

**《实验报告》**

实验题目 抽取Hugging Face的数据工具

学院系别 信息学院

专业名称 人工智能

学生姓名 郑凯航

学生学号 37220222203885

任课教师 曹冬林

2024 年 3 月 14 日

# 实验目的

列举实验要达到的几个目的。

抽取Hugging Face下列信息：

数据简介

数据页面链接

下载次数

数据大小

数据数目

# 实验环境

列举实验机器环境，配置环境。

os: Windows

platform Windows-10-10.0.22621-SP0

version: 10.0.22621

python: 3.8.18

requests==2.27.1

bs4==0.0.2

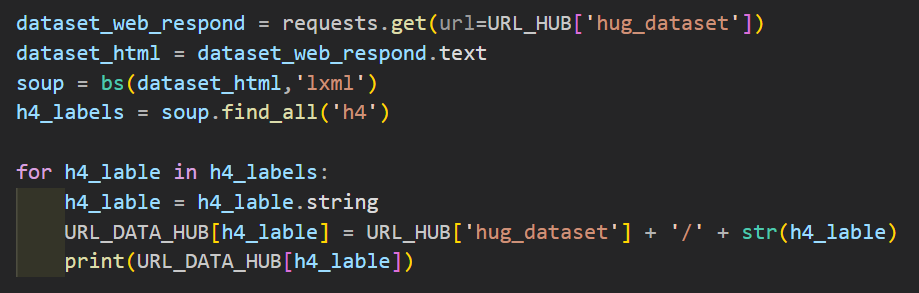
lxml==4.9.3

# 实验步骤

详细说明所做实验的基本实验步骤及流程图。

* 爬取网页，获得数据集相关路径

爬取huggingface/datasets界面，获取数据界面的链接



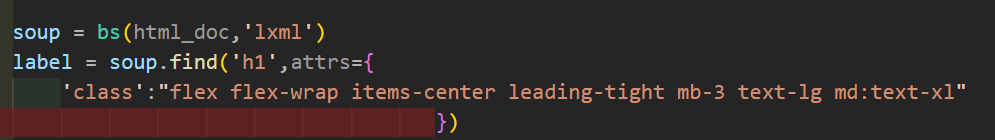
* 爬取数据集网页

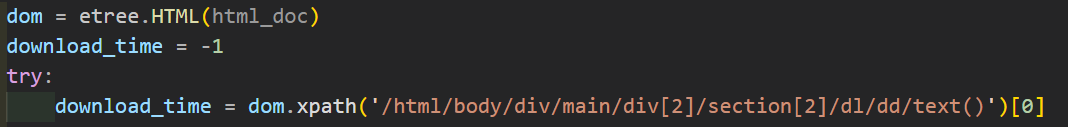
通过各个数据集的url，将html文件下载到本地，由于短时间访问次数过多会导致爬虫失败，所以设置了时间间隔，如果发现本地已经下载了相关文件，则不重复下载

