2019年信息检索（2019.12.26）

一、填空题

总共46个空，每空一分（太多了，默写不下来）

二、名词解释

（1）爬虫

（2）正则表达式

（3）HMM

（4）TF/IDF

（5）LSA

（6）N-gram语言模型

三、简答题

（1）简述布尔模型及其特点

（2）SimHash是什么算法，简述其编码步骤

（3）简述常用的网页排序算法

（4）信息检索的主要评价指标是什么

2019年信息检索知识点（根据老师串讲录音整理，所有考题全在上面）

第一讲 搜索

1. IR（信息检索是什么样的学科）

实质上是融合了文本及多媒体检索、数据挖掘、机器学习和自然语言处理的综合学科

1. 为什么要进行信息检索？信息过载

3、搜索做了什么操作？（搜索的过程）

从大规模非结构化数据(通常是文本)的集合(通常保存在计算机上)中找出满足用户信息需求的资料(通常是文档)的过程

4、IR作为一门学科，是研究信息的**获取**(acquisition)、**表示**(representation)、**存储**(storage)、**组织**(organization)和**访问**(access)的一门学问

1. 信息检索本质

给定一个查询*Q*，从文档集合*C*中，计算每篇文档*D*与*Q*的相关度，并排序(Ranking)

6、确定文档和查询之间的**相关度**是IR的核心问题

7、什么是相关度

相关度是一个查询和文档相关的程度，形式上说，信息检索中的**相关度**是一个函数*f*，输入是查询*Q*、文档*D*和文档集合*C*，返回的是一个实数值 R, *R* = *f*(*Q*,*D*,*C*)

8、**相关度(relevance)不同于相似度(Similarity)**

(1)相关取决于用户的判断，是一个主观概念

(2)不同用户做出的判断很难保证一致

(3)即使是同一用户在不同时期、不同环境下做出的判断也不尽相同

9、信息检索模型

描述信息检索中的文档、查询和它们之间关系(匹配函数)的数学模型

10、信息检索主要技术

（1）文本分析（NLP）

（2）建立索引

（3）查询，包括查询分析（NLP），相关度计算(和信息检索模型相关)

（4）排序（实验室评价）

1. 搜索引擎的工作原理

（1）爬行和抓取

（2）建立索引（可能会考的知识点：蜘蛛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，这个过程即是[索引](http://baike.baidu.com/view/262241.htm)（index).搜索[引擎](http://baike.baidu.com/view/53607.htm)的核心数据结构为[倒排文件](http://baike.baidu.com/view/228996.htm)（也称[倒排索引](http://baike.baidu.com/view/676861.htm)））

（3）搜索词处理 （4）排序

1. 搜索引擎评价

（1）覆盖面 （2）更新周期 （3）响应速度 （4）排序结果是否满足用户的查询要求

1. 课程安排

爬虫

**第2讲 网络爬虫技术** **第3讲 网页分析技术**

文本预处理

**第4讲 词项词典 第5讲 中文分词**

建立索引和查询

相关度计算 **第6讲 布尔检索 第7讲 倒排索引 第8讲 向量空间模型**

排序

**第9讲 检索排序** **第10讲 搜索引擎优化** **第11讲 信息检索的评价**

用户反馈

**第12讲 相关反馈及查询扩展**

第二讲 网络爬虫技术

1. 爬虫（会考一个名词解释）

一种自动获取网页内容的程序，从一个或若干初始网页的[URL](http://baike.baidu.com/view/1496.htm)开始，获取并解析它们，提取它们指向的URL，将提取的url放在队列中，获取队列中的每个URL并重复此过程，直到满足系统的一定停止条件

2、爬虫必须具有的功能

**礼貌性** 只爬允许爬的内容、尊重 robots.txt

3、搜索策略：深度优先, 广度优先

4、URL 判重

利用哈希法，URL经过哈希函数得到哈希码，判断是否已经在散列中来判断是否爬取过

5、常见的开源爬虫 Nutch Heritrix

第三讲 网页分析技术

1、正则表达式的定义

正则表达式是对[字符](http://baike.baidu.com/view/263416.htm)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。

2、基于正则表达式的信息提取的步骤

（1）在获取数据前应尽量去除无用部分（2）提取网页内的链接 （3）提取网页标题（4）提取网页内的文本

3、正则表达式的工具有哪些

Java java.util.regex包 Python的 re模块

4、正则表达式匹配特点是什么

（1）正则表达式匹配速度快，

（2）但表达能力较弱，只具有正规文法的表示能力。

（3）在对网页内容的信噪比要求不高的情况下可以使用基于正则表达式匹配的爬取程序

（4）受网页噪音影响较大

5、什么叫做DOM

文档对象模型（document object model，DOM）,DOM将一个XML文档转换成一个对象集合，然后可以任意处理该对象模型。

6、开源HTML解析器（能够列出一两种即可）

（1）JAVA：HTMLParser,jsoup

（2）C/C++：htmlcxx

（3）Python：Beautiful Soup

7、两种方法比较

正则表达式匹配

（1）正则表达式匹配速度快，但表达能力较弱，只具有正规文法的表示能力。

（2）在对网页内容的信噪比要求不高的情况下可以使用基于正则表达式匹配的爬取程序

HTML DOM树

（1）提取HTML DOM树提取在解析HTML时速度较慢，但其表达能力相当于上下文无关文法。

（2）在网页自动分类等需要进行网页去噪处理的情况时使用基HTMLDOM树的爬取程序

8、元搜索引擎（只需要知道概念）

（1）元搜索引擎又[称多](http://baike.baidu.com/view/703714.htm)搜索[引擎](http://baike.baidu.com/subview/53607/5078350.htm)

（2）通过一个统一的用户界面帮助用户在多个搜索引擎中选择和利用合适的（甚至是同时利用若干个）搜索引擎来实现检索操作，是对分布于网络的多种检索工具的全局控制机制

9、网站防爬取措施有哪些及爬虫是如何模拟浏览器来应对这些措施

（1）Robot 协议（知道文件名是robots.txt即可）

（2）IP屏蔽 网站识别Useragent，例如识别出是JavaClient1.6就封IP，爬虫应对措施：伪造Useragent，同一 IP 访问频繁也封IP，爬虫应对措施：连接代理服务器， 多IP并行，增大爬取时间间隔

