2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)

Data Analysis and Crawler Application Implementation Based on Python

Hejing Wu, Fang Liu, Long Zhao, Yabin Shao
East University of Heilongjiang
Heilongjiang, 150086, China
E-mail: 499917928@qq.com

Abstract—**In this age of information explosion, how to find the data we want efficiently from various miscellaneous data and extract them from the network in batches has become a key problem. And sometimes the data not processed itself may be confusing for people, through what kind of technical means, how to get to the complex data processing, finally become a kind of intuitive figures, or the trend that people can directly extract information from is also a very important topic to study in the era of the data. This topic will choose Steam online game platform as the research object. Steam is an online game retail platform launched by Valve company in the United States in 2003. Under the real circumstance, to explore how to develop a complete crawler method based on Scrapy frameworks for Steam top sell list and publishers, developers and stores page, to crawl the various data of all works of developers and publishers under the page on Steam platform. Based on the crawled data, use basic data analysis to analysis user's favorite game types in the top sell list, the total number of releases of game platforms of certain developers and publishers, the proportion of favorable comments, etc. and extract useful information through the data analysis process. To draw the conclusion and make summary in the final. In short, at first, this paper will explore how to develope a crawler with scontrolable and automatic crawling abilitys which can crawl specific target; Then the data that is crawled will be analysised and visualized by using Pandas library and Matplotlib library, the useful information will be extracted from the data analysis and visualization process, so as to complete.**

Keywords-Python Crawler; Scrapy Framework; The Selenium; Steam Platform

I. INTRODUCTION

The 21st century is a book written by information. With the rapid development of information technology, today's society has become a huge information polymer, and there are various kinds of data in this huge polymer. Data is an embodiment of information. Most of the time, the data is contained in the Internet, and the data is huge and jumbled. It is very difficult to use traditional manual means to separate these jumbled data from the Internet and summarize the useful information. This is why we need to use the means of computer crawler to efficiently and automatically crawl the information existing in the Internet, and use technical means to analyze the data and summarize the rules.

With the rapid development of the Internet in recent decades, while the number and difficulty of crawling and the types of crawling are increasing, the types and forms of today's reptiles have also been greatly expanded in the process of development. Modern web crawlers generally use libraries to develop, many languages have their own crawler libraries, which are different and have their own advantages in different crawling functions. Python related crawlers are widely used in various fields. The common crawler frameworks in Python are: scratch, Crawley, pysipider, cola, demiurge, robobrowser, etc. This paper will focus on the research of the crawler framework based on python.

II. RELEVANT TECHNOLOGIES AND FRAMEWORKS

In the aspect of reptile framework, this paper mainly selects scrapy as the framework of crawler, and uses beautifulsoup as the crawler analysis library of this topic. On the basis of scrapy, selenium is added to help crawl dynamic pages. The basic principle of the scrapy framework is shown in Figure 1.

Scrapy framework is very popular in the application of general web crawlers. Its first version was released in 2008, and now it is quite mature as a crawler framework. In the aspect of data analysis, the project mainly uses Python's pandas library and Matplotlib visualization library to perform basic data analysis and data visualization on the crawled data.

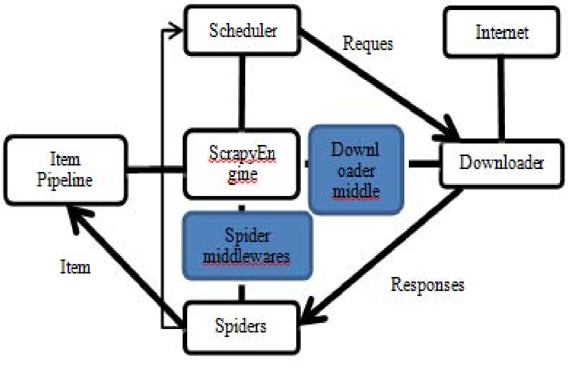


Figure 1. Basic principles of Scrapy frame

III. DESIGN OF CRAWLER

The basic idea is to start from the hot sale page, traverse the list, crawl and store it, then enter each link to read the sub page of the product, grab all the required information and record it, and pass it to the next method to crawl the manufacturer's page, crawl the manufacturer's basic

978-1-7281-7083-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCNEA50255.2020.00086

information and the manufacturer's game store link, and finally crawl each manufacturer through the game store link Basic information of all games under the banner.

