中国科学院大学
University of Chinese Academy of Sciences

试 题 专 用 纸

开课编号：**201M4006H**
课程名称：信息检索导论
任课教师：李波 林政 古晓艳

学生姓名：_____ 学号:_____ 培养单位：_____ 分数：____

一、选择题（单选，每题 2 分，共 20 分）

1. 关于布尔检索，以下说法正确的是（ ）

A) 在布尔检索系统中，进行词干还原从不降低正确率

B) 在布尔检索系统中，进行词干还原从不降低召回率

C) 词干还原会增加词项词典的大小

D) 词干还原应该在构建索引时调用，而不应在查询处理时调用

2. 关于倒排索引，以下说法错误的是（ ）

A) 词典的开销通常小于倒排记录表的开销 ✓

B) 倒排索引中往往会存储词项的频率 ✓

C) 词典中存储的都是词

D) 倒排索引可以加快搜索的速度

3. 关于向量空间模型的特点，下面说法不正确的是（ ）

A) 支持部分匹配和近似匹配，结果可以排序 ✓

B) 理论上不够严谨，往往基于直觉的经验性公式 ✓

C) 词项之间的独立性假设与实际不符 ✓

D) 应用于检索问题时，效果不如概率检索模型

4. 关于tf-idf，以下说法不正确的是（ ）

A) 一个罕见词的idf往往很高 B) idf的大小是无限的

C) 词项的tf-idf权重可以超过1 ✓ D) 词项t在所有文档中出现，则权重取值很小 ✓

5. 关于隐式相关反馈，下面说法错误的是（ ）

A) 不需要用户显式参与，减轻用户负担 ✓

B) 对行为分析有较高要求 ✓

C) 一定能提升检索准确率

D) 某些情况下需要增加额外设备 ✓

6. SVM中的正则化因子（也称为惩罚因子）是指（　）

A) 交叉验证的次数　　B) 用到的核函数

C) 分类准确性和模型复杂度之间的权衡　　D) Lagrange乘子

7. 把一个连接的Web网络看作一个蝴蝶结结构，其中20%的页面为IN，60%页面在SCC区域，20%的页面为OUT。假设从3个起始页面开始爬取网页，目标是尽可能多地访问此Web中的页面。请问从下面哪一个选项中选取三个网页作为初始种子网页能爬取最多的网页？（　）

A) 从IN、OUT各选一页　　B) SCC选一页，IN选两页

C) SCC选两页，IN选一页　　D) 从IN中选三页

8. 关于聚类结果评价的外部准则，不包括以下哪种方法（　）

A) 兰迪指数（Rand Index）　　B) 残差平方和（RSS）

C) 归一化互信息（NMI）　　D) 纯度（Purity）

9. 假设有两个Web搜索引擎A和B，现从各自索引中随机均匀地生成大量网页。A的30%网页出现在B的索引中，而B的50%网页出现在A的索引中。请问，相对于B搜索引擎，A索引的页数是多少？（　）

A)　|A| = 5/3 |B|　　B)　|A| = 3/5 |B|

C)　|A| = 2/3 |B|　　D)　|A| = 3/2|B|

10. 关于原始PageRank算法，以下说法不正确的是（　）

A) PageRank是基于链接分析的排序算法 ✓

B) 节点的PageRank值依赖于Web图的链接结构 ✓

C) 网页的PageRank与用户输入的查询相关

D) PageRank在随机游走过程中访问越频繁的网页越重要 ✓

二、判断题，将 √ 或 × 填在括号内（每题 2 分，共 20 分）

1. 在信息检索系统中，词条的个数往往不小于词项的个数。（）✓
2. BIM概率模型中，没有考虑词项在文档中出现的次数。（）
3. 非线性分类器的效果一定优于线性分类器。（）✗
4. 贝努利朴素贝叶斯算法考虑了词项在文档中出现的次数。（）✓
5. 原始词项-文档矩阵经过LSI变换以后会得到一个维度更小的词项-文档矩阵。（）✓
6. 文本分类算法中的参数可以在测试集上调优。（）
7. 统计语言模型的阶数越高(即n-gram中n更大)，需要估计的参数也越多。（）
8. K-均值聚类算法选用某个初始聚类种子集合，有可能不收敛。（）✗

