org/downloads/>, which is the official website of python. Our group used the version of Python3.7.

1.2 Additional Python Modules

The implementation of this example code uses a number of third-party modules, the main modules are as follows:

- Jieba. It is the most widely used Chinese word segmentation component. Official address: <https://pypi.python.org/pypi/jieba/> Use command `pip install jieba/pip3 install jieba` to install the package.
- Translate. Translate is a simple but powerful translation tool written in python with support for multiple translation providers. Official address: <https://pypi.org/project/translate/> Use command `pip install translate/pip3 install translate` to install the package
- Wordcloud. A small word cloud generator in Python. Official website: <https://pypi.org/project/wordcloud/> Use command `pip install wordcloud/pip3 install wordcloud` to install the package. This package depends on the `numpy` and `pillow`.

2 Data Acquisition

2.1 Obtaining an API Key

We have registered an API Key from official Marvel API website.

(public_key="d39dc67a9bbc2ff4c0505a7cd0e1ae87",private_key="68b08096d3088751561f83cbac4ad08e843b6bab")

Or the users can register their own API keys on website of

<https://www.marvel.com/signin?referer=https%3A%2F%2Fdeveloper.marvel.com%2Faccount> and replace the keys in the module of Marvelapi.py.

2.2 Obtain Comments Data from Maoyan.com

We have scrapped comments for 12 movies. Each movie has a unique ID in the Maoyan Movie website. To save time, we include one comments' data scraping program in the main program as an example. Other comments data can be obtained by modifying the ID. For example, the website of Deadpool1 is

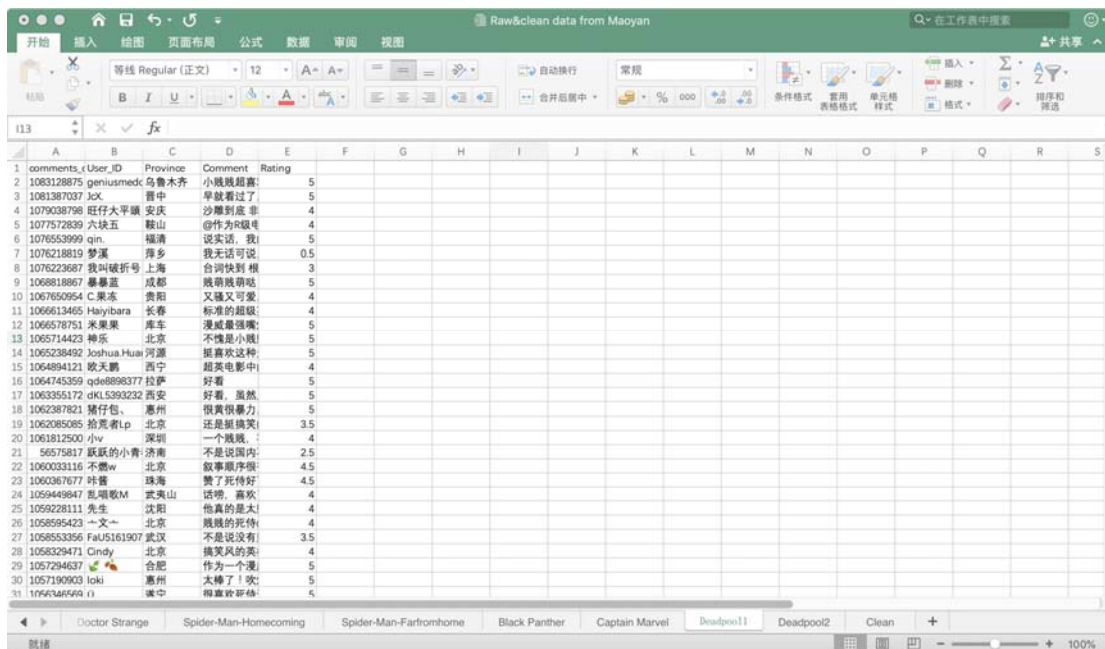
<http://m.maoyan.com/mmdb/comments/movie/246127.json> and **246127** is the ID for the Deadpool1. There are 3 places where we use the ID. They locate in **line 20,39,49**.

