个人周工作报告

|  |  |  |
| --- | --- | --- |
| 姓名：徐贝澄 | 项目名称：大数据实训第二周 | 填报时间：2020/11/29 |

1.本周工作内容

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目名称** | **项目阶段** | **起始时间** | **本周工作简述** | **任务是否完成** |
| 大数据实训 | 第二周 | 11月23日 | 爬取数据并处理，  集群环境运行 | 是 |
| **其它工作内容** | | | | |
| 1、第一次课，老师讲解了爬虫的基本用法，并且使用jupyter notebook进行在线交互式编程。具体步骤如下：  1. 进入目标网站，分析url。在同级之间互相比对，分析出url具体哪部分对应什么查询参数。  2. 根据对应情况，设计一个函数，函数的参数就是查询的关键词，然后根据之前得到的对应关系，使用python的.format组成最终需要的查询url。  3. 设计发送请求函数get\_url\_json() 使用requests 模块获取网站的数据。注意编码格式设置为GBK，以防中文乱码。  4. 对获取的json数据进行解析。首先使用json解析器展现数据的完整结构组成，然后找到所需要的数据的“key”，然后各自将这些数据保存到变量中。  5. 将数据一行行写入 txt 文件中。 | | | | 是 |
| 2、第二次课，使用Map-Reduce处理之前爬取的数据。目标是获取相应工作经验所能得到的工资范围。  1. 首先Map阶段拆分数据，然后使用正则表达式匹配出对应的工作经验和工资两项。  2. 在Reduce阶段，对收到工资数组的每个数据进行拆分取出最大和最小，得到整个数组的最大值和最小值。就得到了该工作经验所对应的工资范围，可写入文件中。 | | | | 是 |
| 3、 | | | | 是 |

|  |  |
| --- | --- |
| **存在问题** | **改进措施及预期结果** |
| 1、在使用Map-Reduce处理工资的时候，很多数据并不是标准形式，导致得到的结果失真。具体问题如下：  （1）起初对于一行数据，拆分成数组后var后，取var[1]为工作经验，var[2]为工资并用正则判断是否含有数据。但实际上很多公司名称中间就有空格，这样判断就会舍弃那些数据。也有数据公司名称的第二部分包含数字，这样就把工作经验当成了工资。  （2）得到的工资单位不同，有月薪，有年薪，有的单位是万，有的单位是千。而不做处理得到的结果必然是错误的。  （3）有些数据过度偏移，大大影响了结果。 | （1）进行更精确的查找。分析得到工作经验数据的特征是包含“经验”或“本科”或“硕士”；薪资的特点是包含“/月”或“/年”，这里略去“/天”，因为那样的工作每月的工作天数较少或不固定。然后使用正则表达式从数组中匹配出对应的数据。  （2）在reduce阶段用正则匹配，将薪资统一转换为k/month单位。  （3）例如“建投物联股份有限公司 10年以上经验 1.5-2千/月 南昌-青山湖区 研发总工”这样的数据，过度偏离，就手动删除。 |
| 2、在master虚拟机中，打开DataNode进程后，进程就立即自己结束。 |  |

2.下周工作安排

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **项目名称** | **项目状态** | **起始时间** | **本周工作简述** | **任务检核内容** |
| 大数据实训 | 第三周 | 11月30日 | 开始小组项目 | 完成初期目标 |
| **其它工作内容** | | | | |
| 1、小组商讨所需要的数据 | | | |  |
| 2、从网站上爬取招生数据 | | | |  |
| 3、完成Map设计 | | | |  |