# Application Research of Crawler and Data Analysis Based on Python

Wu Hejing
East University of Heilongjiang
Heilongjiang, 150086
E-mail: 499917928@qq.com,

Zhao Long
East University of Heilongjiang
Heilongjiang, 150086

Shao Yabin
East University of Heilongjiang
Heilongjiang, 150086

Liu Fang
East University of Heilongjiang
Heilongjiang, 150086）

Cui Ran
East University of Heilongjiang
Heilongjiang, 150086

***Abstract*—Combined with the actual situation, this paper explores how to develop a crawler method based on the specific framework for the complete interface of steam manufacturers and stores, which should be able to automatically and efficiently crawl the data of specific targets, analyze the dynamic pages, and complete the data cleaning, downloading, saving and other operations, explore the methods of general data analysis, and Analyze the downloaded data, extract useful information from it, analyze and summarize the specific crawler method and data analysis method through practical application.**

***Keywords-Python; Scrapy; Selenium; BeautifulSoup***

## 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 a kind of embodiment of information. In this era of information explosion, how to efficiently find the data we want from all kinds of miscellaneous data and extract them from the network in batches has become a key problem. However, sometimes the unprocessed data itself may be confusing for people. How to process the huge and complex data obtained through what kind of technical means, and finally become an intuitive number, or trend, and become the information that people can obtain intuitively is also a very important topic to be studied in this data age.

## II. STATISTICAL INVESTIGATION ON THE PREFERENCE SALES VOLUME

In this project, the American Steam online game platform mall is selected as the research object of the crawler. By setting a specific game company as a search keyword in steam's online mall, the data of all works of the company in steam mall are crawled, and the useful information is extracted by analyzing the basic data of each manufacturer's preference for game production type, series sales volume, and praise In addition, the game manufacturers are comprehensively scored and evaluated.

DOI: 10.21307/ijanmc-2020-018

## III. RELEVANT TECHNOLOGY AND FRAMEWORK

This project will use the scrapy framework based on Python language to crawl steam website. Python as a language has the advantages of lightweight, simplicity, wide range of application and so on. At present, various crawler frameworks and application libraries based on Python have been very mature, among which the crawler 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. The basicprinciple of the scrapy framework is shown in Figure 1.

