准备工作

目标：抓取豆瓣电影Top250 电影信息

1.引入常用的第三方包stringr 和 rvest 。

```
install.packages("rvest")  # 处理HTML
install.packages("stringr") # 处理字符串
library(rvest)
library(stringr)
```

2.测试XPath 并获取每个电影的具体信息

- 获取电影名称

```
web=read_html("https://movie.douban.com/top250?start=0&filter=")  # 读取页面
web %>% html_nodes(xpath="//span[@class='title'][1]") %>% html_text()
```

- 结果如下：

```
 [1] "肖申克的救赎"         "霸王别姬"           "阿甘正传"           "这个杀手不
太冷"
 [5] "泰坦尼克号"           "美丽人生"           "千与千寻"           "辛德勒的名
单"
 [9] "盗梦空间"             "忠犬八公的故事"     "楚门的世界"         "星际穿越"
[13] "海上钢琴师"           "三傻大闹宝莱坞"     "机器人总动员"       "放牛班的春
天"
[17] "无间道"               "大话西游之大圣娶亲" "疯狂动物城"         "熔炉"
[21] "教父"                 "当幸福来敲门"       "龙猫"               "控方证人"
[25] "怦然心动"
```

- 获取某电影其他元素的XPath如下：

```
# 获取 导演，主演，年份，国家，电影分类
web %>% html_nodes(xpath="//div[@class='info']//div[@class='bd']//p[@class]
[1]") %>% html_text()
# 获取 电影评分
web %>%
html_nodes(xpath="//div[@class='info']//div[@class='bd']//div//span[@class='r
ating_num']") %>% html_text()
# 获取 电影评价人数
web %>%
html_nodes(xpath="//div[@class='info']//div[@class='bd']//div//span[last()]")
%>% html_text()
# 获取 电影详情链接
web %>% html_nodes(xpath="//div[@class='info']//div[@class='hd']//a[1]") %>%
html_attr('href')
```

开始抓取数据

```
if (!require("stringr"))
  install.packages("stringr")
if (!require("rvest"))
  install.packages("rvest")

df_douban_250 = data.frame()
for(page_number in 0:9){
  content = read_html(str_c(
    "https://movie.douban.com/top250?start=", 25*page_number, "&filter="))

  middle_text = content %>% html_nodes(
    xpath="//div[@class='info']//div[@class='bd']//p[@class][1]") %>%
html_text()
  middle_text_line_1 = sapply(str_split(str_trim(middle_text), '\n'), "[[",
1)  #[1] "导演: 弗兰克·德拉邦特 Frank Darabont   主演: 蒂姆·罗宾斯 Tim Robbins
/..."
  middle_text_line_2 = sapply(str_split(str_trim(middle_text), '\n'), "[[",
2)  #[1] "1994 / 美国 / 犯罪 剧情"

  m_director = sapply(str_split(middle_text_line_1, "[\u00A0]{3}"), "[[", 1)
# 注意:   字符需要用unicode去匹配
  m_actor = str_split(middle_text_line_1, "[\u00A0]{3}")
  for(i in 1:length(m_actor)){
    if(length(m_actor[[i]])<2){m_actor[[i]][2]=''}  # 补齐操作
  }
```

```
  m_actor = sapply(m_actor, "[[", 2)

  list_year = str_split(middle_text_line_2, "/")
  for(i in 1:length(list_year)){
    if(length(list_year[[i]])>3){  # 处理No.054的异常数据
      list_year[[i]][3] = list_year[[i]][length(list_year[[i]])]
      list_year[[i]][2] = list_year[[i]][length(list_year[[i]])-1]
    }
  }
  m_year = str_trim(sapply(list_year, "[[", 1))
  m_country = str_trim(sapply(list_year, "[[", 2))
  m_type = str_trim(sapply(list_year, "[[", 3))

  m_name = content %>% html_nodes(
    xpath="//span[@class='title'][1]") %>% html_text()
  m_score = content %>% html_nodes(

xpath="//div[@class='info']//div[@class='bd']//div//span[@class='rating_num']
") %>% html_text()
  m_commit_count = content %>% html_nodes(
    xpath="//div[@class='info']//div[@class='bd']//div//span[last()]") %>%
html_text()
  m_commit_count = str_sub(m_commit_count, end=-4)
  m_detail_url = content %>% html_nodes(
    xpath="//div[@class='info']//div[@class='hd']//a[1]") %>%
html_attr('href')

  df_one_page = data.frame(
    m_name,  # 电影名称
    m_director,  # 导演
    m_actor,  # 主演
    m_year,  # 上映年份
    m_country,  # 国家
    m_type,  # 所属分类
    m_score,  # 评分
    m_commit_count,  # 评论人数
    m_detail_url  # 详细页面
  )
  df_douban_250 = rbind(df_douban_250, df_one_page)
}
```

