
南京财经大學

# 《电子商务导论》 课程实验
# 报告 1

2021 年 4 月 25 日

教师：朱桂祥

学生：王文龙
学号：2120191796

# Problem 1

**1. 选择的爬虫工具是什么？该工具具有什么特点？**

**Solution:**

Scrapy 框架是基于 python 的一个库，为了抓取网页数据、提取结构性数据而编写的应用框架，该框架是封装的，包含 request （异步调度和处理）、下载器（多线程的 Downloader）、解析器（selector）和 twisted（异步处理）等。对于网站的内容爬取，其速度非常快捷。优点：通过管道的方式存入数据库，灵活，可保存为多种形式。缺点：无法用它完成分布式爬取 [1]。MRR@20 是准确推荐项目的排序倒数平均值，该指标衡量的是模型推荐项目的排序性能 [2]。直观地说，在实践中推荐准确的项目排序得越高越好。MRR@20 的定义如下：

$$\text{MRR@20} = \frac{1}{|\mathcal{T}|} \sum_{u \in \mathcal{T}} \frac{1}{R_{u,g_u}}, \tag{1}$$

其中，如果 $R_{u,g_u} \geq 20$，排序的倒数值将设置为 0。

Table 1: 购买预测数据中冷启动用户的比例

| 数据集 | # 冷启动用户 | 比例 (%) |
|---|---|---|
| $D_1$ | 225,749 | 52.34 |
| $D_2$ | 245,703 | 60.96 |
| $D_3$ | 121,006 | 55.61 |

# Problem 2

**2. 爬虫工具环境部署成功截图。**

**Solution:**

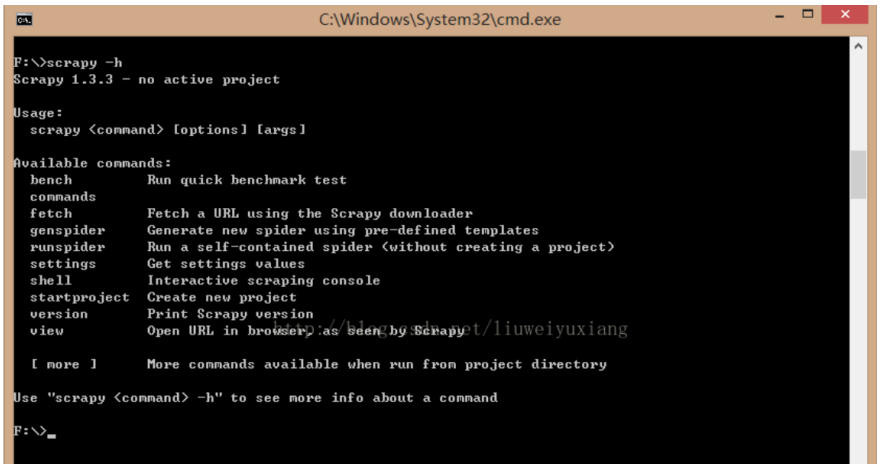


Figure 1: Scrapy 环境部署成功

# Problem 3

**3. 爬取豆瓣电影 Top250 页面 [3]，开始的 URL：https://movie.douban.com/top250，获取每部电影的序号、片名、导演、编剧、主演、类型、制作国家/地区、语言、上映日期、片长、又名、豆瓣评分和剧情简介等内容，将数据存入本地 txt 或者 xlsx 文件。**

**Solution:**

(1) 打开 url 并返回 BeautifulSop 对象：

```python
# -*- coding:utf-8 -*-

from urllib.request import urlopen
from bs4 import BeautifulSoup
from collections import defaultdict
import pandas as pd
import time
import re


class DoubanMovieTop():
    def __init__(self):
        self.top_urls = ['https://movie.douban.com/top250?start={0}&filter='.format(x*25) for x in range(10)]
        self.data = defaultdict(list)
        self.columns = ['title', 'link', 'score', 'score_cnt', 'top_no', 'director', 'writers', 'actors', 'types',
                        'edit_location', 'language', 'dates', 'play_location', 'length', 'rating_per', 'betters',
                        'had_seen', 'want_see', 'tags', 'short_review', 'review', 'ask', 'discussion']
        self.df = None

    def get_bsobj(self, url):
        html = urlopen(url).read().decode('utf-8')
        bsobj = BeautifulSoup(html, 'lxml')
        return bsobj
```