This chart shows IDs for 12 movies we used.

Movie Name	Unique ID
Avengers: Endgame	248172
Black Panther	341138
Doctor Strange	246124
Deadpool1	246127

Deadpool2	345808
Ant-Man	78392
Ant-Man and the Wasp	343208
Spider-Man:Far from Home	1198925
Spider-Man:Homecoming	334620
Guardians of the Galaxy	78336
Guardians of the Galaxy: Vol.2	248683
Captain Marvel	341139

We have stored ALL the data of comments results in **comment.xlsx**. One of the results is shown below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	comments	User_ID	Province	Comment	Rating														
2	1063128875	geniusmedc	乌鲁木齐	小贱贱超赞	5														
3	1081387037	JK	晋中	早就看过了	5														
4	1079038798	旺仔大平头	安庆	沙雕到底	4														
5	1077572839	六块五	鞍山	@作为R级电	4														
6	1076553999	qin	福清	说实话, 我	5														
7	1076218819	梦溪	萍乡	我无话可说	0.5														
8	1076223687	我叫破折号	上海	台词快到根	3														
9	1068818867	暴暴蓝	成都	贱萌贱萌哒	5														
10	1067650954	C果冻	贵阳	又骚又可爱	4														
11	1066613465	Hayibara	长春	标准的超级	4														
12	1066578751	米果果	库车	漫威最强嘴	5														
13	1065714423	神乐	北京	不愧是小孩	5														
14	1065238492	Joshua.Huai	河源	挺喜欢这种	5														
15	1064894121	欧天鹅	西宁	超英电影中	4														
16	1064745359	qde8898377	拉萨	好看	5														
17	1063355172	dKL5393232	西安	好看, 虽然	5														
18	1062387821	猪仔包	惠州	很赞很暴力	5														
19	1062085085	拾荒者Lp	北京	还是挺搞笑	3.5														
20	1061812500	小v	深圳	一个贱贱	4														
21	56575817	威威的小青	济南	不是说国内	2.5														
22	1060033116	不赞w	北京	叙事顺序很	4.5														
23	1060367677	吐露	珠海	赞了死侍好	4.5														
24	1059449847	乱唱歌M	武夷山	诶呀, 喜欢	4														
25	1059228111	先生	沈阳	他真的是太	4														
26	1058595423	→文←	北京	贱贱的死侍	4														
27	1058553356	FaU5161907	武汉	不是说没有	3.5														
28	1058329471	Cindy	北京	搞笑风的奥	4														
29	1057294637		合肥	作为一个漫	5														
30	1057190903	loki	惠州	太棒了! 实	5														
31	1056346569	()	汉中	很喜欢这种	5														

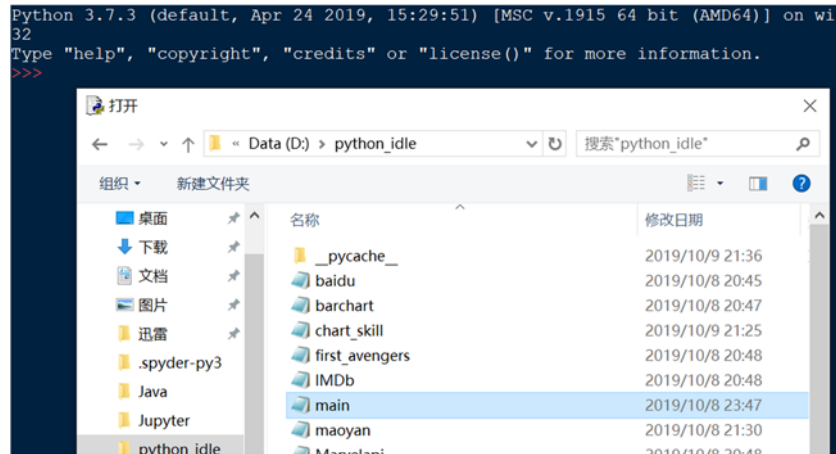
Comment Data of Deadpool1

Note:

Some processes of data scraping may take a long time, so we put our previously-scraped copies of data into the folder.

