依赖库配置

In [1]:
```
import pandas as pd
import numpy as np
```

加载爬取的所有数据，进行预处理，保证数据的干净，尽量没有噪音

初步查看，数据，数据含义见表头

In [2]:
```
data = pd.read_csv('E:\爬虫\jjwx\jjwxc.csv') # 需要正确的路径
data
```

Out[2]:

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
0	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	原创-言情-近代现代-爱情	轻松	完结	372127	15400449024	2018-07-24 23:08:19	NaN
1	Zoody	oneauthor.php?authorid=3310540	云边咖啡馆	onebook.php?novelid=5095598	原创-言情-近代现代-爱情	轻松	完结	87661	384190016	2020-08-16 06:46:34	package
2	竹已	oneauthor.php?authorid=1938546	难哄	onebook.php?novelid=4001734	原创-言情-近代现代-爱情	轻松	完结	389894	14779252736	2019-01-21 00:29:54	NaN
3	栖见	oneauthor.php?authorid=585107	白日梦我	onebook.php?novelid=3525826	原创-言情-近代现代-爱情	正剧	完结	427269	9028750336	2018-03-16 13:21:32	NaN

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
4	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	原创-言情-近代现代-爱情	轻松	完结	372127	15391267840	2018-07-24 23:08:19	vip
...	...	...	...	...	...	...	...	...	...	...	...
127081	河鳝的脑花	oneauthor.php?authorid=2213495	【全职/恋与】叶修x白起 等你在风里	onebook.php?novelid=3500790	衍生-纯爱-近代现代-东方衍生	正剧	连载	317075	12327059	2018-02-21 17:10:15	free
127082	未有雨	oneauthor.php?authorid=2037159	心于长熙	onebook.php?novelid=4940602	原创-纯爱-近代现代-游戏	正剧	完结	40439	14887185	2020-06-28 22:51:22	free
127083	明蒿	oneauthor.php?authorid=1872296	乞人轶事	onebook.php?novelid=2952730	原创-纯爱-近代现代-爱情	正剧	完结	95528	7041752	2016-10-10 21:31:57	free
127084	聪明机智菜菜菜	oneauthor.php?authorid=1351301	西游记之九世一生	onebook.php?novelid=2507295	衍生-纯爱-古色古香-古典衍生	正剧	完结	282775	39535896	2015-07-16 11:56:54	free

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
127085	残阳落暖	oneauthor.php?authorid=642847	我有特殊的侧写技巧	onebook.php?novelid=3554034	衍生-纯爱-近代现代-西方衍生	正剧	暂停	119208	19726530	2018-04-10 22:56:55	free

127086 rows × 11 columns

上表可见，data的shape 为 127086X11,说明有127086条数据，11个字段

预处理第一步：去重（爬取的作品中含有大量重复的作品，因为有些小说既是原创、也是言情）

In [3]:
```
data = data.drop_duplicates()
data = data.drop(index=57295) ## 与表头相同的列，需要删除
data
```

Out[3]:

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
0	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	原创-言情-近代现代-爱情	轻松	完结	372127	15400449024	2018-07-24 23:08:19	NaN
1	Zoody	oneauthor.php?authorid=3310540	云边咖啡馆	onebook.php?novelid=5095598	原创-言情-近代现代-爱情	轻松	完结	87661	384190016	2020-08-16 06:46:34	package

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
2	竹已	oneauthor.php?authorid=1938546	难哄	onebook.php?novelid=4001734	原创-言情-近代现代-爱情	轻松	完结	389894	14779252736	2019-01-21 00:29:54	NaN
3	栖见	oneauthor.php?authorid=585107	白日梦我	onebook.php?novelid=3525826	原创-言情-近代现代-爱情	正剧	完结	427269	9028750336	2018-03-16 13:21:32	NaN
4	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	原创-言情-近代现代-爱情	轻松	完结	372127	15391267840	2018-07-24 23:08:19	vip
...	...	...	...	...	...	...	...	...	...	...	...
127078	千佑Alleycat	oneauthor.php?authorid=2417786	嫌疑人每天都会把侦探气个半死	onebook.php?novelid=6273634	衍生-无CP-近代现代-轻小说	轻松	连载	61162	8173361	2021-10-18 12:58:04	free
127080	闲云向晚	oneauthor.php?authorid=2415488	我做妖王迷弟的那些年	onebook.php?novelid=4011416	衍生-纯爱-古色古香-东方衍生	轻松	连载	163686	23800602	2019-01-27 23:53:35	free