Figure 2: 返回 BeautifulSop 对象

(2) 解析并获取目标对象：

```python
def get_info(self):
    for url in self.top_urls:
        bsobj = self.get_bsobj(url)
        main = bsobj.find('ol', {'class': 'grid_view'})

        # 标题及链接信息
        title_objs = main.findAll('div', {'class': 'hd'})
        titles = [i.find('span').text for i in title_objs]
        links = [i.find('a')['href'] for i in title_objs]

        # 评分信息
        score_objs = main.findAll('div', {'class': 'star'})
        scores = [i.find('span', {'class': 'rating_num'}).text for i in score_objs]
        score_cnts = [i.findAll('span')[-1].text for i in score_objs]

        for title, link, score, score_cnt in zip(titles, links, scores, score_cnts):
            self.data[title].extend([title, link, score, score_cnt])
            bsobj_more = self.get_bsobj(link)
            more_data = self.get_more_info(bsobj_more)
            self.data[title].extend(more_data)
            print(self.data[title])
            print(len(self.data))
            time.sleep(1)
```

Figure 3: 解析并获取目标对象

(3) 爬虫数据存为 csv 格式文件：

```
def dump_data(self):
    data = []
    for title, value in self.data.items():
        data.append(value)
    self.df = pd.DataFrame(data, columns=self.columns)
    self.df.to_csv('douban_top250.csv', index=False)
```

Figure 4: 爬虫结果存储

(4) 运行执行函数：

```
if __name__ == '__main__':
    douban = DoubanMovieTop()
    douban.get_info()
    douban.dump_data()
```

