

实例 2：电影信息提取

在“电影.txt”文件中，包含电影排名、电影名称、评分、类别、演员等信息。虽然该文件中数据杂乱，不能很清晰地了解全部数据信息，但是每种数据都有相对应的标签，例如 title 标签对应着电影名称、rating 标签对应着电影评分、rank 标签对应着电影排名。为了能够提取指定的数据信息，可以使用正则表达式。图 1 所示为“电影.txt”文件中数据。



图1 电影.txt

本实例要求编写程序，实现提取排名前 20 的电影名称与评分信息的功能。

实例目标

- 掌握 re 模块中 compile() 函数的使用
- 掌握 re 模块中 findall() 方法的使用

实例分析

在使用正则表达式匹配之前，我们需要先读取“电影.txt”文件中的数据，将读取的数据作为正则表达式待匹配的目标文本对象。由于实例要求提取排名前 20 的电影名称及评分，所以需要编写符合要求的正则表达式，具体如下：

- 电影名称对应的正则表达式为 `title": "(.*)"。`
- 电影评分对应的正则表达式为 `rating": \["(.*)", "(d+)"。`
- 电影排名对应的正则表达式为 `rank": (\d+)。`

代码实现

```
import re

data = open("电影.txt", 'r', encoding='utf-8').read()

# 定义正则表达式分别匹配电影名称/评分/排名

title = r'"title": "(.*?)"'

rating = r'"rating": \["(.*?)", "\d+"\]'

rank = r'"rank": (\d+)'

# 预编译正则表达式

pattern_title = re.compile(title)

pattern_rating = re.compile(rating)

pattern_rank = re.compile(rank)

# 查找全部匹配的数据（返回列表）

data_title = pattern_title.findall(data)

data_rating = pattern_rating.findall(data)

data_rank = pattern_rank.findall(data)

for i in range(20):

    print("排名: ", data_rank[i] + "\t\t" + "电影名: " + data_title[i]

        + "\t\t" + "评分: " + data_rating[i])
```

以上代码首先导入了 `re` 模块，打开“电影.txt”文件并将读取的数据赋值给 `data`，然后编写了分别匹配电影名称、电影评分、电影排名的正则表达式 `title`、`rating`、`rank`，使用 `compile()` 函数预编译正则表达式，通过 `findall()` 方法查找匹配的内容，最后遍历输出前 20 条数据，即排名前 20 的电影信息。

代码测试

运行代码，控制台输出结果如下：

排名: 1	电影名: 肖申克的救赎	评分: 9.6
排名: 2	电影名: 霸王别姬	评分: 9.6
排名: 3	电影名: 控方证人	评分: 9.6
排名: 4	电影名: 伊丽莎白	评分: 9.6
排名: 5	电影名: 美丽人生	评分: 9.5

排名： 6	电影名： 辛德勒的名单	评分： 9.5
排名： 7	电影名： 这个杀手不太冷	评分： 9.4
排名： 8	电影名： 阿甘正传	评分： 9.4
排名： 9	电影名： 十二怒汉	评分： 9.4
排名： 10	电影名： 泰坦尼克号 3D 版	评分： 9.4
排名： 11	电影名： 背靠背，脸对脸	评分： 9.4
排名： 12	电影名： 灿烂人生	评分： 9.4
排名： 13	电影名： 茶馆	评分： 9.4
排名： 14	电影名： 十二怒汉	评分： 9.4
排名： 15	电影名： 巴黎圣母院	评分： 9.4
排名： 16	电影名： 控方证人	评分： 9.4
排名： 17	电影名： 罗密欧与朱丽叶	评分： 9.4
排名： 18	电影名： 盗梦空间	评分： 9.3
排名： 19	电影名： 泰坦尼克号	评分： 9.3
排名： 20	电影名： 千与千寻	评分： 9.3