本版本只能保证题目正确性，不对答案的正确性进行保证。
答案中不仅包含粗略的答案，还有部分相关内容一并进行了整理，所以是杂乱无章的，仅作参考。
如有疑问，与本人无关。
如有更好答案，求来一份。

1. 简述搜索引擎的工作原理

1）收集因特网上几千万到几十亿个网页并对网页中的每一个词(即关键字)进行索引，建立索引数据库的全文搜索引擎。

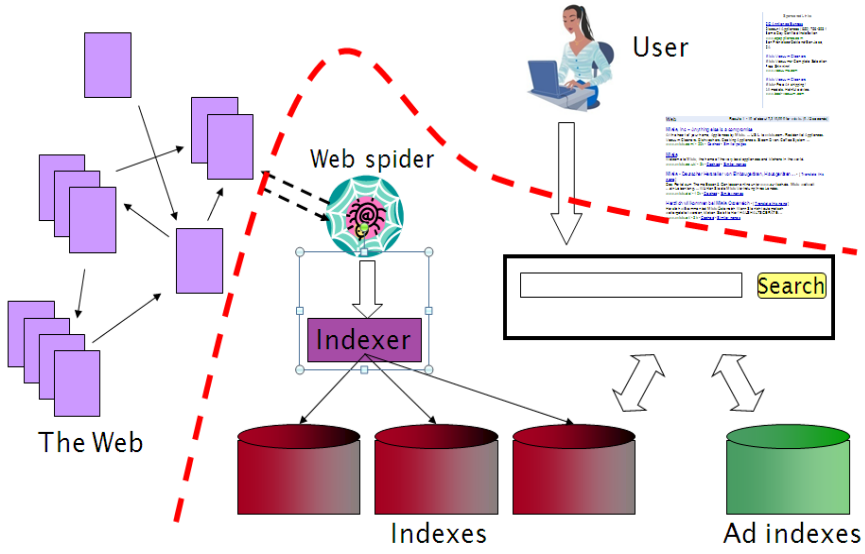
2）当用户查找某个关键词的时候，所有在页面内容中包含了该关键词的网页都将作为搜索结果被搜出来。

3）在经过复杂的算法进行排序后，这些结果将按照与搜索关键词的相关度高低，依次排列。

2. 简述信息检索技术有哪些，流程如何

网页爬取、网页预处理、文本处理、建立索引、查询、Rank、用户反馈

搜索引擎从已知的数据库出发，就像正常用户的浏览器一样访问这些网页并抓取文件。将爬虫抓取的页面文件分解、分析，并以巨大表格的形式存入数据库，建立索引。用户在搜索引擎界面输入关键词，单击“搜索”按钮后，搜索引擎程序即对搜索词进行处理。对搜索词进行处理后，搜索引擎便开始工作，从索引数据库中找出所有包含搜索词的网页，并且根据排名算法计算出哪些网页应该排在前面，然后再按照一定格式返回到“搜索”页面。利用用户的相关反馈信息对查询进行修改。



3. 简述爬虫程序主要完成什么工作

详细版本：从已知的 URL 开始，就像正常用户的浏览器一样访问这些网页并抓取网页文件。对网页进行简单分析，提取 URL，消除噪音，区分 URL 是否抓取，存入 URL 队列。将根据一定的搜索策略(深度优先、广度优先、最佳优先)从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。

简单版本：从一个或若干初始网页的 URL 开始，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止条件。

4. 简述网站防爬取策略有哪些？对这些策略爬虫如何应对的？

网站防爬取措施：Robot 协议、IP 屏蔽、登录、JavaScript 渲染

Robot 协议：网站通过其告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取。Robot.txt 文件是一个文本文件。当一个搜索蜘蛛访问一个站点时，它会首先检查该站点根目录下是否存在 robots.txt，如果存在，搜索机器人就会按照该文件中的内容来确定访问的范围。

