In [5]:

```
## 数据操作
import numpy as np
import pandas as pd
pd.set_option('display.max_rows', 100)
pd.set_option('display.unicode.ambiguous_as_wide', True)
pd.set_option('display.unicode.east_asian_width', True)

from tqdm import tqdm
```

In [4]:

```
# 读取文件
data = pd.read_csv('E:\爬虫\jjwx\无差别爬取数据\无差别爬取.csv')
data
```

```
D:\anaconda2022\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns (0,3,4,5,6,9,13,14,15,16,17,19,20,24) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[4]:

	a_作品字数	a_作品状态	a_作者	a_作者所发送红包数	a_小说完本数	a_小说暂停数	a_小说连载数	a_最后更新时间	a_最近更新作品	a_被收藏数	...	b_总点击量	b_文章名称	b_文章当前被收藏数	b_文章积分	b
0	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	39657	[综]尤莉卡的秘密恋爱关系	865	98595808	很生
1	60744	连载	一宫兴子	20	18	0	25	2020-12-13 15:34:56	《[综英美]而那个成功的反派却想要回到过去》	1530	...	22671	[综英美]而那个成功的反派却想要回到过去	591	9550995	很生

	a_作品字数	a_作品状态	a_作者	a_作者所送包数	a_小说完本数	a_小说暂停数	a_小说连载数	a_最后更新时间	a_最近更新作品	a_被收藏数	...	b_总点击量	b_文章名称	b_文章当前被收藏数	b_文章积分	b
2	60744	连载	一宫兴子	20	18	0	25	2020-12-13 15:34:56	《[综英美]而那个成功的反派却想要回到过去》	1530	...	123219	[TSN/ME]当你的脑子里多了一个人的正确解决方法	904	15546017	衍生
3	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	10135	[猫鼠游戏]圆桌骑士和亚瑟王的圣杯	110	90343528	衍生
4	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	17097	[奇幻贵公子/恶灵猎人]生当复来归	263	93998864	衍生
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
320276	146993	连载	写文字的静一	109	1	0	1	2021-06-07 10:40:42	《替嫁假千金每天都想休夫》	52	...	13368	醉扶青山	309	16599650	原创

	a_作品字数	a_作品状态	a_作者	a_作者所送包数	a_小说完本数	a_小说暂停数	a_小说连载数	a_最后更新时间	a_最近更新作品	a_被收藏数	...	b_总点击量	b_文章名称	b_文章当前被收藏数	b_文章积分	b
320277	659250	连载	香酥栗	217	12	0	3	2022-03-14 19:39:39	《重回九零拆迁前》	13225	...	1151192	幼稚园全都重生了，除了.....[九零]	19660	531464704	月创情
320278	2259421	连载	甘缘茶	2792	4	0	7	2022-03-14 23:10:56	《外挂回收临时工[快穿]》	1715	...	217760	穿越未来之这特么是金手指么 摔	1446	53495640	月创爱
320279	86170	连载	天明未羽	0	2	0	3	2013-01-17 18:48:48	《{倩女幽魂}穿成书生的弱鸡生活》	32	...	97016	{倩女幽魂}穿成书生的弱鸡生活	476	6851500	衍生爱香
320280	10732	连载	月落十三月	0	0	0	26	2021-08-20 16:27:38	《cos成真的我有了复数的欧豆豆》	195	...	136023	本该叫做虎杖悠仁的我在失忆期间被冠上了两面宿傩的名字后...	1556	4091129	衍生CP

320281 rows × 27 columns

In [13]:
```
# 删除重复的列，和与表头相同的列
data = data.drop_duplicates().drop(index=22)
data.shape
```

Out[13]:
```
(318676, 27)
```

对比原数据，数据量 320281->318676

In [14]:

```
## 字段，nunique等基本信息
def get_info(df, head=10):
  print(df.shape)
  print('\n', '----------------各列特征信息如下------------------')
  stats = []
  for col in df.columns:
    stats.append(
      (col, df[col].nunique(),
       round(df[col].isnull().sum()*100 / df.shape[0], 3),
       round(df[col].value_counts(normalize=True, dropna=False).values[0]*100, 3),
       df[col].dtype
      )
    )
  stats_df = pd.DataFrame(stats, columns=['特征', '属性个性', '缺失值占比', '最大属性占比', '特征
  print(stats_df.sort_values('缺失值占比', ascending=False).head(head))
```

In [15]:

```
# 获取数据的基本信息
get_info(data, 100)
```

```
(318676, 27)

