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**	信息检索 Information Retrieval
2.	**授课院系 Originating Department**	计算机科学与工程系 Department of Computer Science and Engineering
3.	**课程编号 Course Code**	CS332
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	春季 Spring
7.	**授课语言 Teaching Language**	中英双语 English & 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）**	宋轩，副教授，计算机科学与工程系，Songx@sustech.edu.cn Xuan Song, Associate Professor, Department of Computer Science and Engineering, songx@sustech.edu.cn
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
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	32	0	32	0	64

Credit Hours					

12.	先修课程、其它学习要求 **Pre-requisites or Other Academic Requirements**	CS203 数据结构与算法分析 Data Structures and Algorithm Analysis
13.	后续课程、其它学习规划 **Courses for which this course is a pre-requisite**	无 none
14.	其它要求修读本课程的学系 **Cross-listing Dept.**	无 none

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

With the development of modern internet usage and extreme large amount of metadata produced each day, the need to search for relevant information becomes a vital part of technology. Search engines such as Google and Bing, are utilizing methods in information retrieval to bring to users fast and accurate search results. In this course, we will also learn about the basics in information retrieval and its web-based and text-based applications.

伴随着现代互联网的使用和每天产生的大规模数据，如何能更快的搜索到关联性更高的信息变得愈发重要。像谷歌和必应这些搜索引擎，正在使用信息检索技术向用户们提供更快、更准确的搜索结果。在信息检索这门课中，我们将会学习到信息检索的基础原理以及其在 Web 和文本检索方面的应用。

16. 预达学习成果 Learning Outcomes

After this course, students should be able to:

1. To gain an understanding of the basic concepts and techniques in Information Retrieval;
2. To understand how statistical models of text can be used to solve problems in IR, with a focus on how the vector-space model and the language model can be applied to the document retrieval problem;
3. To understand how statistical models of text can be used for other IR applications, for example clustering;
4. To appreciate the importance of data structures such as an index to allow efficient access to the information in large bodies of text;
5. To have experience of building a document retrieval system, through the practical sessions, including the implementation of a relevance feedback system.

学习了这门课后，学生们应该能够做到：

1：掌握信息检索的基础概念和相关技术。

2：理解基于文本的概率模型是如何用于解决信息检索领域的问题，同时也能掌握向量空间模型和语言模型以及它们在在文档检索方面的应用。

3：理解概率模型在信息检索的其他方面的应用，比如文本聚类。

4：能够懂得数据结构的重要性，比如索引在大规模的文本查询中的作用。

5：通过一系列实验课程，包括实验课上对相关反馈系统的实现等方面，能掌握如何构建一个文档检索系统。

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.)

共计 64 小时，每周两小时理论课及两小时实验课。

第一周： 信息检索绪论

实验课 1：Python 介绍和 IDE 安装

第二周：布尔检索及倒序索引

实验课 2：实现倒排索引

第三周：词汇表和倒排记录表

实验课 3：实现布尔搜索

第四周：词典和容错式检索

实验课 4：使用一种方式实现拼写校正

第五周：索引的构建

实验课 5：实现内存式单遍扫描索引构建算法

第六周：索引的压缩

实验课 6：实现变长编码和解码算法

第七周：文档评分及向量空间模型

实验课 7：对文档进行 Jaccard 系数和 tf-idf 权重计算

第八周：一个完整的检索系统

实验课 8：使用 tf-idf 权重计算方式实现简单的向量空间模型

第九周： 前沿文献阅读汇报和讨论

实验课 9： 前沿文献方法实现

第十周：检索系统的评价

实验课 10：对搜素系统进行评价

第十一周： 相关反馈、查询扩展和 XML 检索

实验课 11：XML 检索

第十二周：概率检索模型和语言检索模型

实验课 12：选择一种模型实现检索

第十三周：文本分类

实验课 13：选择一种方法进行文本分类

第十四周：文本聚类

实验课 14：选择一种方法实现文本聚类

第十五周：网络爬虫和链接分析

实验课 15：选择一种方法进行文本爬虫

第十六周：项目最终答辩

实验课 16：项目程序演示和验收

64 hours in total. 2 hours lecture and 2 hours lab for each week.

Week1: Introduction to Information Retrieval

Lab1: Introduction to Python and Installation of the Python IDE

Week2: Boolean Retrieval and Inverted Indices

Lab2: Implement of Inverted Indices

Week3: Term Vocabulary and Posting List

Lab3: Implement of Boolean query

Week4: Dictionary and Tolerant Retrieval

Lab4: Use one method to realize spelling correction

Week5: Index Construction

Lab5: Implement of Single-pass in-memory indexing algorithm

Week6: Index Compression

Lab6: Implement of Variable length encoding and decoding algorithm

Week7: Scoring and Vector Space Model

Lab7: Computing Jaccard coefficient and tf-idf weighting for documents

Week8: A Complete Search System

Lab8: Use tf-idf weighting to realize a simple Vector Space Model

Week9: Mid-term Exam

Lab9: Review

Week10: Evaluation

Lab10: Evaluation for a search system

Week11: Relevance Feedback and Query Expansion, and XML Information Retrieval

Lab11: XML Retrieval

Week12: Probabilistic Information Retrieval and Language models for Information Retrieval

Lab12: Choose one of models to realize retrieval

Week13: Text Classification

Lab13: Choose one of methods to realize text classification

Week14: Text Clustering

Lab14: Choose one of methods to realize text clustering

Week15: Web Crawling and Link analysis

Lab15: Implement a crawling of text

Week16: Final Exam

Lab16: Review

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

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/

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				
课堂表现 Class Performance		10%		随机随堂测验 Radom in-class quizzes
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		40%		平时上机实验 Lab Assignments
期中考试 Mid-Term Test		20%		前沿文献讨论和报告 Paper Reading, Discussion and Presentation
期末考试 Final Exam				
期末报告 Final Presentation		30%		课程项目最终报告、答辩和程序验收 Final Project Presentation

其它（可根据需要改写以上评估方式） **Others (The above may be modified as necessary)**				

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