（3）访问限制

网站措施：a.交互登陆，提交用户名，口令b.JavaScript渲染,AJAX, Cookie，JSON

c.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

爬虫应对：模拟浏览器工作，用HTTP分析工具，分析HTTP 传递的口令

10、掌握http分析工具和浏览器自动化测试框架（掌握一个即可，但是要知道流程，这个不太可能会考）

11、验证码思路和模块（待查）

第四讲 词项词典

（会出简答题，有问的比较细的，例如比较两种方法优缺点，还有比较粗的例如课程涉及到哪些技术，还有如何建立词项词典，就是下面除去第七个的其他内容）

1、如何建立词项词典？

一、文档解析(Parsing a document)

二、词条化 (Tokenization)

三、词项归一化 (Normalization)

四、词干还原 (Stemming)

五、词形归并 (Lemmatization)

六、去掉停用词 (Stop Words)

2、词条化 (Tokenization)（可能会考名词解释）

将给定的字符序列拆分成一系列子序列的过程，其中每一个子序列称之为一个“词条”Token。（就是分词过程）

3、词项归一化 (Normalization)

将文档和查询中的词条“归一化”成一致的形式（希望USA和U.S.A.之间也能形成匹配 ）

归一化的结果： 在IR系统的词项词典中，形成多个近似词项的一个等价类

4、拼写错误要知道对于英文的拼音算法（找到拼写错误的）SoundEx 算法

5、词干还原

含义: 通常指去除单词两端词缀的启发式过程

e.g., automate(s), automatic, automation 🡪 automat

作用：词干还原能够提高召回率，但是会降低准确率

6、词干还原有一个算法是Porter算法

7、词性归并：利用词汇表和词形分析来减少屈折变化的形式，将其转变为基本形式

e.g.

am, are, is 🡪 be car, cars, car's, cars' 🡪car

the boy's cars are different colors 🡪the boy car be different color

词形归并可以减少词项词典中的词项数量

8、词形归并(Lemmatization)和词干还原有什么区别

（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）假如给定词条 saw，

* + 词干还原过程可能仅返回 s，
  + 而词形归并过程将返回see或者saw，
  + 具体返回哪个词取决于当前上下文中saw是动词还是名词。

（3）两个过程的区别还在于：

* + 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。

9、停用词定义

没有区分度的词就是停用词

10、确定停用词的两种方法

a.查表法

b.基于文档中词项的词频tf,特别大的即为停用词

11、消除停用词的优缺点

优点：

（1）停用词消除可以减少term的个数

（2）缩小搜索范围，

（3）提高搜索的效率

（4）机器学习文本分类算法的文档的预处理

缺点：有时消除的停用词对检索是有意义的。

“的士”、“to be or not to be”

第5讲 中文分词

1、中文分词定义

指的是将一个汉字序列切分成一个一个单独的词，分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

2、中文分词的两种方法

方法1：基于字符串匹配的分词方法

* + 按照扫描方向：正向匹配和逆向匹配
  + 按照扫描长度：最大匹配和最小匹配

方法2：基于统计的分词方法

3、基于字符串匹配的分词方法的优缺点

优点：（1）程序简单易行，开发周期短；

（2）仅需很少的语言资源（词表）

（3）不需要任何词法、句法、语义资源

（4）可以自定义词库，增加新词

缺点：

（1）Out of Vocabulary未登录词

(2)歧义消解能力差；

(3切分正确率不高，一般在95%左右。

4、基于统计的分词方法的优缺点

优点：

（1）分词准确度高；

（2）能够平衡地看待词表词和未登录词的识别问题。

缺点：

（1）局限性，会经常抽出一些共现频度高、但并不是词的常用字组；

例如“这一”、“之一”、“有的”、“我的”、“许多的”等，

（2）对常用词的识别精度差，时空开销大；

（3）学习算法的复杂度往往较高，计算代价较大，依赖手工定义的特征工程。

5、基于统计的分词方法中掌握HMM**隐马尔可夫模型**(Hidden Markov Model，HMM)

6、HMM作用：用来描述一个含有隐含未知参数的马尔可夫过程。

7、HMM模型是一个五元组:

StatusSet: 状态值集合

ObservedSet: 观察值集合

TransProbMatrix: 转移概率矩阵

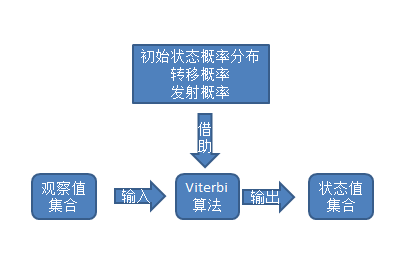
EmitProbMatrix: 发射概率矩阵

InitStatus: 初始状态分布

8、HMM用来分词的话采用的是Viterbi算法，只考到这个算法的宏观描述

* 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

知道哪几个参数来预测哪个参数，



* ObservedSet 输入的句子
* StatusSet 输出的分词结果

9、常见开源分词软件

* StandardAnalyzer
* IKAnalyzer
* 庖丁解牛分词
* Jieba库

第六讲 布尔模型与倒排索引

1、什么是信息检索模型（按照这里的来，更详细）

**信息检索模型**（IR model），依照用户查询，对文档集合进行相关排序的一组前提假设和算法。IR模型可形式地表示为一个四元组< D, Q, F, R(qi,dj) >

D是一个文档集合，Q是一个查询集合，R(qi,dj) 是一个排序函数，它给查询qi和文档 dj 之间的相关度赋予一个排序值，F是一个框架,用以构建文档,查询以及它们之间关系的模型

2、基于内容的信息检索模型有哪些？

* 集合论模型：布尔模型、模糊集合模型、扩展布尔模型
* 代数模型： 向量空间模型、广义向量空间模型、潜在语义标引模型、神经网络模型
* 概率模型： 经典概率论模型、推理网络模型、置信（信念）网络模型
* 深度学习模型

3、布尔模型是什么（都开始自己总结了，肯定是考题）

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:

