中国科学技术大学

University of Science and Technology of China

Lab 2

Python 爬虫实验

姓名：高茂航

学号：PB22061161

日期：2024.3.21

Report 2

1 Task1

1.1 Algorithm Description

使用`urllib.request.urlopen().read()`提取网页的 html 文件，然后用`decode()`函数将 html 文件转为文本文件并写入`page.txt`文件。

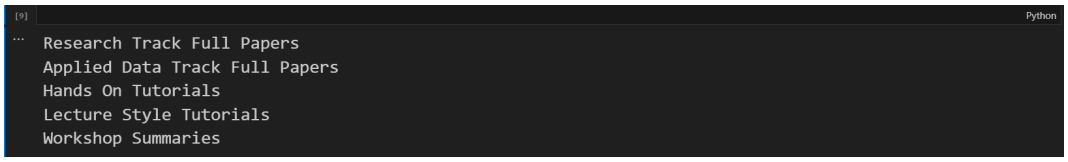
1.2 Results

2 Task2

2.1 Algorithm Description

使用`re.findall()`找到所有的`<h2 id="*">` 和`</h2>` 之间的字符串（也即是 track 的名称）并输出。

2.2 Results



3 Task3

3.1 Algorithm Description

通过正则表达式从文本文件中提取出”Research Track Full Papers” 和”Applied Data Track Full Papers” 两个部分的内容。接着划分其中所有相邻`<cite`和`</cite>`之间的内容 (即一篇文章的范围) 储存到一个字符串元组中，元组的长度即文章的篇数。然后定义一个

函数extract_paper_info() 运用正则表达式提取每篇论文的作者、标题和起始页码、结束页码。随后将提取到的论文信息存储在一个包含字典的列表 tracks 中，每个字典包含跟踪名称及其对应的论文列表。最后，使用json.dump()将 tracks 列表转换为 JSON 格式，写入kdd23.json文件。

3.2 Results

```
[1]                                                                    Python
...  Number of research papers:  313
     Number of applied papers:  183
```

4 Task4

4.1 Algorithm Description

使用正则表达式识别作者链接、orcID(有多个就爬取所有的)、论文信息。提取 Research Track Full Papers 和 Applied Data Track Full Papers 两部分的内容，并找到前 10 个persistent URL:之间（即一篇文章）的内容。遍历每篇文章，使用requests.get(link)访问作者链接。提取<li class="underline" title="jump to the 2020s"> 和

<li class="underline" title="jump to the 2010s"> 之间的内容（即 2020 年以后发表的文章），并根据</cite>划分各篇文章的内容，用正则表达式提取年份，接着去除所有尖括号之间的内容，把剩下的字符串存为一个字符串元组，依次提取作者和题目信息，剩下的所有内容就是出版信息。最后将作者、orcID、出版信息和年份以字典形式存储存储到researchers.json文件中。

另外，由于爬虫时频繁请求服务器，曾被网站认定为攻击行为并报错：

ConnectionResetError: [WinError 10054] 远程主机强迫关闭了一个现有的连接

故采取访问后关闭响应和在各个请求之间添加随机延时等待这两个措施来避免报错。

4.2 Results

researchers.json > {} 1


```
[
  {
    "researcher": "Florian Adriaens",
    "orcID": [
      "0000-0001-7820-6883"
    ],
    "papers": [
      {
        "authors": [
          "Florian Adriaens",
          "Honglian Wang",
          "Aristides Gionis"
        ],
        "title": "Minimizing Hitting Time between Disparate Groups with Shortcut Edges.",
        "publishInfo": "KDD 2023: 1-10",
        "year": 2023
      },
      {
        "authors": [
          "Florian Adriaens",
          "Simon Apers"
        ],
        "title": "Testing Cluster Properties of Signed Graphs.",
```