# **Pandas数据处理案例|一个简单的电影推荐系统**

这里主要根据电影主题实现简单电影推荐

## 1.数据集

本文使用的数据集来自于https://grouplens.org/datasets/movielens/，上面有多种版本，主要区别在于数据集的大小（即收录的电影和用户信息的多少），这里使用的是最小的一个数据集：



## 2.数据初探

数据集里包含了4个文件，本文只用到了其中的两个:电影信息movies.csv,用户对电影的评分ratings.csv,先来看看这两个文件长什么样：

import pandas as pd

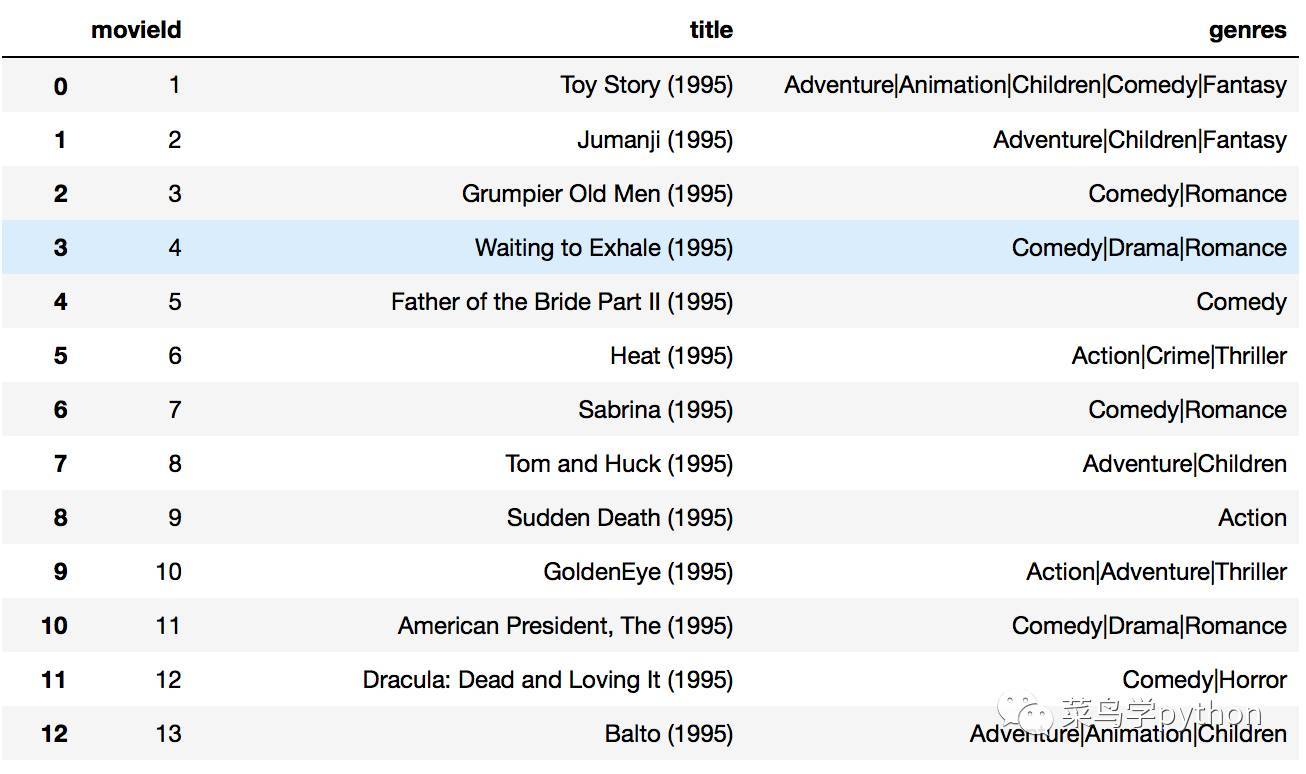
movies\_data=pd.read\_csv('movies.csv')

ratings\_data=pd.read\_csv('ratings.csv')