IP 屏蔽：网站查看 Useragent，如果不是浏览器的，就封 IP；如果同一 IP 频繁访问同一网站，同样封 IP。 **应对措施**：伪造 Useragent；连接代理服务器、多 IP 并行、增大爬取时间间隔。

访问限制：交互登陆，提交用户名，口令。JavaScript 渲染,AJAX, Cookie，JSON。动态网页，数据在后台数据库，通过 GET（POST)参数，后台 PHP 程序生成的网页。 **应对措施**：模拟浏览器工作(HTTP 分析工具可以分析 HTTP 传递的口令)

验证码 应对措施：脚本或人工对其图片进行爬虫遍历，然后将所有的图片保存后与关键字进行对比并关联入库(分割验证码图像、丢进百度识图 API 函数、返回百度识图结果)。

5. 简述正则表达式方法和 DOM 树分析方法的异同点

正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。

DOM 将 HTML 视为树状结构的元素，所有元素以及他们的文字和属性可通过 DOM 树来操作与访问。

异点：1)正则表达式匹配速度快，而 DOM 树分析在解析 HTML 时速度较慢 2)正则表达式表达能力弱，只具有正规文法的表示能力，而 DOM 树分析的表达能力相当于上下文无关文法。3)正则表达式适用于对网页内容的信噪比要求不高的情况，而 DOM 树分析适用于需要进行网页去噪处理的情况。

同点：1)都是对 HTML 页面进行预处理，去除噪音、网页结构，保留 URL 和文本内容(瞎编的，不要相信)

6. 简述词干还原(Stemming)和词形归并(Lemmatization)的异同点

词干还原通常指去除单词两端词缀的启发式过程。其能够提高召回率，但是会降低准确率。

词形归并利用词汇表和词分析来减少屈折变化的形式，将其转变为基本形式。其可以减少词项词典中的词项数量。

异点：1)代表意义不同：Stemming 通常指很粗略的去除单词两端词缀的启发式过程。Lemmatization 通常指利用词汇表和词形分析来去除屈折词缀，从而返回词的原形或词典中的词的过程。2)词干还原在一般情况下会将多个派生相关词合并在一起，而词形归并通常只将同一词元的不同屈折形式进行合并。

同点：1)都体现了不同语言之间的差异性，包括：不同语言之间的差异，特殊专业语言与一般语言的差异。2)词干还原或者词形归并往往通过在索引过程中增加插件程序的方式来实现

7. 什么是 HMM？简述基于 HMM 的中文分词方法

HMM，Hidden Markov Model，隐马尔可夫模型，是一个统计模型，用来描述一个含有隐含未知参数的马尔可夫过程。

HMM 是一个五元组：

StatusSet(状态值集合)：输出的分词结果，状态值集合为(B, M, E, S)：{B:begin, M:middle, E:end, S:single}

ObservedSet(观察值集合)：输入的句子

TransProbMatrix(转移概率矩阵)：一个 4×4 的矩阵，表示不同状态之间转移的概率

EmitProbMatrix(发射概率矩阵)：表示在某一状态下对应到某字的概率

InitStatus(初始状态分布)：表示句子的第一个字属于{B, M, E, S}这四种状态的概率

Viterbi 算法，一种动态规划算法，它用于寻找最有可能产生观测事件序列的维特比路径——隐含状态序列。对应于中文分词，它用来寻找最有可能产生某一句子的 BEMS 状态值序列。

8. 简述布尔模型模型及其特点

布尔模型是一种简单的检索模型，建立在经典的集合论和布尔代数的基础上。遵循两条基本规则: 每个索引词在一篇文档中只有两种状态：出现或不出现，对应权值为 0 或 1。

优点：1)查询简单，容易理解。2)通过使用复杂的布尔表达式，可方便地控制查询结果 3)相当有效的实现方法 4)经过训练的用户可以容易地写出布尔查询式 5)布尔模型可以通过扩展来包含排序的功能

缺点：1)是精确匹配，信息需求的能力表达不足，不能输出部分匹配的情况。2)无权重设计，无法排序 3)用户必须会用布尔表达式提问，一般而言，检出的文档或者太多或者太少。4)很难进行自动的相关反馈。

