数据分析及实践

Analysis and Practice of the Data

实验课

刘 淇

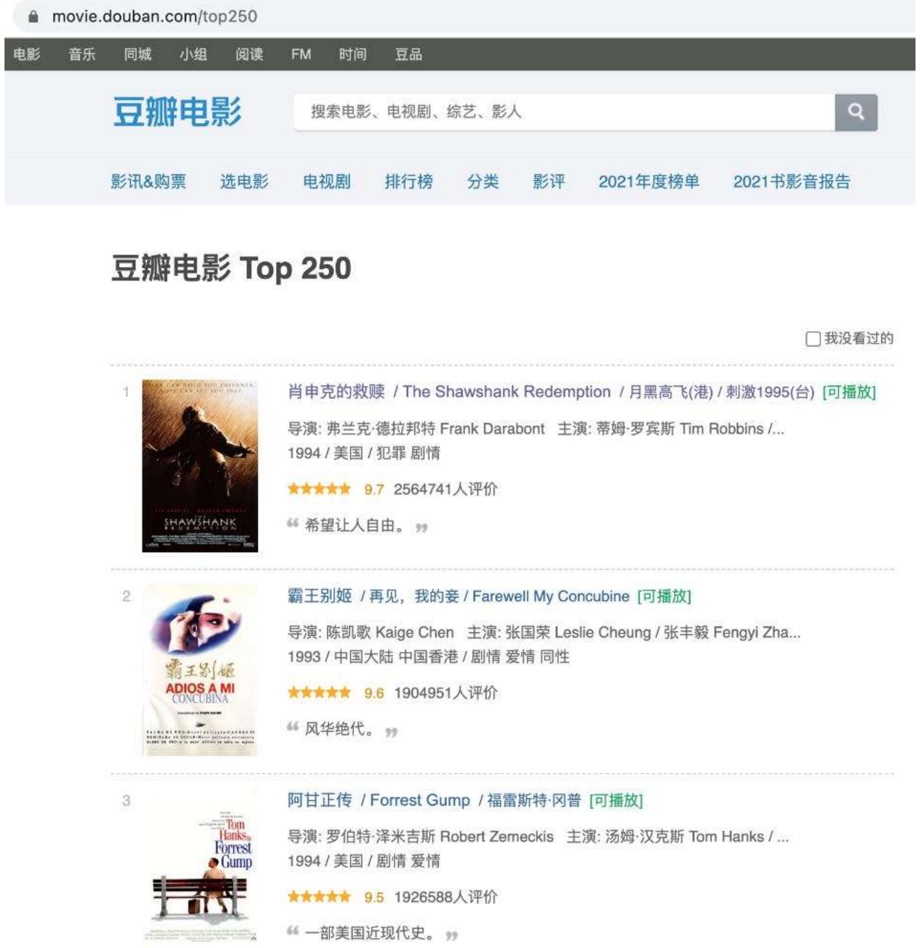
Email: qiliuql@ustc.edu.cn

课程主页：

http://staff.ustc.edu.cn/~qiliuql/AD2022.html

数据获取与管理实验

- 从以下两个实验任意选择一项完成
  - 豆瓣网站https://movie.douban.com的电影详细信息爬取
  - POJ网站http://poj.org/problemlist的题目详细信息爬取



movie.douban.com/top250

电影 音乐 同城 小组 阅读 FM 时间 豆品
豆瓣电影
搜索电影、电视剧、综艺、影人
影讯&购票 选电影 电视剧 排行榜 分类 影评 2021年度榜单 2021书影音报告

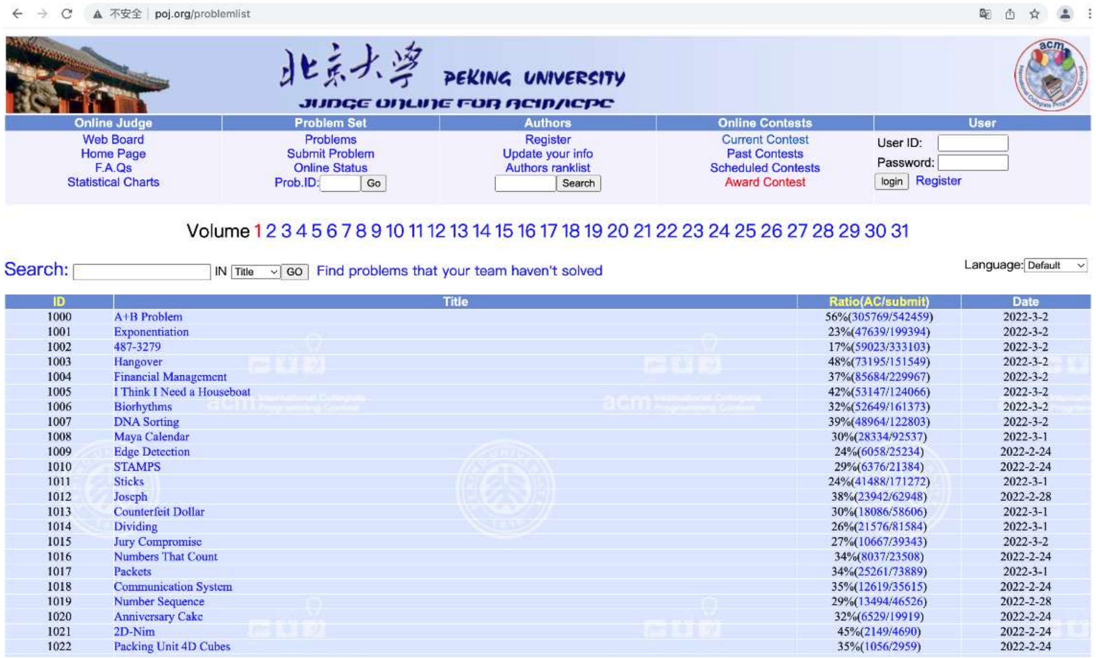
豆瓣电影 Top 250

□ 我没看过的

1 肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]
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 2564741人评价
“希望让人自由。”

2 霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]
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 1904951人评价
“风华绝代。”

3 阿甘正传 / Forrest Gump / 福雷斯特·冈普 [可播放]
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 1926588人评价
“一部美国近现代史。”