3 Running

Note: Our application does not have an error-handling function for inputs other than numbers. So please only enter a number for each input when you run the main method. Our product's functions and interface is based on the mind map we drew, which is put in the folder.(mindmap.png)



Open IDLE - File - Open - Open the main module, the main module will pop up and you can click the "Run - Run Module". Our main interface will then be shown. You can enter the option number to choose either data scraping or analysis.

```
WHO WILL BE THE NEXT AVENGERS?
Enter '1' to start Data Scraping
Enter '2' to start Multi-dimentional Character Analysis
Enter '3' to start New Avengers Prediction
Enter '0' to exit.
Please enter the number: |
```

- 1) In command line, when you enter '1', five different options of data scraping will be shown and you can choose one of them to scrape.

```
Enter '1' to scrape Marvel API
Enter '2' to scrape IMDb
Enter '3' to scrape Maoyan
Enter '4' to scrape Power Grip
Enter '5' to scrape Relationships
Please enter the number: |
```

a)Input '1': Scrape data from MarvelAPI

Note 1: Issue of Marvel internal server. It takes a long time to scrape the data. Also due to the issues of internal server error of Marvel, it may need to try many times to get the whole data.

Note 2: Go Back to main menu automatically. Every time a scraping process finishes, it will go back to the initial command line. So please input '1' again to start another option of scraping.

Note 3: Files of data scraping. If the scraping process succeeds, outputs will be saved to the default file of IDLE. Names of the output files are:

1. Marvel API: marvel_characters.csv
2. IMDb: Result_IMDb.csv
3. Maoyan: comments.txt
4. Power Grip: baidu.csv

5. Relationships: message.csv,
names_message.csv,
relation_message.csv

```
Enter '1' to scrape Marvel API
Enter '2' to scrape IMDB
Enter '3' to scrape Maoyan
Enter '4' to scrape Power Grip
Enter '5' to scrape Relationships
Please enter the number: 1
0
1
2
3
4
5
6
7
8
9
10
11
12
Due to the issues of internal server error of Marvel API, please try again later
.
Go back to MENU.

      WHO WILL BE THE NEXT AVENGERS?

Enter '1' to start Data Scraping
Enter '2' to start Multi-dimentional Character Analysis
Enter '3' to start New Avengers Prediction
Enter '0' to exit.
```

b)Input '1'->'2': Scrape data from IMDB

c) Due to the issue of encoding, '1'->'3' will not work here. Scrape data from maoyan

Note 4: Issue of encoding for IDLE. When we scrap the data from maoyan.com, the IDLE has something wrong with the encoding of Chinese characters. So we need to use other IDEs (e.g.PyCharm) to scrape the data from maoyan. In the screen recording we used Pycharm to show that our function works well.

