《自然语言处理》课程教学大纲（2020 版）

<table>
<tr><th colspan="6">课程基本信息（Course Information）</th></tr>
<tr><td>课程代码
(Course Code)</td><td>FL3308</td><td>*学时（Credit Hours）</td><td>32</td><td>*学分
(Credits)</td><td>2</td></tr>
<tr><td rowspan="2">*课程名称
(Course Name)</td><td colspan="5">（中文）自然语言处理</td></tr>
<tr><td colspan="5">（英文）Natural Language Processing</td></tr>
<tr><td>课程类型
(Course Type)</td><td colspan="5">必修课
compulsory course</td></tr>
<tr><td>授课对象
(Target Audience)</td><td colspan="5">英语(语言学)本科大三学生
Third-grade English undergraduates (Linguistics Major Seniors)</td></tr>
<tr><td>授课语言
(Language of Instruction)</td><td colspan="5">双语
Bilingual</td></tr>
<tr><td>*开课院系
(School)</td><td colspan="5">外国语学院
School of Foreign Languages</td></tr>
<tr><td>先修课程
(Prerequisite)</td><td>语料库语言学导论
An Introduction to Corpus Linguistics</td><td>后续课程
(post)</td><td colspan="3"></td></tr>
<tr><td>*课程负责人
(Instructor)</td><td>管新潮</td><td>课程网址
(Course Webpage)</td><td colspan="3"></td></tr>
<tr><td>*课程简介（中文）
(Description)</td><td colspan="5">（中文 300-500 字，含课程性质、主要教学内容、课程教学目标等）
本课程旨在培养和提升外语类学生在面对与本学科相关的数据类型时的技术逻辑思维能力和运用水平，专注于探索语言知识与技术的融合性分析路径，所涉技术为 Python 编程语言。Python 的优势在于文本处理，如句法结构分析、语料库文本标注、语境识别、文本摘要、文本聚类、文本情感分析、相似性度量、语义分析、主题建模、语料库语言学多变量设置等。学习目标是利用 Python 获取更多可应用于描述学科研究目的的数据信息。与课程相关的学习目标是以语料库挖掘手段，通过词法、句法和语义分析，从语言数据中获取研究用统计信息，理解并掌握语言数据分析的统计原理。教学方式以案例讲解为主，其中的工具案例用于描述技术工具的适用性和可靠性，解决技术应用之前有关编程技术的知识问题；语言学路径案例则紧密结合语言学/翻译学知识探索如何以技术手段解决教学科研中的相关问题。</td></tr>
<tr><td>*课程简介（英文）
(Description)</td><td colspan="5">（英文 300-500 字）
This course is aimed to improve foreign language students' technical logical thinking ability and application level when facing data types related to this discipline, and to focus on doing an integration analysis of language knowledge and technology. The technology involved is Python</td></tr>
</table>

	programming language. Python's advantage lies in text processing, such as syntactic structure analysis, text tagging, context recognition, text summarization, text clustering, text sentiment analysis, similarity measurement, semantic analysis, topic modeling, corpus linguistics multivariate setting and so on. The goal is to use Python to obtain more data information that can be used to describe the research purpose of the discipline. The learning goal related to the curriculum is to obtain statistical information for research from language data by means of corpus mining and lexical, syntactic and semantic analysis, and to understand and master the statistical principle of language data analysis. The teaching method is mainly based on different cases, in which tool cases are used to describe the applicability and reliability of technical tools and solve the knowledge problems about programming technology before technology application; the cases about different linguistic paths closely combine different knowledge from linguistics or translation studies to explore how to solve related problems in teaching and scientific research.

课程目标与内容（Course objectives and contents）

*课程目标 (Course Object)	1.能了解自然语言处理在本学科的基本应用方法，认识到技术应用有其独特的语言学意义，提升专业自信心。（A3） 2.能以创新方式将自然语言处理技术应用于语言学问题的处理，掌握并提升解决相关问题的能力。（B1） 3.善于发现语言学与大数据相结合所产生的新问题，并为此应用合适的自然语言处理技术，养成独立解决问题的能力。（B3） 4.能开展自然语言处理技术的语言学适应性验证，以确定相应技术是否适宜于为语言学目的所用。（B2） 5.能通过语言知识与技术的融合性学习，解决涉及语言产品的细腻性问题，适应社会对高素质语言专业人才的需求。（B5） 6.能善于思考语言与技术融合过程中的算法问题，认识到语言学在现代社会中的价值与意义。（C3）

	章节	教学内容（要点）	学时	教学形式	作业及考核要求	课程思政融入点	对应课程目标
*教学内容进度安排及对应课程目标（Class Schedule & Requirements & Course Objectives）	示例：						
	1	自然语言处理概述 内容说明：本节概述内容涉及：1. 数据与数据特征 2. 自然语言 3. 语言句法和结构 4. 文本语料库及其标注 5. 自然语言处理 6. 语境识别 7. 文本摘要 8. 文本聚类 9. 文本分析 10. 机器学习 11. 深度学习	2	以讲述为主	阅读文章	认识到与语言学相结合的自然语言处理其社会价值和意义，提升专业热情	1，2
	2	Python 数据结构	2	案例讲	复杂数据	认识到数据结	1，2