不安全 | poj.org/problemlist

Volume 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

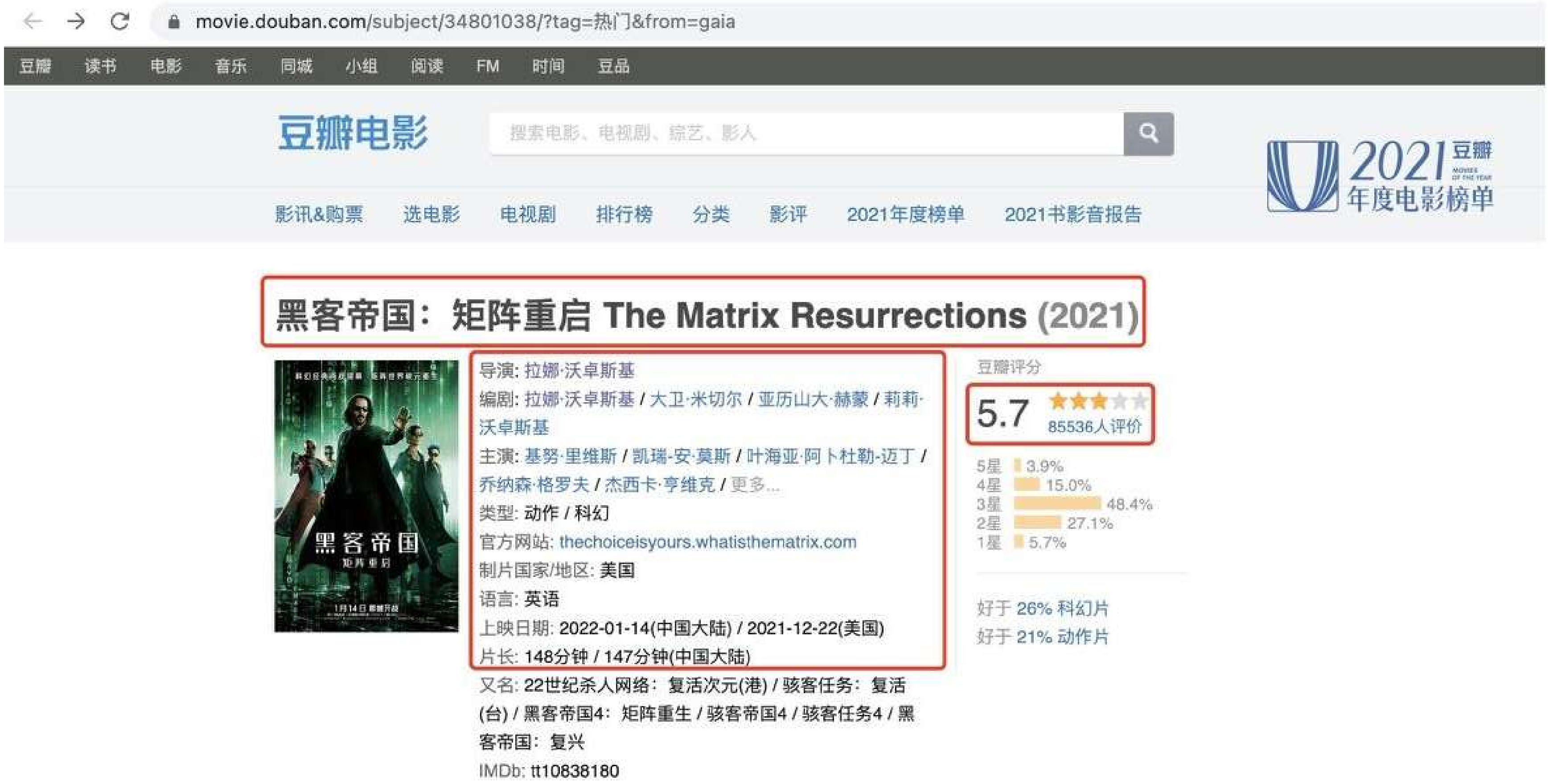
Search: IN Title GO Find problems that your team haven't solved

Language: Default

ID	Title	Ratio(AC/submit)	Date
1000	A+B Problem	56%(305769/542459)	2022-3-2
1001	Exponentiation	23%(47639/199394)	2022-3-2
1002	487-3279	17%(59023/333103)	2022-3-2
1003	Hangover	48%(73195/151549)	2022-3-2
1004	Financial Management	37%(85684/229967)	2022-3-2
1005	I Think I Need a Houseboat	42%(53147/124066)	2022-3-2
1006	Biorhythms	32%(52649/161373)	2022-3-2
1007	DNA Sorting	39%(48964/122803)	2022-3-2
1008	Maya Calendar	30%(28334/92537)	2022-3-1
1009	Edge Detection	24%(6058/25234)	2022-2-24
1010	STAMPS	29%(6376/21384)	2022-2-24
1011	Sticks	24%(41488/171272)	2022-3-1
1012	Joseph	38%(23942/62948)	2022-2-28
1013	Counterfeit Dollar	30%(18086/58606)	2022-3-1
1014	Dividing	26%(21576/81584)	2022-3-1
1015	Jury Compromise	27%(10667/39343)	2022-3-2
1016	Numbers That Count	34%(8037/23508)	2022-2-24
1017	Packets	34%(25261/73889)	2022-3-1
1018	Communication System	35%(12619/35615)	2022-2-24
1019	Number Sequence	29%(13494/46526)	2022-2-28
1020	Anniversary Cake	32%(6529/19919)	2022-2-24
1021	2D-Nim	45%(2149/4690)	2022-2-24
1022	Packing Unit 4D Cubes	35%(1056/2959)	2022-2-24

实验二-Douban Part1

- □ 实验要求Part1
  - □ 给定网站: https://movie.douban.com，需要设计一个网站遍历策略，爬取每部电影的相关信息，记录于json文件中。部分信息标于红框中：



movie.douban.com/subject/34801038/?tag=热门&from=gaia

豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 豆品
豆瓣电影
搜索电影、电视剧、综艺、影人
影讯&购票 选电影 电视剧 排行榜 分类 影评 2021年度榜单 2021书影音报告
2021 豆瓣 年度电影榜单

黑客帝国：矩阵重启 The Matrix Resurrections (2021)