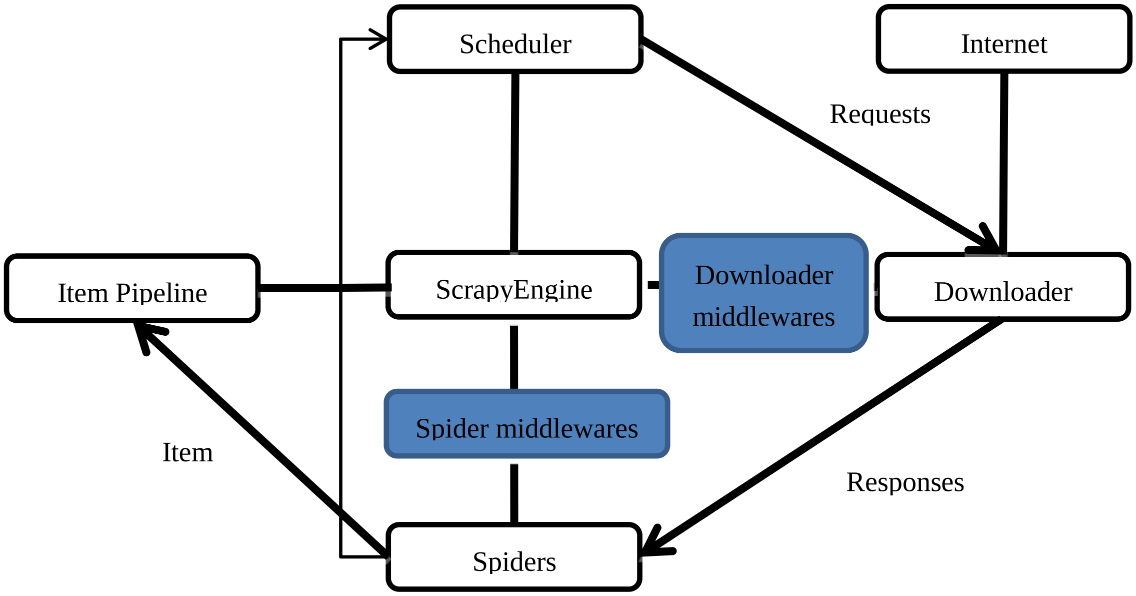


Figure 1. Basic principles of Scrapy frame

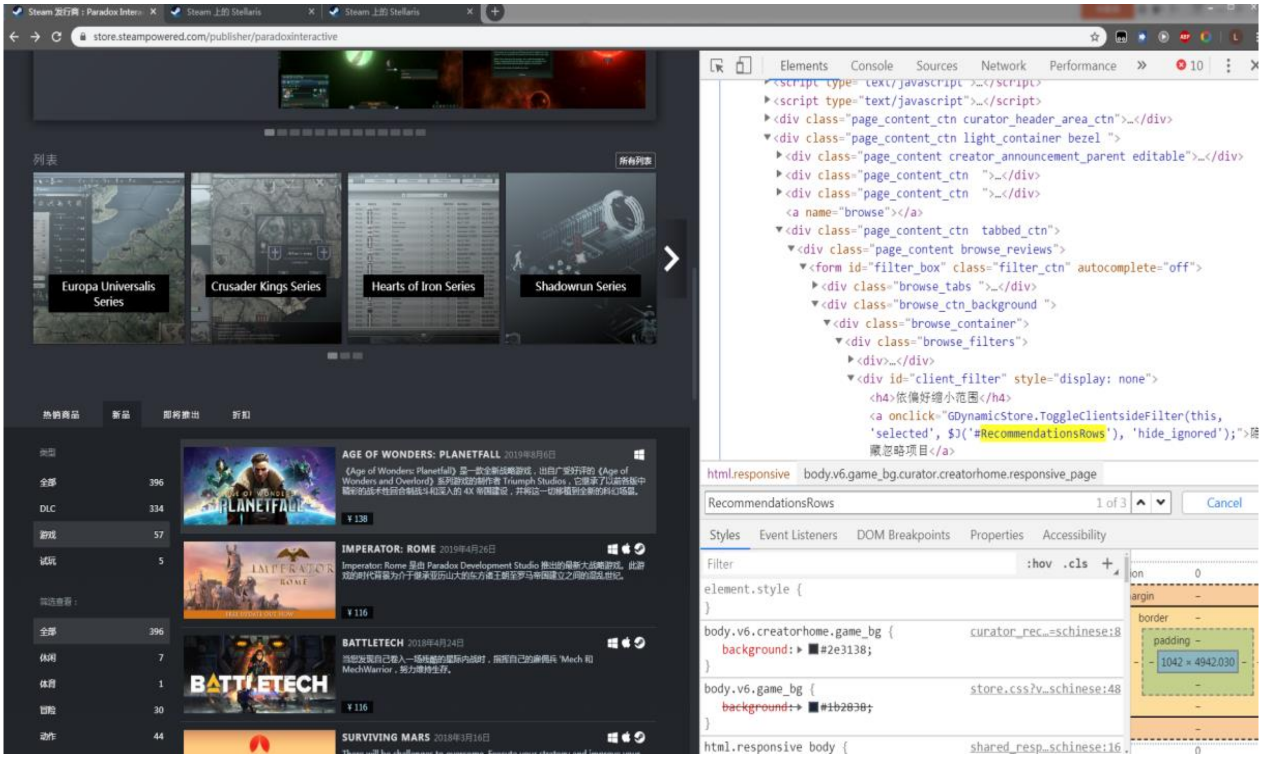
## IV. DESIGN OF CRAWLER

### A. General design idea

The process of crawler itself is actually to simulate the user's operation on the browser with a program. First of all, the starting point and range of crawling need to be specified. As the target of crawling is for manufacturers and their works, the interface of manufacturers is taken as the starting point. For example, the page of paradox, a manufacturer, first analyzes the entire manufacturer's page, and finds that the page links and information of all games or game related DLC downloads of the manufacturer are stored in the recommendation div framework of each sub recommendation of recommendations rows, as shown in Figure 2

### B. Design and implementation of reptile functions

The crawler architecture is composed of items, spiders, piplings and middleware. Among them, items are mainly used to define the items to be crawled, spiders are responsible for defining the whole process of crawling, what means to crawl, pipes are responsible for the basic operations such as data cleaning and saving, middleware can be responsible for the bridge service of scratch and other plug-ins or architectures.



```
<script type="text/javascript">…</script>
<div class="page_content_ctn curator_header_area_ctn">…</div>
<div class="page_content_ctn light_container bezel ">
  <div class="page_content creator_announcement_parent editable">…</div>
  <div class="page_content_ctn  ">…</div>
  <div class="page_content_ctn  ">…</div>
  <a name="browse"></a>
  <div class="page_content_ctn  tabbed_ctn">
    <div class="page_content browse_reviews">
      <form id="filter_box" class="filter_ctn" autocomplete="off">
        <div class="browse_tabs ">…</div>
        <div class="browse_ctn_background ">
          <div class="browse_container">
            <div class="browse_filters">
              <div>…</div>
              <div id="client_filter" style="display: none">
                <h4>依偏好缩小范围</h4>
                <a onclick="GDynamicStore.ToggleClientsideFilter(this,
                'selected', $J('#RecommendationsRows'), 'hide_ignored');">隐
                藏忽略项目</a>
```

html.responsive body.v6.game_bg.curator.creatorhome.responsive_page

RecommendationsRows 1 of 3 Cancel

Styles Event Listeners DOM Breakpoints Properties Accessibility

```
element.style {
}
body.v6.creatorhome.game_bg {                curator_rec…=schinese:8
  background: ▸ ■#2e3138;
}
body.v6.game_bg {                            store.css?v…schinese:48
  ~~background: ▸ ■#1b2838;~~
}
html.responsive body {                       shared_resp…schinese:16
```

