


# 实验二

中国科大2024年春季学期“数据分析及实践”课程 - 实验二说明文档。

## 任务概述

DBLP（主页：<https://dblp.uni-trier.de>）是计算机领域学术研究的一个英文文献集成数据库系统，在学术界有很好的声誉。用户可以在搜索栏输入关键词（如论文名称、作者名、会议名称）以获取相关文献的元数据（如标题、作者、发表日期等）。



**дблп**  
computer science bibliography

Stop the war!

SCHLOSS DAGSTUHL  
Leibniz Center for Informatics

home | browse | search | about | nfdi

search dblp

Welcome to dblp

> Home

browse authors | editors

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

browse journals

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z by publisher

browse conferences | workshops

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

browse series

CoRR LNCS CEUR-WS LNEE IFIP LNI EPTCS LIPICS other

browse monographs

books & theses reference works edited collections

About dblp

The *dblp computer science bibliography* provides open bibliographic information on major computer science journals and proceedings. Originally created at the University of Trier, dblp is now operated and further developed by Schloss Dagstuhl. For more information check out our F.A.Q.

dblp statistics


unable to load statistics

# of authors: 3,455,213

# of conferences: 6,595

# of journals: 1,860

# of publications by year of publication:



more statistics

XML data

dblp blog

2024-01-01: 7 million publications [News]

read as PDF

(read full post)

2023-05-22: DTD update May 2023 [News]

(updated 2023-06-28) A few days ago, we discussed the new dataset publications in dblp. As a preparation for more and more detailed datasets we slightly modify the DTD that defines the structure of our XML data export. A quick reminder: you can download the dblp dataset as a single XML [...]

(read full post)


现欲通过 DBLP 获取数据挖掘领域顶级学术会议 KDD (ACM SIGKDD Conference on Knowledge Discovery and Data Mining) 在 2023 年的论文收录和相关作者信息，请你按要求编写 Python 代码实现任务列表中的内容。

## 任务列表

1. (15%) 进入 DBLP 主页，通过搜索功能，打开罗列 KDD 2023 所有会议文献的页面。

Q. 读取整个页面的 html 内容并解码为文本串（可使用urllib.request的相应方法），将其以UTF-8编码格式写入page.txt文件，留待后续处理。


2. (15%) 本页面展示了 KDD 2023 会议文献在不同 Track 下的论文收录情况。



дблп

computer science bibliography

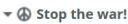
a service of



SCHLOSS DAGSTUHL

Leibniz Center for Informatics

[home](#) | [browse](#) | [search](#) | [about](#) | [nfdi](#)



29th KDD 2023: Long Beach, CA, USA

[+](#)
[-](#)

[Home](#) > [Conferences and Workshops](#) > [KDD](#)

Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, Jieping Ye:

Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023. ACM 2023

Refine list

showing all ?? records

Research Track Full Papers

Florian Adriaens, Honglian Wang, Aristides Gionis:

Minimizing Hitting Time between Disparate Groups with Shortcut Edges. 1-10

Rishi Advani, Paolo Papotti, Abolfazl Asudeh:

Maximizing Neutrality in News Ordering. 11-24

Amine Allouah, Christian Kroer, Xuan Zhang, Vashist Avadhanula, Nona Bohanon, Anil Dania, Caner Gocmen, Sergey Pupyrev, Parikshit Shah, Nicolás Stier Moses, Ken Rodríguez Taarup:

Fair Allocation Over Time, with Applications to Content Moderation. 25-35

Mario Almagro, Emilio J. Almazán, Diego Ortego, David Jiménez:

LEA: Improving Sentence Similarity Robustness to Typos Using Lexical Attention Bias. 36-46

Amel Awadelkarim, Arjun Seshadri, Itai Ashlagi, Irene Lo, Johan Ugander:

Rank-heterogeneous Preference Models for School Choice. 47-56

Jiaxin Bai, Chen Luo, Zheng Li, Qingyu Yin, Bing Yin, Yangqiu Song:

Knowledge Graph Reasoning over Entities and Numerical Values. 57-68

- Q. 打开page.txt文件，观察 Track 名称、论文标题等关键元素的组成规律。从这个文本串中提取各 Track 的名称并输出（可利用字符串类型的split()和strip()方法）。
- 3. (25%) 可以看到，"Research Track Full Papers" 和 "Applied Data Track Full Papers" 中的论文占据了绝大多数，为更好地跟进数据挖掘领域学术前沿，现欲收集这两个 Track 下的论文信息。
- Q. 提取这两个 Track 下的所有论文信息（包含作者列表authors、论文标题title、收录起始页startPage与终止页endPage），并按照以下格式存储到一个字典列表中：

```
[
  {
    "track": "Research Track Full Papers",
    "papers": [
      {
        "authors": [
          "Florian Adriaens",
          "Honglian Wang",
          "Aristides Gionis"
        ],
        "title": "Minimizing Hitting Time between Disparate Groups with
Shortcut Edges.",
        "startPage": 1,
        "endPage": 10
      },
      ...
    ]
  },
  {
    "track": "Applied Data Track Full Papers",
    "papers": [
      ...
    ]
  }
]
```

```
}  
]
```

- Q. 基于上述结果，输出这两个 Track 各自包含的论文数量。
4. (10%) JSON是一种轻量级的数据交互格式，常用于存储和表示数据。
- Q. 将上一步获得的字典列表转化为 json 对象（可使用json包的相应方法），并以 2 字符缩进的方式写入kdd23.json文件中。
5. (35%) 在论文作者条目中，作者姓名可超链接到其过往发表的论文列表页面，如第一篇论文第一作者的过往发表论文信息页面如下图所示。

The screenshot displays the dblp profile for Florian Adriaens. The main content area lists publications from 2020 to the present. The right sidebar includes a 'Refine list' section with a bar chart showing the distribution of publications by year (2014 to 2024) and options to refine the list by search term, type, and coauthor.

- Q. 现要求基于之前爬取的页面文本，分别针对这两个 Track 前 10 篇论文的所有相关作者，仿照上述步骤爬取他们的以下信息：（1）该研究者的学术标识符orcID；（2）该研究者从 2020 年至今发表的所有论文信息（包含作者authors、标题title、收录信息publishInfo和年份year）。相应存储格式为：

```
[  
  {  
    "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
  },
  ...
]
},
...
]
```

请将最终结果转化为 json 对象，并以 2 字符缩进的方式写入 `researchers.json` 文件中。

## 格式要求

1. 请按具体任务分步编写代码，存储于 `.ipynb` 格式文件中用于复现，必要时可增加注释。
2. 本实验可以使用 Python 自带标准库中的所有方法实现，无需局限于任务要求中指定的方法。
3. 实验报告必须涵盖任务列表中的所有内容和相应结果，并请存储于 `.pdf` 格式文件中。

## 参考资料

以下资料可能会对你顺利完成实验有所帮助。

1. 使用 Conda 配置虚拟环境与管理安装包：[点击这里](#)
2. Conda 轻量级版本 Miniconda 的安装地址：[点击这里](#)
3. 在 VSCode 中使用 Jupyter Notebook 进行代码实现：[点击这里](#)
4. Python 官方教程（版本：3.10.13）：[点击这里](#)