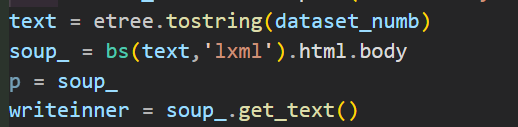


* 分析页面，抓取信息

使用xpath，bs，lxml等包分析获取的html文件，找到其中的数据集名称，下载量，数据大小等信息

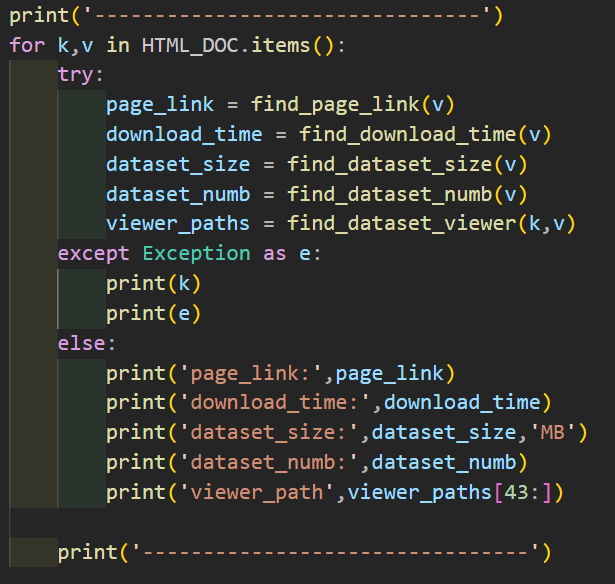






* 展示信息

遍历下载文档，找出相关信息并

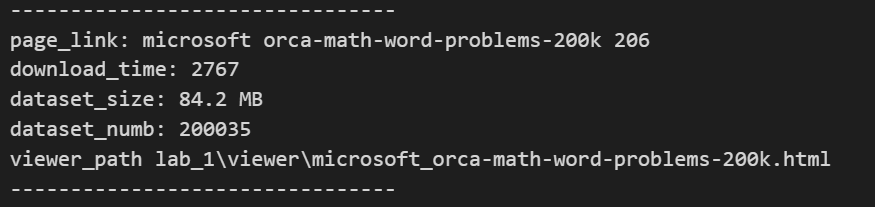


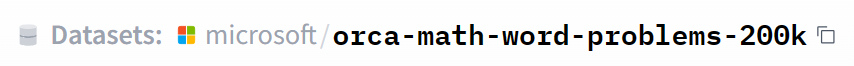
# 实验分析

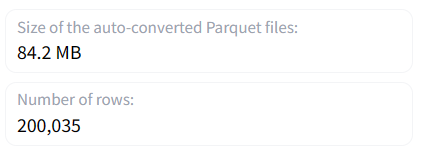
详细分析实验结果，包括程序中使用数据的基本情况、程序的正确性验证、实验的效果对比。

