大作业项目总结报告

题目：A

小组成员：

边一宸（21373444）

王士举（21373455）

潘卓然（21371249）

陈绍锋（21373325）

2023年7月

一、功能简介（简述项目拥有的功能）

1. 用户的登录、注册、选择学习偏好
2. 根据用户偏好自动生产json格式的知识图谱文件
3. 导入json格式的neo4j知识图谱文件
4. 根据知识图谱文件中的结点信息，在各大课程平台上进行资源抓取
5. 在网页上可视化知识图谱
6. 拖曳知识节点改变导图构型，放大、缩小页面（鼠标滚轮）
7. 左击节点，边栏弹出对应课程资源，每种课程资源来源下属课程依据不同指标的排序列表
8. 右击节点，选择删除、增加、突出显示结点三种功能按钮
9. 左侧历史记录栏显示最近访问的结点
10. 左上搜索框可键入关键词进行课程资源搜索
11. 左下保存按钮可将json格式的知识图谱文件导出（可在下次登录时上传）

二、已完成任务

必做任务完成情况（4/4）

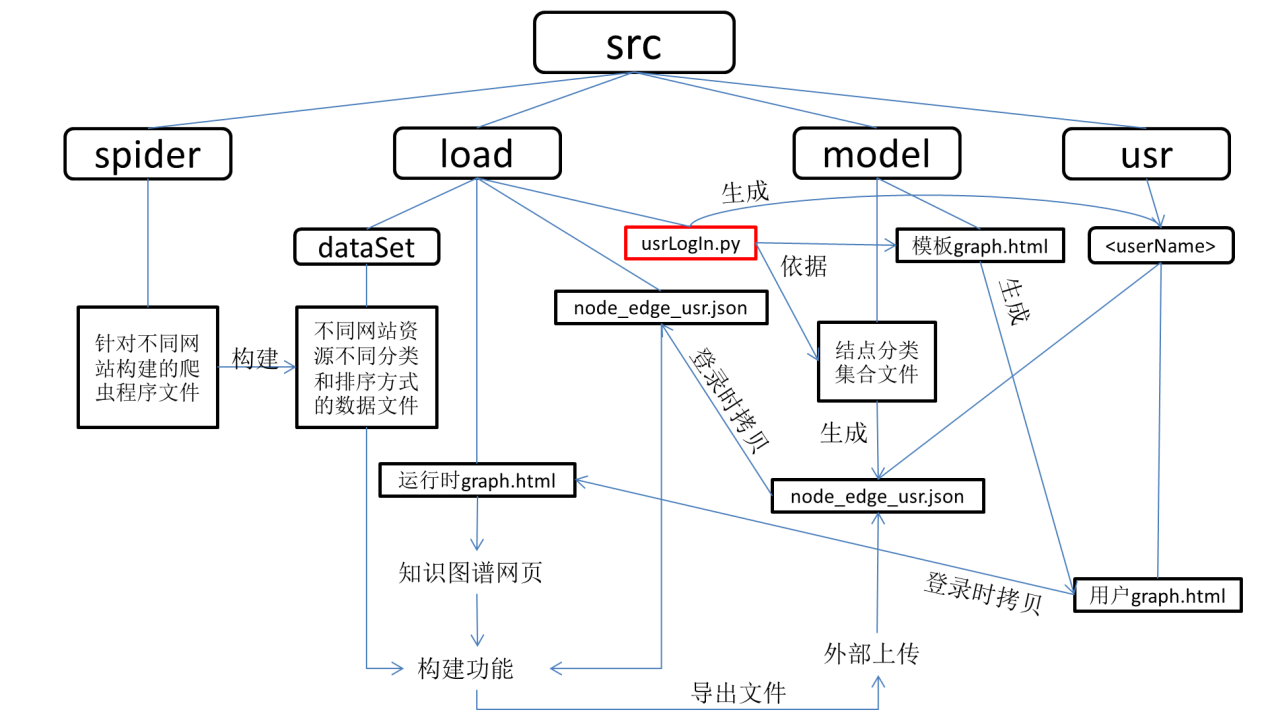
1. 开发了一个从课程学习网站中抓取课程资源的爬虫工具
2. 导入json格式的neo4j知识图谱文件并可视化。同时，提供删除、新增、突出显示知识图谱节点的方法
3. 左击节点，边栏弹出对应课程资源，每种课程资源来源下属课程依据不同指标的排序列表
4. 实现了一个用户系统，包括注册、登录和用户仪表板。知识图谱资源根据不同的用户设置的偏好或上传的自定义知识图谱文件进行区分