整理数据

异常数据处理：

- No.054《大闹天宫》年份数据异常`1961(中国大陆) / 1964(中国大陆) / 1978(中国大陆) / 中国大陆 / 剧情 动画 奇幻 古装`

  需要从右向左按 / 字符split 3个字符串，才能符合其他的电影信息的格式。

  由于stringr文档中的str_split不支持从右向左，因此仅取了第一个年份，丢掉了第二和第三个。

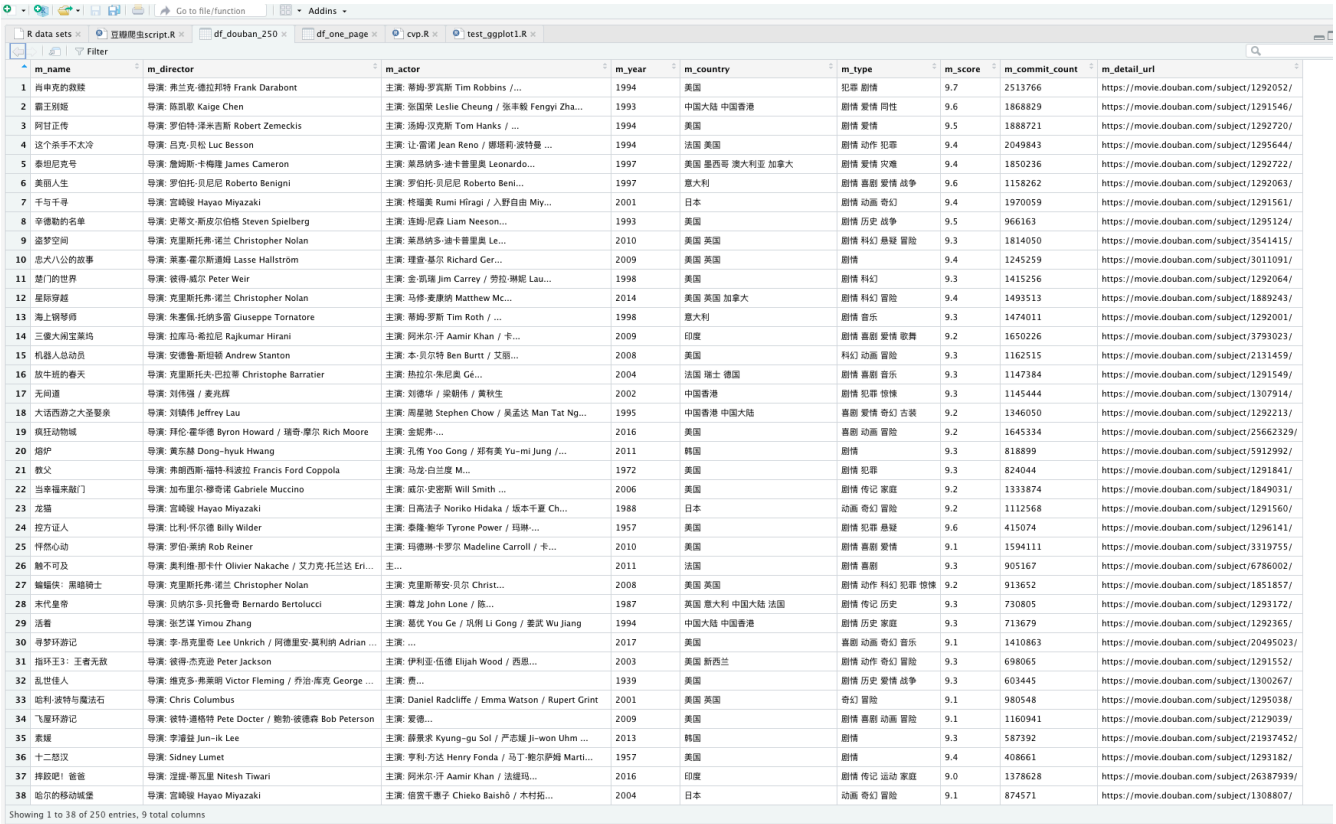
需要补充的数据：

- No.063《窃听风云》 没有主演数据`导演: 弗洛里安·亨克尔·冯·多纳斯马尔克 Florian Henckel von Donnersmarck  &n...`
- No.235《黑客帝国2》没有主演数据`导演: 拉娜·沃卓斯基 Lana Wachowski / 莉莉·沃卓斯基 Lilly Wachowski  ...`
- No.207《初恋这件小事》 没有主演数据`导演: 普特鹏·普罗萨卡·那·萨克那卡林 Puttipong Promsaka Na Sakolnakorn / 华森·波克彭...`