Figure 2. Investigation of HTML page structure of steam manufacturers by using viewers

First, the items to be crawled are defined in the items file. Finally, these items may be submitted to the analysis part for data analysis. The specific design and implementation code is:

```
import scrapy
class SteamDevItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    qry_nam = scrapy.Field()
    if_dev = scrapy.Field()
    pub_sum = scrapy.Field()
    pub_gam_sum = scrapy.Field()
    pub_dlc_sum = scrapy.Field()
    dev_nam = scrapy.Field()
    pub_nam = scrapy.Field()
    gam_title = scrapy.Field()
    res_date = scrapy.Field()
    gam_type = scrapy.Field()
    gam_tag = scrapy.Field()
    if_muti = scrapy.Field()
    gam_score = scrapy.Field()
    gam_score_sum = scrapy.Field()
    gam_score_ratio = scrapy.Field()
  pass
```

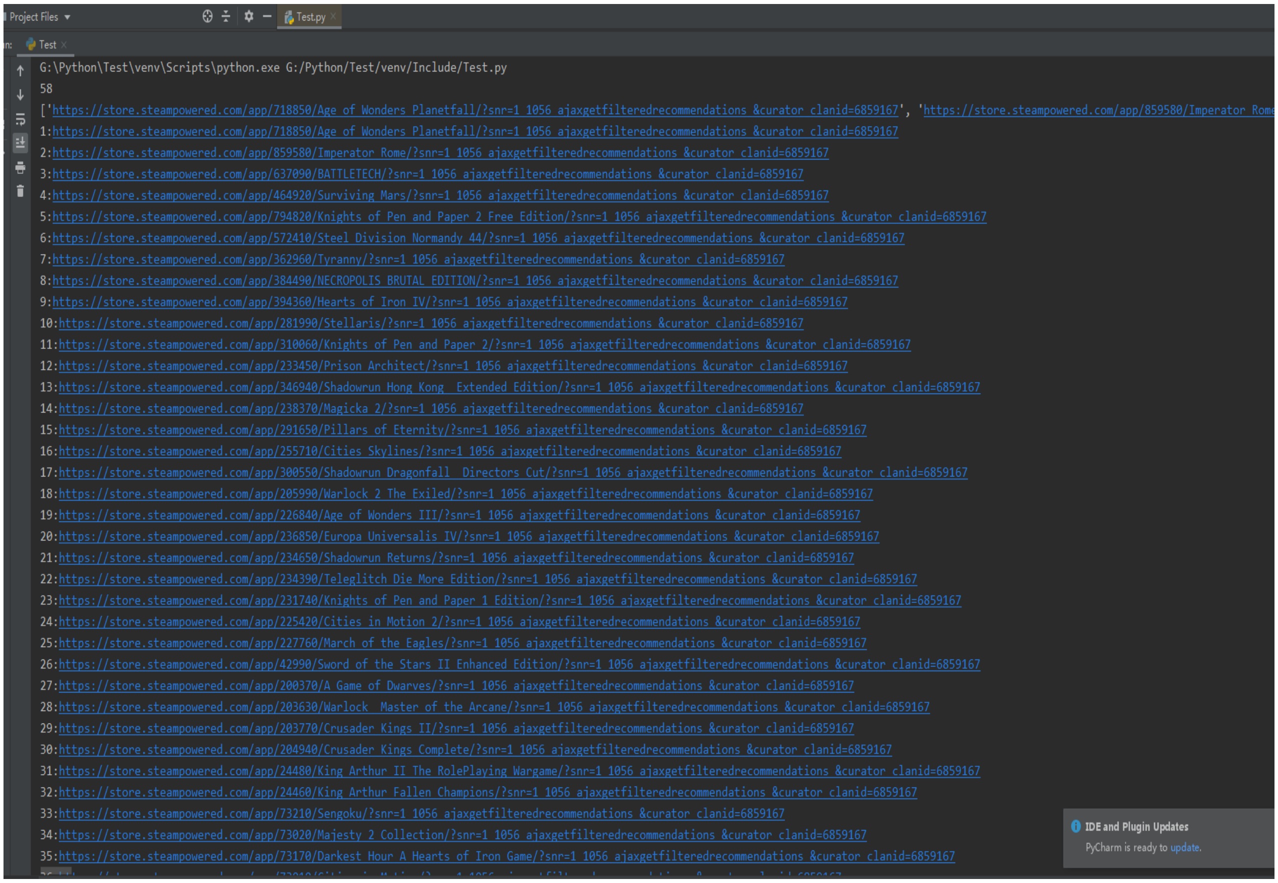
### *C. Spider design*

The design of spider is the key point of this project. Whether the initial dynamic page connection or the last static page information crawling mode will be defined in this file. In this project, spider will be named steam, and some key implementation codes will be pasted here, with running results and some notes attached. First, introduce start_ the design method of dynamic page crawling of selenium in requests method:

```
chrome_opt = webdriver.ChromeOptions()
    prefs = {
"profile.managed_default_content_settings.images": 2,
        'permissions.default.stylesheet': 2
    }
    chrome_opt.add_experimental_option("prefs", prefs)
    browser = webdriver.Chrome(options=chrome_opt)
    browser.get("https://store.steampowered.com/" + Qry_sta + "/" + Qry_Target)
    bs = BeautifulSoup(browser.page_source, 'html.parser')     #Beautiful Soup
```