The difficulty of the whole design lies in that part of the store pages are Ajax dynamic asynchronous loading pages. If only scrapy is used for static page crawling, some data will not be loaded, so selenium is also needed to simulate user operation to achieve dynamic page crawling. And steam's dynamic vendor page is still in the beta stage, not all vendors have dynamic loading home page, so how to judge whether the vendor's page is a dynamic page also needs attention.

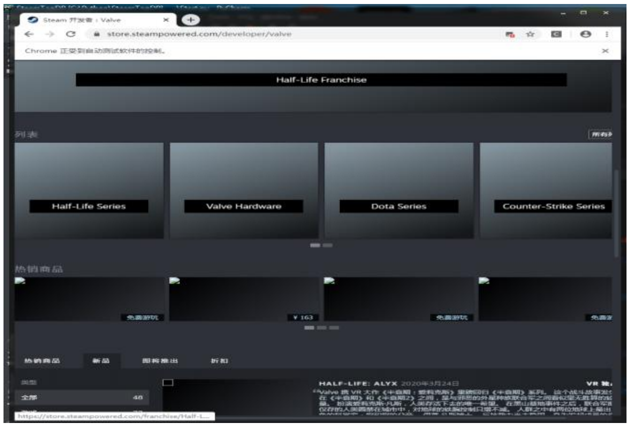


Figure 2. Selenium calls Google browser to crawl the dynamic page

The Scrapy crawler architecture itself consists mainly of items, spiders, piplines, middlewares. Items are mainly used to define the Items to be crawled, and the spider is responsible for defining the whole crawling process and the means of crawling.

Pipeline is responsible for some basic operations such as data cleaning and saving. You can define an output crawl result here. Middleware can be responsible for bridging services for Scrapy and other plug-ins or architectures.

In addition, the Scrapy crawler architecture provides Settings files where users can control the use of cookies, crawl speed limits, declare items added in pipelines, and so on as needed.

A. item design

In this procedure of item design, item of devitem developer, item of pubitem publisher and item of gamitem game are defined respectively.

TABLE I. DEVITEM DEVELOPER INFORMATION SHEET

Number	Dev_Item	
	Item	***annotation***
1	Nam	Name of developer
2	Pub_sum	Total number of developer platform releases
3	Gam_sum	Number of platform games released by developers
4	Dlc_sum	Developer platform DLC release number
5	follower	Number of developer followers

TABLE II. PUBITEM PUBLISHER INFORMATION SHEET

Number	Pub_Item	
	Item	***annotation***
1	Nam	Name of developer
2	Pub_sum	Total number of developer platform releases
3	Gam_sum	Number of platform games released by developers
4	Dlc_sum	Developer platform DLC release number
5	follower	Number of developer followers

TABLE III. GAMITEM GAME INFORMATION SHEET

Number	Gam_Item	
	Item	***annotation***
1	Nam	Name of publisher
2	Pub_sum	Total number of publisher platform releases
3	Gam_sum	Number of publisher platform game releases
4	Dlc_sum	Publisher platform DLC release
5	follower	Number of Publisher followers
1	Nam	Name of publisher

The Items that need to be crawled, as defined here, are also determined based on the requirements of the final data analysis to be performed. For example, the Nam vendor name of the developer and publisher and the Gam_title of the game list can be set to Index as a unique identifier for finding a particular piece of data.