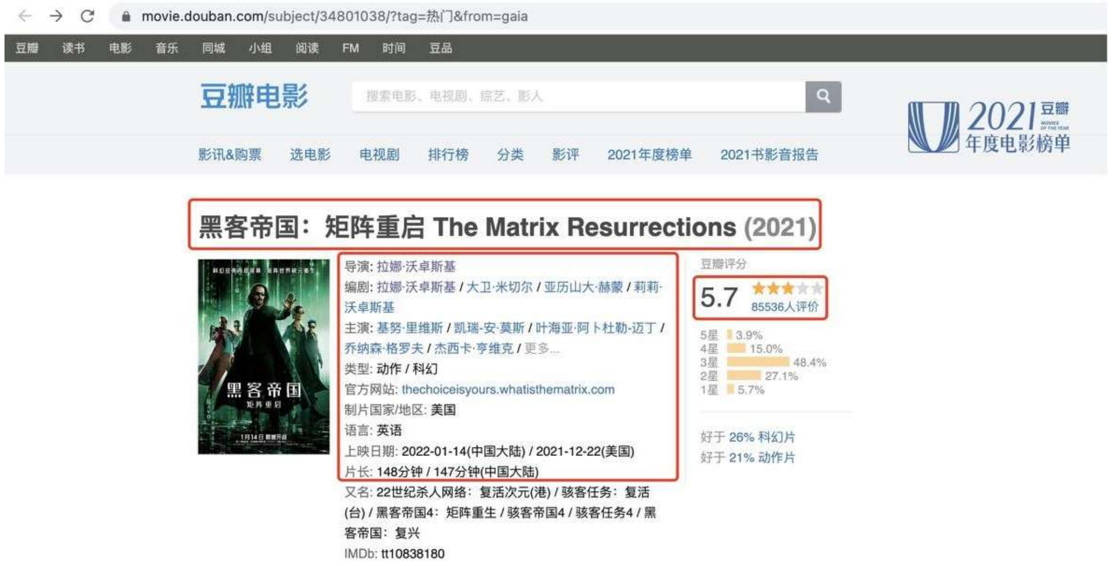
导演: 拉娜·沃卓斯基
编剧: 拉娜·沃卓斯基 / 大卫·米切尔 / 亚历山大·赫蒙 / 莉莉·沃卓斯基
主演: 基努·里维斯 / 凯瑞-安·莫斯 / 叶海亚·阿卜杜勒-迈丁 / 乔纳森·格罗夫 / 杰西卡·亨维克 / 更多...
类型: 动作 / 科幻
官方网站: thechoiceisyours.whatisthematrix.com
制片国家/地区: 美国
语言: 英语
上映日期: 2022-01-14(中国大陆) / 2021-12-22(美国)
片长: 148分钟 / 147分钟(中国大陆)
又名: 22世纪杀人网络：复活次元(港) / 骇客任务：复活(台) / 黑客帝国4：矩阵重生 / 骇客帝国4 / 骇客任务4 / 黑客帝国：复兴
IMDb: tt10838180

豆瓣评分
5.7 ★★★☆☆
85536人评价

5星	3.9%
4星	15.0%
3星	48.4%
2星	27.1%
1星	5.7%

好于 26% 科幻片
好于 21% 动作片

实验二-Douban Part1


movie.douban.com/subject/34801038/?tag=热门&from=gaia
豆瓣 读书 电影 音乐 同城 小组 阅读 FM 时间 豆品
豆瓣电影
搜索电影、电视剧、综艺、影人
影讯&购票 选电影 电视剧 排行榜 分类 影评 2021年度榜单 2021书影音报告
2021 豆瓣 年度电影榜单
黑客帝国 矩阵重启

黑客帝国：矩阵重启 The Matrix Resurrections (2021)

导演: 拉娜·沃卓斯基
编剧: 拉娜·沃卓斯基 / 大卫·米切尔 / 亚历山大·赫蒙 / 莉莉·沃卓斯基
主演: 基努·里维斯 / 凯瑞-安·莫斯 / 叶海亚·阿卜杜勒-迈丁 / 乔纳森·格罗夫 / 杰西卡·亨维克 / 更多...
类型: 动作 / 科幻
官方网站: thechoiceisyours.whatisthematrix.com
制片国家/地区: 美国
语言: 英语
上映日期: 2022-01-14(中国大陆) / 2021-12-22(美国)
片长: 148分钟 / 147分钟(中国大陆)
又名: 22世纪杀人网络：复活次元(港) / 骇客任务：复活(台) / 黑客帝国4：矩阵重生 / 骇客帝国4 / 骇客任务4 / 黑客帝国：复兴
IMDb: tt10838180

