|  |  |  |  |
| --- | --- | --- | --- |
| 课题名称 | 第8章 文本数据分析 | 计划学时 | 4 课时 |
| 内容分析 | 在自然语言处理领域中，文本类型的数据占据着很大的市场，由于其自身具有半结构的特点，且自然语言的分类繁多，所以针对不同的语言Python分别提供了相依的库来处理，最常见的处理英文的NLTK库，它自带的语料都是英文的，由于中文比英文的结构复杂的多，不适合NLTK进行处理，所以提供了jieba库来更好地处理中文。 |
| 教学目标及基本要求 | 1. 文本数据分析工具
2. 文本预处理
3. 文本情感分析
4. 文本相似度
5. 文本分类
6. 完成案例—商品评价分析
 |
| 教学重点 | 1. 文本数据分析工具
2. 文本预处理
3. 文本情感分析
4. 文本相似度
5. 文本分类
 |
| 教学难点 | 1. 文本预处理
2. 文本情感分析
 |
| 教学方式 | 教学采用教师课堂讲授为主，使用教学PPT讲解 |
| 教学过程 | **第一课时****（NLTK与jieba概述、安装NLTK和下载语料库、jieba库的安装）****一、创设情境，导入函数的概念**1. 教师根据课件，引出文本数据分析工具。

例如问题是：假设在电商平台中购买一件商品，如何通过商品的评价来确定这件商品是否值得买？并根据学生的回答进行总结，从而引出文本数据分析工具。本课时接下来针对文本数据分析工具进行介绍。1. 明确学习目标
* 要求学生了解NLTK与jieba
* 要求学生掌握安装安装NLTK和下载语料库
* 要求学生掌握jieba库的安装

**二、进行重点知识的讲解**1. 教师根据课件，讲述NLTK与jieba。

NLTK全称为Natural Language Toolkit，它是一套基于Python的自然语言处理工具包，可以方便地完成自然语言处理的任务，包括分词、词性标注、命名实体识别（NER）及句法分析等。jieba库主要针对中文分析而开发出文本分析库，其功能包括分词、添加自定义词典、关键词提取、词性标注、并行分词等等。1. 教师根据课件，讲述如何安装NLTK和下载语料库。

NLTK库可使用命令行下载，也可以在jupyter notebook中下载。1. 教师根据课件，讲述jieba库的安装。

安装jieba的方式比较简单，可以直接使用pip命令进行安装。**三、归纳总结，布置作业/随堂练习**1. 回顾上课前的学习目标，并对本节课的内容进行总结。

教师总结本节课需要掌握的知识点，包括NLTK与jieba的概述、安装NLTK和下载语料库、jieba库的安装。1. 布置随堂练习，检查学生掌握情况。

根据在线教学平台和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。1. 使用日照职业技术学院在线教学平台下发课后作业。

**第二课时****（文本预处理的流程、分词、词性标注）****一、回顾上节课的内容，继续讲解本课时的知识**1. 教师对学生们的疑问进行统一答疑。
2. 回顾总结上节课内容，继续介绍本课时的内容。

上节课介绍了文本数据分析工具，接下来，本节课继续介绍文本预处理。1. 明确学习目标
* 要求学生掌握预处理的过程
* 要求学生掌握分词
* 要求学生掌握词性标注

**二、进行重点知识的讲解**1. 教师根据教材中图8-3所示，讲解文本预处理的流程。

文本预处理一般包括分词、词性归一化、删除停用词。1. 教师根据课件，讲解文本预处理中的分词。

分词是指将由连续字符组成的语句，按照一定的规则划分成一个个独立词语的过程。不同的语言具有不同的语法结构，以常见的英文和中文举例，英文的句子中是以空格为分隔符的，所以可以指定空格为分词的标记，而中文并没有一个形式上的分界符，它只有字、句和段能通过明显的分界符来简单地划分。因此，中文分词要比英文分词困难很多。1. 教师根据课件，讲解什么是词性标注。

词性是对词语分类的一种方式。词性标注，又称词类标注，是指为分词结果中的每个单词标注一个正确的词性，也就是说确定每个单词是名词、动词、形容词或其它词性的过程。比如在“I love itcast”中，“I”为人称代词，“love”为动词，“itcast”为名词。**三、归纳总结，布置作业**1. 回顾学习目标，对本节课的内容进行总结。