抓取信息与网站内容相同

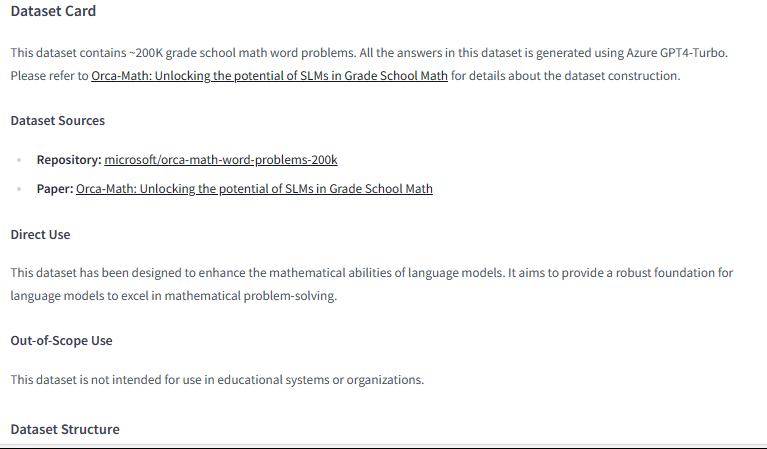
数据页面链接，下载次数，数据大小，数据数目

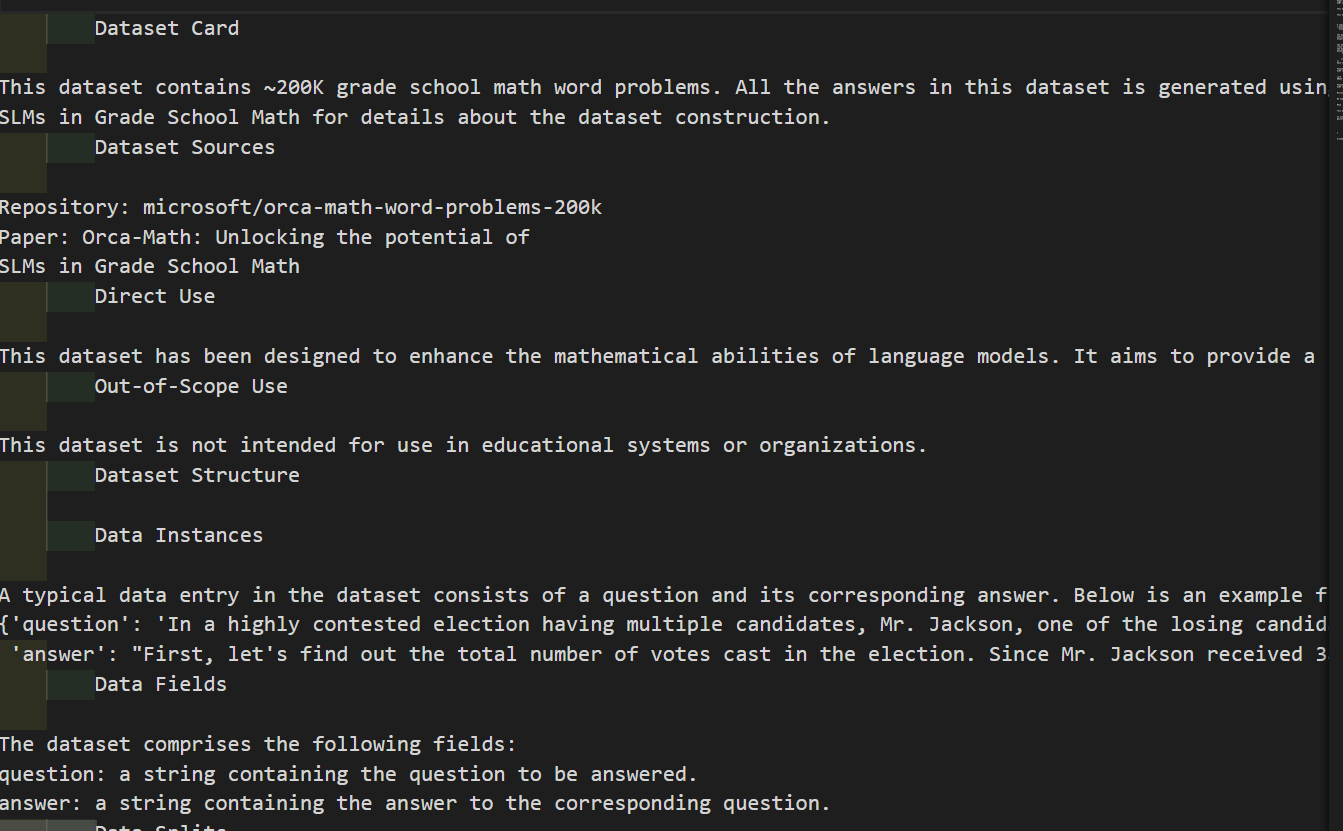




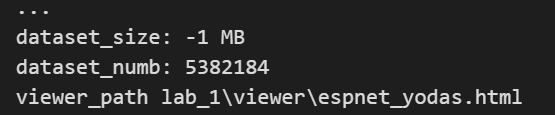


数据简介





部分网页无下载量，数据大小等，统一使用-1标识



# 问题和解决方案

详细列举实验中出现的问题以及自己所给出的解决方法。

* 短时间多次发出请求会引起报错HTTPSConnectionPool
  + 设置等待间隔，失败后重新尝试
* Html文档的处理
  + 使用BeautifulSoup，lxml等库，字符串操作等
* Html文件的保存
  + 以二进制编码形式写入，避免特殊字符无法保存
* 文档结构不统一，部分操作会报错
  + Try except else finally屏蔽报错，并使用其他处理方式

# 实验总结

总结实验效果，还有哪些地方可以继续改进。

总结：

1. 爬虫使用时可以选择一些方便框架，如bs4等，部分难定位的数据也可使用字符串和正则处理
2. 由于网页规格不完全统一，同一种搜索方式可能会不适用，可以多检查几个网页，分别找出不同的搜索方法加以使用

改进：

1. 继续爬取数据标签，
2. 文档处理方式不够通用
3. 部分文档定位不够准确