  还有一些电影的主演在排行榜页面上没有，需要进一步完善。

查看抓取的数据：

使用`View(df_douban_250)`命令查看如下



	m_name	m_director	m_actor	m_year	m_country	m_type	m_score	m_commit_count	m_detail_url
1	肖申克的救赎	导演: 弗兰克·德拉邦特 Frank Darabont	主演: 蒂姆·罗宾斯 Tim Robbins /...	1994	美国	犯罪 剧情	9.7	2513766	https://movie.douban.com/subject/1292052/
2	霸王别姬	导演: 陈凯歌 Kaige Chen	主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...	1993	中国大陆 中国香港	剧情 爱情 同性	9.6	1868829	https://movie.douban.com/subject/1291546/
3	阿甘正传	导演: 罗伯特·泽米吉斯 Robert Zemeckis	主演: 汤姆·汉克斯 Tom Hanks / ...	1994	美国	剧情 爱情	9.5	1888721	https://movie.douban.com/subject/1292720/
4	这个杀手不太冷	导演: 吕克·贝松 Luc Besson	主演: 让·雷诺 Jean Reno / 娜塔莉·波特曼 ...	1994	法国 美国	剧情 动作 犯罪	9.4	2049843	https://movie.douban.com/subject/1295644/
5	泰坦尼克号	导演: 詹姆斯·卡梅隆 James Cameron	主演: 莱昂纳多·迪卡普里奥 Leonardo...	1997	美国 墨西哥 澳大利亚 加拿大	剧情 爱情 灾难	9.4	1850236	https://movie.douban.com/subject/1292722/
6	美丽人生	导演: 罗伯托·贝尼尼 Roberto Benigni	主演: 罗伯托·贝尼尼 Roberto Beni...	1997	意大利	剧情 喜剧 爱情 战争	9.6	1158262	https://movie.douban.com/subject/1292063/
7	千与千寻	导演: 宫崎骏 Hayao Miyazaki	主演: 柊瑠美 Rumi Hīragi / 入野自由 Miy...	2001	日本	剧情 动画 奇幻	9.4	1970059	https://movie.douban.com/subject/1291561/
8	辛德勒的名单	导演: 史蒂文·斯皮尔伯格 Steven Spielberg	主演: 连姆·尼森 Liam Neeson...	1993	美国	剧情 历史 战争	9.5	966163	https://movie.douban.com/subject/1295124/
9	盗梦空间	导演: 克里斯托弗·诺兰 Christopher Nolan	主演: 莱昂纳多·迪卡普里奥 Le...	2010	美国 英国	剧情 科幻 悬疑 冒险	9.3	1814050	https://movie.douban.com/subject/3541415/
10	忠犬八公的故事	导演: 莱塞·霍尔斯道姆 Lasse Hallström	主演: 理查·基尔 Richard Ger...	2009	美国 英国	剧情	9.4	1245259	https://movie.douban.com/subject/3011091/
11	楚门的世界	导演: 彼得·威尔 Peter Weir	主演: 金·凯瑞 Jim Carrey / 劳拉·琳妮 Lau...	1998	美国	剧情 科幻	9.3	1415256	https://movie.douban.com/subject/1292064/
12	星际穿越	导演: 克里斯托弗·诺兰 Christopher Nolan	主演: 马修·麦康纳 Matthew Mc...	2014	美国 英国 加拿大	剧情 科幻 冒险	9.4	1493513	https://movie.douban.com/subject/1889243/