Figure 5: 爬虫执行的主函数

(5) 爬虫结果展示：

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 中文名 | 外文名 | 电影链接 | 图片链接 | 评分 | 评价人数 | 概评 | 概述 | | | | | | | | | |
| 2 | 肖申克的救 | TheShaws | https://mov | https://img | 9.7 | 2271492 | 希望让人自 | 导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗宾斯TimRobbins...1994美国犯罪剧情 | | | | | | | | | |
| 3 | 霸王别姬 | | https://mov | https://img | 9.6 | 1686271 | 风华绝代。 | 导演:陈凯歌KaigeChen主演:张国荣LeslieCheung张丰毅FengyiZha...1993中国大陆中国香港剧情爱情同性 | | | | | | | | | |
| 4 | 阿甘正传 | ForrestGu | https://mov | https://img | 9.5 | 1710247 | 一部美国近 | 导演:罗伯特·泽米吉斯RobertZemeckis主演:汤姆·汉克斯TomHanks...1994美国剧情爱情 | | | | | | | | | |
| 5 | 这个杀手不 | Léon | https://mov | https://img | 9.4 | 1893186 | 怪蜀黍和小 | 导演:吕克·贝松LucBesson主演:让·雷诺JeanReno娜塔莉·波特曼...1994法国美国剧情动作犯罪 | | | | | | | | | |
| 6 | 泰坦尼克号 | Titanic | https://mov | https://img | 9.4 | 1669190 | 失去的才是 | 导演:詹姆斯·卡梅隆JamesCameron主演:莱昂纳多·迪卡普里奥Leonardo...1997美国剧情爱情灾难 | | | | | | | | | |
| 7 | 美丽人生 | Lavitaèbell | https://mov | https://img | 9.5 | 1056697 | 最美的谎言 | 导演:罗伯托·贝尼尼RobertoBenigni主演:罗伯托·贝尼尼RobertoBeni...1997意大利剧情喜剧爱情战争 | | | | | | | | | |
| 8 | 千与千寻 | 千と千尋の | https://mov | https://img | 9.4 | 1787954 | 最好的宫崎 | 导演:宫崎骏HayaoMiyazaki主演:柊瑠美RumiHiiragi入野自由Miy...2001日本剧情动画奇幻 | | | | | | | | | |
| 9 | 辛德勒的名 | Schindler's | https://mov | https://img | 9.5 | 873227 | 拯救一个人 | 导演:史蒂文·斯皮尔伯格StevenSpielberg主演:连姆·尼森LiamNeeson...1993美国剧情历史战争 | | | | | | | | | |
| 10 | 盗梦空间 | Inception | https://mov | https://img | 9.3 | 1658977 | 诺兰给了我 | 导演:克里斯托弗·诺兰ChristopherNolan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 | | | | | | | | | |
| 11 | 忠犬八公的 | Hachi:ADo | https://mov | https://img | 9.4 | 1135721 | 永远都不能 | 导演:莱塞·霍尔斯道姆LasseHallström主演:理查·基尔RichardGer...2009美国英国剧情 | | | | | | | | | |
| 12 | 星际穿越 | Interstellar | https://mov | https://img | 9.3 | 1331082 | 爱是一种力 | 导演:克里斯托弗·诺兰ChristopherNolan主演:马修·麦康纳MatthewMc...2014美国英国加拿大剧情科幻冒险 | | | | | | | | | |
| 13 | 海上钢琴师 | Laleggend | https://mov | https://img | 9.3 | 1349887 | 每个人都是 | 导演:朱塞佩·托纳多雷GiuseppeTornatore主演:蒂姆·罗斯TimRoth...1998意大利剧情音乐 | | | | | | | | | |
| 14 | 楚门的世界 | TheTruman | https://mov | https://img | 9.3 | 1244054 | 如果再也不 | 导演:彼得·威尔PeterWeir主演:金·凯瑞JimCarrey劳拉·琳妮Lau...1998美国剧情科幻 | | | | | | | | | |
| 15 | 三傻大闹宝 | 3Idiots | https://mov | https://img | 9.2 | 1512164 | 英俊版憨豆 | 导演:拉库马·希拉尼RajkumarHirani主演:阿米尔·汗AamirKhan卡...2009印度剧情喜剧爱情歌舞 | | | | | | | | | |
| 16 | 机器人总动 | WALL·E | https://mov | https://img | 9.3 | 1066994 | 小瓦力，方 | 导演:安德鲁·斯坦顿AndrewStanton主演:本·贝尔特BenBurtt艾丽...2008美国科幻动画冒险 | | | | | | | | | |
| 17 | 放牛班的春 | Leschorist | https://mov | https://img | 9.3 | 1049666 | 天籁一般的 | 导演:克里斯托夫·巴拉蒂ChristopheBarratier主演:热拉尔·朱尼奥Gé...2004法国瑞士德国剧情音乐 | | | | | | | | | |
| 18 | 大话西游之 | 西遊記大鬧 | https://mov | https://img | 9.2 | 1216024 | 一生所爱。 | 导演:刘镇伟JeffreyLau主演:周星驰StephenChow吴孟达ManTatNg...1995中国香港中国大陆喜剧爱情奇幻古装 | | | | | | | | | |
| 19 | 疯狂动物城 | Zootopia | https://mov | https://img | 9.2 | 1471339 | 迪士尼给我 | 导演:拜伦·霍华德ByronHoward瑞奇·摩尔RichMoore主演:金妮弗...2016美国喜剧动画冒险 | | | | | | | | | |
| 20 | 无间道 | 無間道 | https://mov | https://img | 9.2 | 1006599 | 香港电影我 | 导演:刘伟强麦兆辉主演:刘德华梁朝伟黄秋生2002中国香港剧情犯罪悬疑 | | | | | | | | | |
| 21 | 熔炉 | 도가니 | https://mov | https://img | 9.3 | 742870 | 我们一路都 | 导演:黄东赫Dong-hyukHwang主演:孔侑YooGong郑有美Yu-miJung...2011韩国剧情 | | | | | | | | | |
| 22 | 教父 | TheGodfat | https://mov | https://img | 9.3 | 742747 | 千万不要让 | 导演:弗朗西斯·福特·科波拉FrancisFordCoppola主演:马龙·白兰度M...1972美国剧情犯罪 | | | | | | | | | |
| 23 | 当幸福来敲 | ThePursuit | https://mov | https://img | 9.1 | 1216322 | 平民励志片 | 导演:加布里尔·穆奇诺GabrieleMuccino主演:威尔·史密斯WillSmith...2006美国剧情传记家庭 | | | | | | | | | |
| 24 | 龙猫 | となりの | https://mov | https://img | 9.2 | 1014045 | 人人心中都 | 导演:宫崎骏HayaoMiyazaki主演:日高法子NorikoHidaka坂本千夏Ch...1988日本动画奇幻冒险 | | | | | | | | | |

Figure 6: CSV 保存的爬虫数据

# Problem 4

**4. 附加题：在上述的爬虫程序中将获取的数据直接传入数据库，选择的什么数据库，导入数据的代码的截图和最终数据查询的截图。**

**Solution:**

空

# References

[1] Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):809–821, 2015.

[2] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 26th ACM on Conference on Information and Knowledge Management*, pages 1419–1428. ACM, 2017.

[3] Long Guo, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, and Bin Cui. Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1984–1992, 2019.