B. Spider design

Spider design is the key point of this project. It defines how to grab the entities of items. No matter how to get the information from the manufacturer's product store on the dynamic page of the initial manufacturer, or how to get the information from the static page of the last store or the game list on the top sell list, it will be defined in this file. In this project, spider is mainly set in spider class Count start_ Requests method, top_ sell_ Parse method, store_ Parse method, DP_ Parse method and Gam_ The parse method implements the above function of grabbing items.

The start_requests method is responsible for reading the start_URL hot list url defined in the Spider class and passing its Request to the next top_sell_PARS method.

Top_sell_parse for top lists crawl method, is mainly responsible for climb top lists list, mainly use the beautifulsoup to parse the access list of sell like hot cakes search_resultsRows anchor tags in the links in the content, and pass the list of links with the request to the next way to parse, which can be modified. Find_all limit of the field of sell like hot cakes top how much investigation control, defines the limit of 15 in this program, that is investigating the Top15 in the hit list of manufacturers.

Store_parse receives a request from the above method and parses and crawls the store page for each item in the top sellers list. This method can classify the developer and publisher pages in the crawled store page. Since the link information of the developer and publisher is stored in the A TAB of the page details_block, the words developer and Publisher need to be determined from the obtained URL to determine whether the crawl target is the developer page or the publisher page.

Finally, the method circulates to get the Top_dev_list developer list, Top_pub_list publisher list, outputs the obtained results, and informs the following DP_parse vendor page whether the request returned by the store_parse method crawls a developer page or a publisher page by passing a flag.

In the DP_parse method, DP is Developer & Publisher, which is responsible for crawling the developer and publisher pages and returning developer and publisher information. Dp_parse first use of if statements to determine passed to his request, decided to instantiate the developer information object or issuer information object, because the developers page has both static pages and dynamic pages to the issuer, in determining whether the incoming object for developers or publishers, dp_parse method will continue to analyze the incoming object connection, check whether contain '/ search/field is dynamic or static page, determine whether to crawl dynamic pages using Selenium.

Steam business dynamic pages will business total number of all kinds of product release in the platform to write on the web page, a list of your products to crawl at the beginning of the page load will only load 10 products, so need to Selenium on web pages to simulate human drop-down operation, for these items to load, but each time the drop-down will only on the basis of the original in this load 10 entries, needs to read to crawl the total number of items of the project, divided by the number of more than 10 and take down, so you can set according to the total number of entries drop-down list number, if the total number of not more than 11, direct reading list. Beautifulsoup reads a loaded dynamic page and crawls the list of urls before closing the browser.

The browser is set to not load images and CSS mode to save system resources, to achieve a more efficient crawler. Each item of the specific store connection existed among the items one by one a anchor tag, use a loop to read these connections to the list of defined link_list, complete list crawl, but sometimes in the entry of words and images may contain a label, and will point to the same page, if direct application may cause repeated crawl, use the if not on the list to heavy in statement cycle.

The gam_parse method receives links to all game stores from each merchant page passed in from the DP_PARSE method, crawls the final game store page from each vendor, and returns the basic information entities of each game defined in Items, such as game name, media score, player media reputation, and so on. In the process of crawling, some old games without marked game type or developer, or some data without rating and evaluation were completed. That is the end of the Spider.

C. Pipeline design

In Scrapy, a pipeline is a pipeline that deals with captured Items. Generally responsible for the completion of crawling data cleaning and saving operations. The Pipeline in this crawler project is mainly responsible for the output of Items and creation and writing of CSV files. DevItem_to_CSV, PubItem_to_CSV and GamItem_to_CSV classes are defined in the Pipeline. They are responsible for exporting developers, publishers and game information Items to different CSV files in the same folder. All three classes are designed to be similar to the methods in the class.

