《人工智能软件开发与实践》

（ 2023 学年 秋季 学期）

作

业

报

告

学　　号：___

姓　　名：_____

班　　级：____

任课教师：_____

作业报告

实验名称： **数据下载及 JSON 封装**

成绩：

实验类别： 验证/综合型实验 **实验要求：** 1 人 1 组 **时间：** 2023 年 8 月 28 日

一、 实验目的

使用爬虫从网络上爬取 Html 源代码，解析并抽取指定信息，并封装为 Json 格式。

二、 实验内容

1. 获取网页 request

```
import requests
response = requests.get('http://www.hit.edu.cn')
print(response.content)
```

2. 提取信息 BeautifulSoup

```
import bs4
from bs4 import BeautifulSoup

# 根据html网页字符串创建BeautifulSoup对象
html_doc = """<html><head><title>The Dormouse's story</title></head><body><p
class="title"><b>The Dormouse's story</b></p><p class="story">Once upon a time there were
three little sisters; and their names were<a href="http://example.com/elsie" class="sister"
id="link1">Elsie</a>,<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and<a
href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p><p class="story">...</p>"""

soup = BeautifulSoup(html_doc,' 'html.parser'')
print(soup.prettify())
# 访问 html 节点
print(soup.title)
print(soup.title.name)
print(soup.title.string)
print(soup.title.parent.name)

print(soup.p)
print(soup.p['class'])

#指定定tag、class或id
print(soup.find_all('a'))
print(soup.find('a'))
print(soup.find(class_='title'))
print(soup.find(id="link3"))
print(soup.find('p',class_='title'))
```

```
#从文档中找到所有标签的链接
for link in soup.find_all('a'):
    print(link.get('href'))
#从文档中获取所有文字内容
print(soup.get_text())
#正则匹配
link_node = soup.find('a',href=re.compile(r"til"))
print(link_node)
```

3. 讲抽取的信息转换为 json 格式， json

```
import json
# 定义一个 Dict
my_dict = {"name": "John", "age": 30, "city": "New York"}
# 将 Dict 转换为 JSON
json_str = json.dumps(my_dict)
# 输出 JSON 字符串
print(json_str)
```

4. 以 www.hit.edu.cn 为初始种子，将本页面内含有的超链接和起文本名字，用 json 的格式返回。

三、使用的算法名称（若无，可以不填）

四、程序源码（拷贝至此处，同时作为附件和报告再一份单独的程序）

```
import requests
import re
response = requests.get('http://www.hit.edu.cn')

from bs4 import BeautifulSoup

soup = BeautifulSoup(response.content,'html.parser')
#print(soup.prettify())

import json
dic = {}
for link in soup.find_all('a'):
#    print(link.get('href'),link.get('title'))
    dic[link.get('title')] = link.get('href')
json_str = json.dumps(dic,ensure_ascii=False)
print(json_str)
```