9. 伪相关反馈一定能提高每个查询的检索效果。（） ✗

10. 倒排索引中的VB编码压缩方法是一种无损压缩方法。（）✓

三、计算题（每题 8 分，共 32 分）

1. 给定矩阵C，$C=\begin{pmatrix}1 & 1\\ 0 & 1\\ 1 & 0\end{pmatrix}$，利用SVD分解，得到分解结果为$U=\begin{pmatrix}-0.816 & 0.000\\ -0.408 & -0.707\\ -0.408 & 0.707\end{pmatrix}$，

$\sum=\begin{pmatrix}1.732 & 0.000\\ 0.000 & 1.000\end{pmatrix}$，$V^T=\begin{pmatrix}-0.707 & -0.707\\ 0.707 & -0.707\end{pmatrix}$计算矩阵$C$的1-秩逼近$C_1$，并给出该逼近下的F范数误差值。

2. 已知采用欧式归一化方法对表1中三个文档的tf值进行归一化，得到Doc1、Doc2和Doc3的$\sqrt{\sum_{i=1}^{M}V_i^2(d)}$值分别是30.56、46.84和41.3，

	Doc1	Doc2	Doc3
北京	27	4	24
中国	3	33	0
天安门	0	33	29
科技	14	0	17

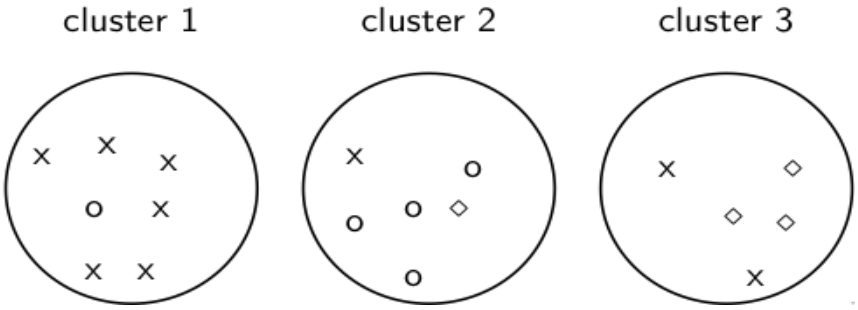
表1 tf值

a) 请计算tf归一化的结果，填入下表：

	Doc1	Doc2	Doc3
北京	0.884	0.085	0.581
中国	0.098	0.705	0
天安门	0	0.705	0.702
科技	0.458	0	0.412

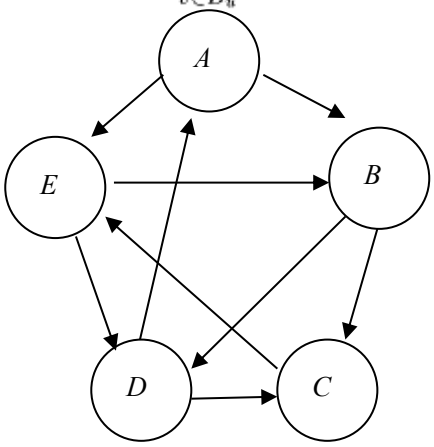
b) 已知Doc1、Doc2和Doc3的静态得分分别是0.25，0.5和0.8，画出当使用静态得分与欧式归一化tf值求和的结果进行排序的倒排记录表。

3. 有18篇已经标注好的文档，所属类别分别用×、○、◇表示。通过聚类算法得到了三个簇，如下图所示，请计算兰迪指数（Rand Index），并附计算过程。



4. 假设五个网页*A*、*B*、*C*、*D*、*E*构成如下的链接关系，试采用如下公式(*u*指当前页面，B_u是所有链接到*u*的页面集合，N_v是页面*v*的出链(outlink)总数，*R*(*u*)和*R*(*v*)分别是*u*和*v*的PageRank值，为计算方便，假定*c*=1)计算每个网页的归一化PageRank值，即五个网页的PageRank总和为1。

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$



四、简答题（共 28 分）

1. 通过对信息检索的结果进行聚类，可以为用户提供更有效的展示。若采用 K-均值聚类算法，为了达到较好的聚类效果，请描述下可能采用的初始质心向量（种子）的选择方法。（8 分）
2. 为了提高查询召回率，请设计一种不带相关反馈的查询扩展方法。请详细描述你的思路和做法。（10分）
3. 请设计一种语义检索模型，即查询和文档之间用词可能完全不同，但是语义近似。请详细描述你的思路和做法。（10分）