Initialization in the init method defines the output directory of the file, use the w + file write mode, each time it is covered with the output of the rewrite the existing files, call the write row (write) function to the first line of the file to each column of the project, defines the output of a particular Item in process_Item method, using the if the instance to determine whether to output Item, because every Item output of different files, different Item entry is not identical also, so the need for judgment. Finally, close the file when the spider closes. In addition to the pipeline of the Items output file, pipline and data analysis pipeline are defined for output folder detection. Output folder detection pipeline will be executed before the output pipeline of the previous file to detect whether the OUTPUT folder of CSV file exists. If the folder does not exist, it will be created by itself. The data analysis pipeline is responsible for the data analysis of the saved data after the Spider completes the crawl and writes the file to CSV file at last. Only the definition of data analysis method in pipeline is listed here, and the code and explanation of specific data analysis method will be carried out after this chapter.

The data analysis method class is instantiated and the CSV_reader method in the data analysis method is called to begin the data analysis. After the Pipeline is written, the re-settings file declares each Pipeline class and sets the running priority of each class to complete the definition of the output class of the file in the Pipeline.

D. HeaderRdm design

It is worth noting that the website may suspect an attack and ask for Headers when doing a lot of crawling. The Headers request is the data that will be passed into the web browser and system when browsing the web in a normal browser. Sometimes the server USES this to determine if it is a scripting operation.

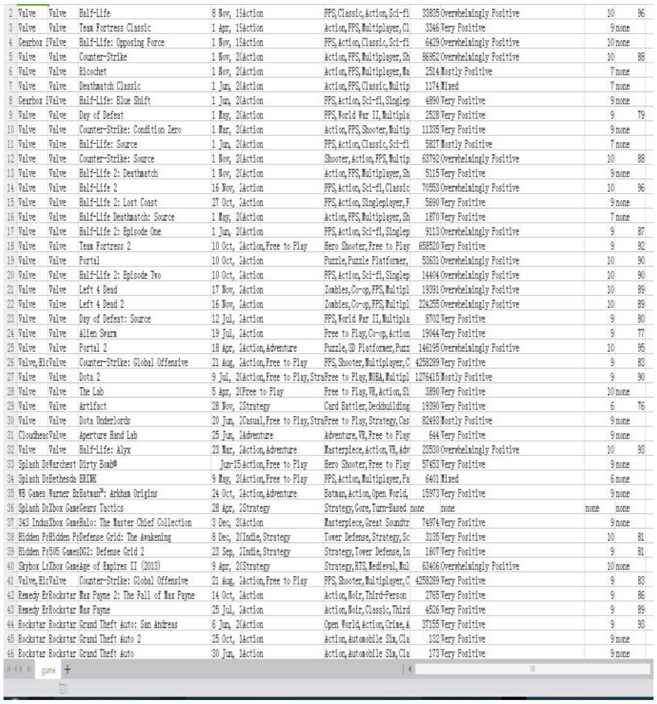
In addition, due to the need to input age when visiting specific pages sometimes, these age data will be saved in cookies after input. Although this is not true for every page, if such a page is encountered in the crawler process, an error will be reported, leading to the inability to crawl. Therefore, the HeaderRdm method provides a list of request headers and cookies parameters, and USES the random library provided by Python, using the headers['user-agent']= Random. Choice (USER_AGENTS) statement to return a random request header each time the method in this class is called, so as to achieve a certain purpose of anti-crawler.

E. Crawl results

This program has written a py file named start, and directly runs the Scrapy program in the IDE by calling the function of the CmdLine library, which avoids many troubles.

According to Scrapy return logs, 724 pages were taken, and the fastest speed was 96 pages per minute, for a total of 17 minutes and an average of 40 pages per minute. Due to the small size of some vendors, there is no vendor page, so the follower data is none. In addition, due to the early release time of some games, the media on the store page does not have the rating at that time, so some games are also rated as none. See the figure below for the part of specific results.