（1）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。

（2）每篇文档：索引词（0或1）的集合

进行查询的时候，用布尔表达式进行匹配，计算二值的相关度。

4、什么是bag of words 模型

在信息检索中，Bag of words model假定

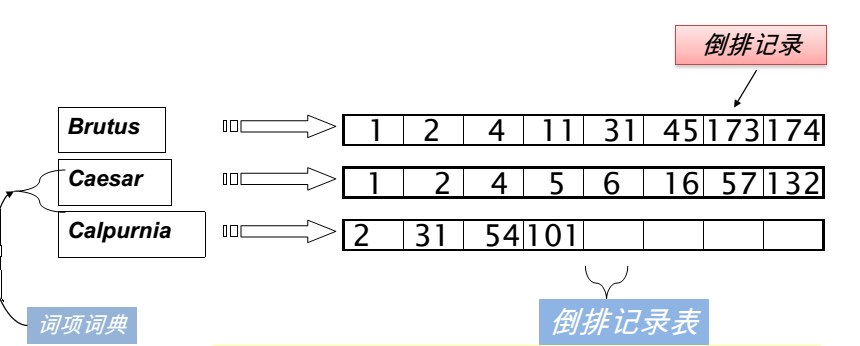
（1）对于一个文本，忽略其词序和语法，句法，将其仅仅看做是一个词集合，或者说是词的一个组合，

（2）文本中每个词的出现都是独立的，不依赖于其他词是否出现，在任意一个位置选择一个词汇都不受前面句子的影响而独立选择的。

5、搜索[引擎](http://baike.baidu.com/view/53607.htm)的核心数据结构为[倒排文件](http://baike.baidu.com/view/228996.htm)（Inverted Files）（也叫倒排索引）

6、什么是倒排索引

有词项和倒排记录组成，词项词典：对于每一个词项，存储所有包含这个词项的文档的一个列表。倒排记录表：一个文档用一个序列号docID来表示。



7、包含位置信息的倒排记录表

下面的介绍要理解，不要求背

* 倒排记录表数据结构的扩展形式
* 用于短语查询（Phrase Query）
  + 例如
    - 用户希望将类似“stanford university”的查询中的二个词看成是一个整体。
    - 类似“I want to university at stanford”这样的文档是不会被匹配的。
  + 大部分的搜索引擎都支持双引号的短语查询，这种语法很容易理解并被用户成功使用。
  + 有很多查询在输入时没有加双引号，其实都是隐式的短语查询(如人名)。
  + 要支持短语查询，只记录<term : docs> 这样的条目是不能满足用户需要的。

8、上面的包含位置信息的倒排记录表有两种实现方式

方法1：二元词索引(Biword indexes)

将文档中每个连续词对看成一个短语，其中的每一个二元词对都将作为词典中的词项，经过上述的处理，此时可以处理二个词构成的短语查询。

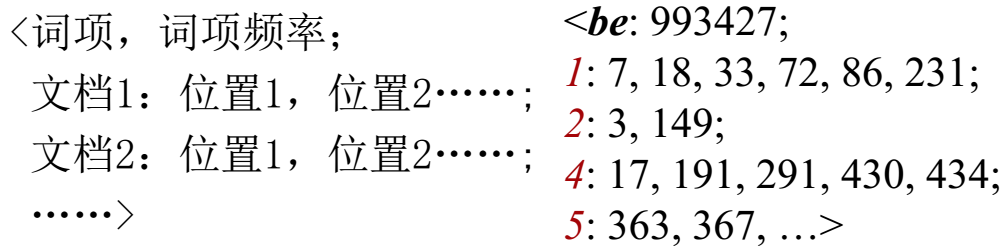
例如，文本“Friends, Romans, Countrymen”将生成如下的二元连续词对：

* + friends romans
  + romans countrymen

方法2：位置信息索引(Positional indexes)

在倒排记录表中记录词项的位置信息。

（下面的帮助理解）在这种索引中，对每个词项，采取以下方式存储倒排表记录：



对于短语查询，仍然采用合并算法，查找符合的文档不只是简单的判断二个词是否出现在同一文档中，还需要检查它们出现的位置情况

9、布尔检索模型的特点是什么

优点：（1）查询简单，因此容易理解（下面的具体说明理解即可）

* + - 布尔模型也许是IR系统中的最简单的模型
    - 是近30年来最主要的商业搜索工具
    - 当前使用的很多系统依然是使用的布尔模型
      * 电子邮件，图书馆分类系统，mac osx的spotlight

（2）通过使用复杂的布尔表达式，可方便地控制查询结果

* + - 同义关系 电脑 OR 计算机
    - 词组 数据 AND 挖掘

缺点 （1）准确匹配，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

第七讲 向量空间模型

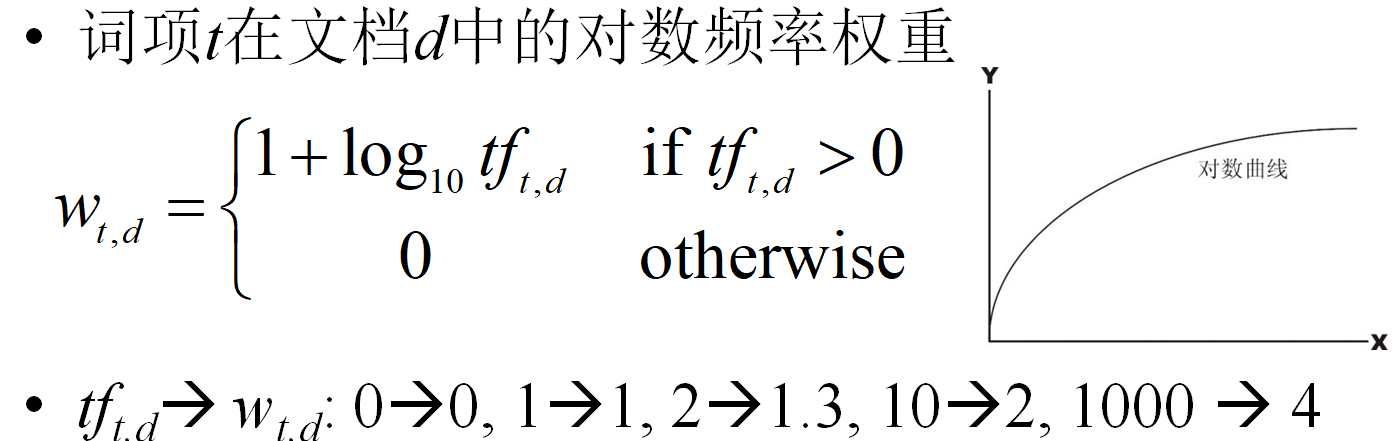
1、Jaccard 系数

* 一种常用的衡量两个**集合**A,B**重叠度**的方法
  + Jaccard(A,B) = |A ∩ B| / |A ∪ B|（回答这个公式即可）
  + Jaccard(A,A) = 1
  + Jaccard(A,B) = 0 if A ∩ B = 0
* 集合A和B不需要具有同样的规模