9. 简述向量空间模型及其特点

向量空间模型：每篇文档表示成一个基于 tf-idf 权重的实值向量$\in \mathbf{R}^{|V|}$(V 是词项集合，|V|表示词项个数)。文本内容的处理简化为向量空间中的向量，以空间上的相似度表达语义的相似度。

特点：1)维度非常高：特别是互联网搜索引擎，空间可能达到千万维或更高 2)向量空间非常稀疏：对每个向量来说大部分都是 0

优点：1)帮助改善了检索结果 2)部分匹配的文档也可以被检索到 3)可以基于向量 cosine 的值进行排序，提供给用户。

缺点：1)这种方法假设标记词是相互独立的，但实际可能不是这样，如同义词、近义词等往往被认为是不相关的词。

10. 简述非精确 top K 检索的策略

找一个文档集合 A，K < |A| << N，利用 A 中的 top K 结果代替整个文档集的 top K 结果。即给定查询后，A 是整个文档集上近似剪枝得到的结果。

策略一：索引去除(Index elimination)，对于一个包含多个词项的查询来说，很显然可以仅仅考虑那些至少包含一个查询词项的文档（进一步拓展思路：只考虑那些词项的 idf 值超过一定阈值的文档，只考虑包含多个查询词项）

策略二：胜者表(Champion list)，对于词典中的每个词项 t，预先计算出 r 个最

高权重的文档。词项 t 所对应的 tf 值最高的 r 篇文档构成 t 的胜者表，也称为优胜表(fancy list)或高分文档(top doc)。其中 r 的值需要在索引建立时给定。

策略三：静态得分，希望排序靠前的文档既是相关的又是权威的。相关性通过余弦相似度得分来判断；权威性是与 query 无关的文档本身的属性决定的。

策略四：影响度(Impact)排序，多个 term 对应的文档次序不是统一的，即多种顺序(文档内容相关的排序方式)。将词项 t 对应的所有文档 d 按照 $tf_{t,d}$ 值降序排列(不同的文档对不同的 t 具有不同的顺序)。

策略五：簇剪枝方法——预处理，随机选 $\sqrt{N}$ 篇文档作为先导者。对于其它文档，计算和它最近的先导者。这些文档依附在一个先导者上面，称为追随者(follower)。这样一个先导者平均大约有~$\sqrt{N}$ 个追随者。给定查询 q，通过与先导者计算余弦相似度，找出和它最近的一个先导者 L，候选集合 A 包括 L 及其追随者，然后对 A 中的所有的文档计算余弦相似度。

11. 简述常用的网页排序算法

PageRank 算法：在随机游走过程中访问越频繁的网页越重要。基于“从许多优质的网页链接过来的网页，必定还是优质网页”的回归关系，来判定所有网页的重要性。

公式：

$$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+\mathrm{L}\ +PR(T_n)/C(T_n))=(1-d)+\sum_{i=1}^{n}PR(T_i)/C(T_i)$$

PR(A)表示页面 A 的级别，页面 T_i 链向页面 A，$C(T_i)$是页面 T_i 链出的链接数量。d 在 0-1 之间，成为阻尼系数，1-d 是页面本身所具有的网页级别。

Hilltop 算法：指导思想与 PageRank 一致，主题相关网页之间的链接对于权重计算的贡献比主题不相关的链接价值要更高。

Hilltop 算法定义一个网站与其它网站的相关性，作为识别跨站点的链接交换干扰与识别相似链接的技术，以杜绝那些想通过任意链接来扰乱排名规则、那些想通过增加无效链接来提高网页 PageRank 值的做弊行为。

HITS 算法：超链导向的主题搜索(Hyperlink - Induced Topic Search)，对每个网页都要计算两个值:权威值(authority)与中心值(hub)。

权威网页：一个网页被多次引用，则它可能是很重要的；一个网页虽然没有被多次引用，但是被重要的网页引用，则它也可能是很重要的；一个网页的重要性被平均的传递到它所引用的网页。