Figure 3. The three files Output in the Output folder



2	Valve	Valve	Half-Life	8 Nov, 1	Action	FPS,Classic,Action,Sci-fi	33835	Overwhelmingly Positive	10	96
3	Valve	Valve	Team Fortress Classic	1 Apr, 1	Action	Action,FPS,Multiplayer,Cl	3346	Very Positive	9	none
4	Gearbox S	Valve	Half-Life: Opposing Force	1 Nov, 1	Action	FPS,Action,Classic,Sci-fi	6429	Overwhelmingly Positive	10	none
5	Valve	Valve	Counter-Strike	1 Nov, 2	Action	Action,FPS,Multiplayer,Sh	86952	Overwhelmingly Positive	10	88
6	Valve	Valve	Ricochet	1 Nov, 2	Action	Action,FPS,Multiplayer,Ma	2514	Mostly Positive	7	none
7	Valve	Valve	Deathmatch Classic	1 Jun, 2	Action	Action,FPS,Classic,Multip	1174	Mixed	7	none
8	Gearbox S	Valve	Half-Life: Blue Shift	1 Jun, 2	Action	FPS,Action,Sci-fi,Singlep	4890	Very Positive	9	none
9	Valve	Valve	Day of Defeat	1 May, 2	Action	FPS,World War II,Multipla	2528	Very Positive	9	79
10	Valve	Valve	Counter-Strike: Condition Zero	1 Mar, 2	Action	Action,FPS,Shooter,Multip	11335	Very Positive	9	none
11	Valve	Valve	Half-Life: Source	1 Jun, 2	Action	FPS,Action,Classic,Sci-fi	5827	Mostly Positive	7	none
12	Valve	Valve	Counter-Strike: Source	1 Nov, 2	Action	Shooter,Action,FPS,Multip	63792	Overwhelmingly Positive	10	88
13	Valve	Valve	Half-Life 2: Deathmatch	1 Nov, 2	Action	Action,FPS,Multiplayer,Sh	5115	Very Positive	9	none
14	Valve	Valve	Half-Life 2	16 Nov, 2	Action	FPS,Action,Sci-fi,Classic	70553	Overwhelmingly Positive	10	96
15	Valve	Valve	Half-Life 2: Lost Coast	27 Oct, 2	Action	FPS,Action,Singleplayer,F	5690	Very Positive	9	none
16	Valve	Valve	Half-Life Deathmatch: Source	1 May, 2	Action	Action,FPS,Multiplayer,Sh	1870	Very Positive	7	none
17	Valve	Valve	Half-Life 2: Episode One	1 Jun, 2	Action	FPS,Action,Sci-fi,Singlep	9113	Overwhelmingly Positive	9	87
18	Valve	Valve	Team Fortress 2	10 Oct, 2	Action,Free to Play	Hero Shooter,Free to Play	658520	Very Positive	9	92
19	Valve	Valve	Portal	10 Oct, 2	Action	Puzzle,Puzzle Platformer,	53631	Overwhelmingly Positive	10	90
20	Valve	Valve	Half-Life 2: Episode Two	10 Oct, 2	Action	FPS,Action,Sci-fi,Singlep	14404	Overwhelmingly Positive	10	90
21	Valve	Valve	Left 4 Dead	17 Nov, 2	Action	Zombies,Co-op,FPS,Multipl	19391	Overwhelmingly Positive	10	89
22	Valve	Valve	Left 4 Dead 2	16 Nov, 2	Action	Zombies,Co-op,FPS,Multipl	224255	Overwhelmingly Positive	10	89
23	Valve	Valve	Day of Defeat: Source	12 Jul, 2	Action	FPS,World War II,Multipla	8702	Very Positive	9	80
24	Valve	Valve	Alien Swarm	19 Jul, 2	Action	Free to Play,Co-op,Action	19044	Very Positive	9	77