The specific store connections of each product exist in the a anchor label of each entry, and these connections are read to the defined links using the loop_ In the list list, crawling of the list is completed, but sometimes the text and picture in the entry may contain a tag, and they all point to the same page. If direct application may cause repeated crawling, a loop is used here, and if not in statement is used to de duplicate the list.

After using the print statement to verify the function of the module, the verification results are shown in Figure 3.

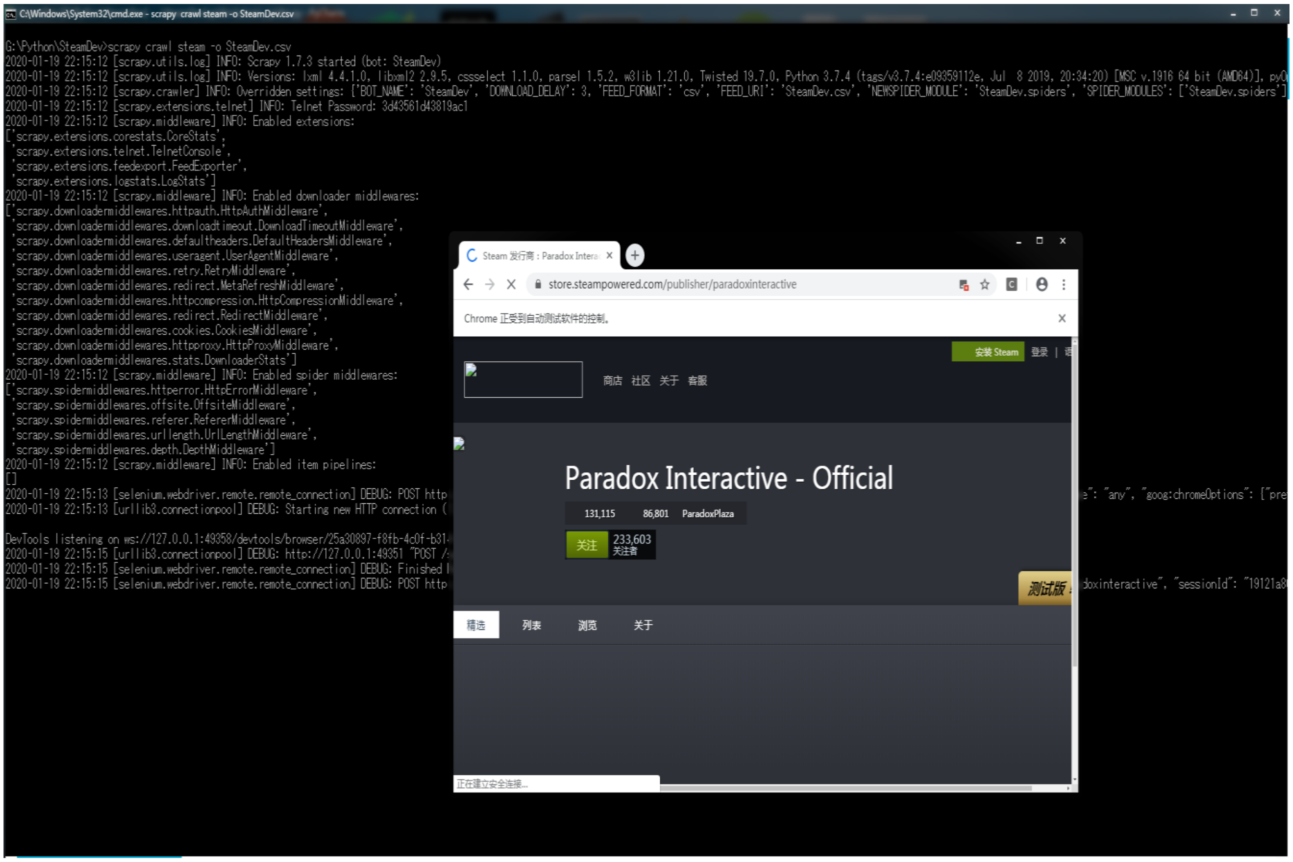


```
G:\Python\Test\venv\Scripts\python.exe G:/Python/Test/venv/Include/Test.py
58
['https://store.steampowered.com/app/718850/Age of Wonders Planetfall/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167', 'https://store.steampowered.com/app/859580/Imperator Rome
1:https://store.steampowered.com/app/718850/Age of Wonders Planetfall/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
2:https://store.steampowered.com/app/859580/Imperator Rome/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
3:https://store.steampowered.com/app/637090/BATTLETECH/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
4:https://store.steampowered.com/app/464920/Surviving Mars/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
5:https://store.steampowered.com/app/794820/Knights of Pen and Paper 2 Free Edition/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
6:https://store.steampowered.com/app/572410/Steel Division Normandy 44/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
7:https://store.steampowered.com/app/362960/Tyranny/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
8:https://store.steampowered.com/app/384490/NECROPOLIS BRUTAL EDITION/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
9:https://store.steampowered.com/app/394360/Hearts of Iron IV/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
10:https://store.steampowered.com/app/281990/Stellaris/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
11:https://store.steampowered.com/app/310060/Knights of Pen and Paper 2/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
12:https://store.steampowered.com/app/233450/Prison Architect/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
13:https://store.steampowered.com/app/346940/Shadowrun Hong Kong  Extended Edition/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
14:https://store.steampowered.com/app/238370/Magicka 2/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
15:https://store.steampowered.com/app/291650/Pillars of Eternity/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
16:https://store.steampowered.com/app/255710/Cities Skylines/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
17:https://store.steampowered.com/app/300550/Shadowrun Dragonfall  Directors Cut/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
18:https://store.steampowered.com/app/205990/Warlock 2 The Exiled/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
19:https://store.steampowered.com/app/226840/Age of Wonders III/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
20:https://store.steampowered.com/app/236850/Europa Universalis IV/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
21:https://store.steampowered.com/app/234650/Shadowrun Returns/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
22:https://store.steampowered.com/app/234390/Teleglitch Die More Edition/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
23:https://store.steampowered.com/app/231740/Knights of Pen and Paper 1 Edition/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
24:https://store.steampowered.com/app/225420/Cities in Motion 2/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
25:https://store.steampowered.com/app/227760/March of the Eagles/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
26:https://store.steampowered.com/app/42990/Sword of the Stars II Enhanced Edition/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
27:https://store.steampowered.com/app/200370/A Game of Dwarves/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
28:https://store.steampowered.com/app/203630/Warlock  Master of the Arcane/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
29:https://store.steampowered.com/app/203770/Crusader Kings II/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
30:https://store.steampowered.com/app/204940/Crusader Kings Complete/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
31:https://store.steampowered.com/app/24480/King Arthur II The RolePlaying Wargame/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
32:https://store.steampowered.com/app/24460/King Arthur Fallen Champions/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
33:https://store.steampowered.com/app/73210/Sengoku/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
34:https://store.steampowered.com/app/73020/Majesty 2 Collection/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
35:https://store.steampowered.com/app/73170/Darkest Hour A Hearts of Iron Game/?snr=1 1056 ajaxgetfilteredrecommendations &curator clanid=6859167
```

