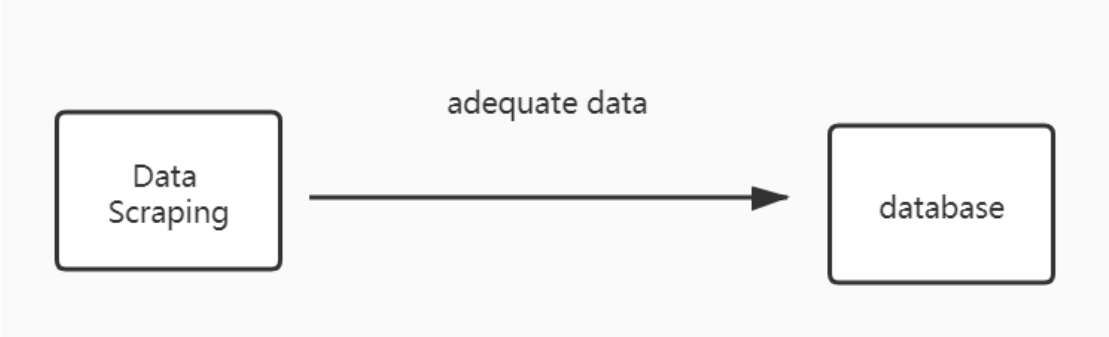
Data Scaping Document

1.Introduction

In order to achieve the function of recommending their favorite songs to users, we decided to get a lot of song information from the music website. After analyzing the user characteristics and information acquisition difficulty of several major music websites, we decided to obtain the data we needed from Netease cloud music website. What we need is song information, but only song list information is available on Netease cloud website. We need to get the song information we need from each song list. This step is the difficulty of program design, we decided to adopt the method of simulator, that is, simulate to open each song list, and then get the information.



2. Data scaping program design

2.1 Data acquisition method

Because the information of songs is nested, it is no longer suitable to use XPath to get data after getting the source code. In this system, selenium and chromdriver are used to obtain data. This is because the requests module is a module that does not completely simulate the browser behavior. It can only crawl to the HTML document information of the web page, and cannot parse and execute CSS and JavaScript code. Therefore, we need to make human judgment. The essence of selenium module is to drive the browser, fully simulate the browser's operation, such as jump, input, click, drop-down, etc., to get the results of web page rendering, and can support a variety of browsers; because selenium parses and executes CSS and JavaScript, its performance is relatively low compared with requests.

1. Selenium installation

   pip install selenium

2. Chromdriver installation

   Download chromdriver.exe , move to the scripts directory in the python installation path.

   Note: the version of chromedriver should correspond to the version of chrome.

3. Selenium selector

   The steps to simulate the browser are as follows:

Request ---> display page ---> search tag --->click the tag, so the key of selenium is how to find the tag in the page, and then trigger the tag event.

(1)Positioning by tag ID attribute:

browser.find_element(By.ID,").send_keys("")

browser.find_element_by_id(").send_keys(")

(2) Positioning by tag name attribute:

browser.find_element_by_name("").send_keys("")

browser.find_element(By.NAME,").send_keys("")

(3) Positioning by tag name

browser.find_element_by_tag_name("").send_keys("")

browser.find_element(By.TAG_NAME, ").send_keys(")

(4) Positioning through CSS search

browser.find_element(By.CSS_SELECTOR, ").send_keys(")

browser.find_element(By.CSS_SELECTOR, ").send_keys(' ')

4. Wait for the element to be loaded

Selenium only simulates the behavior of the browser. However, it takes time for the browser to parse the page (execute CSS, JS). Some elements may take some time to load. In order to ensure that the elements can be found, we must wait.

There are two ways to wait:

Explicit wait: specifies to wait for a tag to finish loading

Implicit wait: wait for all tags to load

2.2 Data content

After discussion in this group, we decided to get the information we need from Netease cloud music. In order to implement the recommendation system, the following information is important:

Song title: as a result of recommendation to users

Songwriter: used to match users' favorite musicians

Duration: show song details

Song list: recommended for users

2.3 Results

Some data are as follows:

	歌名	时间	歌手	专辑名字	歌单名称
2	Fashion Bl	4:37	RHYME SC	Fashion Bl	【日语】听这些就可以走路带风
3	Comme D	3:01	Rina Sawa	SAWAYAM	【日语】听这些就可以走路带风
4	Transcend	3:34	Ovall	Ovall Rewo	【日语】听这些就可以走路带风
5	御伽の街	3:23	DAOKO	御伽の街	【日语】听这些就可以走路带风
6	MAIGO	3:52	SIRUP/Joe	CIY	【日语】听这些就可以走路带风
7	Lost (Fresh	3:03	End of the	Lost (Fresh	【日语】听这些就可以走路带风
8	RUNAWAY	3:45	Nao Kawar	RUNAWAY	【日语】听这些就可以走路带风
9	In Your Arr	3:07	Aiobahn/R	In Your Arr	【日语】听这些就可以走路带风
10	Hurly Burly	5:12	Perfume	Spending a	【日语】听这些就可以走路带风
11	呼吸	4:57	蔡健雅	Tanya 蔡健	你的声音连同气息 穿过秋天漫长的电话线
12	你的样子	5:48	刘莱斯	你的样子	你的声音连同气息 穿过秋天漫长的电话线
13	是想你的声	3:54	傲七爷	是想你的声	你的声音连同气息 穿过秋天漫长的电话线
14	永不失联的	4:19	周兴哲	如果雨之后	你的声音连同气息 穿过秋天漫长的电话线
15	你还好吗	4:34	吴大文	你还好吗	你的声音连同气息 穿过秋天漫长的电话线
16	看见你的声	4:15	陈零九	看见你的声	你的声音连同气息 穿过秋天漫长的电话线
17	心领神会	4:16	莫文蔚	我们在中场	你的声音连同气息 穿过秋天漫长的电话线
18	或是一首等	4:34	田馥甄	或是一首等	你的声音连同气息 穿过秋天漫长的电话线
19	多远都要在	3:37	G.E.M.邓紫	新的心跳	你的声音连同气息 穿过秋天漫长的电话线
20	秋海棠	3:44	激激limpic	秋海棠	你的声音连同气息 穿过秋天漫长的电话线
21	Dance Like	3:02	Iggy Azale	Dance Like	街头扮酷指南	小心别被节奏带跑偏
22	imma	2:03	bbno$/Ler	imma	街头扮酷指南	小心别被节奏带跑偏
23	Baggin'	3:17	Marshmell	Baggin'	街头扮酷指南	小心别被节奏带跑偏
24	Endorphin	3:25	tobi lou	Endorphin	街头扮酷指南	小心别被节奏带跑偏
25	LOCKED U	3:23	6ix9ine/Ak	TattleTales	街头扮酷指南	小心别被节奏带跑偏
26	Lucky Mist	2:55	Vincent/Ali	Lucky Mist	街头扮酷指南	小心别被节奏带跑偏
27	Lie to Me	2:54	BLOWFEVE	Lie to Me	街头扮酷指南	小心别被节奏带跑偏
28	Ring	2:54	T.I./Young	Ring (feat.	街头扮酷指南	小心别被节奏带跑偏
29	Kobe	2:51	Dame D.O	Kobe (feat.	街头扮酷指南	小心别被节奏带跑偏
30	99 Problen	2:17	Hugo	Old Tyme	街头扮酷指南	小心别被节奏带跑偏

3. Conclusion

Using selenium method to obtain data is less efficient than other methods, but it does not need to consider the website protection mechanism. Any information we see on the web can be obtained in this way.