

数据分析及实践 实验二 实验报告

崔士强 PB22151743

2024 年 3 月 28 日

任务 1

读取后得到的 page.txt 文件如图所示:

[illegible]

图 1: page.txt 文件内容

任务 2

通过观察可以发现，track 名称跟在<h2 id=后，对应设计正则表达式：

```
1 title_pattern = '<h2 id={1,100}>.{1,100}</h2></header>'
```

输出 5 个 track 名称:

- Research Track Full Papers
- Applied Data Track Full Papers
- Hands On Tutorials
- Lecture Style Tutorials
- Workshop Summaries

图 2: 各个 track 名称

任务 3

首先根据任务 2 中的规律，将字符串按 track 进行初步分割，再按照论文进一步分割。对于目标元素，观察网页代码发现如下规律：

1. 作者姓名在``之后
2. 论文标题在``之后
3. 收录起始页与终止页在``之后

对应设计正则表达式之后利用`split()`方法取得目标字符串。对论文进行计数的结果如图所示

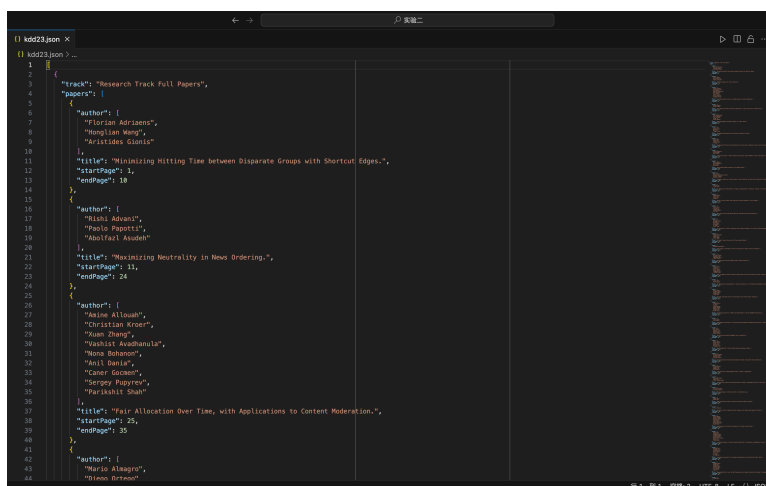
```
print(len(track1dict["papers"]))
print(len(track2dict["papers"]))

[8]
... 313
    183
```

图 3: 两个 track 的论文计数

任务 4

存入`kdd23.json`文件后如图所示



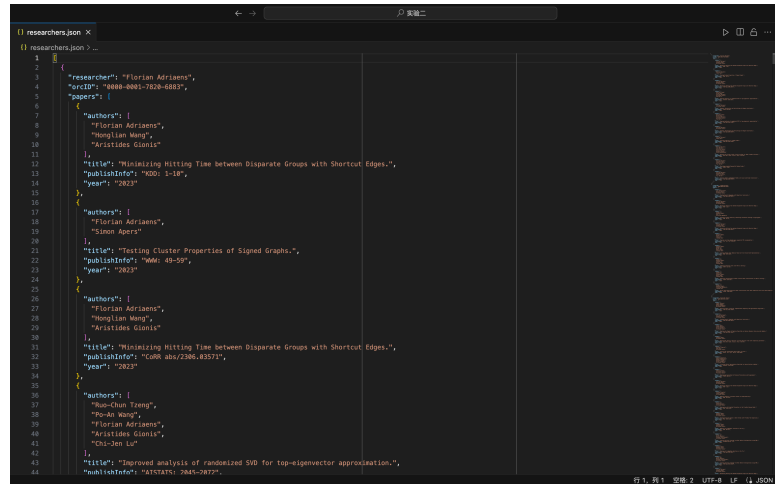
```
{
  "track": "Research Track Full Papers",
  "papers": [
    {
      "author": [
        "Florian Adams",
        "Honglan Wang",
        "Aristides Gionis"
      ],
      "title": "Minimizing Hitting Time between Disparate Groups with Shortcut Edges.",
      "startPage": 3,
      "endPage": 18
    },
    {
      "author": [
        "Rishabh Adams",
        "Paola Papotti",
        "Abdellatif Assouli"
      ],
      "title": "Maximizing Neutrality in News Ordering.",
      "startPage": 21,
      "endPage": 24
    },
    {
      "author": [
        "Rishabh Adams",
        "Christian Kroger",
        "Kuan Zhang",
        "Rishabh Assouli",
        "Nemo Bohannon",
        "Neil Deneke",
        "Casper Gohm",
        "Sergey Poryvov",
        "Rishabh Assouli"
      ],
      "title": "Fair Allocation Over Time, with Applications to Content Moderation.",
      "startPage": 25,
      "endPage": 35
    },
    {
      "author": [
        "Maria Almagro",
        "Dionis Stenon"
      ],
      "title": "Fair Allocation Over Time, with Applications to Content Moderation.",
      "startPage": 36,
      "endPage": 45
    }
  ]
}
```

图 4: `kdd2023.json`

任务 5

第一步是获取每个作者主页的链接，方法与任务 3 中大致相同。之后利用`request`库逐个读取链接。观察网页代码可以发现，`publishInfo`可能包括: 1). `volume` 2). `issue` 3). `pagination`。因此逐个使用`try...except` 语句查找，之后进行连接。

最后将信息写入`researchers.json`文件，如图所示：



```

1 {
2   "researcher": "Florian Adriaens",
3   "arxivId": "0000-0001-7820-6883",
4   "papers": [
5     {
6       "authors": [
7         "Florian Adriaens",
8         "Honglan Wang",
9         "Aristides Gionis"
10      ],
11       "title": "Minimizing Hitting Time between Disparate Groups with Shortcut Edges.",
12       "publishInfo": "QOS 1-18",
13       "year": "2023"
14     },
15     {
16       "authors": [
17         "Florian Adriaens",
18         "Simon Aperi"
19      ],
20       "title": "Testing Cluster Properties of Signed Graphs.",
21       "publishInfo": "MM: 49-59",
22       "year": "2023"
23     },
24     {
25       "authors": [
26         "Florian Adriaens",
27         "Honglan Wang",
28         "Aristides Gionis"
29      ],
30       "title": "Minimizing Hitting Time between Disparate Groups with Shortcut Edges.",
31       "publishInfo": "CoRR abs/2306.03571",
32       "year": "2023"
33     },
34     {
35       "authors": [
36         "Run-Chun Tieng",
37         "Honglan Wang",
38         "Florian Adriaens",
39         "Aristides Gionis",
40         "Chi-Jen Lu"
41      ],
42       "title": "Improved analysis of randomized SVD for top-eigenvector approximation.",
43       "publishInfo": "ATOTAT: 2845-2872"
44     }
45   ]
46 }

```

图 5: name