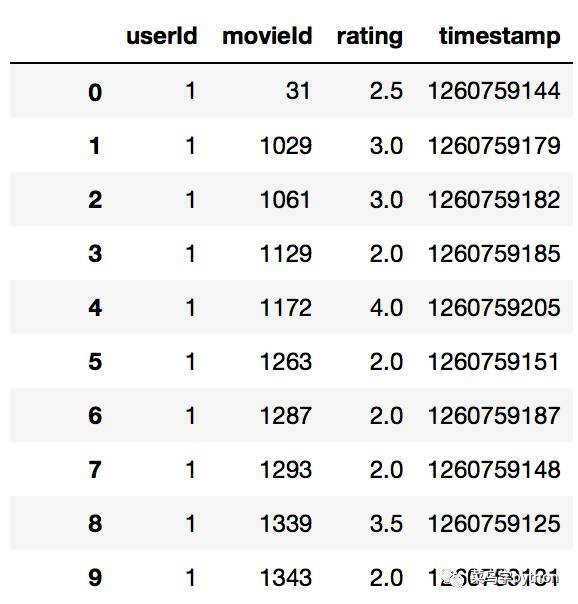
### 1).看一下movies\_data里的数据

print (movies\_data)

如何取前面top5，如何取？



### .看一下用户评分的数据：



显然，如果单纯从单个文件里来分析的话似乎得不到什么有用或者是直观的信息，但是我们可以把这两个数据帧（DataFrame）结合起来，合成新的，把用户评分和电影信息结合起来，方便我们后续的分析处理。

请大家思考，如何合并？

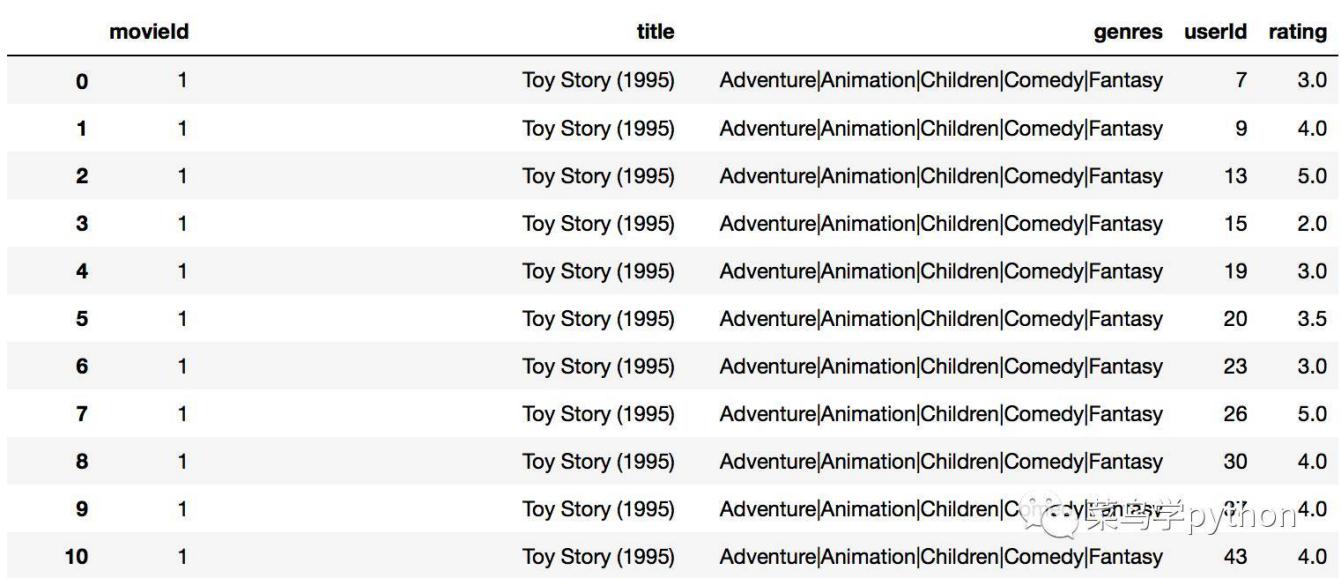
### 3).合并两个数据集

#合并

data=pd.merge(movies\_data,ratings\_data)

#删除列，可写可不写，不会产生影响

data.drop('timestamp',1,inplace=True)