2、什么是词项频率tf (Term frequency )（了解：越大说明在这篇文档中越重要）

词项*t*在文档*d*中出现的次数，记为*tft,d）*

3、一种替代原始tf的方法: 对数词频，其公式是什么



回答时用文字描述出来，即原始的词频tf以10为底取对数再加一

4、什么是idf

文档频率 (Document frequency，df)

* + **文档频率**:出现词项的文档数目
  + *dft* 文档集合中包含t的文档数目
    - 与词项*t*包含的信息量成反比
    - *dft* <= *N* (*N*是文档的总数)

idf (inverse document frequency)逆文档频率

* + *idft* = log10(*N*/*dft*)
    - *idft* 是反映词项*t*的信息量的一个指标
    - 用log (*N*/*dft*) 代替*N*/*dft* 来抑制idf的作用

回答时：是逆文档频率，*idft* = log10(*N*/*dft*)，df是文档频率，指出现词项的文档数目

5、tf-idf是什么

是信息检索中最著名的权重计算方法，表示t对于文档d的重要程度，词项*t*的tf-idf 由它的tf和idf组合而成 *wt,d*=(1+log *tft,d*) × log10(*N*/*dft*)

（理解一下和重要程度是否符合：tf-idf值随着词项在单个文档中出现次数(tf)增加而增大，tf-idf值随着词项在文档集中数目(df)增加而减小）

6、Query的最终文档排序( Query词)（q和文档d的相关程度）

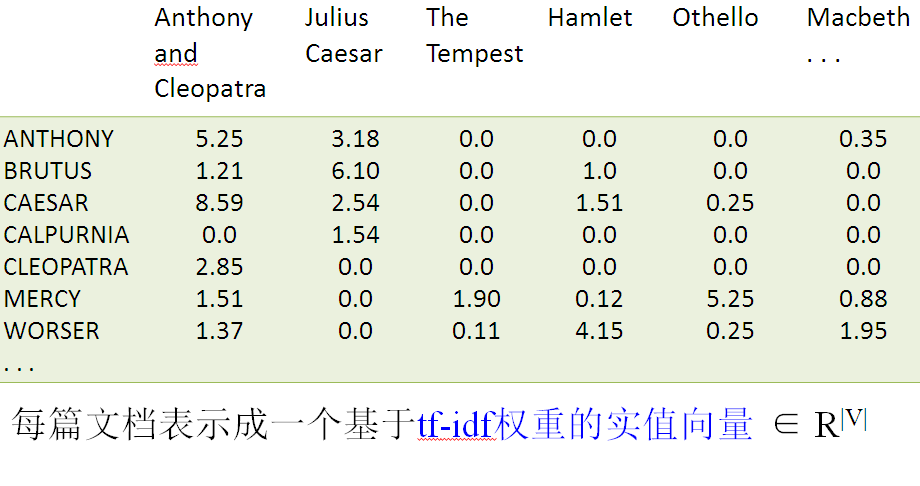


这个I是交集符号

7、什么叫做向量空间模型

是一个|*V*|维实向量空间（*V*是词项集合，|*V*|表示词项个数），空间的每一维都对应一个词项，

每篇文档表示成一个基于tf-idf权重的实值向量，向量的维度是词项的个数，文档是空间中的点或者向量，这就是向量空间模型（用下面的矩阵来理解，行是词项，列是文档）



8、计算查询和文档两个向量之间的相似度（注：用向量来表示文档后，求查询和文档之间的相关度就是计算这两个向量的相关度）

余玄相似度：（认为cos(di,q) > cos(dj,q)，夹角更小，所以di比dj与q更相关）

R(d,q) = cos(d,q) = d·q/|d|×|q|



9、向量空间模型特点

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

（2）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

（3）向量空间非常稀疏：对每个向量来说大部分都是0

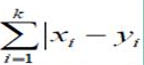
第八讲 相似度计算

1、欧式距离

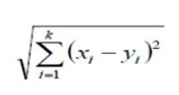
**范数Norm**：Lr范数就是x,y每个维度差距上取r次方加和后再开r次方根

其中三种距离：

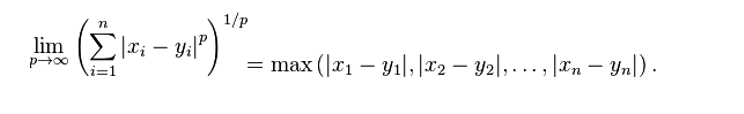
曼哈顿距离 Manhattan distance：L1范数：向量各个元素绝对值之和。



欧几里得距离 Eucledian Distance：L2范数：向量各个元素的平方求和然后求平方根，也叫欧式范数、欧氏距离。

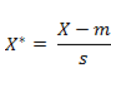


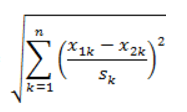
切比雪夫距离Chebyshev distance：L∞范数，向量各个元素求绝对值，最大那个元素的绝对值。也就是x，y在任意维度上差距最大的那个值



2、标准化欧氏距离 (Standardized Euclidean Distance)

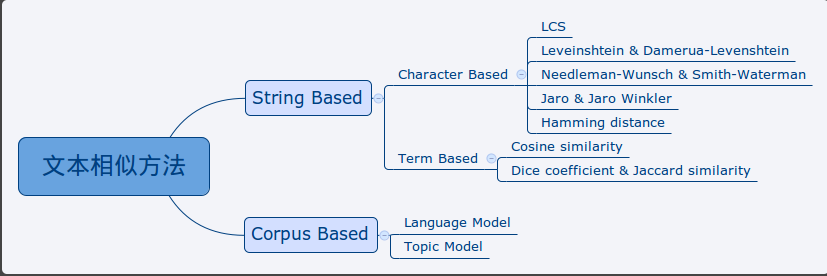
标准欧氏距离的思路：既然数据各维分量的分布不一样，那先将各个分量都“标准化”到均值、方差相等，来避免不同维度上分布不均匀对其造成的影响