		内容说明：设想从语言数据中获取有价值的语言信息并用于教学或科研活动，首先必须把文本形式（以文本语料为例）的语言数据转换为适宜于 Python 处理的特定数据结构。特定的教学或科研活动需要特定的语言信息，只有以特定的数据结构呈现的语言信息方能满足特定的教学或科研目的。换言之，若能将语言数据转换为复杂的语言数据结构，便能获取复杂的语言信息。语言数据结构越复杂，所获取的语言信息就越能表示语言数据本身所蕴含的信息和意义。		解和讨论	结构的转换	构在获取语言数据信息中的意义和作用，培养严谨认真的专业态度	
	3	文本数据清洗 内容说明：本节课试图对不同的语言数据清洗方法进行归纳总结，以求具有针对性地说明数据清洗的复杂和繁琐。通过学习清洗方法和案例，可掌握较为系统地清洗语言数据的方法，也能根据特定的语言数据运用合适的清洗方法。	2	案例讲解和讨论	Ngram 数据清洗	不同语言数据的不同清洗方式，培养严谨认真的专业态度	1，2
	4	编程中的正则表达式 内容说明：本节内容涉及：（1）正则表达式的基础内容如常见实例、特殊字符、重复字符、选择性字符、指定字符等。(2) re 模块如相关函数、标志位等。（3）应用案例，如连续与非连续结构、成语匹配、指定术语等。	2	案例讲解和讨论	提取首字母是元音的所有单词	掌握信息提取的重要手段，培养严谨认真的专业态度	1，2
	5	短语数据处理工具 内容说明：作为局部语法研究内容之一的短语学，	2	案例讲解和讨论	从语料文本中提取有效的	掌握短语数据的语言学意义，认知专业	3，4，5，6

		产生于计算语言学信息处理的现实需求，现已被确立为语言学的一个专门学科，并应用于自然语言处理等领域。局部语法涵盖词汇、句法、语义、语用等内容，融合形式、意义、功能于一体，是语料库语言学的新发展，为短语学提供了一条新的研究路径。从技术发展的现实角度出发，所述短语学是指语料库短语学。语料库短语学是以单语或双语中短语意义单位为基元，基于语料库研究范式进行语言学的相关研究。			N-grams	知识的社会意义和价值	
	6	N-grams 分析应用 内容说明：N-grams 在各类研究中有多种用处，在本次论文解读中用于预示语言学研究趋势。无论应用在何处，其关键是如何提取有效的 N-grams。本节课从学术论文解读开始，了解不同工具提取多连词的价值和意义（结合学术论文解读分析 NLTK 和 spaCy）。	2	案例讲解和讨论	学术文本模糊短语的弱化表述手段	区分短语学不同应用可能性，认知专业知识的社会意义和应用价值	3，4，5，6
	7	情感分析工具 内容说明：情感分析工具可分为英文类、中文类、混合类三种，能够实现影评、产品评价、公众舆论、政治、预测等方面的简易、复杂、高级情感分析，一般有极性分析和词表分析两类。将情感分析工具用于相关分析时，须思考工具包的组合应用效力。	2	案例讲解和讨论	不同工具的应用价值区分	识别具体工具的应用可能性，认知专业知识的社会意义和价值	3，4，5，6
	8	情感分析理论与应用 内容说明：情感是指人与	2	案例讲解和讨	情感分析与传统民	情感分析工具在话语建构中	3，4，5，6

		人之间以及人与特定对象之间的连接关系和精神依赖，是构成社会归属关系的重要维系纽带。而情感分析旨在以量化方式为体现这一社会纽带关系的情感给出特定的数值，以作为情感评价的判断基础，进而实现人与人、人与特定对象之间关系的精准描述，为可接受的行为给出定位取向。情感分析技术已在市场营销、贸易和公共部门等领域广为接受，它可以识别出人类语言的叙述特点，或者利用公共数据资源来评估和预测公共反馈意见且具有良好的精准度，或者通过分析社交媒体或在线论坛上的信息情感能够为商业企业创造不可计数的商业价值。		论	意调查比较	的作用，认知专业知识促进话语权的社会意义和价值	
	9	相似性度量工具 内容说明：相似性度量工具可分为词汇类、句子类、语篇类、词汇与句子类四种，旨在区分不同应用类型下的具体特征。词汇相似性度量与句子类、语篇类区别较大，但其也是后两者的基础。词汇相似性度量不仅涉及知识库方法如 WordNet 方法，也列举了时下多有应用的词向量方法如 spaCy 方法，同时也描述了传统的互信息方法。句子类与语篇类相似性度量颇为相似如 gensim 方法或 spaCy 方法，涉及度量部分的代码是相同的，区别仅在于语料文本的加载方法。	2	案例讲解和讨论	相似性度量三种方法比较	掌握知识库的重要作用，认知专业知识的社会意义和价值	3，4，5，6