选做任务完成情况（3/3）

1. 实现知识图的 CRUD(创建、读取、更新、删除)操作和历史记录
2. 关键词检索功能，方便在过大的知识图谱中快速寻找资源（自定义）
3. 利用爬虫从四个网站（Bilibili，中国MOOC，好大学，iCourse）获取资源

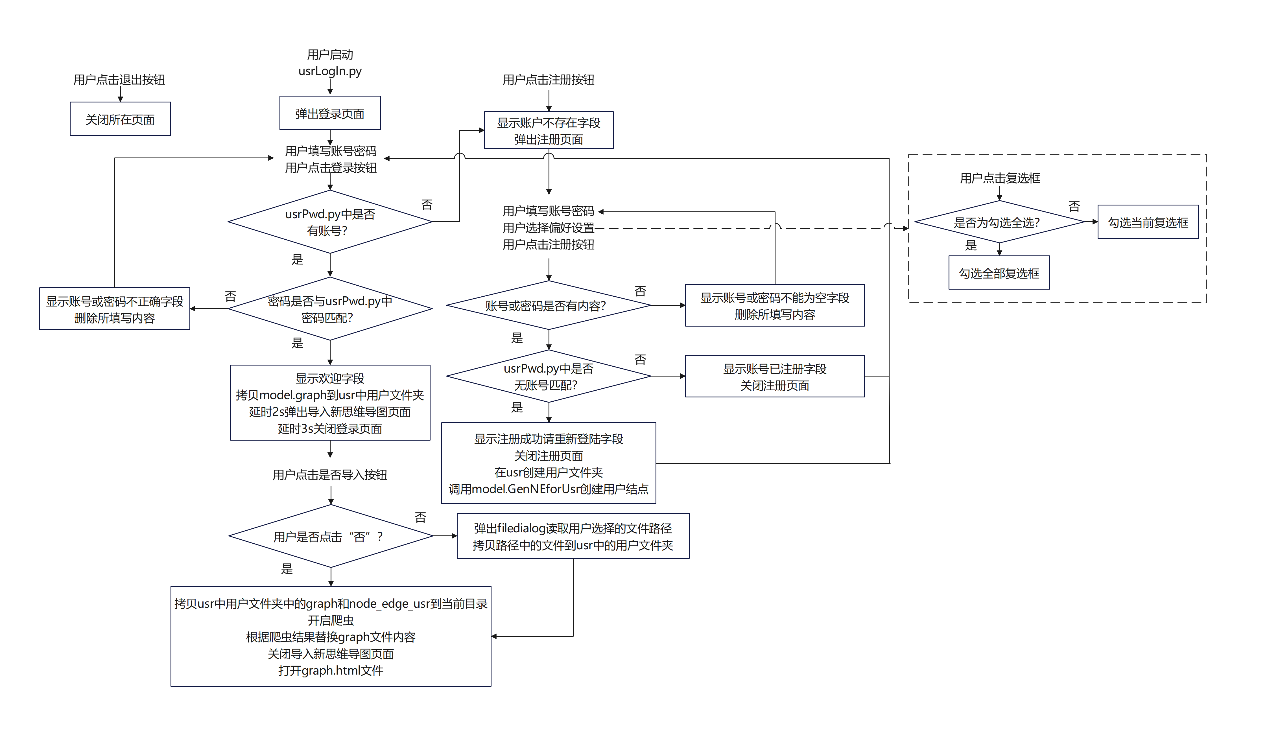
三、总体设计方案

具体实现总览：



1. 登录、注册、偏好设置、知识图谱文件上传

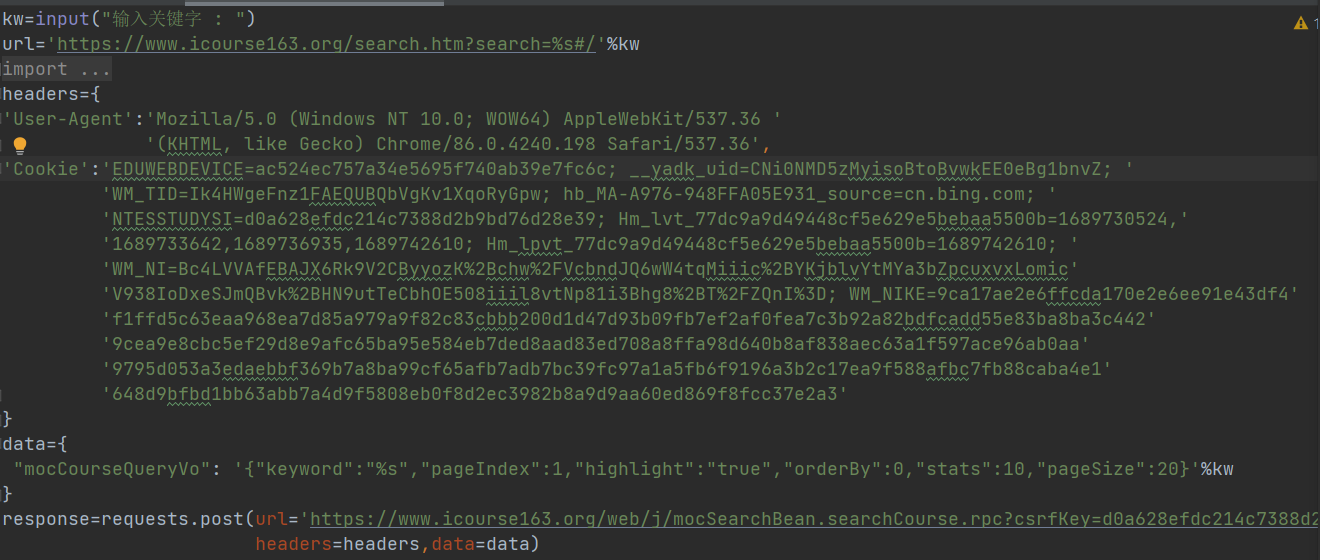
通过用户运行src/model/usrLogIn.py启动程序并弹出登录窗口。伴随用户不同点击行为可以进行登录、注册、偏好设置、知识图谱上传，具体流程如下：



其中，GUI利用ttkbootstrap库进行简要美化，注册的账号密码利用同目录下的usrPwd.json辅助储存并检查，文件上传路径通过tkinter.filedialog实现可视化文件选择，打开html文件涉及webbrowser、http.server、socketserver。

1. 课程资源爬取

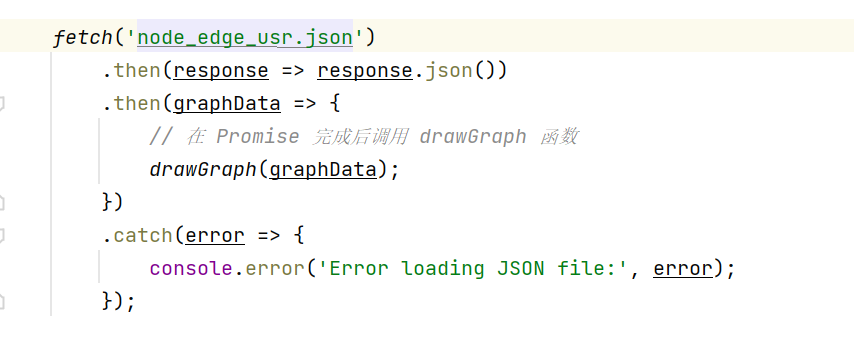
根据不同网站的特征，选择了不同的数据爬取方式。大部分网站采取了动态抓包的方法，寻找ajax请求来完成数据爬取。



Bilibili可以采用输入关键词+标号(1~5，分别对应综合，最多播放，最新发布，最多弹幕和最多收藏)的方式，爬取到相应分类标准的视频名称及url。中国大学Mooc可以输入关键词，得到相关课程的信息（url，开设学校和老师，是否国家精品等）。iCourse，好大学用法同Mooc，输入关键词即可得到课程的名称，url，开设学校和老师等信息。在每次登录时会调用爬虫模块对未提前存储的信息进行爬取，用户点击知识节点和侧边栏即可看到相关课程信息。

1. 知识图谱可视化

部署在python提供http库的构建的虚拟服务器上的graph.html文件读取load/edge\_node\_usr.json文件，根据其中的结点和边信息在页面上绘制动态知识图谱：



1. 点击显示课程资源

点击节点触发“弹出右边栏”事件，在右边栏上根据不同网站显示课程资源按钮，点击按钮触发“加载资源”事件，将load/dataset下之前爬取的课程资源信息根据分类显示在边栏上（以Bilibili资源为例）：