Figure 3. List of URLs obtained by selenium and beautiful soup

### *D. Start directional climbing*

After designing and debugging the spider, run the CMD command window of the system, open the root directory of the crawler file, and input the crawler stream-o SteamDev.csv , crawl the target website. Input - O SteamDev.csv The purpose is to let the crawler save the last crawled data in the form of CSV table. The saved data appears in the project root. See Figure 4 for the climbing process.



C:\Windows\System32\cmd.exe - scrapy crawl steam -o SteamDev.csv

```
G:\Python\SteamDev>scrapy crawl steam -o SteamDev.csv
2020-01-19 22:15:12 [scrapy.utils.log] INFO: Scrapy 1.7.3 started (bot: SteamDev)
2020-01-19 22:15:12 [scrapy.utils.log] INFO: Versions: lxml 4.4.1.0, libxml2 2.9.5, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0, Twisted 19.7.0, Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2019, 20:34:20) [MSC v.1916 64 bit (AMD64)], pyO
2020-01-19 22:15:12 [scrapy.crawler] INFO: Overridden settings: {'BOT_NAME': 'SteamDev', 'DOWNLOAD_DELAY': 3, 'FEED_FORMAT': 'csv', 'FEED_URI': 'SteamDev.csv', 'NEWSPIDER_MODULE': 'SteamDev.spiders', 'SPIDER_MODULES': ['SteamDev.spiders']
2020-01-19 22:15:12 [scrapy.extensions.telnet] INFO: Telnet Password: 3d43561d43819ac1
2020-01-19 22:15:12 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2020-01-19 22:15:12 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2020-01-19 22:15:12 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2020-01-19 22:15:12 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2020-01-19 22:15:13 [selenium.webdriver.remote.remote_connection] DEBUG: POST http                                                  e": "any", "goog:chromeOptions": {"pre
2020-01-19 22:15:13 [urllib3.connectionpool] DEBUG: Starting new HTTP connection (

DevTools listening on ws://127.0.0.1:49358/devtools/browser/25a30897-f8fb-4c0f-b31
2020-01-19 22:15:15 [urllib3.connectionpool] DEBUG: http://127.0.0.1:49351 "POST /
2020-01-19 22:15:15 [selenium.webdriver.remote.remote_connection] DEBUG: Finished 
2020-01-19 22:15:15 [selenium.webdriver.remote.remote_connection] DEBUG: POST http                                                  doxinteractive", "sessionId": "19121a8
```