好了，完成了这一步，我们把重要的信息都结合在一起了，可以开始我们真正的数据分析或探索了。

接着看看那些电影收到的评分最多，这里注意-可以使用value\_counts()的pandas函数。

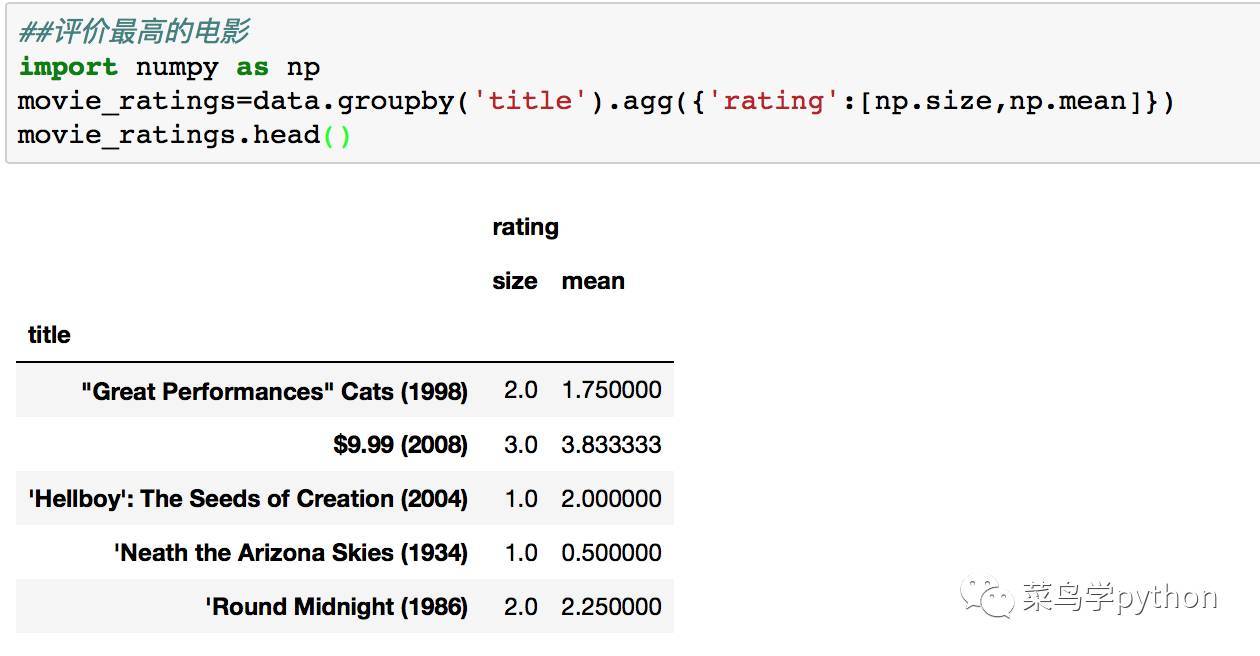
### 4).看看评价最多的20名电影



阿甘正传、低俗小说、肖申克的救赎、沉默的羔羊等电影

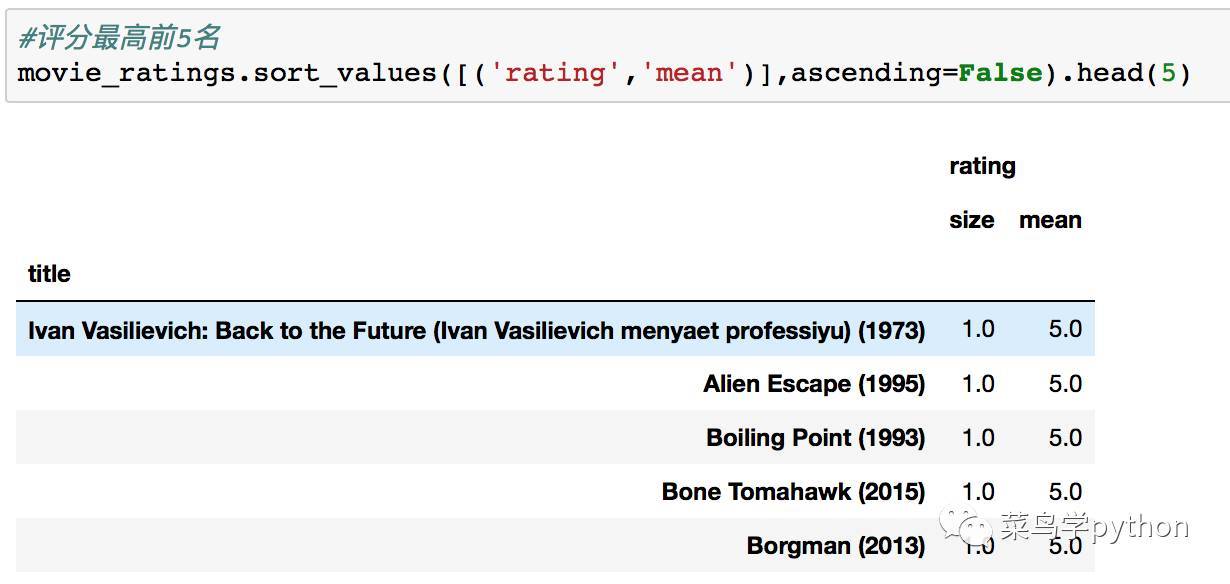
看完评价最多的电影，但大多数人更加关心的是电影究竟好不好看，于是乎，我们就把目光转向评分这里. 那接着就来统计下用户对每一部电影的平均评分吧