	10	相似性度量与文本分析 内容说明：相似性度量方法的应用旨在发现文本所特有的规律性，即译本是具有偏向机器翻译的特征还是人工翻译的特点，或者同一词汇在不同文本语境下其搭配概念是否一致，或者以文本相似性聚类方法是否可以判断语料库构成的平衡性。这一方法的应用可能性应该是多种多样的，既可进行纯粹的相似性度量，也可结合其他技术，主要在于应用场景是否适宜于技术的应用。	2	案例讲解和讨论	多译本相似性度量	掌握相似性度量方法的机器翻译应用，认知专业知识的社会实践意义	3，4，5，6
	11	语义分析工具 内容说明：可供语义分析的工具有多种，如信息贡献度方法、语义网资源、语义网络分析等，还有词向量（词嵌入）模型如 Word2Vec 模型和 FastText 模型、向量模型如 LSI 模型、LDA 模型。	2	案例讲解和讨论	语义分析模型的区分	掌握四类模型的效果知识，认知专业知识的社会意义和价值	3，4，5，6
	12	语义分析与相关模型及其分析路径 内容说明：无论是局部描述，还是系统呈现，语义分析可在文本挖掘分析中具有巨大潜力和各种应用可能性。本节课将以多样性、系统性、针对性为视角展开文本的语义分析，旨在研究语义分析工具与语言学或翻译学之间的可融合性，为提取更多有效的语言信息探索一条可行之路。语义分析的相关模型为语义迁移与分布式词向量、语义主题词与信息贡献度、语义关系与语义网、文本语	2	案例讲解和讨论	著作权法/版权法概念 copyright 词向量关联性	了解新技术对专业学科的发展所起的作用，认知专业发展的意义	3，4，5，6

		义与语义网络分析。					
	13	主题建模工具 内容说明：本节课选用 gensim 和 sklearn 工具包等的相关主题建模工具进行描述，旨在探究具体工具的应用效果。应用主题建模工具的关键是主题数的设置，但无任何其他信息可供参照的情况下，主题数的设置可能是一种盲从，必须付出极大的工作量才会有所收获，或者颗粒无收。所以，了解并掌握具体工具的基本功能（无论是多主题数还是一维主题），才能有效结合其他技术手段，去实现语料文本的深层次主题挖掘。	2	案例讲解和讨论	区分主题建模中的 gensim 和 sklearn 方法	掌握主题建模技术的作用及其局限性，认知专业知识的社会意义和价值	3，4，5，6
	14	主题建模中的主题挖掘 内容说明：本节课以学术文本、新闻文本、法律法条文本为主体建模的语料，旨在探索主题建模方法对不同体裁的适用性。这种适用性不仅体现为语料的区别，也表现在不同主题建模技术的应用上。如何实现不同体裁语料和不同主题建模技术的有效结合，是主题建模应用的关键。本节将以三个案例说明这种结合的实际意义和作用。	2	案例讲解和讨论	话语分析中的主题建模适用性	掌握话语分析中的技术应用，认知专业知识的社会意义和价值	3，4，5，6
	15	变量设置工具 内容说明：本节课尝试以改进或创新方式从三个层面引入变量设置工具，即词汇层面、句子层面和语篇层面。变量设置工具以词汇层面的居多，这与词汇是句子和语篇的基	2	案例讲解和讨论	语料库语言学变量与概率及其分布	掌握概念的统计学知识，认知专业知识的社会意义和价值	3，4，5，6

		本构成单位和基础不无关系。只有充分设置了词汇层面的工具，才有机会上升至其他层面。					
	16	变量设置的学理意义 内容说明：语料库语言学以真实语言数据为研究对象，凭借计算机技术，采用数据驱动的实证主义研究方法，从宏观的角度对大数量的语言事实、对语言交际和语言学习的行为规律进行多层面或建模研究，尤其是提供有关语言使用的概率及其分布信息，这就为语言学研究提供了新途径、新方法，必将加深对语言本质的理解。其中的概率和概率分布是变量设置的关键。	2	案例讲解和讨论	词汇复杂性/成熟度的教材词汇评估应用	掌握变量设置的任务属性，认知专业知识的社会意义和价值	3，4，5，6
	注 1：建议按照教学周周学时编排。 注 2：相应章节的课程思政融入点根据实际情况填写。						
*考核方式 (Grading)	示例： （1）平时作业 30 分 （2）期末大论文 70 分						
*教材或参考资料 (Textbooks & Other Materials)	（必含信息：教材名称，作者，出版社，出版年份，版次，书号） 《语料库与 Python 应用》，管新潮，上海交通大学出版社，2018，第 1 版，书号 978-7-313-19748-1 Text Analytics with Python (2019)，D. Sarkar，APRESS/Springer，2019，第 2 版，书号 978-1-4842-4353-4 《Python 3：语料库技术与应用》，陆晓蕾、倪斌，厦门大学出版社，2021，第 1 版，书号 978-7-5615-7727-1						
其它（More）							
备注（Notes）							

备注说明：

1．带*内容为必填项。

2．课程简介字数为 300-500 字；课程大纲以表述清楚教学安排为宜，字数不限。