	a_作者	a_作者_url	b_作品	b_作品_url	c_类型	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型
127081	河鳝的脑花	oneauthor.php?authorid=2213495	【全职/恋与】叶修x白起等你在风里	onebook.php?novelid=3500790	衍生-纯爱-近代现代-东方衍生	正剧	连载	317075	12327059	2018-02-21 17:10:15	free
127084	聪明机智菜菜菜	oneauthor.php?authorid=1351301	西游记之九世一生	onebook.php?novelid=2507295	衍生-纯爱-古色古香-古典衍生	正剧	完结	282775	39535896	2015-07-16 11:56:54	free
127085	残阳落暖	oneauthor.php?authorid=642847	我有特殊的侧写技巧	onebook.php?novelid=3554034	衍生-纯爱-近代现代-西方衍生	正剧	暂停	119208	19726530	2018-04-10 22:56:55	free

61548 rows × 11 columns

去重后，shape变为了 61548X11, 只生下了61658条数据

第二步，将所有类型拆分，方便后面做数据分析，观察小说所属类型,并将原**c_**类型删除，因为该列会变为冗余数据列

In [4]:
```
for i in range(1, 4+1):
    data[f'类型_{i}'] = data['c_类型'].apply(lambda x: x.split(r'-')[i-1])
data = data.drop('c_类型', axis=1)
data
```

Out[4]:

	a_作者	a_作者_url	b_作品	b_作品_url	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2
0	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	轻松	完结	372127	15400449024	2018-07-24 23:08:19	NaN	原创	言情

	a_作者	a_作者_url	b_作品	b_作品_url	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2
1	Zoody	oneauthor.php?authorid=3310540	云边咖啡馆	onebook.php?novelid=5095598	轻松	完结	87661	384190016	2020-08-16 06:46:34	package	原创	言情
2	竹已	oneauthor.php?authorid=1938546	难哄	onebook.php?novelid=4001734	轻松	完结	389894	14779252736	2019-01-21 00:29:54	NaN	原创	言情
3	栖见	oneauthor.php?authorid=585107	白日梦我	onebook.php?novelid=3525826	正剧	完结	427269	9028750336	2018-03-16 13:21:32	NaN	原创	言情
4	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	轻松	完结	372127	15391267840	2018-07-24 23:08:19	vip	原创	言情
...	...	...	...	...	...	...	...	...	...	...	...	...
127078	千佑 Alleycat	oneauthor.php?authorid=2417786	嫌疑人每天都会把侦探气个半死	onebook.php?novelid=6273634	轻松	连载	61162	8173361	2021-10-18 12:58:04	free	衍生	无CP
127080	闲云向晚	oneauthor.php?authorid=2415488	我做妖王迷弟的那些年	onebook.php?novelid=4011416	轻松	连载	163686	23800602	2019-01-27 23:53:35	free	衍生	纯爱
127081	河鳝的脑花	oneauthor.php?authorid=2213495	【全职/恋与】叶修x白起等你在风里	onebook.php?novelid=3500790	正剧	连载	317075	12327059	2018-02-21 17:10:15	free	衍生	纯爱
127084	聪明机智菜菜菜	oneauthor.php?authorid=1351301	西游记之九世一生	onebook.php?novelid=2507295	正剧	完结	282775	39535896	2015-07-16 11:56:54	free	衍生	纯爱
127085	残阳落暖	oneauthor.php?authorid=642847	我有特殊的侧写技巧	onebook.php?novelid=3554034	正剧	暂停	119208	19726530	2018-04-10 22:56:55	free	衍生	纯爱

