

## 湖 南 石 化 职 院 教 案

日期	2020 年 6 月 1 日	周次	第十六周	时数	4 课时
教学单元	使用 Python 统计分析社交数据				
教学目标	<p>知识目标：</p> <ol style="list-style-type: none"> <li>1. 掌握 Python 程序的设计和编程能力；</li> <li>2. 掌握社交网络数据清洗和数据统计分析的方法；</li> <li>3. 掌握使用 Python 语言统计分析社交网络数据。</li> </ol>				
	<p>技能目标：</p> <ol style="list-style-type: none"> <li>1. 学习社交网络中联系人职位、 公司、 年龄等信息的数据清洗和统计分析方法；</li> <li>2. 用 Python 语言中的 Counter、 itemgetter 等模块对 LinkedIn 网站中联系人名单信息进行初步的统计分析。</li> </ol>				
	<p>素质目标：</p> <ol style="list-style-type: none"> <li>1. 树立正确的政治素质（爱党爱国、品行端正、遵纪守法）；</li> <li>2. 培养团队协作意识；培养创新能力；</li> <li>3. 树立正确的职业目标。</li> </ol>				
教学重点	<ol style="list-style-type: none"> <li>1. 掌握 Python 程序的设计和编程能力；</li> <li>2. 掌握社交网络数据清洗和数据统计分析的方法；</li> <li>3. 掌握使用 Python 语言统计分析社交网络数据。</li> </ol>				

教学难点	掌握使用 Python 语言统计分析社交网络数据
教学方法	讲授演示法；任务驱动法
教学手段	云班课信息化教学；多媒体教学；钉钉直播教学
教学过程及内容	
<p>1、新课引入</p> <p>2、知识讲解</p> <p>数据清洗：</p> <p>    数据清洗是指发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等。由于数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把脏数据清除，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再行抽取。不符合要求的数据主要是有不完整的数据、错误的的数据、重复的数据</p>	

三大类。

本实验中使用的数据来源是 LinkedIn 网站中联系人信息，需要清洗的数据主要是由于数据名称不统一造成的。例如，联系人公司中很多都带有后缀 Inc.，Co. 等，联系人职位中很多带有 Prof.，Dr. 等，这些信息在统计时会有干扰作用。例如 IBMInc. 和 IBM 代表的都是 IBM 公司，但程序在进行统计分析时会误认为是两个不同的公司。

部分实现代码如下：

```
fortransformintransforms:  
    companies[i]=companies[i].replace(*transform)
```

在该实验中，学生需用前述的数据清洗方法实现对 LinkedIn 社交网络联系人信息进行数据清洗，并在此基础上，思考如何实现清洗其他社交网络(如新浪微博)联系人信息的方法，记录 Python 代码，并分析实验结果。

Python 语句分析清洗后的社交网站联系人信息时，首先要调用 Counter 模块语句将联系人相关信息进行统计，并调用 PrettyTable 模块将联系人信息存储在表中，最后调用 print 语句按照降序打印用户信息表。

此实验部分实现代码如下

```
pt=PrettyTable(field_names=['Company','Freq'])pt.align='  
c'  
c=Counter(companies)  
[pt.add_row([company,freq])
```

```

for (company, freq) insorted(c.items(), key=itemgetter(1), reverse=True) if freq>0]

print pt

titles=[c['JobTitle'].strip() for c in contacts if c['JobTitle'].strip() != '']

```

在该实验中，学生需用前述的统计分析方法分析 LinkedIn 联系人信息，并在此基础上，思考如何实现分析其他社交网络(微博)联系人信息的方法，记录 Python 代码，并分析实验结果。

### 3、课堂总结

<p>参考资料 及应用</p>	<p>《Python3 网络爬虫开发实战》 主编：崔庆才 出版社：人民邮电出版社</p>
<p>作业</p>	<p>使用 Python 统计分析社交数据</p>
<p>教学反思</p>	