13	海上钢琴师	导演: 朱塞佩·托纳多雷 Giuseppe Tornatore	主演: 蒂姆·罗斯 Tim Roth / ...	1998	意大利	剧情 音乐	9.3	1474011	https://movie.douban.com/subject/1292001/
14	三傻大闹宝莱坞	导演: 拉库马·希拉尼 Rajkumar Hirani	主演: 阿米尔·汗 Aamir Khan / 卡...	2009	印度	剧情 喜剧 爱情 歌舞	9.2	1650226	https://movie.douban.com/subject/3793023/
15	机器人总动员	导演: 安德鲁·斯坦顿 Andrew Stanton	主演: 本·贝尔特 Ben Burtt / 艾丽...	2008	美国	科幻 动画 冒险	9.3	1162515	https://movie.douban.com/subject/2131459/
16	放牛班的春天	导演: 克里斯托夫·巴拉蒂 Christophe Barratier	主演: 热拉尔·朱尼奥 Gé...	2004	法国 瑞士 德国	剧情 喜剧 音乐	9.3	1147384	https://movie.douban.com/subject/1291549/
17	无间道	导演: 刘伟强 / 麦兆辉	主演: 刘德华 / 梁朝伟 / 黄秋生	2002	中国香港	剧情 犯罪 惊悚	9.3	1145444	https://movie.douban.com/subject/1307914/
18	大话西游之大圣娶亲	导演: 刘镇伟 Jeffrey Lau	主演: 周星驰 Stephen Chow / 吴孟达 Man Tat Ng...	1995	中国香港 中国大陆	喜剧 爱情 奇幻 古装	9.2	1346050	https://movie.douban.com/subject/1292213/
19	疯狂动物城	导演: 拜伦·霍华德 Byron Howard / 瑞奇·摩尔 Rich Moore	主演: 金妮弗·...	2016	美国	喜剧 动画 冒险	9.2	1645334	https://movie.douban.com/subject/25662329/
20	熔炉	导演: 黄东赫 Dong-hyuk Hwang	主演: 孔侑 Yoo Gong / 郑有美 Yu-mi Jung /...	2011	韩国	剧情	9.3	818899	https://movie.douban.com/subject/5912992/
21	教父	导演: 弗朗西斯·福特·科波拉 Francis Ford Coppola	主演: 马龙·白兰度 M...	1972	美国	剧情 犯罪	9.3	824044	https://movie.douban.com/subject/1291841/
22	当幸福来敲门	导演: 加布里尔·穆奇诺 Gabriele Muccino	主演: 威尔·史密斯 Will Smith ...	2006	美国	剧情 传记 家庭	9.2	1333874	https://movie.douban.com/subject/1849031/
23	龙猫	导演: 宫崎骏 Hayao Miyazaki	主演: 日高法子 Noriko Hidaka / 坂本千夏 Ch...	1988	日本	动画 奇幻 冒险	9.2	1112568	https://movie.douban.com/subject/1291560/
24	控方证人	导演: 比利·怀尔德 Billy Wilder	主演: 泰隆·鲍华 Tyrone Power / 玛琳·...	1957	美国	剧情 犯罪 悬疑	9.6	415074	https://movie.douban.com/subject/1296141/
25	怦然心动	导演: 罗伯·莱纳 Rob Reiner	主演: 玛德琳·卡罗尔 Madeline Carroll / 卡...	2010	美国	剧情 喜剧 爱情	9.1	1594111	https://movie.douban.com/subject/3319755/
26	触不可及	导演: 奥利维·那卡什 Olivier Nakache / 艾力克·托兰达 Eri...	主...	2011	法国	剧情 喜剧	9.3	905167	https://movie.douban.com/subject/6786002/