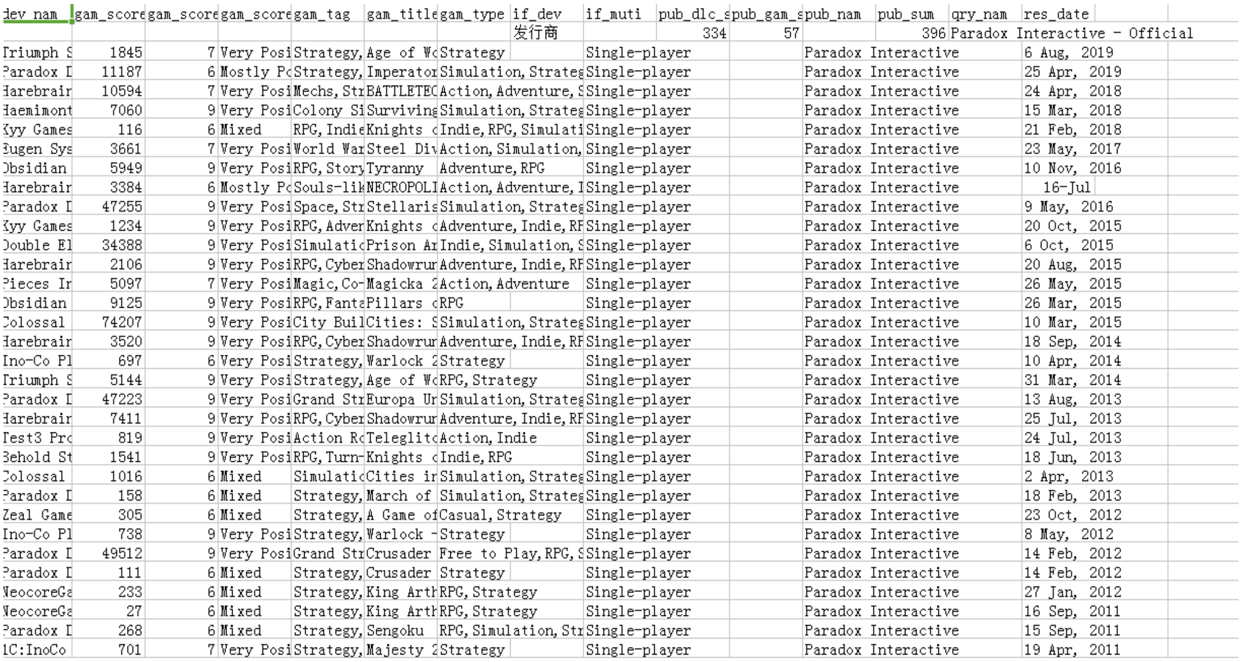
Steam 发行商 : Paradox Interac
store.steampowered.com/publisher/paradoxinteractive
Chrome 正受到自动测试软件的控制。
安装 Steam
登录
商店
社区
关于
客服
Paradox Interactive - Official
131,115
86,801
ParadoxPlaza
关注
233,603
关注者
测试版
精选
列表
浏览
关于
正在建立安全连接...

Figure 4. Executing the start request method selenium pop-up browser to crawl the dynamic page

## V. DATA ANALYSIS

Next, we will perform basic visual operations on the crawled data in the form of operation tables. In the crawler project, we crawled for the Paradox Interactive publisher. The crawled data is presented in the form of CSV tables, as shown in Figure 5.

Through the use of spreadsheets and further collation of the crawled data, the following data are obtained: the publisher has published 396 works in steam platform, of which the majority of DLC has published 334 DLC, most of the games published are single player games, and each game published in its mall has an average of 6800 reviews, of which the proportion of favorable reviews is about 76.4 8%, see the chart below for detailed visual analysis.