d)Input '1'->'4': Scrape data from Baidu

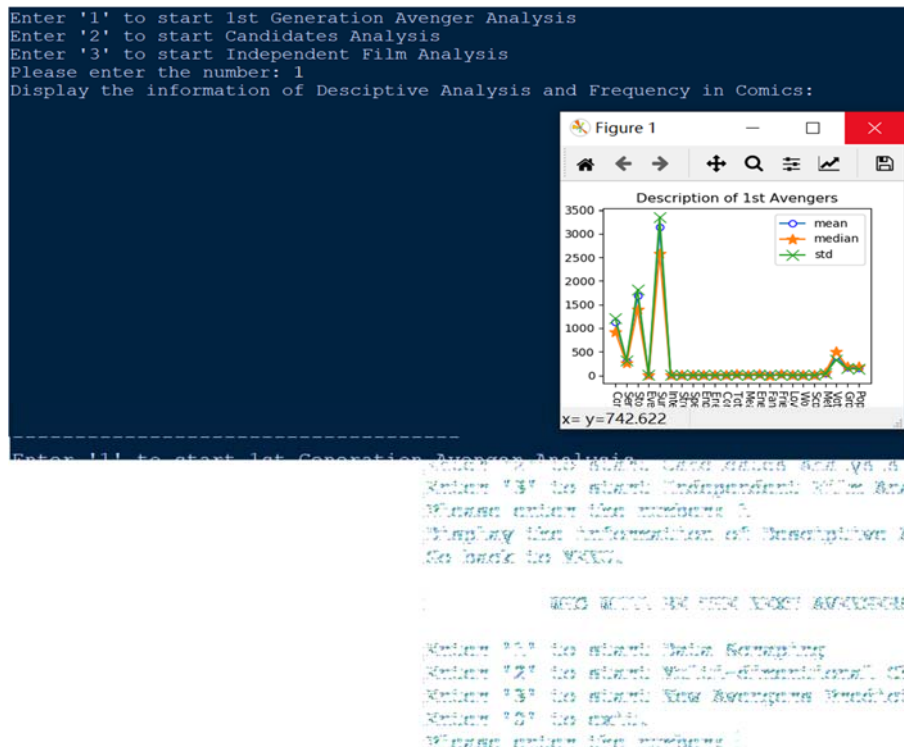
e)Input '1'->'5': Scrape data from Marvel-cinematic-universe

- 2) In command line, when you enter "2", three sub-options related to multi-dimensional character analysis will be shown.

```
-----
Enter '1' to start 1st Generation Avenger Analysis
Enter '2' to start Candidates Analysis
Enter '3' to start Independent Film Analysis
Please enter the number: |
```

a) Input '1': Descriptive analysis for the first avengers

Note 5: Close figure to move on. When an image is shown, click 'close' to show the next image. When all the images are closed, it will then go back to the initial command line. This also applies to the following commands.



b)Input '2'->'2': Analysis for the candidates

c)Input '2'->'3': Analysis for the independent movies

Note 6: Time consuming when generating word cloud figure. Since it takes a long time to translate Chinese to English in word cloud figure, we use 10 words in the demo. In the zip file, we provide a completed word cloud figure for reference.

- 3) In command line, when you enter '3', two sub-options related to new avengers prediction will be shown.

Which method do you want to use? 1:Statistic Method 2:Customized Method

a)Input '1': Statistic method to predict the next avengers

Note 7: We designed to finish the whole process after this option. To switch to the customized method, please run the main module again.

b)Input '3'->'2': Customized method to predict the next avengers

```
Which method do you want to use? 1:Statistic Method 2:Customized Method2
Enter importance degree for 3 factors: official importance, power and popularity.
The sum of these 3 factors should equal to 1.
Input the weights for official importance:0.4
Input the weights for power:0.4
Input the weights for popularity:0.2
Input the amount of team members(from 1 to 10):6
```

i) Input the weight for official importance

ii) Input the weight for power

iii)Input the weight for popularity

iv) Input the number for the next avengers

To exit the whole program, you can enter '0'.

If you enter a wrong number, our system will show a notification and require users to re-enter again.

4 Group Information

Puxing Zhao, puxingz@andrew.cmu.edu

Yingxin Liang, yingxinl@andrew.cmu.edu

Shaoqing Zhang, shaoqinz@andrew.cmu.edu

Jun Yang, juny2@andrew.cmu.edu

Yan Pan, yanp2@andrew.cmu.edu

5 Youtube Hyperlink of the Video

<https://youtu.be/6X-p-oBQT18>