豆瓣评分
5.7 ★★★☆☆ 85536人评价
5星 3.9%
4星 15.0%
3星 48.4%
2星 27.1%
1星 5.7%

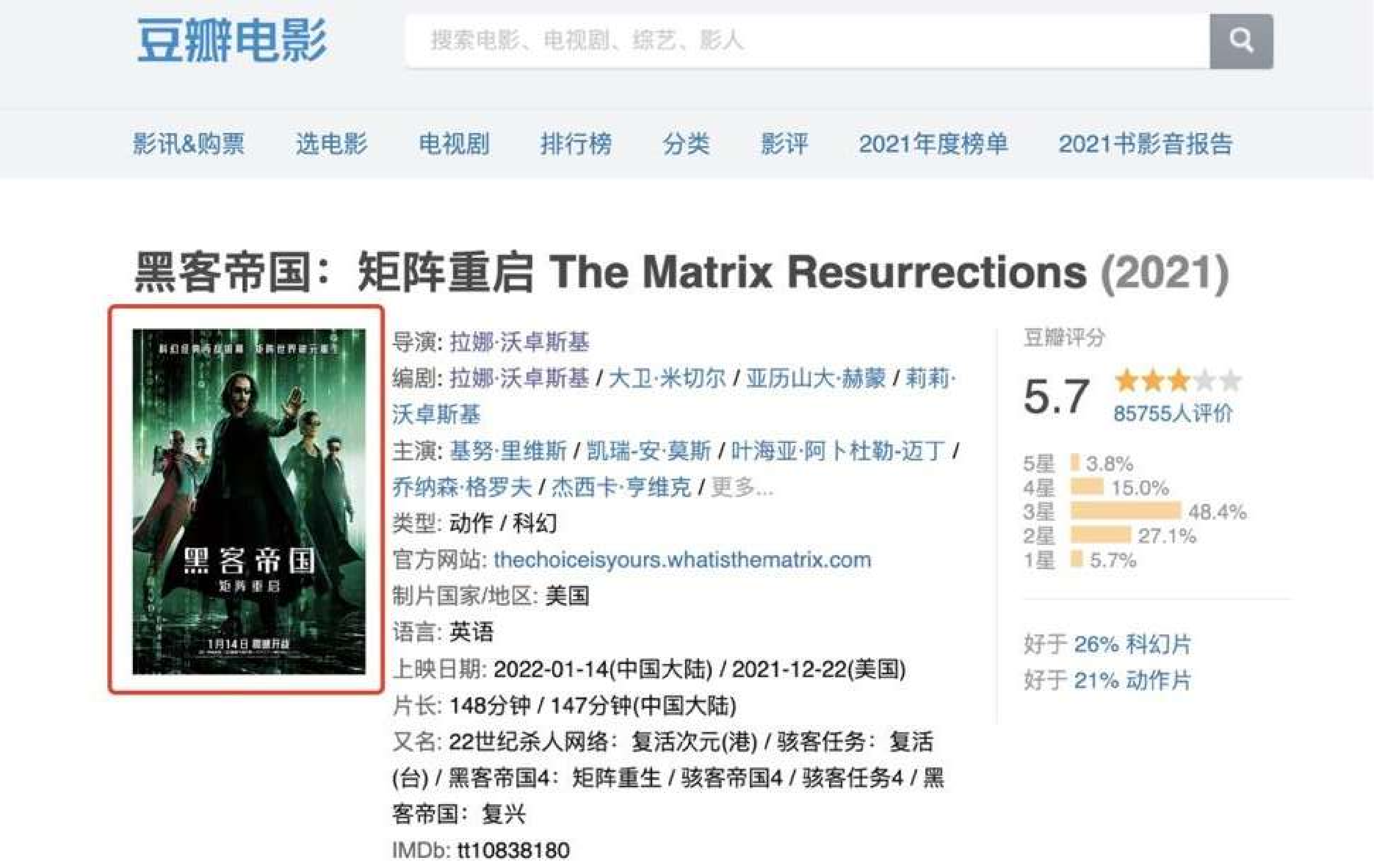
好于 26% 科幻片
好于 21% 动作片

- 样例数据：

```
{
    "片名": "黑客帝国:矩阵重启 The Matrix Resurrections",
    "导演": "拉娜·沃卓斯基",
    "编剧": ["拉娜·沃卓斯基", "大卫·米切尔", "亚历山大·赫蒙", "莉莉·沃卓斯基"],
    "主演": ["基努·里维斯", "凯瑞-安·莫斯", "叶海亚·阿卜杜勒-迈丁", "乔纳森·格罗夫", "杰西卡·亨维克"],
    "类型": ["动作", "科幻"],
    "官方网站":  "thechoiceisyours.whatisthematrix.com",
    "制片国家/地区": "美国",
    "语言": "英语",
    "上映日期": ["2022-01-14(中国大陆)", "2021-12-22(美国)"],
    "片长": ["148分钟", "147分钟(中国大陆)"],
    "评分": 5.7
}
```

实验二 -Douban Part2 （选做）

- 实验要求 Part2
  - 在Part1爬取文本信息的基础上，爬取每部电影对应的图片（红框所示），保存在文件夹中。



黑客帝国：矩阵重启 The Matrix Resurrections (2021)

导演: 拉娜·沃卓斯基
编剧: 拉娜·沃卓斯基 / 大卫·米切尔 / 亚历山大·赫蒙 / 莉莉·沃卓斯基
主演: 基努·里维斯 / 凯瑞-安·莫斯 / 叶海亚·阿卜杜勒-迈丁 / 乔纳森·格罗夫 / 杰西卡·亨维克 / 更多...
类型: 动作 / 科幻
官方网站: thechoiceisyours.whatisthematrix.com
制片国家/地区: 美国
语言: 英语
上映日期: 2022-01-14(中国大陆) / 2021-12-22(美国)
片长: 148分钟 / 147分钟(中国大陆)
又名: 22世纪杀人网络：复活次元(港) / 骇客任务：复活(台) / 黑客帝国4：矩阵重生 / 骇客帝国4 / 骇客任务4 / 黑客帝国：复兴
IMDb: tt10838180

豆瓣评分
5.7 ★★★☆☆
85755人评价

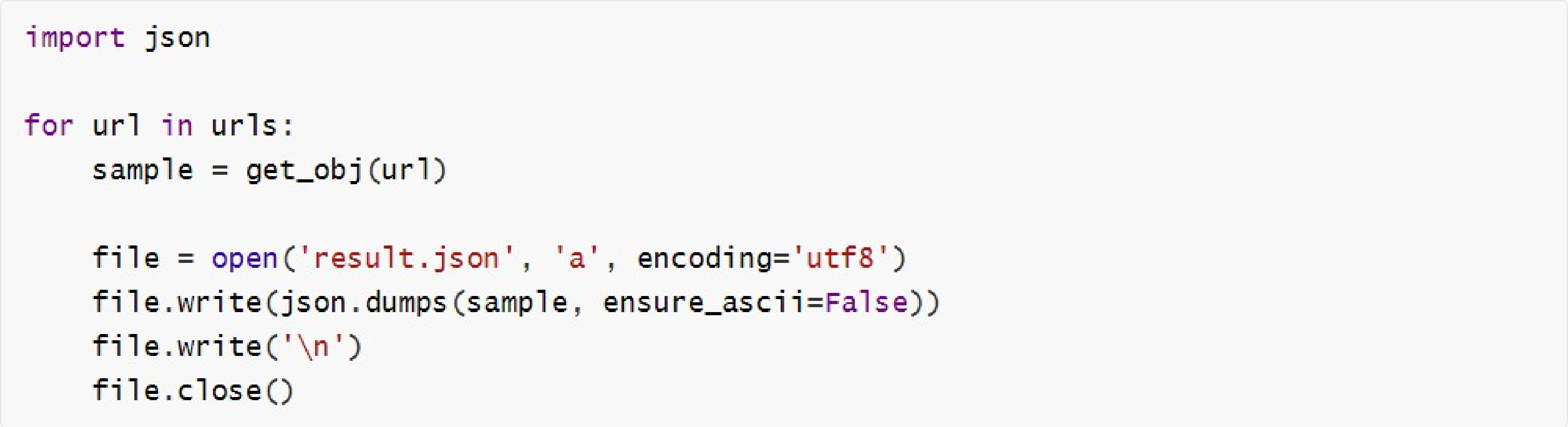
5星	3.8%
4星	15.0%
3星	48.4%
2星	27.1%
1星	5.7%

好于 26% 科幻片
好于 21% 动作片

实验二 -Douban

注意事项

- □ 1. 每位同学爬取至少100部电影的信息，电影种类不限
- □ 2. 保存到json文件的python代码，供参考（ sample 即为你解析得到的一个网页的数据字典）



```
import json

for url in urls:
    sample = get_obj(url)

    file = open('result.json', 'a', encoding='utf8')
    file.write(json.dumps(sample, ensure_ascii=False))
    file.write('\n')
    file.close()
```