假设样本集X的均值(mean)为m，标准差(standard deviation)为s，X的“标准化变量”表示为：

标准化欧氏距离公式：

如果将方差的倒数看成一个权重，也可称之为加权欧氏距离(Weighted Euclidean distance)。

3、文本相似度量方法



String Based Methods：逐个字符比对，看看有多少个字符是一致的

* Character Based Methods 以字符为单位比较的方法
* Term Based Methods 以词为单位的

Corpus Based Methods：基于语料库的方法

4、Character Based Methods

（1）编辑距离(Leveinshtein莱文斯坦 Distance)

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，包括字符的插入 、 删除 或 替换 操作

编辑距离还可用于：拼写纠错（Spell Correction） 命名实体抽取（Named Entity Extraction）

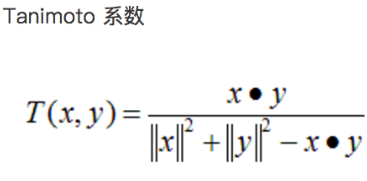
(2)Hamming Distance

Hamming 距离用于 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。Hamming 距离被广泛应用于信息学。

(哈希码比较用的就是海明距离)

5、Term Based Methods

Tanimoto系数（定义：广义Jaccard相似度，元素的取值可以是实数。又叫作谷本系数）



6、文档重复检测

掌握 shingle算法 ，局部敏感哈希（LSH，包括**MinHash，**Simhash算法）

7、shingle算法基本思想

给定正整数*k*及文档*d*的一个词项序列，可以定义文档*d*的*k*-shingle为*d*中所有*k*个连续词项构成的序列，将文件相似性问题转换为集合的相似性问题（用Jaccard 系数来计算集合相似度）

为了让计算速度更快，抽取集合中的子集来比较相似度

8、什么是局部敏感哈希 LSH locality sensitive hash

一种常见的用于处理高维向量的索引办法，利用特定的hash算法，将高维数据映射到低维空间，以较高概率快速寻找相似度高的数据子集，采用过滤一验证的框架，在过滤阶段，LSH利用哈希技术把非相似、不可能成为结果的数据对象过滤掉，过滤之后的数据对象作为候选集，使得相似的数据对象以很高的概率留候选集合中，进而在候选集合上进行实际的距离或者相似性度量计算。

9、LSH有两种

MinHash：可以用来快速估算两个集合的相似度，用在搜索引擎中重复网页检测

要知道基本思想：第一种：使用多个hash函数 第二种：使用单个hash函数

Simhash算法：主要用来文档的相似度计算

simhash算法分为5个步骤：分词、hash、加权、合并、降维

（1）分词

给定一段语句，进行分词，得到有效的特征向量，然后为每一个特征向量设置权重

（权重代表这个单词在整条语句中的重要程度，数字越大代表越重要，权重可以是这个词出现的次数，例如给定一段语句：“CSDN博客结构之法算法之道的作者July”，

* + - * + 1-5等5个级别的权重
        + CSDN(4) 博客(5) 结构(3) 之(1) 法(2) 算法(3) 之(1) 道(2) 的(1) 作者(5) July(5)）

（2） hash

a.选择simhash的位数(综合考虑存储成本以及数据集的大小，比如说32位)

b.通过hash函数计算各个特征向量的hash值

c.hash值为二进制数01组成的n-bit签名。

(比如“CSDN”的hash值Hash(CSDN)为100101，“博客”的hash值Hash(博客)为“101011”)

(3)加权

在hash值的基础上，给所有特征向量进行加权，即W = Hash \* weight，遇到1则hash值和权值正相乘，遇到0则hash值和权值负相乘。

* + - * 例如给“CSDN”的hash值“100101”
      * 加权得到：W(CSDN) = 100101 4 = 4 -4 -4 4 -4 4，
      * 给“博客”的hash值“101011”加权得到：
      * W(博客)=101011 5 = 5 -5 5 -5 5 5

(4)合并

将上述各个特征向量的加权结果累加，变成只有一个序列串。

* + - * 例如“CSDN”的“4 -4 -4 4 -4 4”和“博客”的“5 -5 5 -5 5 5”进行累加，
      * 得到“4+5 -4+-5 -4+5 4+-5 -4+5 4+5”，
      * 得到“9 -9 1 -1 1”。

(5)降维

对于n-bit签名的累加结果，如果大于0则置1，否则置0，从而得到该语句的simhash值

例如把上面计算出来的“9 -9 1 -1 1 9”降维,得到的01串为：“1 0 1 0 1 1”，从而形成它们的simhash签名

相似度判断

* 每篇文档得到SimHash签名值后，接着计算两个签名的海明距离即可。
* 根据经验值，对64位的 SimHash值，海明距离在3以内的可认为相似度比较高

第九讲 检索排序

1、精确top *K* 检索及其加速办法

**(一般)步骤**：对每个文档评分(余弦相似度)，按照评分高低排序，选出前*K*个结果

如何加速：

方法一：快速计算余弦

方法二：堆排序法*N*中选*K*（不对所有文档的评分结果排序而直接选出Top *K*篇文档）

方法三：提前终止计算 （不需要计算所有*Ｎ*篇文档的得分）

2、非精确top *K*检索有哪些策略（每个策略能用自己的话描述即可，具体的算法例子不要求掌握）

非精确topK检索基本思想：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

* 策略一：索引去除(Index elimination)
* 策略二：胜者表(Champion list)
* 策略三：静态得分
* 策略四：影响度(Impact)排序
* 策略五：簇剪枝方法—预处理

3、链接分析算法

Pagerank 算法：强调链接数量与质量整体关系

HITS算法：强调权威页与枢纽页的 相互增强关系

Hilltop算法，知道名字，是前面两种方法的结合即可，由Pagerank 算法和HITS算法融合

Direct Hit 算法（知道名字即可）

4、Pagerank 算法

PageRank的思路：在随机游走过程中访问越频繁的网页越重要

PageRank 的核心思想 :vote

* + 强调链接数量和链接质量的整体关系
  + 参照科技文献重要性的评估方式，核心想法就是“被引用多的就是重要的”
  + PageRank 是基于「**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

