**国 家 开 放 大 学**

**大数据技术综合实训（Python）课程教学大纲**

**（审定版）**

**第一部分 大纲说明**

**一、课程的性质与任务**

大数据技术综合实训（Python）课程是国家开放大学大数据技术专业的综合实践环节课，4学分，72学时，开设一学期。

实训课程的主要内容包括：Python基础演练、网页爬取与内容可视化、基于词频的篇章语义相似度与红楼梦内容分析。

通过本课程的学习与实践，使学生在充分理解数据处理、分析、可视化的相关技术与流程；掌握获取网页数据的基本方法，掌握利用中文分词与词云图可视化分析网页内容的方法，掌握基于词频统计的篇章语义相似度分析方法，掌握Python大数据可视化的常用方法；通过基于Python编程的大数据处理、分析、可视化的实战演练，使学生将所学知识融会贯通，提高学生分析问题和解决问题的能力。整体而言，加强学生实践能力，提高学生的综合素质，以满足社会需要。

**二、与相关课程的关系**

本课程是国家开放大学大数据技术专业的综合实践环节课。其先修课为“Python语言基础”。

**三、课程的教学要求**

1. 理解数据处理、分析、可视化的相关技术与流程；
2. 掌握获取网页数据的基本方法；
3. 掌握利用中文分词与词云图可视化分析网页内容的方法；
4. 掌握基于词频统计的篇章语义相似度分析方法；
5. 掌握Python大数据可视化的常用方法。

**四、课程教学方法和教学形式建议**

1．本课程的特点是：实践性强、涉及面广，因此建议采用在计算机教室进行讲授的教学形式，讲授、实训与课堂讨论相结合。

2．实训1 “Python基础演练”可以独立安排课时，也可以拆分渗透到每个实训步骤的教学环节中。

3．实训5 “综合实训”可以综合利用前面4个实训所学内容，学生自选或由教学点安排相关数据，开展数据的统计、分析、可视化等内容。

4．指导教师应对实训内容的重难点进行集中讲解，并指导学生认真完成实训内容，保证上机机时不少于本教学大纲规定的实训学时。

5．应充分利用网络技术进行授课、答疑和讨论。

**五、课程思政要点**

本课程将课程思政要素融入教学过程，注重强化学生工程伦理教育，培养学生精益求精的大国工匠精神，激发学生科技报国的家国情怀和使命担当。

**第二部分 课程环境与考核要求**

**一、课程学时分配**

本实践环节教学总学时数为72学时，4学分。学时分配如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **章** | **教学内容** | **授课学时** | **实训学时** |
| 实训1 | Python基础演练 | 3 | 6 |
| 实训2 | 网页爬取与内容可视化 | 6 | 12 |
| 实训3 | 新闻语料数据的统计与可视化分析 | 5 | 10 |
| 实训4 | 基于词频的篇章语义相似度与红楼梦内容分析 | 6 | 12 |
| 实训5 | 综合实训 | 4 | 8 |
| 合计 |  | 24 | 48 |

**二．实训环境**

实训环境需要配置Python环境。其中实训2部分需要联通互联网以获取在线网页数据。其中，需要额外安装的Python包有：requests、jieba、tqdm、wordcloud、matplotlib。各教学点可以根据实际情况选用Jupiter Notebook、PyCharm、Visual Studio、Sublime Text等软件作为Python编程的IDE。

**三、考核及要求**

各教学点要对学生完成实训的情况进行考核，由教学点审核、验收。考核建议如下：

1．Python基础演练建议通过考勤、总结报告的形式完成；

2．实训2、实训3与实训4的内容建议通过每一实训的实训报告（实训任务、目的、代码思路、代码内容、执行结果、实训小结等）等形式完成；

3．综合实训考察综合利用前面4个实训所学内容进行大数据处理、分析与可视化的过程，通过实训报告或进行操作演示分享、PPT报告等形式考察。

4．该实训环节为本专业的必修环节，不得免修；

5．未达到该实训环节要求或抄袭作假者，按不及格处理。

6．各教学点应根据本大纲的内容和要求制定该环节的实施细则，以保证本环节的顺利实施。

**第三部分 综合实训内容与要求**

**实训1 Python基础演练（3+6=9学时）**

**一、实训内容**

复习巩固Python语言的基础语法、数据类型、控制逻辑、文件读写、函数等相关知识，为具体的实训操作打基础。

**二、实训要求**

1．了解Python数据类型与表达式；Python中的模块与包。

2．掌握Python数值表示与数值计算；Python字符串及简单字符串处理；列表、元组、字典、集合的简单运算与操作；Python中顺序、分支、循环结构的实现方法；Python基本类型的输入输出；文本文件及格式化JSON文件的读写；函数的定义与调用方法。