### 5).电影的平均评分



其中size是每部电影参评人数，mean是平均分.

### 6).继续找一下评分最高的top5：



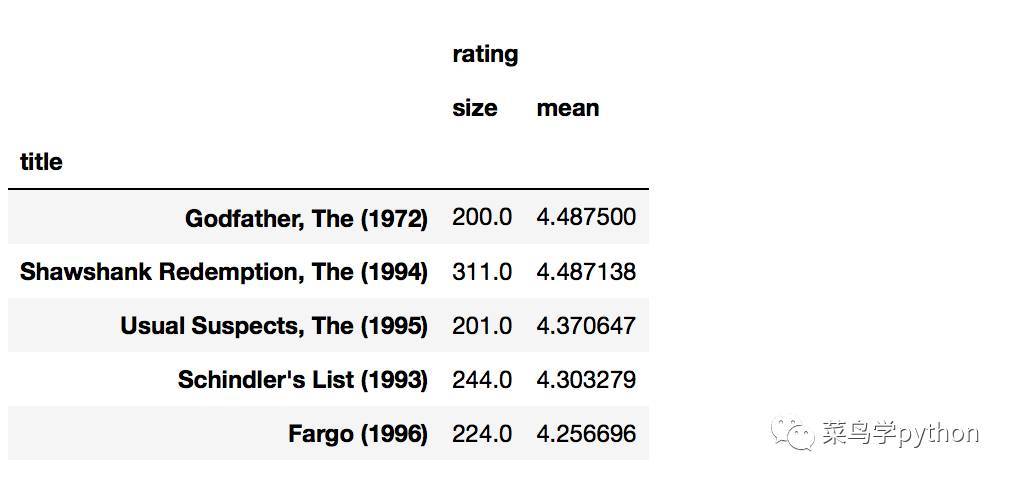
可以看到，虽然评分是5分，但是……评分只有1人啊，这样绝不是我们想看到的，因为不是很客观，于是，我们再来改进下

这次我们把评价人数大于150人的电影找出来，在进行统计：

movie\_ratings[movie\_ratings[ 'rating'][ 'size'] >=150].

sort\_values([( 'rating','mean')] ,ascending=False).head()