1. 增加、删除、突出显示结点

右击节点触发“弹出右边栏”事件，边栏上显示“新增节点”、“删除节点”、“突出显示”按钮，点击触发相应时间，其中增删操作通过修改内存中运行的结点和边信息来完成，突出显示通过改变节点颜色来完成（再次点击突出显示会取消该节点的突出显示）：



1. 历史记录和知识图谱文件导出
2. 历史记录

每次用户点击节点，触发“刷新左边栏”事件，将刚刚访问的结点名称显示在左边栏第一行，之前的记录下移：



1. 文件导出

根据文件中运行的结点和边信息，重新构建node\_edge\_usr.json文件，通过默认浏览器下载：



1. 关键词检索

读取用户输入，在用户键入回车后将目前的字符串作为关键词，在内存中的结点信息中检索，匹配则触发与点击该节点相同的事件：



四、项目运行过程

* 1. 需要安装的库见src/requirements.txt中
  2. 运行时请将src文件夹设置为Sources Root，运行src/load/usrLogIn.py即可
  3. 导入自定义知识图谱文件，请严格参照src/model/node\_edge.json提供的json格式
  4. 由于没有部署在真正的服务器上，在网页中新增节点时，节点视图及时更新，但节点课程资源不会实时更新，请在完成所有增删操作后点击保存，在下一次登录时上传保存的node\_edge\_usr.json文件，即可更新资源
  5. 运行时请勿开启代理，否则资源爬取可能失败

五、项目总结

本项目融合python前端开发、爬虫和javaScript网页开发技术，基本功能全部实现，并在基本功能的基础上做出了实用的扩展功能，为用户构建了一个可本地运行的网页版动态知识图谱系统，有着十分简洁的操作界面和流畅的操作体验，是用户在学习计算机领域知识时的强力工具。

关于改进方面：本项目的操作和显示界面基于的是用户注册时生成、登录时更新的静态html文件，实时性有待提升，后续考虑学习python后端开发的内容，将该工具部署在真正的服务器上，提升即时性，优化操作体验。

六、课程学习总结

**1、课程收获和难点分析（小组成员是否有Python或大作业要求的基础，做完这个大作业自我感觉是否有提高等其他收获，本次项目感觉最困难的地方在哪里）**

小组成员均之前没有python项目开发的基础，在团队紧密合作完成该大作业后，我们小组的收获主要有：

1. python的爬虫工具、GUI构建流程的基本掌握，对之后的学习工作大有裨益
2. javaScript网页开发的初步了解和实践，掌握了一门新语言，了解了网页开发的基本流程
3. 团队开发能力增强，架构设计能力提升，相关开发和版本管理工具的使用熟练度上升

但我们在实现该项目的过程中也遇到了不少困难，例如：

1. 由于浏览器安全的限制，html文件访问本地文件的内容在使用file模式进行打开时不会显示；解决：采用python提供的http和socketserver库构建虚拟服务器和域名，使用http模式打开。
2. 网页窗口不明所以的显示失败；解决：科学安排style规避视图上的遮盖。
3. 注册界面Checkbox复选框关联变量无法传递给触发事件绑定的函数；解决：利用bind()和全局变量关联实现交互。
4. GUI界面事件触发跳转的目标函数无法传参；解决：利用lambda表达式。
5. 爬虫时不同网站的爬取方式并不一样，需要不断摸索相关网站的爬取方式。

**2、教师授课评价（老师上课过程的一些建议，以及希望老师之后能够介绍一些什么东西）**

李老师的课程十分细致专业，对我们这些之前没有接触过python开发的同学也十分友好；希望在之后的课程中可以介绍一些python在最新的人工智能、模式识别等方面的小应用，增强课程的工程性和即时性。

**3、助教评价**

助教学长学姐平时通知消息、布置任务准确及时，在我们开发遇到困难时也会十分耐心地帮助我们，十分感谢！

**4、当前课程教授内容评价与课程进一步改进建议**

当前课程能在短短4个半天的课程中将python基本开发教授给学生是十分惊人的，希望课程平时可以添加一些有趣、有挑战性的作业，满足不同同学的需求。

七、主要参考资料

CSDN python和javaScript开发的基础知识博客

八、项目功能实际展示视频（不超过5min）

详见同文件夹下的演示文件