中心网页：提供指向权威网页的链接集合的 WEB 网页，它本身可能并不重要，或者说没有几个网页指向它，但是它提供了指向就某个主题而言最为重要的站点的链接集合

12. 简述 PageRank 算法和 HITS 算法的异同点

同点：都是基于链接分析的搜索引擎排序算法，并且在算法中两者都利用了特征向量作为理论基础和收敛性依据。

异点：HITS 算法计算的 authority 值只是相对于某个检索主题的权重，因此 HITS 算法也常被称为 Query-dependent 算法；而 PageRank 算法是独立于检索主题，因此也常被称为 Query-independent 算法。

13. 简述 Web spam 的常用技术

针对基于相关性的排序策略的 spam 方法：term spam。在预处理阶段可能形

成的所谓“重要性”因素。顾名思义，既然是在预处理阶段形成的，就是和用户查询无关的。1)保证与查询词相关 2)提高与查询词相关度

针对基于连接分析的排序策略的 spam 方法：link spam。1)创造一个诱饵系统(honey pot，中文是我乱翻的，建议考试时候使用英文)。2)参与链接交换 3)渗透网络目录 4)在博客、未经审核的留言板、留言簿或维基上发布链接 5)购买过期域名

Hiding techniques：1)Terms，内容隐藏：与背景相同的字体，超小字体，单个像素的图片做成的链接。2)Link，Cloaking：对机器人程序和普通用户返回不同内容的页面。识别机器人程序：通过 IP 或者 user-agent；Redirection：载入网页时自动转到另外一个 URL 地址：通过 refresh meta tag 或者 html 的脚本代码。

14. 简述信息检索的主要评价指标

查全率和查准率：

查准率：返回的结果中真正相关结果的比率

查全率：返回的相关结果数占实际相关结果总数的比率

F 值：召回率 R 和查准率 P 的加权调和平均值，F1 标准则综合了精度和查全率，将两者赋予同样的重要性来考虑。

R-查准率：计算序列中第 R 个位置文献的查准率。R 是指与当前查询相关的文档总数。

对多个查询进行查准率评估：1)平均：宏平均(Macro Average): 对每个查询求出某个指标，然后对这些指标进行算术平均；微平均(Micro Average): 将所有查询视为一个查询，将各种情况的文档总数求和，然后进行指标的计算 2)查准率直方图：在多个查询下，分别计算每一查询下的 R-查准率，计算其差值，并用直方图表示。

查准率/查全率曲线：在查全率和查准率间进行权衡。

NDCG：一种总体观察检索排序效果的方法，利用检索序列加和的思路来衡量。

DCG： $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$

NDCG： $nDCG_p = \frac{DCG_p}{IDCG_p}$

面向用户的测度方法：

覆盖率：实际检出的相关文档中，用户已知的相关文档所占的比例。

新颖性：检出的相关文档中，用户未知的相关文档所占的比例。

多样性：检索结果的多样性，检出的相关文档中，不含语义上非常相似或近似 copy 的文档。

15. 什么是 LSA，简述 LSA 的作用

LSA(Latent Semantic Analysis，隐语义分析)：使用统计计算的方法对大量的文本集进行分析，从而提取出词与词之间潜在的语义结构，并用这种潜在的语义结构，来表示词和文本，达到消除词之间的相关性和简化文本向量实现降维的目的。基本观点是：把高维的向量空间模型（VSM）表示中的文档映射到低维的潜在语义空间中。

优势：1)文章和单词都映射到同一个语义空间。2)语义空间的维度明显明显少于源单词-文章矩阵。

应用：1)在低维语义空间可对文档进行比较，进而可用于文档聚类和文档分类。2)在翻译好的文档上进行训练，可以发现不同语言的相似文档，可用于跨语言检索。3)发现词与词之间的关系，可用于同义词、歧义词检测。4)通过查询映射到语义空间，可进行信息检索。5)从语义的角度发现词语的相关性，可用于“选择题回答模型”

