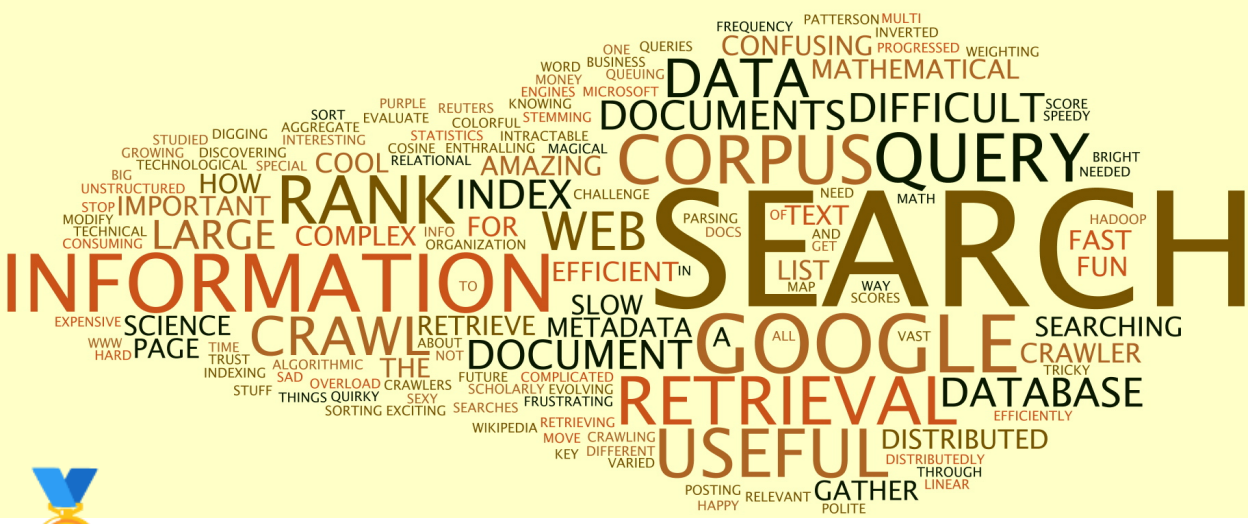

信息检索

一、根据录屏所得

三类题型：

填空题，30空每空1分

5个名词解释，每个8分，会有详细要求（要写公式、工作手段、思想、应用、实现方法等）

3个简答题，每个10分

思路：理解课程思想，每一章工作思想。名词解释和简答题可能没有标准答案，要自己总结。考理解和思想，不考编程。

第一讲 信息检索

第四讲 信息检索应用不考

信息检索用于什么场景：解决信息过载

用于海量的非结构化数据，不想顺序扫描，希望全文搜索。主要功课：建立索引，空间为代价。

定义

信息检索本质（掌握） L2R不考

信息检索模型（掌握）、几个部分，是什么

搜索引擎了解

第二讲 搜索引擎工具

Lucene、相关工具无简答题和名词解释。

Lucene存储内部结构、Lucene操作编程不考

考察：Lucene是什么、Luke是什么、Nutch和Lucene有什么关系

第三讲 词项词典（重要）

一、文档解析和七、开源NLP库要求较低

其余全部要求掌握，不要求有例子

第七部分掌握有哪些开源工具的名字

第四讲 中文分词

分词是什么：（汉字序列）（单独的词）

分词有几种方法，掌握字符串的方法和统计的方法

统计的方法—隐马尔可夫模型（熟练掌握）：

> 隐马尔可夫定义（P22，23）、五元组和三要素、三个基本问题（知道什么求什么，用什么算法能做出来）、解码问题、维特比算法（代码和图示不掌握，理解思想，**描述**）

知道几个开源中文分词软件

第五讲 布尔模型与倒排索引

第四部分 Lucene索引的文件结构要求低不考

第五部分的跳表部分不考

信息检索模型、分类

布尔模型：

> 词包模型概念、布尔操作不考

倒排索引概念、有什么组成、如何构建

5.3掌握：两种查询—二元、位置索引，什么情况选哪个

布尔模型特点掌握

第六讲 向量空间模型

重要

第七讲 概率检索模型

tf-idf计算，知道三要素即可

概率思想，不掌握公式推导，记结论。

BIM模型：信息检索的概率模型，推导不考，记住模型要求RSV，记住需要p和r值，记住怎么得到的p和r值。记住是IDF，通过迭代知道p和r。（理解结论、怎么用、概念）

BM25概念、结论公式（P60）理解、意义。

第八讲 主题模型

去掉一般的矩阵分解、解决PCA问题

EM算法怎么做的（定义不考，中间环节不掌握）知道用什么参数估计什么参数、怎么用、特点

P36例子

迪利克雷先验不考定义

不考概率图

LDA模型中文档的生成方式

哪个工具包帮助完成工作

第九讲 检索排序

L2R不考，其他重点

其他排序要求低

第十讲 搜索引擎优化

第四部分spam 检测不考

第五部分只掌握SEO概念

第十一讲 信息检索评价

一检测基础要求低

二三部分掌握

四部分要求低

第十二讲 相关反馈及查询扩展（无大题）

用户查询过程

去掉质心、Rocchio算法、相关反馈中的假设、相关反馈策略的评价、隐式相关反馈、伪相关反馈

第十三讲 统计语言模型

不考预处理模型（预训练语言模型）和NN统计模型

考统计语言模型