标红加粗的部分是对计算题画的重点

第一讲布尔检索：了解倒排记录表是什么；简单的布尔检索如 and 和 or 如何合并词项。

第二讲词汇表和倒排记录表：词条到词项需要经过词条化处理，如分词、去停用词、词干还原等等，要知道是在做什么；了解跳表指针、双词索引、位置索引的原理。

第三讲词典及容错式检索：通配查询要知道不同索引的处理方法；**编辑距离和 Jaccard 系数的计算。**

第四讲索引构建：BSBI 和 SPIMI 要掌握其原理，有时间的同学建议稍微看细一点。

第五讲索引压缩：有损压缩和无损压缩；词典压缩中的按块存储；**倒排记录表压缩中可变字节编码和γ编码计算。**

第六讲文档评分、词项权重计算及向量空间模型：**tfidf 的计算，权重计算三要素掌握词项频率的 n 和 l、文档频率的 n 和 t 以及归一化方法的 n 和 c。**

第七讲完整搜索系统：**静态得分的使用方法。**

第八讲检索评价：**R-Precision 的定义，F1 值的计算，差值正确率的计算方法，未插值的 AP，宏平均、微平均、MAP 的计算。**

第九讲索引扩展：知道隐式相关反馈是什么，了解什么是显式相关反馈、隐式相关反馈、伪相关反馈；**Rocchio 的计算。**

第十讲概率模型：线性回归、岭回归、Lasso 回归的定义；BIM 模型的大概原理、公式推导的最终结果以及参数的统计方法。

第十一讲语言模型：**利用最大似然估计的语言模型构造。**

第十二讲朴素贝叶斯：朴素贝叶斯的条件独立性和位置独立性假设；**多项式和贝努力 NB 分类器的计算。**

第十三讲向量空间分类：PCA 和 LDA 的基本原理及特点；互信息的计算方法；Rocchio 算法的性质；KNN 分类器原理。

第十四讲 SVM：

第十五讲扁平聚类：掌握 K 均值聚类算法的目标、原理及其特点；**纯度和兰迪指数的计算。**

第十六讲层次聚类：了解单连接、全连接、质心及 GAAC 算法的原理；了解二分 K 均值算法的原理。

第十七讲隐性语义检索：**SVD 低秩逼近的处理方法以及 F 范数的计算**，了解 SVD 为什么有效。

第十八讲 web 搜索：了解 web 图的蝴蝶结形结构；掌握基于重叠区域的相对大小估算算法；了解次高竞标价格拍卖机制；**知道如何通过 shingling 估计两个文档的 Jaccard 相似度。**

第十九讲信息采集：了解采集器的基本的采集过程；了解分布式索引中逻辑文档分区、物理文档分区和词项分区的基本原理。

第二十讲链接分析：**PageRank 的计算；**HITS 算法的基本原理与相关概念。