25	Valve	Valve	Portal 2	18 Apr, 2	Action,Adventure	Puzzle,3D Platformer,Puzz	146195	Overwhelmingly Positive	10	95
26	Valve,Hid	Valve	Counter-Strike: Global Offensive	21 Aug, 2	Action,Free to Play	FPS,Shooter,Multiplayer,C	4258289	Very Positive	9	83
27	Valve	Valve	Dota 2	9 Jul, 2	Action,Free to Play,Stra	Free to Play,MOBA,Multipl	1276415	Mostly Positive	9	90
28	Valve	Valve	The Lab	5 Apr, 2	Free to Play	Free to Play,VR,Action,Si	3890	Very Positive	10	none
29	Valve	Valve	Artifact	28 Nov, 2	Strategy	Card Battler,Deckbuilding	19390	Very Positive	6	76
30	Valve	Valve	Dota Underlords	20 Jun, 2	Casual,Free to Play,Stra	Free to Play,Strategy,Cas	82493	Mostly Positive	9	none
31	Cloudhead	Valve	Aperture Hand Lab	25 Jun, 2	Adventure	Adventure,VR,Free to Play	644	Very Positive	9	none
32	Valve	Valve	Half-Life: Alyx	23 Mar, 2	Action,Adventure	Masterpiece,Action,VR,Adv	23530	Overwhelmingly Positive	10	93
33	Splash Da	Warchest	Dirty Bomb®	Jun-15	Action,Free to Play	Hero Shooter,Free to Play	57453	Very Positive	9	none
34	Splash Da	Bethesda	BRINK	9 May, 2	Action,Free to Play	FPS,Action,Multiplayer,Pa	6401	Mixed	6	none
35	WB Games	Warner Br	Batman™: Arkham Origins	24 Oct, 2	Action,Adventure	Batman,Action,Open World,	15973	Very Positive	9	none
36	Splash Da	Xbox Game	Gears Tactics	28 Apr, 2	Strategy	Strategy,Gore,Turn-Based	none	none	none	none
37	343 Indus	Xbox Game	Halo: The Master Chief Collection	3 Dec, 2	Action	Masterpiece,Great Soundtr	74974	Very Positive	9	none
38	Hidden Pa	Hidden Pa	Defense Grid: The Awakening	8 Dec, 2	Indie,Strategy	Tower Defense,Strategy,Sc	3135	Very Positive	10	81
39	Hidden Pa	505 Games	DG2: Defense Grid 2	23 Sep, 2	Indie,Strategy	Strategy,Tower Defense,In	1607	Very Positive	9	81
40	Skybox La	Xbox Game	Age of Empires II (2013)	9 Apr, 2	Strategy	Strategy,RTS,Medieval,Mul	63406	Overwhelmingly Positive	10	none
41	Valve,Hid	Valve	Counter-Strike: Global Offensive	21 Aug, 2	Action,Free to Play	FPS,Shooter,Multiplayer,C	4258289	Very Positive	9	83
42	Remedy En	Rockstar	Max Payne 2: The Fall of Max Payne	14 Oct, 2	Action	Action,Noir,Third-Person	2765	Very Positive	9	86
43	Remedy En	Rockstar	Max Payne	25 Jul, 2	Action	Action,Noir,Classic,Third	4526	Very Positive	9	89
44	Rockstar	Rockstar	Grand Theft Auto: San Andreas	6 Jun, 2	Action	Open World,Action,Crime,A	37155	Very Positive	9	93
45	Rockstar	Rockstar	Grand Theft Auto 2	25 Oct, 1	Action	Action,Automobile Sim,Cla	132	Very Positive	9	none
46	Rockstar	Rockstar	Grand Theft Auto	30 Jun, 1	Action	Action,Automobile Sim,Cla	173	Very Positive	9	none