27	蝙蝠侠：黑暗骑士	导演: 克里斯托弗·诺兰 Christopher Nolan	主演: 克里斯蒂安·贝尔 Christ...	2008	美国 英国	剧情 动作 科幻 犯罪 惊悚	9.2	913652	https://movie.douban.com/subject/1851857/
28	末代皇帝	导演: 贝纳尔多·贝托鲁奇 Bernardo Bertolucci	主演: 尊龙 John Lone / 陈...	1987	英国 意大利 中国大陆 法国	剧情 传记 历史	9.3	730805	https://movie.douban.com/subject/1293172/
29	活着	导演: 张艺谋 Yimou Zhang	主演: 葛优 You Ge / 巩俐 Li Gong / 姜武 Wu Jiang	1994	中国大陆 中国香港	剧情 历史 家庭	9.3	713679	https://movie.douban.com/subject/1292365/
30	寻梦环游记	导演: 李·昂克里奇 Lee Unkrich / 阿德里安·莫利纳 Adrian ...	主演: ...	2017	美国	喜剧 动画 奇幻 音乐	9.1	1410863	https://movie.douban.com/subject/20495023/
31	指环王3：王者无敌	导演: 彼得·杰克逊 Peter Jackson	主演: 伊利亚·伍德 Elijah Wood / 西恩...	2003	美国 新西兰	剧情 动作 奇幻 冒险	9.3	698065	https://movie.douban.com/subject/1291552/
32	乱世佳人	导演: 维克多·弗莱明 Victor Fleming / 乔治·库克 George ...	主演: 费...	1939	美国	剧情 历史 爱情 战争	9.3	603445	https://movie.douban.com/subject/1300267/
33	哈利·波特与魔法石	导演: Chris Columbus	主演: Daniel Radcliffe / Emma Watson / Rupert Grint	2001	美国 英国	奇幻 冒险	9.1	980548	https://movie.douban.com/subject/1295038/
34	飞屋环游记	导演: 彼特·道格特 Pete Docter / 鲍勃·彼德森 Bob Peterson	主演: 爱德...	2009	美国	剧情 喜剧 动画 冒险	9.1	1160941	https://movie.douban.com/subject/2129039/
35	素媛	导演: 李濬益 Jun-ik Lee	主演: 薛景求 Kyung-gu Sol / 严志媛 Ji-won Uhm ...	2013	韩国	剧情	9.3	587392	https://movie.douban.com/subject/21937452/
36	十二怒汉	导演: Sidney Lumet	主演: 亨利·方达 Henry Fonda / 马丁·鲍尔萨姆 Marti...	1957	美国	剧情	9.4	408661	https://movie.douban.com/subject/1293182/
37	摔跤吧！爸爸	导演: 涅提·蒂瓦里 Nitesh Tiwari	主演: 阿米尔·汗 Aamir Khan / 法缇玛...	2016	印度	剧情 传记 运动 家庭	9.0	1378628	https://movie.douban.com/subject/26387939/
38	哈尔的移动城堡	导演: 宫崎骏 Hayao Miyazaki	主演: 倍赏千惠子 Chieko Baishō / 木村拓...	2004	日本	动画 奇幻 冒险	9.1	874571	https://movie.douban.com/subject/1308807/

Showing 1 to 38 of 250 entries, 9 total columns

导出抓取的数据：

使用`write.save(df_douban_250, 'douban250.csv')`导出CSV文件。

参考列表