中文信息处理（HUM040）课程大纲

1、2019-2020 学年秋季学期 - 2021-2022 学年春季学期

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	**中文信息处理** **Introduction to Chinese Information Processing**
2.	**授课院系** **Originating Department**	人文科学中心 Center for the Humanities
3.	**课程编号** **Course Code**	HUM040
4.	**课程学分 Credit Value**	2 学分 2 Credits
5.	**课程类别** **Course Type**	通识选修课程 General Education (GE) Elective Courses
6.	**授课学期** **Semester**	2019-2020 年秋季 2019-2020 Fall
7.	**授课语言** **Teaching Language**	中文 Chinese
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师）** **Instructor(s), Affiliation& Contact** **（For team teaching, please list all instructors）**	孙顺，讲师，河北师范大学文学院 Sun Shun, Lecturer E-mail: sunshun_2011@163.com
9.	**实验员/助教、所属学系、联系方式** **Tutor/TA(s), Contact**	无 NA
10.	**选课人数限额(可不填)** **Maximum Enrolment（Optional）**	

11.	授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
	学时数 Credit Hours	32				32

12.	先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	无
13.	后续课程、其它学习规划 Courses for which this course is a pre-requisite	无
14.	其它要求修读本课程的学系 Cross-listing Dept.	无

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

1、了解中文信息处理领域的主要工作内容。

2、了解中文信息处理的基础原理。

3、熟悉一些中文信息处理工具和语料库。

4、掌握中文信息处理的一些基础操作。

16. 预达学习成果 Learning Outcomes

1、了解中文信息处理学科的性质和发展史。

2、通过具体的实例操作，理解中文信息处理技术的基础原理，如语料库、匹配、规则、统计、语料标注等。

3、熟悉一些常用语料库，比如国家语委现代汉语语料库、《人民日报》标注语料库等。

4、熟悉 nltk 中文分析工具包、jieba 自动分词工具包等常用中文信息处理工具包。

5、学会利用网络资源，搜寻并运用中文信息处理工具包。

6、学会利用 matlab 和 python 写一些简单的程序。

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

双周上课，每次上 4 课时，共 8 次课。课程内容如下：

第一次课：（4 课时）

第一讲 绪论（2 课时）

课程要求、中文信息处理简介（内涵、发展史）、Matlab 安装、Python 安装。

第二讲 英语人名译名（一）（2 课时）

中文信息处理的基本原理。用 Excel 构建一个语料库，通过 Matlab 编程将一个英文人名自动音译成汉

语，如 Smith 音译为“史密斯”。通过这一操作理解：为什么要构建语料库？何为匹配？

第二次课：（4 课时）

第三讲 英语人名译名（二）（2 课时）

接续上一讲任务。理解：何为穷举？何为规则？两者的优劣？应该构建什么样的语料库？

第四讲 汉字繁简转换（一）（2 课时）

中文信息处理的基本原理。用 Excel 构建一个语料库，通过 Matlab 编程将简体文章转成繁体文章。如：“明月几时有，把酒问青天”转成“明月幾時有，把酒問青天”。通过这一操作理解：单纯匹配的局限。