Figure 4. Screenshot of 692 pieces of data list of game information data crawled

IV. Data Analysis

The data analysis function of this program is to read the data saved in the CSV file by using pandas library and Matplotlib visual library, and perform basic data analysis and visual operation on the crawled data.

The function is realized through CSV_ reader, dir_ finder， get_ date， datanls_ Init consists of four modules and several data analysis modules. get_ Date gets the current system time and returns the combined time of month, month and day in string form. dir_ The finder method is mainly used to detect whether there are Datanls data analysis results output folder in the project directory. If the method of if isdir is used to determine whether the directory has been created, then MKDIR is used to create the directory. In the process, the get_ above is invoked.

The date method gets the current system time, and uses the returned date string to create the related directory. csv_ The reader method first calls dir_ Finder verifies that the output directory exists. And then I used it csv.reader Method reads the data from the three files output by the previous crawler and transfers them to dataframe as parameters to datanls_ Init initialization module.

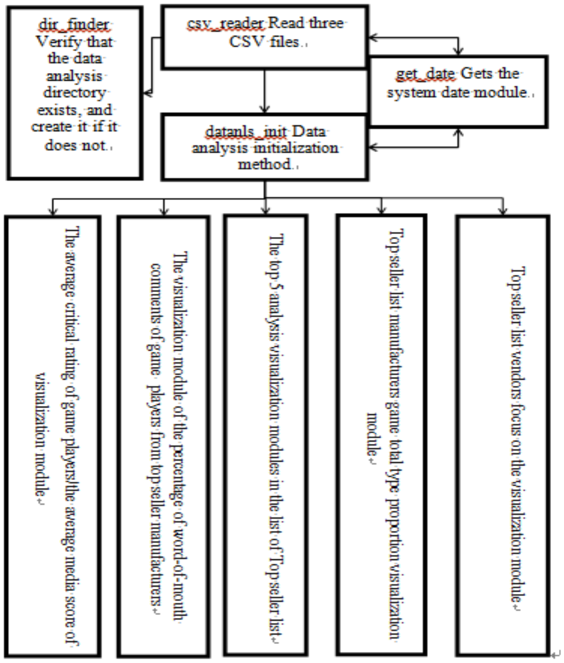

dir_finder Verify that the data analysis directory exists, and create it if it does not.
csv_reader Read three CSV files.
get_date Gets the system date module.
datanls_init Data analysis initialization method.
The average critical rating of game players/the average media score of visualization module
The visualization module of the percentage of word-of-mouth comments of game players from top seller manufacturers
The top 5 analysis visualization modules in the list of Top seller list
Top seller list manufacturers game total type proportion visualization module
Top seller list vendors focus on the visualization module

Figure 5. Data analysis thought map

datanls_ Init module is mainly responsible for initializing each data analysis module, summarizing each information before data analysis, and the module will continue to transfer each data to each data analysis method in the form of parameters.

The analysis visualization module of data contains many sub modules, including Ave_ score_ C manufacturer's work average player good comparison / media average score analysis module, DP_ scorer_ Analysis module of player evaluation proportion of C manufacturer's works, Pt_ flw_ C manufacturer follower ranking module, Pt_ sum_ C distribution ranking module, sum_ tpye_ Game type proportion analysis module of all pie manufacturers.

The first four modules can output different charts according to the input parameters of flag developers or publishers. The following is part of the analysis results.