教师总结本节课需要掌握的知识点，包括预处理的流程、分词、词性标注、。1. 布置随堂练习，检查学生掌握情况。

 根据在线教学平台和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。 1. 使用日照职业技术学院在线教学平台下发课后作业。

**第三课时****（词形归一化、删除停用词、文本情感分析）****一、回顾上节课内容，继续介绍本课时的内容**1. 教师对学生们的疑问进行统一答疑。
2. 回顾上节课的知识，继续讲解本节课的内容。

上节课介绍了文本预处理操作，接下来，本节课继续介绍关于文本预处理的其他操作。1. 明确学习目标
* 要求学生掌握词性归一化
* 要求学生掌握删除停用词
* 要求学生熟悉文本情感分析

**二、进行重点知识的讲解**1. 教师根据课件，讲解词形归一化的使用方法。

在英文中，一个单词常常是另一个单词的变种，比如looking是look这个单词的一般进行式，looked为一般过去式，这些都会影响语料库学习的准确度。一般在信息检索和文本挖掘时，需要对一个词的不同形态进行规范化，以提高文本处理的效率。1. 教师根据课件，讲解删除停用词操作。

停用词是指在信息检索中，为节省存储空间和提高[搜索](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2/1806)效率，在处理[自然语言](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80/4146019)文本之前或之后会自动过滤掉某些没有具体意义的字或词，这些字或词即被称为停用词，比如英文单词“I”、“the”或中文中的“啊”等。停用词的存在直接增加了文本的特征难度，提高了文本数据分析过程中的成本，如果直接用包含大量停用词的文本作为分析对象，则还有可能会导致数据分析的结果存在较大偏差，通常在处理过程中会将它们从文本中删除。1. 教师根据课件，讲解什么是文本情感分析。

文本情感分析，又称为倾向性分析和意见挖掘，是指对带有情感色彩的主管性文本进行分析、处理、归纳和推理的过程。情感分析还可以细分为情感极分析、情感程度分析及主客观分析等。其中，情感极分析的目的在于，对文本进行褒义、贬义、中性的判断，比如对于“喜爱”和“厌恶”这两个词，就属于不同的情感倾向。**三、归纳总结，布置作业**1. 回顾学习目标，对本节课的内容进行总结。
2. 布置随堂练习，检查学生掌握情况。

 根据在线教学平台和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。 （3） 使用日照职业技术学院在线教学平台下发课后作业。**第四课时****（文本相似度、文本分类）****一、回顾上节课内容，继续介绍本课时的内容**1. 教师对学生们的疑问进行统一答疑。
2. 回顾上节课的知识，继续讲解本节课的内容。

上节课介绍了文本预处理操作和文本情感分析，接下来，本节课介绍关于文本相似度和文本分类的操作。1. 明确学习目标
* 要求学生熟悉文本相似度
* 要求学生熟悉文本分类

**二、进行重点知识的讲解**1. 教师根据课件，讲解什么是文本相似度。

在自然语处理中，经常会涉及到度量两个文本的相似性的问题，在诸如信息检索、数据挖掘、机器翻译、文档复制检测等领域中，如何度量句子或短语之间的相似度显得尤为重要。1. 教师根据课件，讲解什么是文本分类。

文本分类是指按照一定的分类体系或标准，用电脑对文本集进行自动分类标记，主要的目的是将文本或文档自动地归类为一种或多种预定义的类别。通俗说，就是拿一篇文章问计算机，这篇文章说的究竟是美食、体育还是政治。**三、归纳总结，布置作业**1. 回顾学习目标，对本节课的内容进行总结。
2. 布置随堂练习，检查学生掌握情况。

 根据在线教学平台和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。1. 使用日照职业技术学院在线教学平台下发课后作业。

**第五课时****（上机练习）**上机练习主要针对本章中需要重点掌握的知识点，以及在程序中容易出错的内容进行练习，通过上机练习可以考察同学对知识点的掌握情况，对代码的熟练程度。**上机一：（练习教材中的示例代码）****形式：单独完成****第六课时****（上机练习）**上机练习主要针对本章中需要重点掌握的知识点，以及在程序中容易出错的内容进行练习，通过上机练习可以考察同学对知识点的掌握情况，对代码的熟练程度。**上机一：（练习教材中的示例代码）****形式：单独完成****上机一：（完成案例—商品评价分析）****形式：单独完成** |
| 思考题和习题 | 见教材第8章配套的习题 |
| 教学后记 |  |