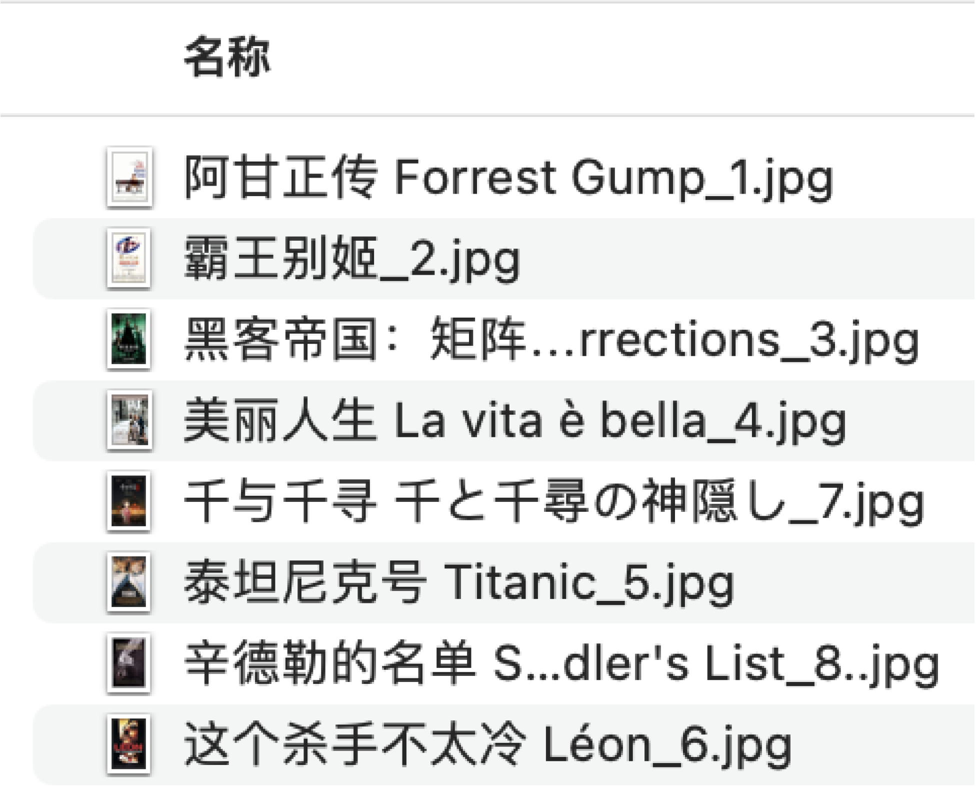
实验二 -Douban

- 3.图片文件命名规则

  以对应的电影名称命名：电影名称_计数.jpg/jpeg/png

  如黑客帝国：矩阵重启 The Matrix Resurrections_3.jpg/jpeg/png

  图片单独存放在一个文件夹里



名称

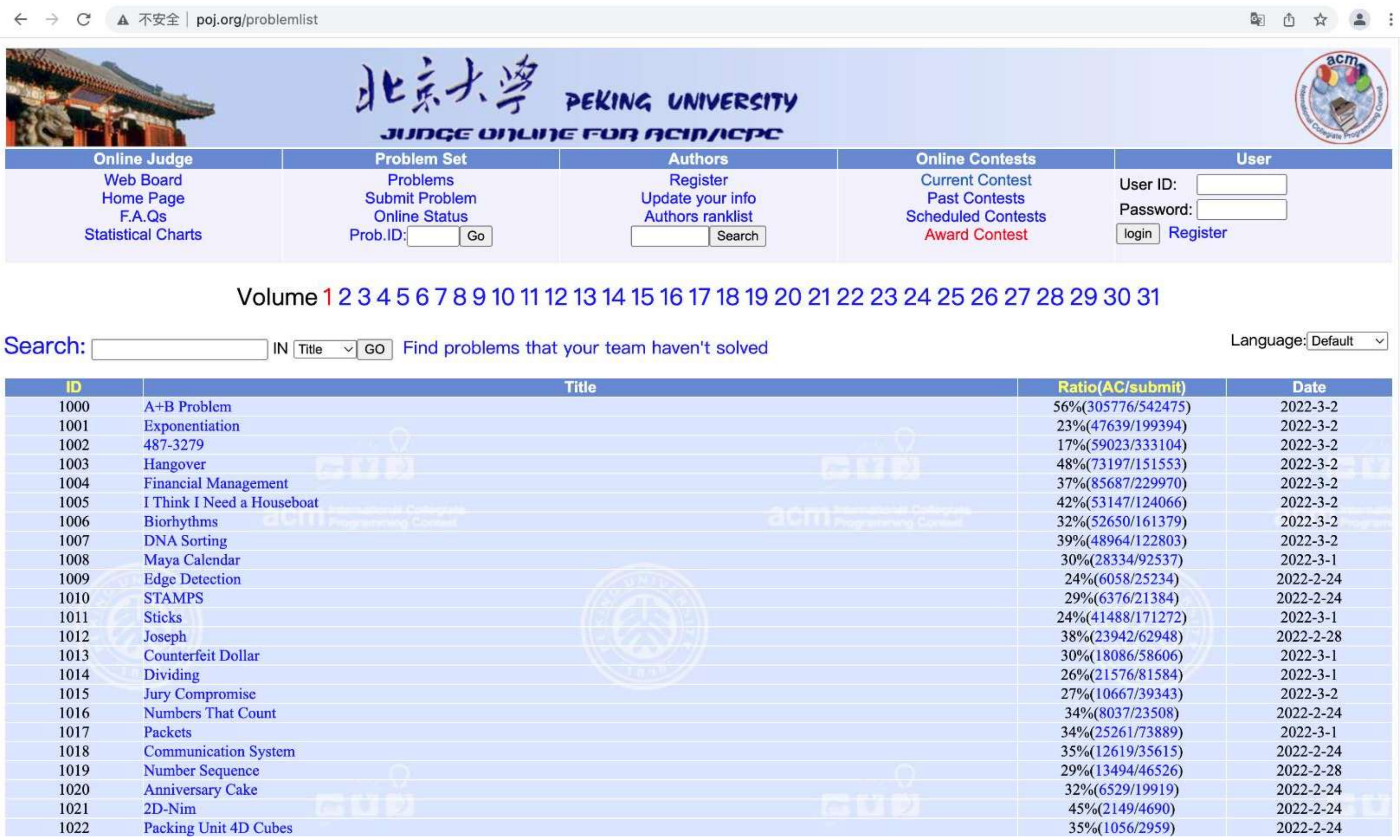
阿甘正传 Forrest Gump_1.jpg
霸王别姬_2.jpg
黑客帝国：矩阵...rrections_3.jpg
美丽人生 La vita è bella_4.jpg
千与千寻 千と千尋の神隠し_7.jpg
泰坦尼克号 Titanic_5.jpg
辛德勒的名单 S...dler's List_8..jpg
这个杀手不太冷 Léon_6.jpg

实验二-Douban

- □ 提交要求
  - □ 将爬虫代码和数据打包成一个压缩文件，发送到助教邮箱：18251859960@163.com
  - □ 邮件标题: 姓名_学号_exp2_douban
    文件命名格式：姓名_学号_exp2_douban.zip
  - □ 截止日期：3月23日
- □ 评分标准：
  - □ 格式是否规范
  - □ 提交是否及时
  - □ 代码是否美观，能否运行