| dev_nam | gam_score | gam_score | gam_score | gam_tag | gam_title | gam_type | if_dev | if_muti | pub_dlc_s | pub_gam_s | pub_nam | pub_sum | qry_nam | res_date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 发行商 | | 334 | 57 | | 396 | Paradox Interactive - Official | |
| Triumph S | 1845 | 7 | Very Posi | Strategy, | Age of W | Strategy | | Single-player | | | Paradox Interactive | | | 6 Aug, 2019 |
| Paradox I | 11187 | 6 | Mostly Po | Strategy, | Imperator | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 25 Apr, 2019 |
| Harebrair | 10594 | 7 | Very Posi | Mechs, Str | BATTLETEC | Action, Adventure, S | | Single-player | | | Paradox Interactive | | | 24 Apr, 2018 |
| Haemimont | 7060 | 9 | Very Posi | Colony Si | Surviving | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 15 Mar, 2018 |
| Kyy Games | 116 | 6 | Mixed | RPG, Indie | Knights c | Indie, RPG, Simulati | | Single-player | | | Paradox Interactive | | | 21 Feb, 2018 |
| Eugen Sys | 3661 | 7 | Very Posi | World War | Steel Div | Action, Simulation, | | Single-player | | | Paradox Interactive | | | 23 May, 2017 |
| Obsidian | 5949 | 9 | Very Posi | RPG, Story | Tyranny | Adventure, RPG | | Single-player | | | Paradox Interactive | | | 10 Nov, 2016 |
| Harebrair | 3384 | 6 | Mostly Po | Souls-lik | NECROPOLI | Action, Adventure, I | | Single-player | | | Paradox Interactive | | | 16-Jul |
| Paradox I | 47255 | 9 | Very Posi | Space, Str | Stellaris | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 9 May, 2016 |
| Kyy Games | 1234 | 9 | Very Posi | RPG, Adver | Knights c | Adventure, Indie, RF | | Single-player | | | Paradox Interactive | | | 20 Oct, 2015 |
| Double El | 34388 | 9 | Very Posi | Simulatic | Prison Ar | Indie, Simulation, S | | Single-player | | | Paradox Interactive | | | 6 Oct, 2015 |
| Harebrair | 2106 | 9 | Very Posi | RPG, Cyber | Shadowrur | Adventure, Indie, RF | | Single-player | | | Paradox Interactive | | | 20 Aug, 2015 |
| Pieces Ir | 5097 | 7 | Very Posi | Magic, Co- | Magicka 2 | Action, Adventure | | Single-player | | | Paradox Interactive | | | 26 May, 2015 |
| Obsidian | 9125 | 9 | Very Posi | RPG, Fanta | Pillars c | RPG | | Single-player | | | Paradox Interactive | | | 26 Mar, 2015 |
| Colossal | 74207 | 9 | Very Posi | City Buil | Cities: S | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 10 Mar, 2015 |
| Harebrair | 3520 | 9 | Very Posi | RPG, Cyber | Shadowrur | Adventure, Indie, RF | | Single-player | | | Paradox Interactive | | | 18 Sep, 2014 |
| Ino-Co Pl | 697 | 6 | Very Posi | Strategy, | Warlock 2 | Strategy | | Single-player | | | Paradox Interactive | | | 10 Apr, 2014 |
| Triumph S | 5144 | 9 | Very Posi | Strategy, | Age of W | RPG, Strategy | | Single-player | | | Paradox Interactive | | | 31 Mar, 2014 |
| Paradox I | 47223 | 9 | Very Posi | Grand Str | Europa Ur | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 13 Aug, 2013 |
| Harebrair | 7411 | 9 | Very Posi | RPG, Cyber | Shadowrur | Adventure, Indie, RF | | Single-player | | | Paradox Interactive | | | 25 Jul, 2013 |
| Test3 Prc | 819 | 9 | Very Posi | Action Rc | Teleglitc | Action, Indie | | Single-player | | | Paradox Interactive | | | 24 Jul, 2013 |
| Behold St | 1541 | 9 | Very Posi | RPG, Turn- | Knights c | Indie, RPG | | Single-player | | | Paradox Interactive | | | 18 Jun, 2013 |
| Colossal | 1016 | 6 | Mixed | Simulatic | Cities ir | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 2 Apr, 2013 |
| Paradox I | 158 | 6 | Mixed | Strategy, | March of | Simulation, Strateg | | Single-player | | | Paradox Interactive | | | 18 Feb, 2013 |
| Zeal Game | 305 | 6 | Mixed | Strategy, | A Game of | Casual, Strategy | | Single-player | | | Paradox Interactive | | | 23 Oct, 2012 |
| Ino-Co Pl | 738 | 9 | Very Posi | Strategy, | Warlock - | Strategy | | Single-player | | | Paradox Interactive | | | 8 May, 2012 |
| Paradox I | 49512 | 9 | Very Posi | Grand Str | Crusader | Free to Play, RPG, S | | Single-player | | | Paradox Interactive | | | 14 Feb, 2012 |
| Paradox I | 111 | 6 | Mixed | Strategy, | Crusader | Strategy | | Single-player | | | Paradox Interactive | | | 14 Feb, 2012 |
| NeocoreGa | 233 | 6 | Mixed | Strategy, | King Arth | RPG, Strategy | | Single-player | | | Paradox Interactive | | | 27 Jan, 2012 |
| NeocoreGa | 27 | 6 | Mixed | Strategy, | King Arth | RPG, Strategy | | Single-player | | | Paradox Interactive | | | 16 Sep, 2011 |
| Paradox I | 268 | 6 | Mixed | Strategy, | Sengoku | RPG, Simulation, Str | | Single-player | | | Paradox Interactive | | | 15 Sep, 2011 |
| 1C:InoCo | 701 | 7 | Very Posi | Strategy, | Majesty 2 | Strategy | | Single-player | | | Paradox Interactive | | | 19 Apr, 2011 |