 ----------------各列特征信息如下------------------
                    特征   属性个性   缺失值占比   最大属性占比 特征类型
11           b_作品视角         7       0.291        55.443   object
12           b_作品风格         7       0.026        58.647   object
18           b_文章名称    305010       0.003         0.061   object
13           b_全文字数    187112       0.001        19.996   object
15           b_总书评数     19960       0.001        20.294   object
25      b_霸王票全站排行    127180       0.001        47.866   object
24           b_营养液数     22149       0.001        31.990   object
23           b_签约状态         2       0.001        86.713   object
22           b_文章进度         3       0.001        49.290   object
21           b_文章类型       452       0.001        21.247   object
20           b_文章积分    277016       0.001         0.072   object
19    b_文章当前被收藏数     33585       0.001         2.338   object
17           b_总点击量    170228       0.001        19.961   object
16        b_总共地雷数量      9512       0.001        37.744   object
26             piece_url    318663       0.001         0.001   object
14  b_前进一名所需地雷数       379       0.001        70.969   object
1            a_作品状态         4       0.000        69.971   object
10           author_url     38072       0.000         0.074   object
9            a_被收藏数      6731       0.000         0.340   object
8         a_最近更新作品     37173       0.000         0.074   object
7         a_最后更新时间     35491       0.000         1.057   object
6           a_小说连载数       150       0.000         7.697   object
5           a_小说暂停数        54       0.000        50.026   object
4           a_小说完本数       185       0.000         6.400   object
3     a_作者所发送红包数      4287       0.000        21.450   object
2                a_作者     38072       0.000         0.074   object
0            a_作品字数     38013       0.000         0.104   object
```

7种作品视角，305010部小说，作品视角缺失率占比0.291，特征类型全部为object,需要合理的转换，比如总书评数，应该为int,整型

In [67]:

```
# 数据转换
data['b_霸王票全站排行'] = data['b_霸王票全站排行'].replace(r'暂无', np.nan)
int_cols = ['a_作品字数', 'a_作者所发送红包数', 'a_小说完本数', 'a_小说暂停数', 'a_小说连载数', 'a_
            'b_总书评数', 'b_总共地雷数量', 'b_总点击量', 'b_文章当前被收藏数', 'b_文章积分', 'b_营
dtypes_ = {i: np.int64 for i in int_cols}
```

```
data = data.astype(dtypes_)
data['a_最后更新时间'] = pd.to_datetime(data['a_最后更新时间'])
```

In [74]:
```
data.dtypes
```

Out[74]:
```
a_作品字数                      int64
a_作品状态                     object
a_作者                         object
a_作者所发送红包数              int64
a_小说完本数                    int64
a_小说暂停数                    int64
a_小说连载数                    int64
a_最后更新时间          datetime64[ns]
a_最近更新作品                 object
a_被收藏数                      int64
author_url                     object
b_作品视角                     object
b_作品风格                     object
b_全文字数                      int64
b_前进一名所需地雷数            int64
b_总书评数                      int64
b_总共地雷数量                  int64
b_总点击量                      int64
b_文章名称                     object
b_文章当前被收藏数              int64
b_文章积分                      int64
b_文章类型                     object
b_文章进度                     object
b_签约状态                     object
b_营养液数                      int64
b_霸王票全站排行               object
piece_url                      object
dtype: object
```

所有字段被转换成了正确的类型，正确的类型有助于数据的分析

In [78]:
```
data.head()
```

Out[78]:

	a_作品字数	a_作品状态	a_作者	a_作者所发送红包数	a_小说完本数	a_小说暂停数	a_小说连载数	a_最后更新时间	a_最近更新作品	a_被收藏数	...	b_总点击量	b_文章名称	b_文章当前被收藏数	b_文章积分	b_文章类型	b_文章进度	b_签约状态	b_营养液数
0	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	39657	[综]尤莉卡的秘密恋爱关系	865	98595808	衍生-言情-架空历史-东方衍生	连载	已签约	6

	a_作品字数	a_作品状态	a_作者	a_作者所发送红包数	a_小说完本数	a_小说暂停数	a_小说连载数	a_最后更新时间	a_最近更新作品	a_被收藏数	...	b_总点击量	b_文章名称	b_文章当前被收藏数	b_文章积分	b_文章类型	b_文章进度	b_签约状态	b_章节
1	60744	连载	一宫兴子	20	18	0	25	2020-12-13 15:34:56	《[综英美]而那个成功的反派却想要回到过去》	1530	...	22671	[综英美]而那个成功的反派却想要回到过去	591	9550995	衍生-纯爱-架空历史-西方衍生	连载	已签约	43
2	60744	连载	一宫兴子	20	18	0	25	2020-12-13 15:34:56	《[综英美]而那个成功的反派却想要回到过去》	1530	...	123219	[TSN/ME]当你的脑子里多了一个人的正确解决方法	904	15546017	衍生-纯爱-近代现代-西方衍生	完结	已签约	29
3	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	10135	[猫鼠游戏]圆桌骑士和亚瑟王的圣杯	110	90343528	衍生-言情-架空历史-西方衍生	完结	已签约	1
4	157033	连载	挖坑不填	40	121	25	173	2022-02-20 22:07:26	《[主名柯]关于我有三话存稿但还是被催稿是不是哪里搞错了》...	25490	...	17097	[奇幻贵公子/恶灵猎人]生当复来归	263	93998864	衍生-言情-架空历史-东方衍生	完结	已签约	4

5 rows × 27 columns

In [76]:

```
# 将结果保存，用于数据分析,并覆盖源数据
data.to_csv('E:\爬虫\jjwx\无差别爬取数据\无差别爬取.csv', index=False)
```