Python程序设计实验报告

实验题目：基于scrapy框架实现北邮、西电、成电招聘网站的信息爬取与数据处理

姓 名：刘帅

学 号：2020212267

日 期：2022年1月1日

1、需求分析

本次作业是基于scrapy框架实现北邮、西电、成电招聘网站招聘信息的爬取和信息处理的任务，因此，首先可将任务分为爬虫和信息处理两个子任务。针对爬虫部分，即spider文件的建立，则首先需要确定网页类型，html结构等。通过分析，本实验对于北邮、西电、成电均依赖selenium API实现动态爬取，进而确定爬取过程中最重要的两个流程：1、请求信息的爬取并由scrapy向网页发起request请求 2、锁定目标数据并添加至本地（items）中。针对信息处理部分，我们依赖csv和openpyxl库，将数据存储至csv，并经过数据处理将其存入至excel文件，进而实现数据的筛选。

2、程序框架

2.1爬虫文件部分

文本

描述已自动生成以北邮爬虫为例（西电、成电框架类似）

文本

描述已自动生成通过lxml的etree类，将网页的html以文本形式显示，提取每个项目的超链接，并利用Actionchain实现“下一页的跳转。”

Parse函数则是对于每个网页进行特征提取，将招聘主题、招聘日期、浏览次数（北邮多一个职位需求）提取并存入item中。

文本

描述已自动生成2.2pipeline部分

其主要功能是将内存数据读入到csv并存储，为后续数据操作进行准备。

文本

描述已自动生成2.3middlewares部分

其中，request函数较为重要，spider文件将请求信息提交至process\_request，利用scrapy框架对每个请求进行打开，进而在spider文件中实现信息的读取。

文本

描述已自动生成2.4begin部分

利用补充模块muti\_crawl进行多线程爬取，上图为程序启动指令。

3、模型分析

当scrapy工程成功的爬取到网页信息并将其存储至csv后，下一步则需要对于爬取到的数据进行数据处理，也分为如下两个子环节：1、csv文件的格式化 2、将csv文件数据进行分析并存入excel。

文本

描述已自动生成3.1 csv文件格式化

利用pandas去除csv文件的空格

文本

描述已自动生成3.2 excel数据处理

3.2.1雇主类型定义（图片请放大查看）

说明：通过观察招聘主题及信息，我们首先人为标注为教育企业、互联网企业、法律企业、通信企业、医疗企业、电子相关企业、电信运营商、互联网企业、环保相关企业、工业、航天企业、汽车企业、工商管理企业、金融企业、土木、建设类企业、其他企业。定义classify函数，对于关键字进行提取，例如出现“学院”“教”“实验室”等信息则判定为“教育企业”

文本

描述已自动生成3.2.2数据排序处理

将csv数据以字典和列表等方式存储并排序，将文件写入excel。

4、数据分析

图形用户界面, 应用程序, 表格, Excel

描述已自动生成表格

描述已自动生成图形用户界面, 应用程序, 表格, Excel

描述已自动生成图形用户界面, 应用程序, 表格, Excel

描述已自动生成4.1预处理完成后的数据

图形用户界面

低可信度描述已自动生成图片包含 散点图

描述已自动生成4.2经数据分析后的数据信息

北邮数据可视化（横坐标为标号，纵坐标为浏览次数）图表, 折线图

描述已自动生成

图表, 直方图

描述已自动生成西电数据可视化（横坐标为标号，纵坐标为浏览次数）

图表

描述已自动生成图表, 直方图

描述已自动生成成电数据可视化（横坐标为标号，纵坐标为浏览次数）

4.3结论

可以看出，互联网行业更加青睐西电学生，而教育行业更加青睐北邮和成电的学生，对于“两电一邮”学校，教育行业和互联网行业对于三所学校学生的需求最大，医疗、法律方面对于学校的需求较少。作为人工智能学院的学生，我认为我们应该更加关心智能医疗领域的教育建设和发展，将医学与科技相结合，同时，建议学校对于人文学院学生加强培养，致力于培养互联网法律相关人才。针对数据分析部分，我们可以看到北邮的最高浏览次数达到1100多次，而浏览次数的平均数在400次浏览左右；西电的最高浏览次数高达2000次，方差较小，数据的平均浏览次数在250次左右；成电的最高浏览次数为1500多次，浏览次数参差不齐，方差较大，平均浏览次数在200次左右。经推断，我认为成电的招生网站因为是研究生网站，同学们对于专业的认知更加清晰，因此对于职位需求的点击量更为有的放矢，个性化更强；而北邮和西电的招聘网站均为本科生招聘网，首先有大量的用户，其次用户的水平残次不齐，招聘信息更加泛化，所以浏览量更为平均。

4、作业小结

本次作业让我了解了scrapy爬虫框架，对前端的相关语言的了解程度加深，对于xpath、css、html均有了一定程度的认识，有助于后续搭建网页的学习。同时，本次大作业更加考核了我的综合能力，使我对python这门语言的掌握应用程度、debug过程和能力均有了提升，更为后续人工智能有关专业课程的学习打下了夯实基础。感谢老师本学期的辛苦付出，本学期python程序设计的学习使我收获颇丰。