**实训2 网页爬取与内容可视化 （6+12=18学时）**

**一、实训内容**

爬取网页数据，提取其中中文内容，分词后绘制词云图，以分析网页内容主题和风格特点。

可选扩展内容：基于BeautifulSoup的网页内容解析、基于宽度优先搜索与网址提取的多网页内容爬取、基于多网页内容爬取和词云图可视化的网站主题分析。

**二、实训要求**

1．了解网页数据的基本格式与网页中文本内容的组织方式。

2．理解Python中中文字符编码范围，能够利用给定公式筛选字符串中的中文内容。

3．掌握基于requests包的网页内容爬取方法；基于jieba工具的中文词切分方法；基于collections.Counter的词频统计方法；基于wordcloud的词云图绘制方法。

**三、实训步骤**

本实训包含4个主要步骤，建议授课学时为6学时，建议实训学时为12学时。每个实训步骤的具体内容建议如下：

步骤1 基于requests的网页数据获取

步骤2 网页数据中的中文文本筛选

步骤3 基于jieba的中文分词与词频统计

步骤4 基于wordcloud的词云图可视化

**实训3 新闻语料数据的统计与可视化分析 （5+10=15学时）**

**一、实训内容**

读取、分析新闻语料数据，进行简单的主题统计、出现特定关键词的篇数统计、词频统计等。并针对统计结果进行可视化展示与分析。

可选扩展内容：基于词频信息与sklearn的新闻主题分类及分类结果可视化，自由设计更多具有实际意义的可视化展示与分析内容。

**二、实训要求**

1．了解大数据统计与数据可视化分析的意义。

2. 理解新闻文本信息统计的常用方法。

3．掌握饼图、柱状图、折线图、词云图等数据可视化方法。

**三、实训步骤**

本实训包含4个主要步骤，建议授课学时为5学时，建议实训学时为10学时。每个实训步骤的具体内容建议如下：

步骤1 新闻语料读取与类型统计分析

步骤2 新闻主题的饼图可视化

步骤3 月份出现篇数的柱状图、折线图可视化

步骤4 科技新闻内容的词云可视化

**四、实施建议**

步骤4开始前建议复习“实训2步骤4 基于wordcloud的词云图可视化”

**实训4 基于词频的篇章语义相似度与红楼梦内容分析 （6+12=18学时）**

**一、实训内容**

基于词频统计与给定公式，完成红楼梦篇章语义相似度计算；基于篇章语义相似度分析红楼梦作者信息；基于matplotlib与wordcloud的词频数据可视化。

可选扩展内容：基于词频信息与sklearn主题模型的篇章语义建模、基于篇章语义相似度与K-means方法的篇章语义聚类分析。

**二、实训要求**

1．了解基于词频的篇章语义相似度计算方法，能够基于给定公式或函数完成篇章语义相似度计算。

2．理解采用篇章语义相似度分析红楼梦作者信息的方法。

3．掌握针对词频数据的柱状图、折线图等数据可视化方法。

**三、实训步骤**

本实训包含3个主要步骤，建议授课学时为6学时，建议实训学时为12学时。每个实训步骤的具体内容建议如下：

步骤1 基于词频的红楼梦篇章语义相似度分析

步骤2 基于篇章语义相似度的红楼梦作者分析

步骤3 基于matplotlib与wordcloud的词频数据可视化

**四、实施建议**

步骤1开始前建议复习“实训2步骤3 基于jieba的中文分词与词频统计”

步骤3开始前建议复习“实训2步骤4 基于wordcloud的词云图可视化”与“实训3步骤3 月份出现篇数的柱状图、折线图可视化”

**实训5 综合实训（4+8=12学时）**

**一、实训内容**

综合利用实训1-4所学内容，学生自选或由教学点安排相关数据，开展统计、分析、可视化等内容。形成报告或进行分享展示。

**二、实训要求**

通过该综合实训，理解大数据处理、分析、可视化的相关技术与流程。使学生将所学知识融会贯通，提高学生分析问题和解决问题的综合能力。