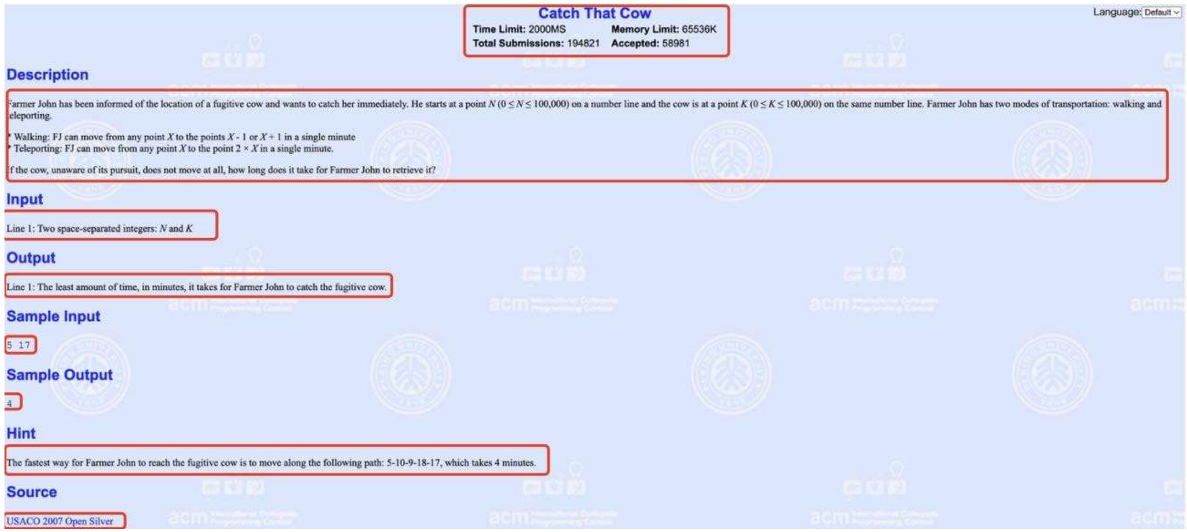
实验二-POJ Part1

- 实验要求Part1
  - 给定网站 http://poj.org/problemlist ，需要设计一个网站遍历策略，爬取网站题目信息。



ID	Title	Ratio(AC/submit)	Date
1000	A+B Problem	56%(305776/542475)	2022-3-2
1001	Exponentiation	23%(47639/199394)	2022-3-2
1002	487-3279	17%(59023/333104)	2022-3-2
1003	Hangover	48%(73197/151553)	2022-3-2
1004	Financial Management	37%(85687/229970)	2022-3-2
1005	I Think I Need a Houseboat	42%(53147/124066)	2022-3-2
1006	Biorhythms	32%(52650/161379)	2022-3-2
1007	DNA Sorting	39%(48964/122803)	2022-3-2
1008	Maya Calendar	30%(28334/92537)	2022-3-1
1009	Edge Detection	24%(6058/25234)	2022-2-24
1010	STAMPS	29%(6376/21384)	2022-2-24
1011	Sticks	24%(41488/171272)	2022-3-1
1012	Joseph	38%(23942/62948)	2022-2-28
1013	Counterfeit Dollar	30%(18086/58606)	2022-3-1
1014	Dividing	26%(21576/81584)	2022-3-1
1015	Jury Compromise	27%(10667/39343)	2022-3-2
1016	Numbers That Count	34%(8037/23508)	2022-2-24
1017	Packets	34%(25261/73889)	2022-3-1
1018	Communication System	35%(12619/35615)	2022-2-24
1019	Number Sequence	29%(13494/46526)	2022-2-28
1020	Anniversary Cake	32%(6529/19919)	2022-2-24
1021	2D-Nim	45%(2149/4690)	2022-2-24
1022	Packing Unit 4D Cubes	35%(1056/2959)	2022-2-24

实验二-POJ Part1



Language: Default

Catch That Cow

Time Limit: 2000MS **Memory Limit:** 65536K
Total Submissions: 194821 **Accepted:** 58981

Description

Farmer John has been informed of the location of a fugitive cow and wants to catch her immediately. He starts at a point N ($0 \le N \le 100,000$) on a number line and the cow is at a point K ($0 \le K \le 100,000$) on the same number line. Farmer John has two modes of transportation: walking and eleporting.

- Walking: FJ can move from any point X to the points $X - 1$ or $X + 1$ in a single minute
- Teleporting: FJ can move from any point X to the point $2 \times X$ in a single minute.

f the cow, unaware of its pursuit, does not move at all, how long does it take for Farmer John to retrieve it?

Input

Line 1: Two space-separated integers: N and K

Output

Line 1: The least amount of time, in minutes, it takes for Farmer John to catch the fugitive cow.

Sample Input

```
5 17
```

Sample Output

```
4
```

Hint

The fastest way for Farmer John to reach the fugitive cow is to move along the following path: 5-10-9-18-17, which takes 4 minutes.

Source

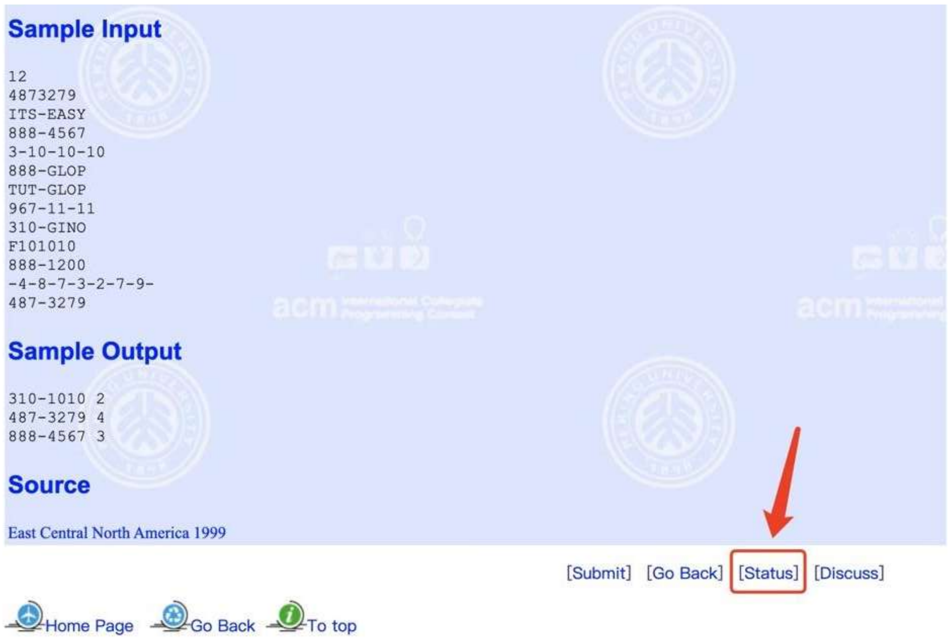
USACO 2007 Open Silver

- 样例数据:

```
[
    {
        "Title": "Catch That Cow",
        "TimeLimit": "2000MS",
        "MemoryLimit": "65536K",
        "TotalSubmissions": "194821",
        "Accepted": "58981",
        "Description": "Farmer John has been informed of the location of a fugitive cow and wants to cat
        "Input": "Line 1: Two space-separated integers: N and K",
        "Output": "Line 1: The least amount of time, in minutes, it takes for Farmer John to catch the f
        "Sample Input": "5 17",
        "Sample Output": "4",
        "Hint": "The fastest way for Farmer John to reach the fugitive cow is to move along the followin
        "Source": "USACO 2007 Open Silver"
    }
]
```