>>



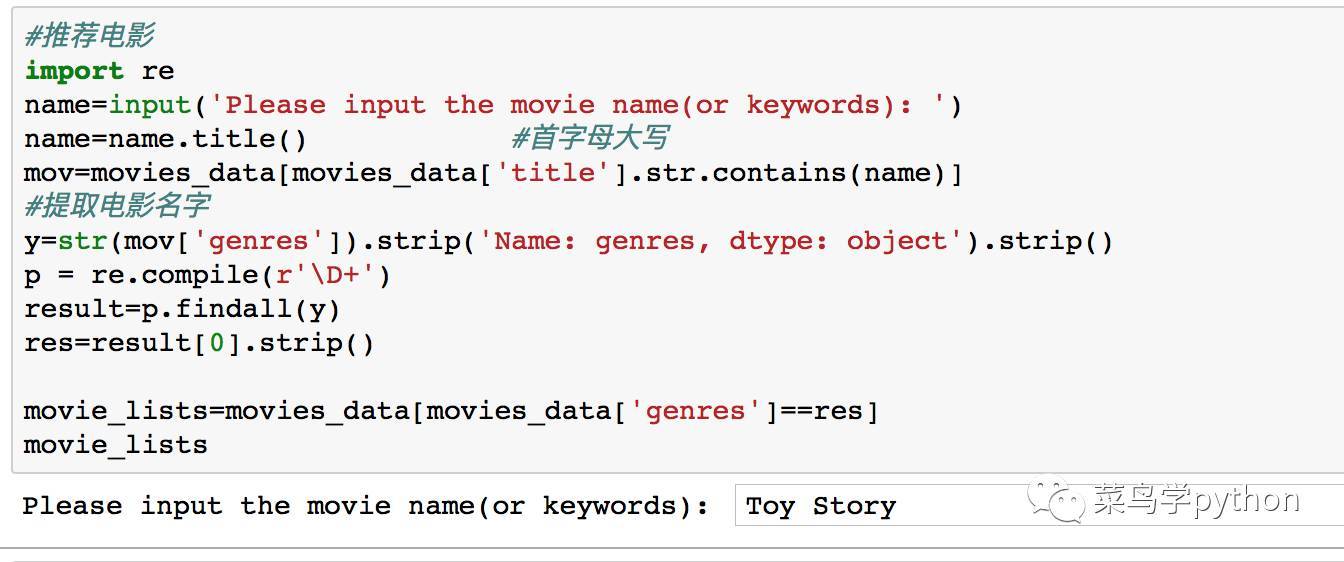
如此，top5分别是：教父、肖申克的救赎、非常嫌疑犯、辛德勒的名单、冰血暴，都是经典电影。

# 3.简单的推荐系统

当我们看完一部觉得好看电影，是不是会有这样的感觉，立马看有没有续集，没有的话就会找相似的电影。

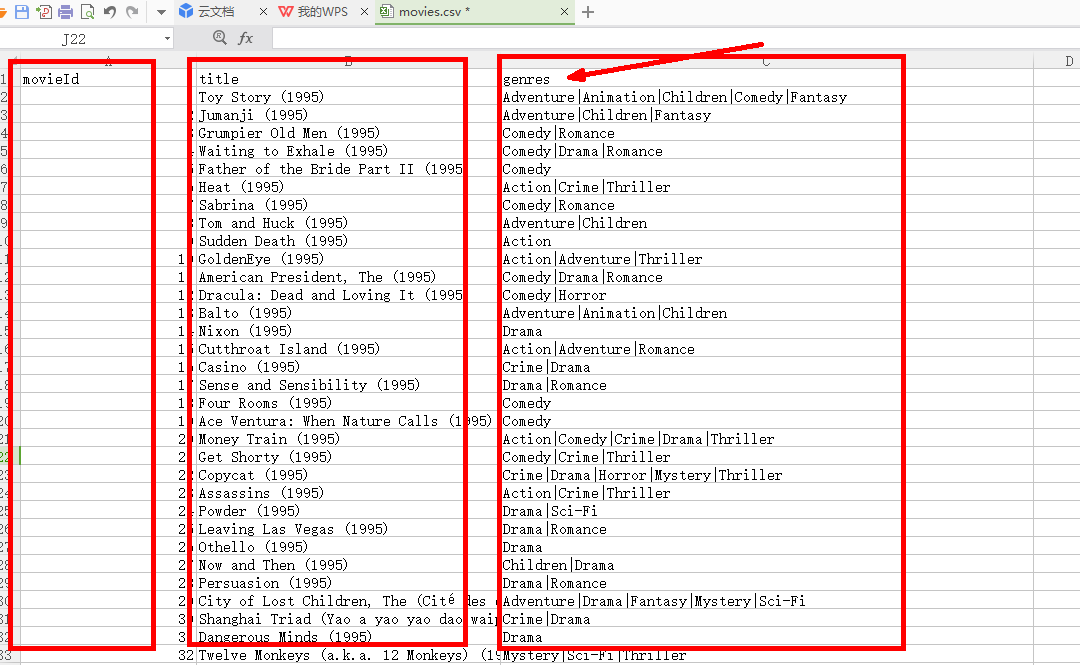
是的，这里是基于这样思想，当我们输入一部电影，就找***同样的类型的电影***，然后返回给我们结果。

