主管领导审核签字

哈尔滨工业大学 2020 学年 春季 季学期

信 息 检 索　　试 题 A

题号	一	二	三	四	五	六	七	总分
得分								
阅卷人								

片纸鉴心 诚信不败

密 封 线

授课教师_____ 姓名_____ 学号_____ 院系_____

一、判断对错（**每题 2 分，共 20 分**）

1. 在布尔检索系统中，进行词干还原不会降低准确率。（　）
2. 词干还原应该在构建索引时调用，而不应该在查询处理时调用。（　）
3. Heaps 定律得出：随着文档数目的增加，词汇的数量会持续增长而不会稳定到一个最大值。（　）
4. 如果一个词项 *t*，在文档集合中的所有文档中都出现了，那么该词项的权值最小。（　）
5. HillTop 算法，认为那些来自具有相同主题的相关链接对于当前页面重要度的贡献会更大。（　）
6. 网络爬虫（Web Crawler）到服务器上抓取网页的时候，可以有选择地遵循该网页所在服务器根目录下的 Robots 协议。（　）
7. 后缀树和后缀数组可以用于大规模序列的检索。（　）
8. 布尔检索模型支持对查询条件的部分匹配。（　）
9. 签名文档中，无法表示词项的频率、位置等信息。（　）
10. 网络爬虫在抓取网页的过程中，采用广度优先搜索策略可以在爬虫程序运行的早期就能捕获页面 PageRank 值较高的网页。（　）

二、简答题（**每题 5 分，共 20 分**）

1. 请给出 HITS 算法的思想简介。
2. 请给出基于 shingle 和超级 shingle 的网页去重方法思想，并给出一种随机算则 shingle 的方法。
3. 请给出网络爬虫的结构，并对每个模块的功能进行说明。
4. 在自动局部分析的隐式相关反馈中，包括两种局部策略：局部聚类、局部上下文分析。请简述局部聚类中关联簇、度量簇、标量簇三种方法的基本思想和特点。

三、文本处理是信息检索系统的一个很重要的内容。请举例说明，在英文文本的处

理过程中，都需要解决哪些问题？并给出简要的解决思路（注：针对每个问题一种解决思路即可）。（**10 分**）

四、假设搜索系统返回来 6 个结果，每个结果与查询之间的相关性分别是 3, 2, 3, 0, 1, 2。请给出该系统的 NDCG@6 的结果。（**15 分**）

n	**doc#**	***rel.***	**CG_n**	***log_2N***	**DCG_n**
1	588	3	3	0	
2	589	2	5	1	
3	576	3	8	1.58	
4	590	0	8	2	
5	986	1	9	2.32	
6	592	2	11	2.58	

NDCG@6=

五、给定一个文本串序列：“AGATACGATATATATAC“，模式串：”ATATA“。请使用串的模式匹配算法 HORSPOOL 算法，在文本串中搜索模式串的所有出现。并按步骤给出详细的匹配过程。（**10 分**）

六、在倒排索引中，索引表所占的空间非常大，因此对倒排索引表进行压缩显得非常重要。针对索引表的压缩，常用的压缩方法有 unary 编码（一元编码），Elias-γ编码和 Elias-δ编码等。（**10 分**）

1. 求数字“9“的 Elias-δ编码结果；
1. 对任意的整数 $x>0$，请给出 Elias-δ编码需要的二进制位数；

七、某公司主要生产路由器相关产品，每天会收到很多关于路由器产品咨询和故障处理相关的问题，公司需要雇佣大量的客服人员来解决这些问题，需要花费大量的人力和物力。为了解决这个问题，公司请你设计一款在线客服系统自动回答用户提出的问题。该系统要满足如下要求：

a) 能够搜集用户常问问题的问题—答案对，构建常问问题库；
b) 对于常问问题库中的问题，能够自动给出回复，并给出相关问题的推荐；
c) 对于不在常问问题库中的问题，要能够从文档集合中进行搜索，由客服人员找出相应的答案返回给用户。

根据上述的要求，请给出该在线客服系统的设计，要求：（**15 分**）

1. 给出该系统的详细设计图；
2. 针对系统中的每一部分功能，给出详细的描述；
3. 针对每一部分的实现，给出详细的实现算法。