第三次课：（4 课时）

第五讲 汉字繁简转换（二）（2 课时）

用 Excel 构建一个语料库，通过 Matlab 编程将简体文章转成繁体文章。通过这一操作理解：为什么要多个语料库衔接？何为算法？

第六讲 语料获取与分析：NLTK 工具包（一）（2 课时）

nltk 工具包安装、nltk.book 语料导入、词汇匹配。

第四次课：（4 课时）

第七讲 语料获取与分析：NLTK 工具包（二）（2 课时）

利用 nltk 工具包计算词频。

第八讲 网页语料获取与分析（一）（2 课时）

利用 urllib 工具包获取网页信息，中文信息的转码。

第五次课：（4 课时）

第九讲 网页语料获取与分析（二）（2 课时）

HTML 网页无效信息的过滤及相关中文信息分析技巧。

第十讲 中文语料自动分词（一）（2 课时）

jieba 工具包、匹配分词、匹配分词的局限。

第六次课：（4 课时）

第十一讲 中文语料自动分词（二）（2 课时）

其他的分词策略。

第十二讲 中文语料自动分词（三）（2 课时）

分词策略的算法实现。

第七次课：（4 课时）

第十三讲 语料库标注（2 课时）

语料库标注的作用、方式，以及标注过的语料库的运用。

第十四讲 中文信息处理语料库介绍（一）（2 课时）

国家语委现代汉语语料库、古代汉语语料库等。

第八次课：（4 课时）

第十五讲 中文信息处理语料库介绍（二）（2 课时）

接续上一讲任务，介绍一些常见的中文数据库以及它们的基本功能。结合一些研究论文，展示研究者如何利用这些中文数据库。

第十六讲 中文信息处理技术进阶管窥（2 课时）

数据挖掘的基本知识、常用的聚类算法，信息处理效果展示。舆情分析、用户画像、机器翻译、语音识别与言语合成等。

期末考核内容：

二选一完成课程项目或者期末报告。

评分标准：

评分等级	具体标准
A+ （97-100）	1、在基础要求上文章具备高度的洞察性和创造性；设计的程序针对中文信息处理领域的前沿问题，程序设计合理，语句简洁； 2、文章善于综合全面地利用各次课程中教授给予的工具和知识； 3、显示出学术或创作上继续发展的高度潜力； 4、无无故缺席。
A （93-96）	1、在基础要求上文章具备较高的洞察性和创造性；设计的程序针对中文信息处理领域的前沿问题，程序设计合理； 2、文章善于综合全面地利用各次课程中教授给予的工具和知识； 3、显示出学术或创作上继续发展的潜力； 4、无无故缺席。
A- （90-92）	1、在基础要求上文章具备一定的洞察性和创造性；设计的程序针对中文信息处理领域的一般性问题，程序设计合理，语句简洁； 2、文章善于综合全面地利用各次课程中教授给予的工具和知识； 3、显示出继续发展的潜力。
B+ （87-89）	1、在基础要求上文章具备一定的洞察性和创造性；设计的程序针对中文信息处理领域的一般性问题，程序设计合理； 2、文章善于利用课程中教授给予的大部分工具和知识。
B （83-86）	1、在基础要求上文章具备某些洞察性和创造性；设计的程序针对中文信息处理领域的一般性问题，程序有一定漏洞； 2、文章善于利用课程中教授给予的一部分工具和知识。
B- （80-82）	在基础要求上文章具备某些洞察性和创造性，或者文章善于利用课程中教授给予的某些工具和知识；设计的程序为课程练习，程序设计合理，语句简洁。
C+ （77-79）	课堂表现，平时作业和大作业均无较大问题，有较多亮点；设计的程序为课程练习，设计合理。
C （77-79）	课堂表现，平时作业和大作业均无较大问题，有部分亮点；设计的程序为课程练习，程序有一定漏洞。
C- （70-72）	课堂表现，平时作业和大作业均无较大问题，亦无亮点；设计的程序为课程练习，程序有较大漏洞。
D+ （67-69）	课堂表现，平时作业和大作业有较大问题但有较多亮点；设计的程序为课程练习，程序有无法执行。
D （63-66）	课堂表现，平时作业和大作业有较大问题但有部分亮点。
D- （60-62）	课堂表现，平时作业和大作业均有较大问题但有亮点。
F （0-59）	未完成基础要求，或课堂表现，平时作业和大作业有严重问题亦无亮点。

18. 教材及其它参考资料 Textbook and Supplementary Readings

教材：

中文信息处理原理及应用（第 2 版），苗夺谦、卫志华、张志飞，清华大学出版社，2015。

参考资料：

1、Matlab 揭秘，David McMahon 著，郑碧波 译；
2、Python 基础教程，（挪威）Megnus Lie Hetland 著，袁忠国 译，人民邮电出版社，2018；
3、树莓派 Python 编程入门与实战(第 2 版)，Richard Blum 等著，人民邮电出版社；
4、数据挖掘导论（Introduction to Data Mining），陈封能 等著，人民邮电出版社。

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Class Performance		20		
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		20		
期中考试 Mid-Term Test				
期末考试 Final Exam				
期末报告 Final Presentation				
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)		50		期末考核：二选一完成课程项目与期末报告。

20. 记分方式 GRADING SYSTEM

√ A. 十三级等级制 Letter Grading
☐ B. 二级记分制（通过/不通过） Pass/Fail Grading

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority

同意开设。

人文中心教学负责人：

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

2、2022-2023 学年秋季学期起

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	**中文信息处理** **An Introduction to Chinese Information Processing**
2.	**授课院系** **Originating Department**	人文科学中心 Center for the Humanities
3.	**课程编号** **Course Code**	HUM040
4.	**课程学分 Credit Value**	2 学分 2 Credits
5.	**课程类别** **Course Type**	通识选修课程 General Education (GE) Elective Courses
6.	**授课学期** **Semester**	春季 / 秋季 Spring / Fall
7.	**授课语言** **Teaching Language**	中文 Chinese
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师）** **Instructor(s), Affiliation& Contact** **（For team teaching, please list all instructors）**	孙顺，助理教授，人文科学中心 Sun Shun, Center for the Humanities, Assistant Professor Email: suns@sustech.edu.cn
9.	**实验员/助教、所属学系、联系方式** **Tutor/TA(s), Contact**	无 NA
10.	**选课人数限额(可不填)** **Maximum Enrolment（Optional）**	