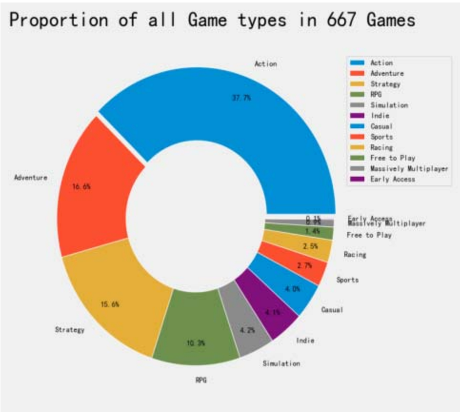


Figure 6. Shows the percentage of all game types output

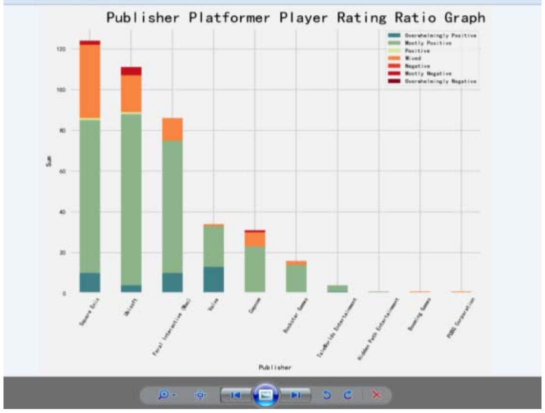


Figure 7. Shows the average player rating ratio of publisher platform games

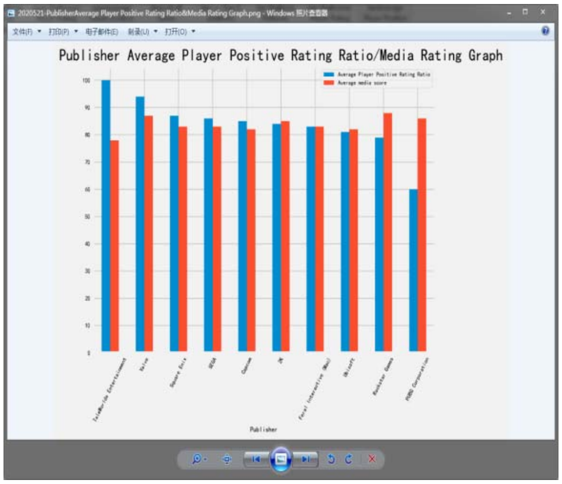


Figure 8. Shows the average player praise & media rating of publisher platform games

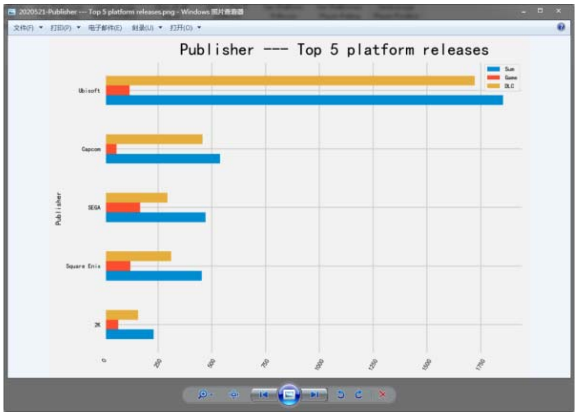


Figure 9. Outputs the top 5 publishers in total

V. SUMMARY

VI. Through the process of crawling to the top sell page of Steam online game store, this paper explores the process of data crawling and basic data analysis of dynamic pages in the use of Selenium library and Python Scrapy framework, and analysis the data in final.

VII. In my opinion, both crawler and each module of data analysis have good expansibility. In the aspect of crawler anti-crawler, Selenium itself has a very good crawler anti-crawler capability. If you want to make further crawler anti-crawler, you can also expand multiple cookies, even set up proxy IP pool and so on. The data analysis section can also use some more complex calculations module, such as recording the daily data and making trend analysis for the crawled data etc.

ACKNOWLEDGMENT

This paper is about the scientific research project of East University of Heilongjiang in 2019, "Implementation of Crawler Based on Python Scrapy Framework", project number HDFKY190109

REFERENCES

[1] Yuhao Fan. Design and implementation of distributed crawler system based on scrapy [J]. IOP Conference Series: Earth and Environmental Science, 2018, 108(4):2-8.

[2] Jing Wang, Yuchun Guo. Scrapy-based crawling and user-behavior characteristics analysis on taobao [P]. Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012 International Conference on, 2012:1-5.

[3] Ryan Mitchell. Python web crawler authority Guide (Second Edition) [M]. Beijing: People's post and Telecommunications Press, 2019:57-70.

[4] Wei Chengcheng. Data information crawler technology based on Python [J]. Electronic world, 2018 (11): 208-209.

[5] Mark. Lutz. Python learning manual (Fifth Edition, Volume I) [M]. Beijing: Mechanical Industry Press, 2019:1-2.