Figure 5. Crawled data list

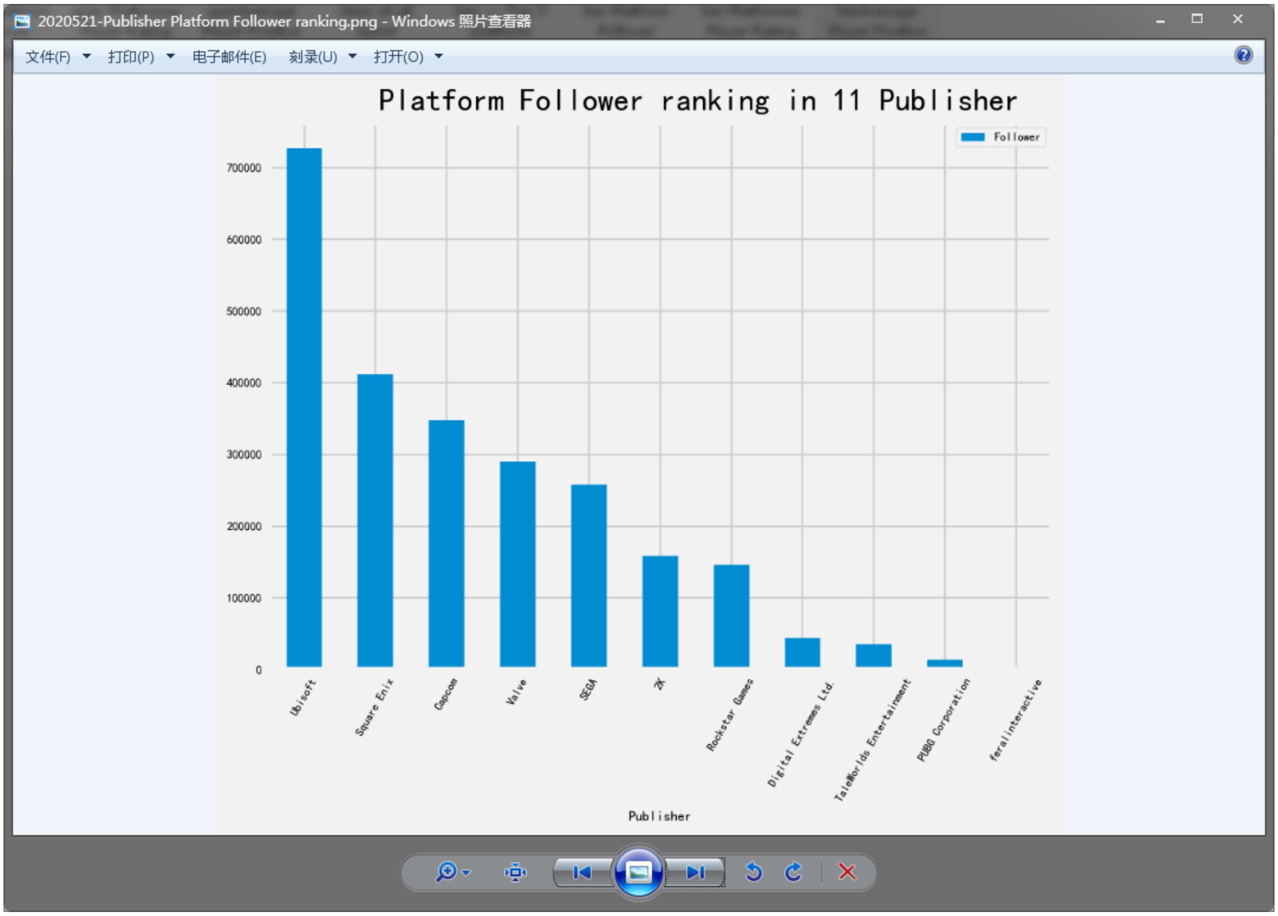


Figure 6. Output the publisher platform follower ranking chart

## VI. CONCLUSION

Through demonstration and part of practice, this paper explores the process of data crawling and basic data analysis of dynamic pages by combining the general Python's story framework with selenium + beautiful soup through crawling the steam online game mall website.

The crawler has good scalability. For example, if you want to compare the crawling data of multiple game manufacturers, you can write a query manufacturer list to get the product URL list from the dynamic web page of the manufacturer list first. In terms of anti-crawler, selenium itself has a very good anti crawler ability. If you want to further anti crawler, you can also expand multiple cookies, and even establish a proxy IP pool.

## ACKNOWLEDGMENT

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

## REFERENCE

[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, 20120:1-5.

[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.

[6] Fan Chuanhui. Python reptile development and project practice [M]. Beijing: Mechanical Industry Press, 2017 (3): 69-72.

[7] Song Yongsheng, Huang Rongmei, Wang Jun. research on Python based data analysis and visualization platform [J]. Modern information technology: 2019 (21): 1-4.

[8] Liu Yuke, Wang Ping. Statistics and graph output of student achievement data based on Python + pandas + Matplotlib [J]. Fujian computer. 2017 (11): 2-6.

[9] Liu Yuke, Wang Ping. Statistics and graph output of student achievement data based on Python + pandas + Matplotlib [J]. Fujian computer. 2017 (11): 2-6.

[10] Long Hu, Yang Hui. Data analysis and visualization in the context of big data [J]. Journal of Kaili University. 2016 (03): 1-3.