61548 rows × 14 columns

第三步：将作品库类型的值，转为同网页一样的命名，方便数据分析,比如 **free** 对应免费作品，**sp** 对应驻站作品

In [5]:

```
# 定义转换字典
zpk_mapping = {
    'vip': 'VIP作品', 'package': '完结半价/包月',
    'sp': '驻站作品', 'scriptures': '经典作品', 'free': '免费作品'
}
data['作品库类型'] = data['作品库类型'].map(zpk_mapping).fillna('完结作品')
data
```

Out[5]:

	a_作者	a_作者_url	b_作品	b_作品_url	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2	类型_3
0	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	轻松	完结	372127	15400449024	2018-07-24 23:08:19	完结作品	原创	言情	近代现代
1	Zoody	oneauthor.php?authorid=3310540	云边咖啡馆	onebook.php?novelid=5095598	轻松	完结	87661	384190016	2020-08-16 06:46:34	完结半价/包月	原创	言情	近代现代
2	竹已	oneauthor.php?authorid=1938546	难哄	onebook.php?novelid=4001734	轻松	完结	389894	14779252736	2019-01-21 00:29:54	完结作品	原创	言情	近代现代
3	栖见	oneauthor.php?authorid=585107	白日梦我	onebook.php?novelid=3525826	正剧	完结	427269	9028750336	2018-03-16 13:21:32	完结作品	原创	言情	近代现代
4	竹已	oneauthor.php?authorid=1938546	偷偷藏不住	onebook.php?novelid=3676002	轻松	完结	372127	15391267840	2018-07-24 23:08:19	VIP作品	原创	言情	近代现代
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127078	千佑Alleycat	oneauthor.php?authorid=2417786	嫌疑人每天都会把侦探气个半死	onebook.php?novelid=6273634	轻松	连载	61162	8173361	2021-10-18 12:58:04	免费作品	衍生	无CP	近代现代
127080	闲云向晚	oneauthor.php?authorid=2415488	我做妖王迷弟的那些年	onebook.php?novelid=4011416	轻松	连载	163686	23800602	2019-01-27 23:53:35	免费作品	衍生	纯爱	古色古香

	a_作者	a_作者_url	b_作品	b_作品_url	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2	类型_3
127081	河鳝的脑花	oneauthor.php?authorid=2213495	【全职/恋与】叶修x白起等你在风里	onebook.php?novelid=3500790	正剧	连载	317075	12327059	2018-02-21 17:10:15	免费作品	衍生	纯爱	近代现代
127084	聪明机智菜菜菜	oneauthor.php?authorid=1351301	西游记之九世一生	onebook.php?novelid=2507295	正剧	完结	282775	39535896	2015-07-16 11:56:54	免费作品	衍生	纯爱	古色古香
127085	残阳落暖	oneauthor.php?authorid=642847	我有特殊的侧写技巧	onebook.php?novelid=3554034	正剧	暂停	119208	19726530	2018-04-10 22:56:55	免费作品	衍生	纯爱	近代现代

61548 rows × 14 columns

第四步： 将数据类型转为正确的数据类型。爬取的数据被默认为字符串类型，因此需要转换

In [6]:
```
data.dtypes
```

Out[6]:
```
a_作者          object
a_作者_url      object
b_作品          object
b_作品_url      object
d_风格          object
e_进度          object
f_字数          object
g_作品积分        object
h_发表时间        object
作品库类型         object
类型_1          object
类型_2          object
类型_3          object
类型_4          object
dtype: object
```

In [7]:
```
data = data.astype({
    'f_字数': np.int32, 'g_作品积分': np.int64
})
```

In [8]:
```
print(f'最大字数为{data[r"f_字数"].max()}, 最小字数为{data["f_字数"].min()}')
```

最大字数为9352212, 最小字数为10019

In [9]:
```
# 将 h_发表时间转为 datetime 类型
data['h_发表时间'] = pd.to_datetime(data['h_发表时间'])
```

```
## 检查改变后的数据类型
data.dtypes
```