16. 什么是 PLSA，简述 PLSA 和 LSA 的异同点

PLSA(Probabilistic Latent Semantic Analysis，概率潜在语义分析)，基于双模式和共现的数据分析方法延伸的经典的统计学方法，以统计学的角度来看待 LSA。PLSA 中生成文档的整个过程便是选定文档生成主题，确定主题生成词。

异点：LSA 隐含高斯分布假设，PLSA 隐含 Multi-nomial 分布假设；PLSA 的优化目标是是 KL-divergence 最小，LSA 依赖于最小均方误差(SVD 是一种最小二乘法)等准则；PLSA 使用的 EM 算法需要反复迭代，这需要很大的计算量。

同点：LSA 和 PLSA 都可以用于文档聚类和文档分类，处理同义词；都是分析文档语料库，以便找到它的新的低维表示；都不能生成新文档的模型。(胡编乱造的)

17. 什么是 LDA，简述 LDA 和 PLSA 的异同点

LDA(Latent Dirichlet Allocation，隐含狄利克雷分布)，不再认为主题分布（各个主题在文档中出现的概率分布）和词分布（各个词语在某个主题下出现的概率分布）是唯一确定的，而是有很多种可能，Dirichlet 先验为某篇文档随机抽取出某个主题分布和词分布。

同点：LDA 和 PLSA 思想上一致：增加了 Dirichlet 先验，全贝叶斯化

异点：PLSA 认为文档 d 产生主题 z 的概率，主题 z 产生单词 w 的概率都是两个固定的值，LDA 不再认为主题分布和词分布是唯一确定的，而是有很多种可能，Dirichlet 先验为某篇文档随机抽取出某个主题分布和词分布。

18. 简述文本相似度量方法

基于字符串的方法：

基于字符的方法：

最长公共子序列：通过计算出两个字符串/序列之间的最长公共子序列，并使用这个子序列的长度来反映两个字符串/序列之间的相似程度。

编辑距离：指两个字串之间，由一个转成另一个所需的最少编辑操作次数

扩展的编辑距离：在思想上与编辑距离一样，只是除插入、删除和替换操作外，还支持 相邻字符的交换 这样一个操作。

Needleman-Wunsch Similarity：对插入错误和删除错误赋予较高的惩罚分数

Smith-Waterman Similarity：是一个局部最优比对方法，它的目的是找出两个序列之间 连续且相同 的子序列。

Jaro Similarity 和 Jaro-Winkler Similarity：考虑两个字符串之间相同字符的顺序位置和个数，适用于像人名这样的较短字符串之间的比较。

Hamming Distance：用于长度相同的序列之间的比较，思想非常简单，就是逐位比较得到的不同次数。

基于项的方法：

余弦相似度：用向量空间中两个向量夹角的余弦值作为衡量两个个体间差

异的大小。

Jaccard Similarity：两个集合 A 与 B 的交集的大小与 A 与 B 并集的大小的比值。

Dice 系数：Dice(s1,s2)=2×comm(s1,s2)/(len(s1)+len(s2))。即 2 倍的两个字符串相同字符的个数与两个字符串长度之和的比值。

基于语料库的方法：通过对大量文档的统计分析得到语义上的相似

语言模型：假设具有相同(或相近)上下文的词，其语义是相近的。

主题模型：通过词与词的共现(Co-occurrence)来反映词与词之间的相似性

19. 简述常用颜色特征选取方法及其主要思想

颜色直方图(Color Histogram)：在颜色空间中采用一定的量化方法对颜色进行量化，然后统计每一个量化通道在整幅图像中所占的比重。

颜色相关图(Color Correlogram)：用颜色对相对于距离的分布来描述信息，它反映了像素对的空间相关性，以及局部像素分布和总体像素分布的相关性。

颜色矩(Color Moment)：在颜色直方图的基础上计算出每个颜色的矩估计，用这些统计量替代颜色的分布来表示颜色特征。

颜色一致性矢量(Color Coherence Vectors, CCV)：本质上是一种引入空间信息改进的直方图算法，统计了图像中各颜色最大区域的像素数量。通过分离开一致性像素和非一致性像素，比直方图算法具有更好的区别效果。