11.	授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
	学时数 Credit Hours	32				32

12.	先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	无
13.	后续课程、其它学习规划 Courses for which this course is a pre-requisite	无
14.	其它要求修读本课程的学系 Cross-listing Dept.	无

教学大纲及教学日历 SYLLABUS

15. **教学目标 Course Objectives**

1、了解中文信息处理领域的主要工作内容。

2、了解中文信息处理的基础原理。

3、熟悉一些中文信息处理工具和语料库。

4、掌握中文信息处理的一些基础操作。

16. **预达学习成果 Learning Outcomes**

1、了解中文信息处理学科的性质和发展史。

2、通过具体的实例操作，理解中文信息处理技术的基础原理，如语料库、匹配、规则、统计、语料标注等。

3、熟悉常见的中文文本信息分析技术，如中文分词、词性标注、句法分析。

4、熟悉常见的中文语音信息分析技术，如语音 FFT 变换、LPC 特征、MFCC 特征、波形合成、参数合成、HMM-GMM 语音识别。

5、熟悉一些常用语料库，比如国家语委现代汉语语料库、《人民日报》标注语料库等。

6、熟悉 nltk 中文分析工具包、jieba 自动分词工具包等常用中文信息处理工具包。

17. **课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）**
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

1-16 周上课，每次 2 课时，共 16 次课。课程内容如下：

第一次课：绪论（2 课时）：
内容简介：主要介绍课程内容及学科发展史。
第二次课：繁简转换（2 课时）：
内容简介：中文繁简转换的匹配问题。

第三次课：中文分词：词典分词（2 课时）：
内容简介：中文规则分词，利用词典进行分词。

第四次课：中文分词：统计分词（2 课时）：
内容简介：中文统计分词，主要介绍隐马模型思路。

第五次课：关键词提取（2 课时）：
内容简介：关键词、停用词及文本距离。

第六次课：词性标注（2 课时）：
内容简介：汉语词类的统计标注。

第七次课：作业报告：中文文本分析（2 课时）：
内容简介：第一次作业讨论。

第八次课：句法分析（2 课时）：
内容简介：句法信息的层次标注，主要介绍 CYK 算法。

第九次课：词向量（2 课时）：
内容简介：词向量与文本距离。

第十次课：语音的数字编码（2 课时）：
内容简介：语音研究的三个角度、数字编码及程序读写。

第十一次课：语音的时域分析（2 课时）：
内容简介：过零率、短时能量与语音的声调分析。

第十二次课：语音的频域分析（2 课时）：
内容简介：语音频域分析的基频原理。

第十三次课：语音的频域参数（2 课时）：
内容简介：频域分析参数，含 FFT、LPC、DOC、MFCC 等。

第十四次课：语音合成（2 课时）：
内容简介：语音合成原理，含波形合成及参数合成。

第十五次课：语音识别（2 课时）：
内容简介：语音识别实例，利用 MFCC 识别元音或发音人。

第十六次课：作业报告：中文语音分析（2 课时）：
内容简介：第二次作业讨论。

18. 教材及其它参考资料 Textbook and Supplementary Readings

教材：
中文信息处理原理及应用（第 2 版），苗夺谦、卫志华、张志飞，清华大学出版社，2015。

参考资料：

1. Daniel Jurafsky，James H.Martin 著，冯志伟 译，自然语言处理综论（第二版），电子工业出版社，2018
2. 涂铭 等，Python 自然语言处理实战：核心技术与算法， 机械工业出版社，2018
3. 吕士楠 等，汉语语音合成:原理和技术，科学出版社，2012
4. 宋知用， MATLAB 语言信号分析与合成（第 2 版），北京航空航天大学出版社，2017
5. 洪青阳，语音识别：原理与应用，电子工业出版社，2020

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		40		两次平时作业，分组进行。
期中考试 Mid-Term Test				
期末考试 Final Exam				
期末报告 Final Presentation				
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)		50		期末考核：课程项目与期末报告二选一。

20. 记分方式 GRADING SYSTEM

√ A. 十三级等级制 Letter Grading
☐ B. 二级记分制（通过/不通过） Pass/Fail Grading

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority

同意开设。

人文中心教学负责人：
年　　月　　日