5、知道PageRank算法是个迭代算法即可，公式要记住



* 其中，PR(A)表示页面A的级别，页面Ti链向页面A，
* L(Ti) 是页面Ti 链出的链接数量
* d取值在0到1之间，d也称为阻尼系数，由于用户不可能无限的单击下去，常常因劳累而随机跳入另一个页面
* 1-d则是页面本身所具有的网页级别。
  + 每一个页面设定的最小值是1-d
  + 没有链入，也不为0

6、HITS算法 Hyperlink- Induced Topic Search (HITS) 超链导向的主题搜索

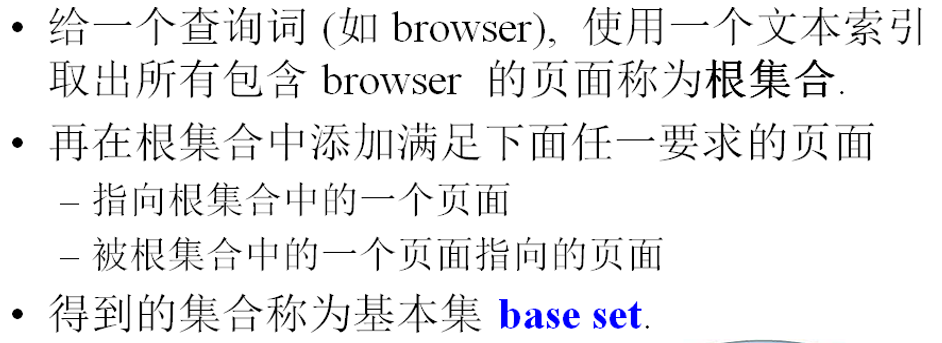
* 在HITS算法中,对每个网页都要计算两个值:权威值(authority)与中心值(hub)

7、HITS和PageRank的区别

a.HITS算法将重要性分为两个值权威值(authority)与中心值(hub)，PageRank只计算一个值

b.HITS和查询有关系，PageRank算法和查询无关

8、HITS基本步骤（基本思想）



对于基本集中的每一个页面*x*计算Hub分*h*(*x*)和Authority分*a*(*x*)

9、机器学习排序

Learning to Rank（L2R）

它有以下3种方法：（计算损失函数的方法，也是构造训练集的方法）

单文档方法

* + - **PointWise Approach**
    - 损失函数评估单个 doc 的预测得分和真实得分之间差异

文档对方法

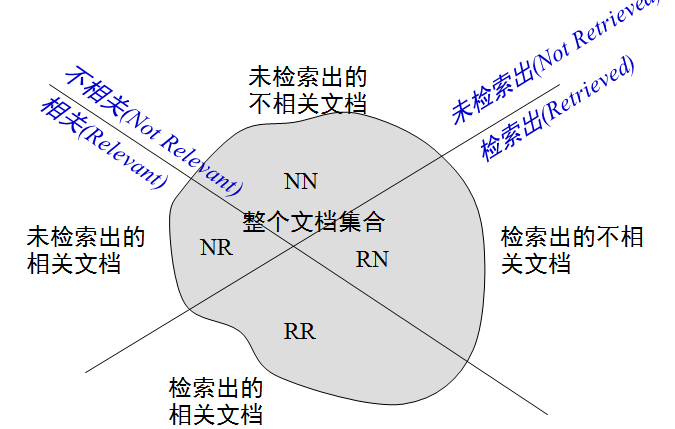
* + - **PairWise Approach**
    - 是判断任意两个文档组成的文档对<D0C1，D0C2>是否满足顺序关系

文档列表方法

* + - **ListWise Approach**
    - 搜索结果列表整体作为一个训练实例

第11讲 信息检索的评价（大考特考）

1、什么是查全率、查准率（这两个加上下面的F值都是对信息检索中无序检索结果的评价）



* **查准率**(Precision):返回的结果中真正相关结果的比率，也称为**查准率**， P∈ [0,1]
* **召回率**(Recall): 返回的相关结果数占实际相关结果总数的比率，也称为**查全率**，R∈ [0,1]

2、*F*值(*F*-measure)（它是评价的第三个指标）

* *F*值(*F*-measure)：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。F1的计算由下面的公式决定（调和平均数）



3、有序检索结果的评价（基本框架如下）

* 2.1 Binary relevance（二元关联）
  + 2.1.1 Precision@K (P@K)
  + 2.1.2 A precision-recall curve
  + 2.1.3 Mean Average Precision (MAP)
  + 2.1.4 Mean Reciprocal Rank (MRR)
* 2.2 Multiple levels of relevance
  + Normalized Discounted Cumulative Gain (NDCG)

4、Precision@K (P@K) R-查准率是什么

* 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.

例子（理解就行了）

* + R=10, R-查准率=4/10;
  + R=3, R-查准率=2/3

5、A precision-recall curve 查准率/查全率曲线

曲线下的面积被称为AP分数（Average precision score）

6、Mean Average Precision (MAP)是什么

* 平均查准率均值
* MAP是多个查询/排名的平均精度
* 在每个相关文档位置上查准率的平均值，被称为**平均查准率** Average Precision (AP)