Out[9]:

```
a_作者                    object
a_作者_url                object
b_作品                    object
b_作品_url                object
d_风格                    object
e_进度                    object
f_字数                     int32
g_作品积分                   int64
h_发表时间          datetime64[ns]
作品库类型                    object
类型_1                    object
类型_2                    object
类型_3                    object
类型_4                    object
dtype: object
```

可以看见，字数、积分、发表时间都转为了正确的数据类型

第五步：将作者的链接和作品的链接提取出来，方便后面爬取信息。并删除，方便做数据分析

In [78]:

```
links = data[['a_作者_url', 'b_作品_url']]
data = data.drop(['a_作者_url', 'b_作品_url'], axis=1)
data
```

Out[78]:

	a_作者	b_作品	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2	类型_3	类型_4
0	竹已	偷偷藏不住	轻松	完结	372127	15400449024	2018-07-24 23:08:19	完结作品	原创	言情	近代现代	爱情
1	Zoody	云边咖啡馆	轻松	完结	87661	384190016	2020-08-16 06:46:34	完结半价/包月	原创	言情	近代现代	爱情
2	竹已	难哄	轻松	完结	389894	14779252736	2019-01-21 00:29:54	完结作品	原创	言情	近代现代	爱情
3	栖见	白日梦我	正剧	完结	427269	9028750336	2018-03-16 13:21:32	完结作品	原创	言情	近代现代	爱情
4	竹已	偷偷藏不住	轻松	完结	372127	15391267840	2018-07-24 23:08:19	VIP作品	原创	言情	近代现代	爱情
...	...	...	...	...	...	...	...	...	...	...	...	...
127078	千佑 Alleycat	嫌疑人每天都会把侦探气个半死	轻松	连载	61162	8173361	2021-10-18 12:58:04	免费作品	衍生	无CP	近代现代	轻小说

	a_作者	b_作品	d_风格	e_进度	f_字数	g_作品积分	h_发表时间	作品库类型	类型_1	类型_2	类型_3	类型_4
127080	闲云向晚	我做妖王迷弟的那些年	轻松	连载	163686	23800602	2019-01-27 23:53:35	免费作品	衍生	纯爱	古色古香	东方衍生
127081	河鳝的脑花	【全职/恋与】叶修x白起 等你在风里	正剧	连载	317075	12327059	2018-02-21 17:10:15	免费作品	衍生	纯爱	近代现代	东方衍生
127084	聪明机智菜菜菜	西游记之九世一生	正剧	完结	282775	39535896	2015-07-16 11:56:54	免费作品	衍生	纯爱	古色古香	古典衍生
127085	残阳落暖	我有特殊的侧写技巧	正剧	暂停	119208	19726530	2018-04-10 22:56:55	免费作品	衍生	纯爱	近代现代	西方衍生

61548 rows × 12 columns

In [87]:
```
author_links = pd.DataFrame()
author_links['author_link'] = links['a_作者_url'].unique()
print(f'共有 {author_links.shape[0]} 个作者')
author_links.to_csv(r'E:\爬虫\jjwx\author_links.csv', index=False)
```

```
共有 19650 个作者
```

In [88]:
```
piece_links = pd.DataFrame()
piece_links['piece_link'] = links['b_作品_url'].unique()
print(f'共有 {piece_links.shape[0]} 个作品')
piece_links.to_csv(r'E:\爬虫\jjwx\piece_links.csv', index=False)
```

```
共有 41608 个作品
```

将处理好的结果保存至文件 data 中

In [15]:
```
data['d_风格'] = data['d_风格'].apply(lambda x: x.strip())
```

In [16]:
```
data.to_csv('E:\爬虫\jjwx\data\作品库.csv', index=False)
```

In [ ]:
```
```