20. 简述常用纹理特征选取方法及其主要思想

基于统计特征的纹理特征：

灰度差分：当图像中采用较小灰度差分值的概率较大时，说明纹理较粗糙；当各级灰度差分值的概率较为平坦时，说明纹理较细。

灰度共现(共生)矩阵：用不同邻接灰度值的相邻状况来刻画图像的纹理特征，并能在此基础上给出更多的统计量，进一步来刻画图像的纹理。

Tamura 纹理特征：Tamura 纹理特征中所有纹理特征都在视觉上有意义，分别对应于心理学角度上纹理特征的六种属性：对比度(contrast)、粗糙度(coarseness)、方向性(directionality)、线像度(1ine likeness)、规整度(regularity)和粗略度(roughness)。

基于信号处理方法描述纹理特征：

傅里叶变换：傅里叶频谱包含非常丰富的图像信息，能粗略描述纹理模式。

Gabor 滤波器：Gabor 小波与人类视觉系统中简单细胞的视觉刺激响应非常相似，它在提取目标的局部空间和频率域信息方面具有良好的特性。

Laws 纹理：一种基于图像能量估计的图像纹理能量转换的纹理特征提取方法，通过使用简单模板处理纹理图像，从而对纹理图像的特征进行描述。

LBP 特征：结合了纹理图像结构和像素统计关系的纹理特征描述方法，主要在于记录像素点与其周围像素点的对比信息，或说差异。

21. 常见的局部形状描述符有哪些？简述其思想

链码：通过一个给定的方向上的单位尺寸的直线片段的序列来描述一条曲线或一个二维形状的边界。

基于网格的方法：将图像形状边界映射到一个标准的网格上，并将该形状边界调整到网格左上角，然后从左向右，从上到下扫描网格，若某个单元格被形状边界全部或者部分覆盖，则赋值 1，否则赋值 0，这样就得到了一个 0.1 组成的串，

用来表征形状特征。

距离直方图：求得形状质心，在边界上均匀取特征点，计算特征点到质心距离，建立起距离直方图。

边界矩：将边界点到质心的距离理解成一个分布，计算分布的矩。将低阶矩作为特征。

傅里叶描述子：假定物体的形状是一条封闭的曲线，沿边界曲线上的一个动点的坐标变化是一个以形状边界周长为周期的函数。这个周期函数可以展开成傅立叶级数形式表示。傅立叶级数中的一系列系数 z(k)是直接与边界曲线的形状有关的，称为傅立叶描述子。高频分量表示形状的细节，而低频分量则表示形状的总体。当系数项取到足够阶次时，它可以将物体的形状信息完全提取并恢复出来。

基于区域的形状描述符：由两个阶段组成：由上而下的，通过对图像进行划分来发现同质的区域；由下而上地，对邻近同质的区域进行合并。

22. 简述协同过滤算法的思路和实现过程

思路：收集用户的历史行为和偏好信息，利用群体智慧给出推荐；朋友之间互相推荐。

Item-based CF：

思路：通过用户对不同 item 的评分来评测 item 之间的相似性，基于 item 之间的相似性做出推荐。

实现过程：首先找到和目标用户兴趣偏好相似的最近邻居，然后根据他的最近邻居对推荐对象的评分来预测目标用户对未评分的推荐对象的评分，选择预测评分最高的若干个推荐对象作为推荐结果反馈给用户。

User-based CF：

思路：通过不同用户对 item 的评分来评测用户之间的相似性，基于用户之间的相似性做出推荐。

实现过程：首先找到目标对象的最近邻居，由于当前用户对最近邻居的评分与对目标推荐对象的评分比较类似，所以可以根据当前用户对最近邻居的评分预测当前用户对目标推荐对象的评分，然后选择预测评分最高的若干个目标对象作为推荐结果呈现给当前用户