7、NDCG是什么

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。

第十二讲 相关反馈及查询扩展  
Relevance Feedback & Query Expansion

1、K-GRAM index

定义：k个连续的**字母**的组合

作用：拼写校正

步骤：(1)给定一个查询query，计算出query所对应的k-gram，

(2)k-gram索引遍历，

（3）对每个单词和query计算Jaccard系数

（4）取得Jaccard系数高出阈值的单词，构成query的相似集合S

（5）计算query与集合S的编辑距离，找最小的，作为纠错词项

2、两种提高召回率的方法是什么（都会降低准确率）

相关反馈：在初始检索结果的基础上，通过用户交互指定哪些文档相关或不相关，然后改进检索的结果

查询扩展：通过在查询中加入同义或者相关的词项来提高检索结果的召回率

第13讲 隐语义空间 Latent Semantic Space

1、LSA的基本思想 **隐语义分析 Latent Semantic Analysis**

（1）建立词频矩阵

（2）计算矩阵的奇异值分解

（3）对于每一个文档d，用排除了SVD中消除后的词的新的向量替换原有的向量

（4）用转换后的矩阵进行文档索引和相似度计算

2、LSA优点

（1）文档和单词都映射到同一个语义空间，所以可以计算文档和文档的相似度，词项和词项的相似度，词项和文档的相似度

（2）语义空间的维度明显明显少于源单词-文章矩阵

3、LSA应用（没怎么提，稍微背一下即可）

（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

4、要知道LSA最大的缺点

LSA的概率模型假设文档和词的分布是服从联合正态分布的，但从观测数据来看是服从泊松分布的。

5、概率潜在语义分析Probabilistic Latent SemanticAnalysis pLSA是一个什么模型？

是一个主题模型，（列出公式）



一个文档包含若干主题，每个主题的概率是p(主题|文档)，在主题中选择一个词，词的概率是p(词语|主题)，二者相乘，遍历每一个主题，求之和得到这篇文档下这个词出现的概率，这就是主题模型

6、如何训练这个主题模型

（1）根据大量已知的文档-词项信息p(w|d) ，

（2）训练出文档-主题p(z|d)和主题-词项p(w|z)

* + - 分布对应了两组多项式分布
    - 需要估计这两组分布的参数

7、pLSA比有LSA什么好处

（1）定义了概率模型，而且每个变量以及相应的概率分布和条件概率分布都有明确的物理解释；

（2）相比于LSA隐含了高斯分布假设，pLSA隐含的多项式分布假设更符合文本特性

8、pLSA用来干什么

（1）文本聚类（2）文本分类 （3）计算文档的相似度（就是之前基于语料库的计算文档相似度的方法）

9、LDA（Latent Dirichlet Allocation）主题模型是什么

隐含狄利克雷分布

10、和前面的pLSA主题模型有什么区别

增加了狄利克雷的先验知识，所有的参数都不是设定的，而是进行了全贝叶斯化，更符合实际的情况

11、Gensim是一个用于从文档中自动提取语义主题的Python库

第14讲 词嵌入word embedding

1、词的表示有哪两种方式，两种表示方法的区别是什么？

One-hot表示

* + - 基于词袋（BOW）
    - 每个词表示为一个很长的向量。
    - 这个向量的维度是词表大小，
    - 其中绝大多数元素为 0，只有一个维度的值为 1
    - 这个维度就代表了当前的词。

分布式表示，

* + 表示的一种低维实数向量
  + 维度以 50 维和 100 维比较常见
  + 每一维可以看成词的语义或者主题信息

2、什么叫做统计语言模型

是对于词的序列的概率分布描述，是自然语言处理的基础，为上下文相关的特性建立数学模型

3、这个统计语言模型的公式是什么？

S 可以表示某一个由一连串特定顺序排列的词而组成的一个有意义的句子。

* + S ：ω1，ω2，...，ωn
  + ***S***在文本中出现的可能性，
    - 即***S***的概率***P(S)***，***P(S)***=***P***(ω1，ω2，...，ωn)
  + 利用条件概率的公式：
    - S 的概率 P(S)等于每一个词出现的概率相乘
    - P(S) =***P***(ω1)•***P***(ω2|ω1)•***P***(ω3|ω1,ω2)•••***P***(ωn|ω1，ω2，...，ωn-1)

4、什么是**n-gram语言模型**

* N-1阶马尔可夫假设:
  + 假定文本中的每个词ωi和前面的N-1个词有关，而与更前面的词无关
* 对应的语言模型称为N元模型(N-Gram Model)。

5、统计语言模型、n-gram语言模型有什么应用

* 文本生成、机器翻译
* 拼写纠错
* 语音识别
* 音字转换
* 分词

6、N-Gram语言模型特点

* + 简单有效
  + 只考虑了词的位置关系，
  + 没有考虑词之间的相似度，词语法和词语义，
  + 还存在数据稀疏的问题

7、word2vec包括两个模型

* CBoW模型（Continuous Bag-of-Words Model）
* Skip-gram模型

8、这两个模型要求就比较高了

每个模型包含了几层，输入的是什么，经过什么操作会有什么输出，用模型的哪一部分来获得向量化过程

CBow模型

* 模型输入：上下文的one hot表示方式
* 输入分别跟同一个VxN的大小的系数矩阵W1相乘得到C个1xN的隐藏层hidden layer，
* 然后C个取平均所以只算一个隐藏层
* 隐藏层跟另一个NxV大小的系数矩阵W2相乘得到1xV的输出层（这个输出层每个元素代表的就是词库里每个词的事后概率）
* 训练结束后的的矩阵，任何一个单词的One-Hot表示乘以这个矩阵W就可以得到其词向量的表示

Skip-gram模型：

模型的本质:

计算输入word的input vector与目标word的output vector之间的余弦相似度，

并进行softmax归一化

input word和output word都是one-hot编码的向量，判断两者是不是邻接关系

模型使用了两种技术：

一种是层次的softmax，另一种是负采样

9、**word2vec**模型训练 Python语言平台 **Gensim**

10、词向量化有什么作用（不就是word to vec吗）

（1）计算两个词的相似程度

（2）计算词和词的语义关系

第15讲 图片检索

1、基于内容的图像检索CBIR

* + Content-based image retrieval
  + 用户输入一张图片，以查找具有相同或相似内容的其他图片。

2、CBIR 的关键技术:图像特征提取和特征匹配

3、图片的特征有颜色特征、形状特征、纹理特征

4、颜色特征的表示有几种

* **1、颜色直方图(Color Histogram)**
* **2、颜色相关图(Color Correlogram)**
* **3、颜色矩(Color Moment)**
* **4、颜色一致性矢量(Color Coherence Vectors, CCV)**

5、基于颜色特征的快速图片检索

其中"感知哈希算法“ Perceptual hash algorithm步骤什么

（1）对每张图片生成一个"指纹"（fingerprint）字符串，也就是图片的特征

（2）然后比较不同图片的指纹，结果越接近，就说明图片越相似（用海明距离来计算）

（之前计算文档相似度的局部敏感哈希也是用hash法，比较哈希码的相似度来判断文档相似程度，都是用海明距离）

那么怎么将图片变为哈希码呢？

（1）均值Hash算法

（2）pHash算法

（3）颜色分布法

6、纹理分析的途径有几种呢？

（1）结构纹理分析 （2）统计纹理分析 （3）模型纹理分析 （4）信号处理纹理分析

7、基于统计的纹理分析中有哪几种呢？

（1）灰度差分统计法

（2）基于灰度共现矩阵的纹理特征

（3）Tamura等定义的6个心理学特征

8、在灰度共现矩阵中提取哪几个特征？

（1）能量 （2）对比度 （3）自相关 （4）熵 （5）逆差矩

9、Tamura纹理特征有哪些？

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)