用代码实现也不多，主要思想是：我们输入电影名或关键词，我们**就从数据帧里匹配，然后找到这部电影的类型，再找出与其相同类型的电影，**最后返回结果，具体代码为：

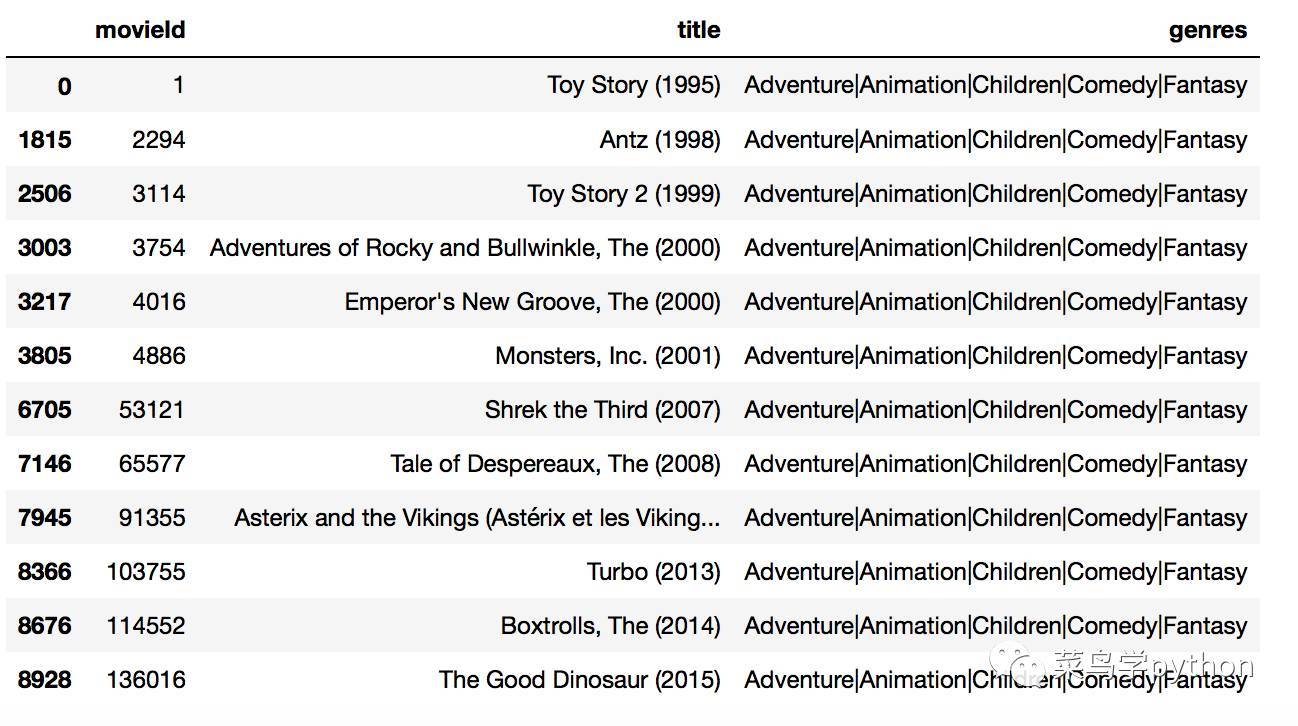


注意：

movieId, title, genres  
movieId:每部电影的id  
title:电影的标题  
genres:电影的类别



>>看看和Toy Story(玩具总动员)类似的动画片有哪些



>>比如找一下星球大战类似的电影：