五、程序运行结果（将程序运行结果的截图，拷贝至此处）


```
{"返回哈尔滨工业大学首页": "/main.htm", "null": "https://beian.miit.gov.cn/#/Integrated/index", " 在逐梦中国式现代化中担起国之重托 哈工大党委书记熊四皓为本科新生讲授“开学第一课”": "https://mp.weixin.qq.com/s/OFs3li6uDaFZ5zPohEbktw", "17202名哈工大新生共赴梦想向未来！ ": "https://mp.weixin.qq.com/s/R4Rczk7-TbNWogiPJBJtig", "探秘！哈工大萌新住的高质量4人间到底长啥样？ ": "https://mp.weixin.qq.com/s/i_tjSQFMcF49M5TZZvwnXg", "哈工大要全面开放校园啦！不用预约，不限名额！ ": "https://mp.weixin.qq.com/s/vymouQtlQrlxI2AqK33TaA", "杰出人才培养的“哈工大之路”": "https://mp.weixin.qq.com/s/wq_ssWk21Rxy145ppRzH5A", "在逐梦中国式现代化中担起国之重托 哈工大党委书记熊四皓为本科新生讲授“开学第一课”": "http://news.hit.edu.cn/2023/0828/c1510a234119/page.htm", "哈工大学生在第七届全国大学生计算机系统能力培养大赛CPU设计赛中荣获团队特等奖": "http://news.hit.edu.cn/2023/0828/c1510a234113/page.htm", "哈工大学子在第五届中国研究生机器人创新设计大赛中创佳绩": "http://news.hit.edu.cn/2023/0827/c1510a234111/page.htm", "哈工大航天馆入选黑龙江省首批哲学社会科学普及示范基地": "http://news.hit.edu.cn/2023/0827/c1510a234110/page.htm", "清华大学校长王希勤一行来校调研": "http://news.hit.edu.cn/2023/0826/c1510a234105/page.htm", "哈工大承办中国生态学学会微生物生态专委会学术年会暨全球华人学者环境科技前沿论坛": "http://news.hit.edu.cn/2023/0826/c1510a234103/page.htm", "近200名专家学者相聚哈尔滨 共话搅拌摩擦焊接与加工技术进展": "http://news.hit.edu.cn/2023/0826/c1510a234107/page.htm", "萌新绽放，哈工大与你一见倾“新”": "https://mp.weixin.qq.com/s/tSJ_9qq0WFvlAwzBSg7Zmg", "3516！哈工大2023级本科新生数据大揭密！ ": "https://mp.weixin.qq.com/s/HAZSixF-cVpDssu1-zDc2w", "哈尔滨工业大学加大社会科普资源供给 学习航天科技知识 激发科学探索热情": "http://paper.people.com.cn/rmrb/html/2023-08/27/nw.D110000renmrb_20230827_4-05.htm", "风雨中的守护，哈工大有力量！ ": "https://wap.peopleapp.com/article/7179952/7025374", "卓越工程师！哈工大探索工科教育的未来→": "https://mp.weixin.qq.com/s/AIY9zjeQraAHOGus31-s8w", "哈工大：动真碰硬解难题 真改实改促发展": "https://wap.peopleapp.com/article/7172692/7018591", "学习贯彻习近平新时代中国特色社会主义思想主题教育": "http://news.hit.edu.cn/xjpxsd", "贺信专题": "http://news.hit.edu.cn/hx/main.htm", "16公里，哈工大行军拉练燃动夜幕下的哈尔滨！ ": "https://mp.weixin.qq.com/s/QhJmWkZcqJ1ZNc04BofJfw", "学在哈工大，享受“私人订制”！ ": "http://tv.hit.edu.cn/2023/0612/c37a433/page.htm", "“在哈工大学三年建筑，感觉看什么都是美的”": "https://mp.weixin.qq.com/s/XM2MGy7HZCul_PKNgPYVNw", "在哈工大，“创”出未来！ ": "https://mp.weixin.qq.com/s/HbjhaqiDuotFPkXuLQve0g", "震撼发布！哈工大2023年招生宣传片《大学之大》": "http://tv.hit.edu.cn/2023/0609/c37a432/page.htm", "同班同学4人考入哈工大，期待未来飞得更高！ ": "https://mp.weixin.qq.com/s/z3oQzmdk_NUqn6FGckjsng", "艾合坦木·艾尼瓦尔：行路不辍、步步生花": "https://mp.weixin.qq.com/s/qgqItY308-Jm5MDPaXvLXw", "张凡池：带领全班“逆袭”的专业第一名": "https://mp.weixin.qq.com/s/TL41y-7EU6-KyVZMlsdfiA", "来自哈工大的他，斩获成都大运会金牌，下个目标是……": "https://mp.weixin.qq.com/s/oi1BTBJ2us1L-CfRIOrtAw", "优秀学子选择与哈工大“双向奔赴”！本科生源质量再创新高！ ": "https://mp.weixin.qq.com/s/rSNpM3ssfdgDsBlaNDJrLg", "程若思：让每一天都闪闪发光": "https://mp.weixin.qq.com/s/pMxW5HeBcHTazCbmZYV6eQ", "王培晨：跑出青春的“加速度”": "https://mp.weixin.qq.com/s/c8o77zp5fw2WzLGsJgqKBg", "生命科学中心李明晖课题组揭示疟原虫多药耐药蛋白结构和调节机制": "http://news.hit.edu.cn/2023/0819/c1510a234083/page.htm", "哈工大与中国移动共建5G应用创新联合研究院": "https://wap.peopleapp.com/article/7163175/7009696", "黑龙江省碳中和产业技术创新联盟在哈工大启
```



```