实验二-POJ Part2 （选做）

- 实验要求Part2
  - 爬取题目对应的状态（status）信息，包括Statistics里的14个字段信息和前20条提交状态信息的user的名字。



Sample Input

```
12
4873279
ITS-EASY
888-4567
3-10-10-10
888-GLOP
TUT-GLOP
967-11-11
310-GINO
F101010
888-1200
-4-8-7-3-2-7-9-
487-3279
```

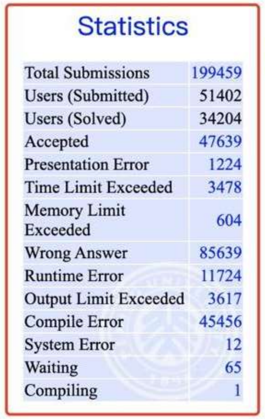
Sample Output

```
310-1010 2
487-3279 4
888-4567 3
```

Source

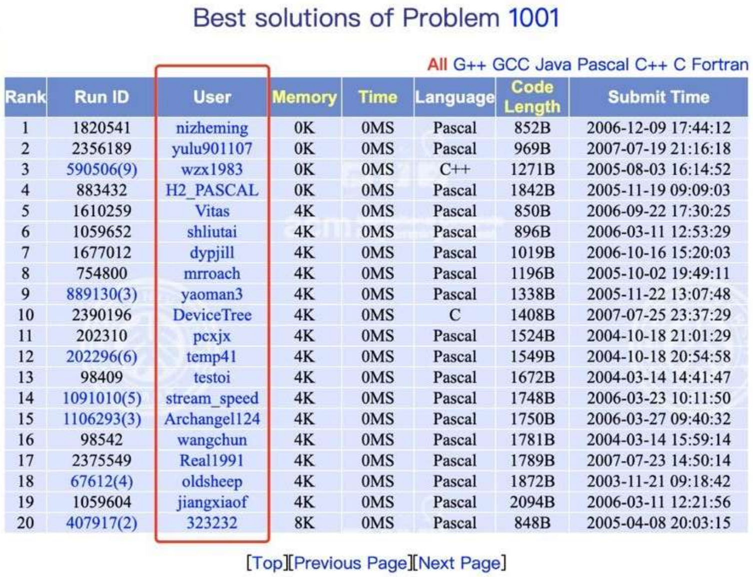
East Central North America 1999

[Submit] [Go Back] [Status] [Discuss]



Statistics

Total Submissions	199459
Users (Submitted)	51402
Users (Solved)	34204
Accepted	47639
Presentation Error	1224
Time Limit Exceeded	3478
Memory Limit Exceeded	604
Wrong Answer	85639
Runtime Error	11724
Output Limit Exceeded	3617
Compile Error	45456
System Error	12
Waiting	65
Compiling	1



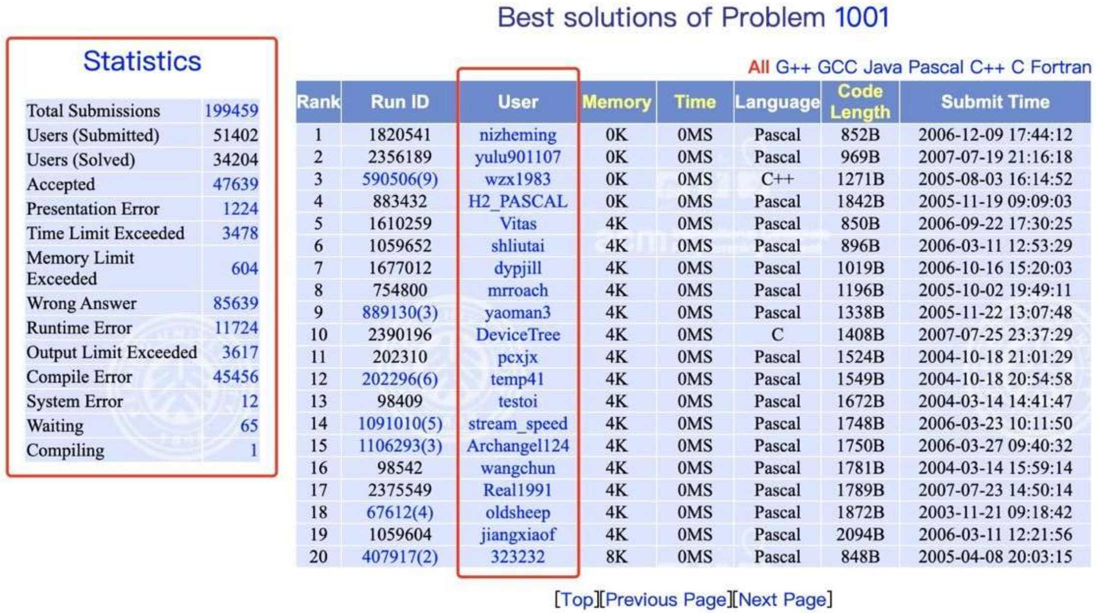
Best solutions of Problem 1001

All G++ GCC Java Pascal C++ C Fortran

Rank	Run ID	User	Memory	Time	Language	Code Length	Submit Time
1	1820541	nizheming	0K	0MS	Pascal	852B	2006-12-09 17:44:12
2	2356189	yulu901107	0K	0MS	Pascal	969B	2007-07-19 21:16:18
3	590506(9)	wzx1983	0K	0MS	C++	1271B	2005-08-03 16:14:52
4	883432	H2_PASCAL	0K	0MS	Pascal	1842B	2005-11-19 09:09:03
5	1610259	Vitas	4K	0MS	Pascal	850B	2006-09-22 17:30:25
6	1059652	shliutai	4K	0MS	Pascal	896B	2006-03-11 12:53:29
7	1677012	dypjill	4K	0MS	Pascal	1019B	2006-10-16 15:20:03
8	754800	mrroach	4K	0MS	Pascal	1196B	2005-10-02 19:49:11
9	889130(3)	yaoman3	4K	0MS	Pascal	1338B	2005-11-22 13:07:48
10	2390196	DeviceTree	4K	0MS	C	1408B	2007-07-25 23:37:29
11	202310	pcxjx	4K	0MS	Pascal	1524B	2004-10-18 21:01:29
12	202296(6)	temp41	4K	0MS	Pascal	1549B	2004-10-18 20:54:58
13	98409	testoi	4K	0MS	Pascal	1672B	2004-03-14 14:41:47
14	1091010(5)	stream_speed	4K	0MS	Pascal	1748B	2006-03-23 10:11:50
15	1106293(3)	Archangel124	4K	0MS	Pascal	1750B	2006-03-27 09:40:32
16	98542	wangchun	4K	0MS	Pascal	1781B	2004-03-14 15:59:14
17	2375549	Real1991	4K	0MS	Pascal	1789B	2007-07-23 14:50:14
18	67612(4)	oldsheep	4K	0MS	Pascal	1872B	2003-11-21 09:18:42
19	1059604	jiangxiaof	4K	0MS	Pascal	2094B	2006-03-11 12:21:56
20	407917(2)	323232	8K	0MS	Pascal	848B	2005-04-08 20:03:15