10、基于信号处理方法描述纹理特征中有几种方法？

（1）傅里叶频谱法

（2）Gabor变换

（3）小波(Wavelet)变换

11、LBP是什么

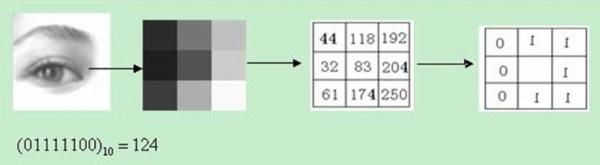
局部二值模式 Local Binary Patterns，结合了纹理图像结构和像素统计关系的纹理特征描述方法

12、LBP显著的特点

* + 对光照具有不变性。
  + 具有旋转不变性
  + 灰度不变性等

13、LBP怎么构造

* LBP算子定义为在3\*3的窗口内，
* 以窗口中心像素为阈值，将相邻的8个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素 点的位置被标记为1，否则为0。
* 3\*3邻域内的8个点经比较可产生8位二进制数（通常转换为十进制数即LBP码，共256种），即得到该窗口中心像 素点的LBP值，并用这个值来反映该区域的纹理信息。



14、LBP有哪些应用

纹理分类、人脸分析

15、形状特征分为哪两类

轮廓 区域

16、基于轮廓的形状描述符有哪几种呢？

链码 傅里叶描述子 基于网格的方法 距离直方图 边界矩

17、大津法（用于寻找最佳阈值，即使得类中差异最小，类间差异最大的阈值）

* ["大津法"](http://en.wikipedia.org/wiki/Otsu's_method)（Otsu's method）
  + 灰度值小于阈值的像素为 n1 个，均值为μ1
  + 大于等于阈值的像素为 n2 个，均值为μ2
  + w1 和 w2 表示这两种像素各自的比重
  + w1 = n1 / n w2=n2/n
  + 类内差异 = w1(σ1的平方) + w2(σ2的平方)
  + 类间差异 = w1w2(μ1-μ2)^2

采用穷举法来寻找使得类间差异最大的阈值

18、图像的局部特征（大考特考）

有三种 HOG特征描述、SIFT特征描述、LBP特征描述

19、HOG特征是什么

方向梯度直方图（Histogram of Oriented Gradient, HOG），一种在计算机视觉和图像处理中用来进行物体检测的特征描述子，通过计算和统计图像局部区域的梯度方向直方图来构成特征。

20、HOG有什么应用？行人检测

21、HOG特征的提取步骤

（1）计算图像每个像素的梯度

（2）将图像划分成小cells

（3）统计每个cell的梯度直方图

（4）将每几个cell组成一个block

（5）将图像image内的所有block的HOG特征descriptor串联起来就可以得到该image（你要检测的目标）的HOG特征descriptor了

22、HOG特征的优点（理解着背）

由于HOG是在图像的局部方格单元上操作，所以它对图像几何的和光学的形变都能保持很好的不 变性，这两种形变只会出现在更大的空间领域上。其次，在粗的空域抽样、精细的方向抽样以及较强的局部光学归一化等条件下，只要行人大体上能够保持直立的姿 势，可以容许行人有一些细微的肢体动作，这些细微的动作可以被忽略而不影响检测效果。因此HOG特征是特别适合于做图像中的人体检测的

23、SIFT特征是什么

尺度不变特征转换，Scale-invariant feature transform或SIFT，在空间尺度中寻找极值点，并提取出其位置、尺度、旋转不变量。

24、SIFT特征和HOG特征好处

SIFT特征不只具有尺度不变性，即使改变旋转角度，图像亮度或拍摄视角，仍然能够得到好的检测效果，Hog没有旋转和尺度不变性

25、SIFT有哪几个步骤

* + 步骤一：建立尺度空间
    - 即建立高斯差分(DoG)金字塔
  + 步骤二：在尺度空间中检测极值点，并进行精确定位和筛选
  + 步骤三：特征点方向赋值，
    - 完成此步骤后，每个特征点有三个信息：位置、尺度、方向
  + 步骤四：计算特征描述子

26、SIFT特征的匹配是暴力匹配

27、图像检索算法作用

基于内容的图像检索(Content-based Image Retrieval, CBIR)方法利用从图像提取的特征来进行检索。

28、encoding 把图像的各个局部特征形成编码，有哪三种方法？

* + [**BoW**](http://yongyuan.name/blog/Bag%20of%20visual%20words%20model:%20recognizing%20object%20categories) **(Bag of visual word) BOF**
  + [**VLAD**](https://hal.inria.fr/inria-00633013/document) **(Aggregating local descriptors)局部聚合向量**
  + [**FV**](https://hal.inria.fr/hal-00830491/document) **(Fisher Vector)**

29、Bag-of-Featrures

* **BOF图像检索算法流程：（用自己的话描述即可，改成自己理解的即可）**
  + 1.用surf算法生成图像库中每幅图的特征点及描述符。
    - surf算法是关键点计算和描述算法，作用和SIFT相似。
  + 2.再用k-means算法对图像库中的特征点进行训练，生成类心。
  + 3.生成每幅图像的BOF，
    - 判断图像的每个特征点与哪个类心最近，最近则放入该类心，最后将生成一列频数表，即初步的无权BOF（直方图向量）。
  + 4.通过tf-idf对频数表加上权重，生成最终的bof。
    - 因为每个类心对图像的影响不同。比如超市里条形码中的第一位总是6，它对辨别产品毫无作用，因此权重要减小。
    - TF/IDF
  + 5.对查询图像也进行3.4步操作，生成该图的直方图向量BOF。
  + 6.将查询图像的Bof向量与图像库中每幅图的Bof向量计算相似度
    - 求夹角余弦。

2019.12.26