题组揭示疟原虫多药耐药蛋白结构和调节机制": "http://news.hit.edu.cn/2023/0819/c1510a234083/page.htm", "哈工大与中国移动共建5G应用创新联合研究院": "https://wap.peopleapp.com/article/7163175/7009696", "黑龙江省碳中和产业技术创新联盟在哈工大启动 哈电集团-哈工大绿色低碳能源创新研究院成立": "http://news.hit.edu.cn/2023/0810/c1510a234037/page.htm", "哈工大：在“大有可为”的新时代创造“大有作为”的新业绩": "http://news.hit.edu.cn/2023/0703/c1510a233733/page.htm", "仪器学院李浩宇教授团队突破高通量超分辨显微成像难题 研究成果在《自然光子学》发表": "http://news.hit.edu.cn/2023/0615/c1510a233600/page.htm", "生命科学中心何元政课题组揭示欧米茄3（Omega-3）脂肪酸受体4（FFAR4）激活和G蛋白偶联选择性结构基础": "http://news.hit.edu.cn/2023/0614/c1510a233574/page.htm", "哈工大牵头的科技创新2030-“新一代人工智能”重大项目“群体智能自主作业智慧农场”春季现场会暨中期检查会召开": "http://news.hit.edu.cn/2023/0609/c1510a233514/page.htm", "航天学院董永康教授团队在微波光子频率测量领域取得重要突破": "http://news.hit.edu.cn/2023/0608/c1510a233500/page.htm", "@HITers，快来晒出你的暑假清单！ ": "https://mp.weixin.qq.com/s/YlZaJyortkytmCTcfXSHxA", "学食住行，哈工大给你全新体验！ ": "https://mp.weixin.qq.com/s/742OSPHFBLbpVzOGyHDrXQ", "端午佳节，和哈工大龙舟队一起劈波斩浪！ ": "https://mp.weixin.qq.com/s/aS0qJj5rCEWv2ELJIxF0Uw", "登上《新闻联播》！哈工大火炎焱燚了！ ": "https://mp.weixin.qq.com/s/IOvEjkou-SG9Nmb-DwAtiw", "超燃！多图直击哈工大田径运动会！ ": "https://mp.weixin.qq.com/s/i3S6ky0SE8h9lt0NgHo5_A", "“方大曾校园行”特别讲座《与小方同行》走进哈工大": "http://news.hit.edu.cn/2023/0527/c1510a233361/page.htm", "校医院": "http://hityy.hit.edu.cn/", "设备共享平台": "http://hitcam.hit.edu.cn/", "哈工大报": "http://hit.ihwrm.com/", "网络电视": "http://tv.hit.edu.cn/", "学报编辑部": "http://hit.alljournals.cn/home/default.aspx", "哈工大学报（社科版）": "http://hit.alljournals.cn/home/#auditing_article_list_simple_ui", "--党群机构--": "#", "学校办公室": "http://office.hit.edu.cn/", "组织部": "http://zzb.hit.edu.cn/", "宣传部/教师工作部": "http://news.hit.edu.cn/", "博物馆": "http://www.museum.hit.edu.cn/", "统战部": "http://tzb.hit.edu.cn/", "纪委办公室": "http://qingfeng.hit.edu.cn/", "学生工作部": "http://xg.hit.edu.cn/", "保卫部": "http://bwc.hit.edu.cn/", "工会": "http://gh.hit.edu.cn/", "团委": "https://xg.hit.edu.cn/", "机关党委": "http://jgdw.hit.edu.cn", "--管理与服务机构--": "#", "本科生院": "http://hituc.hit.edu.cn/", "本科生招生办公室": "http://zsb.hit.edu.cn/", "研究生院": "http://hitgs.hit.edu.cn/", "学科建设办公室": "http://xkb.hit.edu.cn/", "科学与工业技术研究院": "http://keyan.hit.                    处": "http://xg.hit.edu.cn/", "人事处": "http://rsc.hit.edu.cn/", "国际合作部": "http://www.internatio                    计划财务处": "http://cwc.hit.edu.cn", "监察处/行政效能投诉中心": "http://qingfeng.hit.edu.cn", "审计处                    edu.cn/", "国有资产管理处": "http://gzc.hit.edu.cn/", "资产经营公司": "http://aim.hit.edu.cn/", "总务处": "http://hqjt.hit.edu.cn/", "基建处": "http://cco.hit.edu.cn/", "保卫处": "http://bwc.hit.edu.cn/", "保密处": "http://bmc.hit.edu.cn/", "离退休工作处": "http://ltxc.hit.edu.cn/", "网络安全和信息化办公室": "http://ca.hit.edu.cn/wxbgw/index", "实验室与设备管理处": "/_redirect?siteId=238&columnId=11332&articleId=218727", "--教学与科研机构--": "#", "航天学院": "http://sa.hit.edu.cn/", "电子与信息工程学院": "http://seie.hit.edu.cn/", "机电工程学院": "http://sme.hit.edu.cn/", "材料科学与工程学院": "http://mse.hit.edu.cn/", "能源科学与工程学院": "http://power.hit.edu.cn/", "电气工程及自动化学院": "http://hitee.hit.edu.cn/", "仪器科学
```

六、心得体会和遇到的困难

对 Python 语言还很陌生，对 conda 环境配置和包的安装比较陌生，造成了很多困难。学会利用互联网搜索解决问题，理解实验报告上每一个步骤对于实验目的的完成起到的作用很重要。