Python第二轮考核

- 知识点
- 推荐教程
- 初学者作业
- 有基础者作业
- 额外作业
- 作业要求
- 有想学人工智能学习的同学注意
- 预习下一轮
- 考核截止日期
- 提交方式

知识点

1. 网页架构(html+css+css选择器)
2. 网页抓包工具的使用
3. 网页请求(requests库的使用，请求头，请求参数，代理)
4. 数据提取(正则表达式，xpath(推荐)，bs4)
5. json，解码，模拟登录，反爬，简单的js逆向(可选)
6. selenium的使用
7. 数据库的使用，推荐使用mysql
8. scrapy框架的使用

推荐教程

1. b站视频 https://www.bilibili.com/video/BV1Yh411o7Sz (前面比较推荐，后面异步协程多任务有兴趣的可以学)
2. Python3 网络爬虫开发实战教程 https://cuiqingcai.com/5052.html (不要全看,重点requests，xpath，ajax，其他太杂不需要)
3. devtools https://learn.microsoft.com/zh-cn/microsoft-edge/devtools-guide-chromium/elements-tool/elements-tool (重点**元素工具和网络工具**，控制台工具，应用程序工具也比较有用，源代码工具js逆向时会用到)
4. xpath教程 https://www.runoob.com/xpath/xpath-syntax.html
5. selenium https://blog.csdn.net/IT_LanTian/article/details/122986725
6. mysql下载安装 https://blog.csdn.net/weixin_63294643/article/details/127176401 (可以下载Navicat for MySQL获取可视化页面)
7. python操作mysql https://www.runoob.com/python3/python3-mysql.html
8. github https://github.com/

初学者作业

1. 爬取福大教务通知 https://jwch.fzu.edu.cn/jxtz.htm。

   要求:

   - 获取教务通知(最近100条即可，但需要获取总页数或条数)
   - 提取通知信息中的“通知人”(如：质量办、计划科)、标题、日期、详情链接。
   - 爬取通知详情的html，可能存在“附件”，提取附件名，附件下载次数，附件链接吗，有能力请尽可能将附件爬取下来。

- 上述内容一律要去除回车、括号等无用符号
- 把除附件外爬取到的数据存入数据库中

2. 爬取百度百科历史上的今天 https://baike.baidu.com/calendar/。

   要求:

   - 获取一年内每天的历史上的今天发生了什么，包括年份，事件类型(birth、death等)，标题，简要内容
   - 上述内容一律要去除回车、括号等无用符号
   - 把爬取到的数据存入数据库中

有基础者作业

1. 用scrapy框架爬取b站评论，不要用selenium https://www.bilibili.com/。

   要求:

   - 获取视频的投币、点赞、收藏及评论总数
   - 获取该视频下的所有评论和子评论！所有评论和子评论！所有评论和子评论！(不要只有三条子评论的那种，要所有子评论)，包括评论用户，评论内容，评论时间，评论点赞数
   - 把爬取到的数据存入数据库中，注意区分是否为子评论

额外作业(给想学更多的同学的，对接下来的学习也有帮助)

1. 做一个简单的页面，可以选择复刻 http://news.fznews.com.cn/guonei/20221024/6356459aa9b15.shtml的主体部分，不要求页头和页尾，不要求js，自己找别的类似的页面进行复刻也可以。
   - 知识点：html+css
   - 推荐教程：菜鸟教程, w3cschool
2. 试用numpy、pandas等库分析初学者作业第一题中：(对人工智能感兴趣的同学尽可能尝试)
   要求:
   - 附件下载次数与通知人是否关系，若有，有什么联系?
   - 统计每天的通知数，分析哪段时间通知比较密集?
   - 作业提交请附上报告(代码运行结果及其分析)
3. 试用selenium驱动浏览器对豆瓣(https://movie.douban.com/top250)做一个长截图并保存在本地（请将源代码和截图一起提交）
4. 学习使用用matplotlib画图

   要求:

   1. 用matplotlib画正态分布曲线，其中平均值为0，标准差为1，需要画出的要点为：标题，xy轴标签，概率曲线的标签，双y轴
   2. 用matplotlib画三位圆锥，其解析式为z=-np.sqrt(x ** 2 + y ** 2)，xy的范围都为[-5,5],步进0.5，需要画出的要点为：标题，xy轴标签，三位平面，colorbar，色彩映射使用viridis

作业要求

1. 不要抄袭哦
2. 遇到不会的时候先自己去网上找资料，实在找不到再来问，用搜索引擎解决问题的能力非常重要
3. 有基础者的只需要做有基础者作业即可
4. 额外作业想做就做，也可以不做

有想学人工智能学习的同学注意

1. 继续学习机器学习基础知识

2. 额外作业第二题必须要做，也算是对数据分析和处理的熟悉过程。
3. 试着了解机器学习的一部分基础知识概念
   比如它能解决什么问题
   能分辨出需要解决的问题是什么类型
   什么是训练集 什么是测试集

预习下一轮

1. 了解js的基本知识
2. 了解一下flask库/fastapi库 (两者选一个即可)
3. 尝试用flask库/fastapi库写一个todoList
4. 了解RESTful API规范
5. 熟悉数据库的使用
6. 注册GitHub账号，学习git的使用
7. flask教程 https://dormousehole.readthedocs.io/en/latest/
8. fastapi教程 https://fastapi.tiangolo.com/

考核截止日期

2022/12/4

提交方式

请将各题文件命名为题目类型+题号（如，初学1.py， 额外1.py）后打包为压缩包

上传地址https://www.wjx.cn/vm/eIWARIm.aspx#