[Top][Previous Page][Next Page]

实验二-POJ Part2 （选做）



Best solutions of Problem 1001

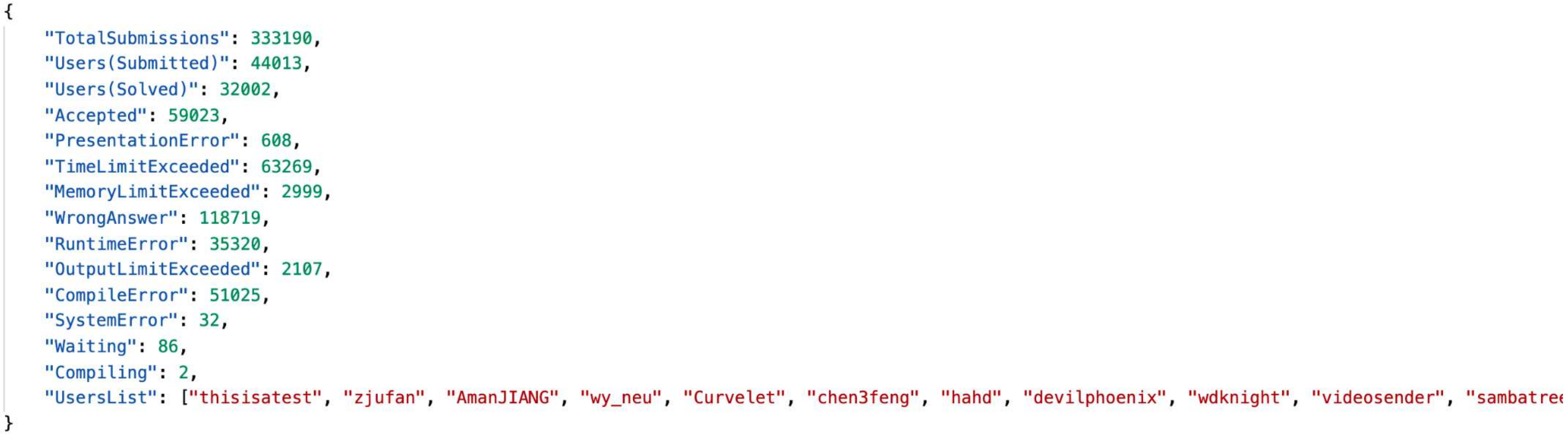
Statistics	
Total Submissions	199459
Users (Submitted)	51402
Users (Solved)	34204
Accepted	47639
Presentation Error	1224
Time Limit Exceeded	3478
Memory Limit Exceeded	604
Wrong Answer	85639
Runtime Error	11724
Output Limit Exceeded	3617
Compile Error	45456
System Error	12
Waiting	65
Compiling	1

All G++ GCC Java Pascal C++ C Fortran

Rank	Run ID	User	Memory	Time	Language	Code Length	Submit Time
1	1820541	nizheming	0K	0MS	Pascal	852B	2006-12-09 17:44:12
2	2356189	yulu901107	0K	0MS	Pascal	969B	2007-07-19 21:16:18
3	590506(9)	wzx1983	0K	0MS	C++	1271B	2005-08-03 16:14:52
4	883432	H2_PASCAL	0K	0MS	Pascal	1842B	2005-11-19 09:09:03
5	1610259	Vitas	4K	0MS	Pascal	850B	2006-09-22 17:30:25
6	1059652	shliutai	4K	0MS	Pascal	896B	2006-03-11 12:53:29
7	1677012	dypjill	4K	0MS	Pascal	1019B	2006-10-16 15:20:03
8	754800	mrroach	4K	0MS	Pascal	1196B	2005-10-02 19:49:11
9	889130(3)	yaoman3	4K	0MS	Pascal	1338B	2005-11-22 13:07:48
10	2390196	DeviceTree	4K	0MS	C	1408B	2007-07-25 23:37:29
11	202310	pcxjx	4K	0MS	Pascal	1524B	2004-10-18 21:01:29
12	202296(6)	temp41	4K	0MS	Pascal	1549B	2004-10-18 20:54:58
13	98409	testoi	4K	0MS	Pascal	1672B	2004-03-14 14:41:47
14	1091010(5)	stream_speed	4K	0MS	Pascal	1748B	2006-03-23 10:11:50
15	1106293(3)	Archangel124	4K	0MS	Pascal	1750B	2006-03-27 09:40:32
16	98542	wangchun	4K	0MS	Pascal	1781B	2004-03-14 15:59:14
17	2375549	Real1991	4K	0MS	Pascal	1789B	2007-07-23 14:50:14
18	67612(4)	oldsheep	4K	0MS	Pascal	1872B	2003-11-21 09:18:42
19	1059604	jiangxiaof	4K	0MS	Pascal	2094B	2006-03-11 12:21:56
20	407917(2)	323232	8K	0MS	Pascal	848B	2005-04-08 20:03:15

[Top][Previous Page][Next Page]

- □ 样例数据



```
{
    "TotalSubmissions": 333190,
    "Users(Submitted)": 44013,
    "Users(Solved)": 32002,
    "Accepted": 59023,
    "PresentationError": 608,
    "TimeLimitExceeded": 63269,
    "MemoryLimitExceeded": 2999,
    "WrongAnswer": 118719,
    "RuntimeError": 35320,
    "OutputLimitExceeded": 2107,
    "CompileError": 51025,
    "SystemError": 32,
    "Waiting": 86,
    "Compiling": 2,
    "UsersList": ["thisisatest", "zjufan", "AmanJIANG", "wy_neu", "Curvelet", "chen3feng", "hahd", "devilphoenix", "wdknight", "videosender", "sambatre
}
```

实验二 - POJ

注意事项

- □ 1. 豆瓣项目与POJ项目**任选一个**完成即可
  - □ Part 2为选做题，供感兴趣的同学选做
- □ 2. 每位同学爬取 100道题目详细信息，类别不限
- □ 3. 每道题目只需要选择前20个user名即可，存放在UserList里

实验二- POJ

□ 提交要求

  □ 将爬虫代码和数据打包成一个压缩文件，发送给助教：18251859960@163.com

  □ 邮件标题: 姓名_学号_exp2_POJ
  文件命名格式：姓名_学号_exp2_POJ.zip

  □ 截止日期：3月23日

□ 评分标准：

  □ 格式是否规范

  □ 提交是否及时

  □ 代码是否美观，能否运行

实验二-参考资料

- □ request库、正则表达式、beautifulsoup库、 Scrapy库等。
- □ 可以看相关博客入门，也可以阅读参考书籍：