1. 搜索
2. 为什么搜索？（搜索和推荐的意义）
   1. 信息过载，数据爆炸性的增长，而人的处理能力有限；
   2. 网页数量近千亿，数据总量10万亿GB；
3. 什么是搜索？
   1. 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
   2. 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
4. 信息检索的本质：
   1. 给定一个查询*Q*，
   2. 从文档集合*C*中
   3. 计算每篇文档*D*与*Q*的相关度
   4. 并排序(Ranking)。
5. 什么是相关度？
   1. 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*)
   2. 相关度通常只有相对意义，
      1. 对一个*Q*，不同文档的相关度可以比较，
      2. 而对于不同的*Q*的相关度不便比较
   3. **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
6. 什么是信息检索模型？
   1. 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。
7. 信息检索的主要技术
   1. 文本分析
      1. NLP
   2. 建立索引
   3. 查询
      1. 查询分析 NLP
      2. 相关度计算
         1. 和信息检索模型相关
   4. 排序
      1. 实验室评价
8. 搜索引擎的主要技术
   1. 爬虫
   2. 文本分析
      1. NLP
   3. 索引
   4. 查询
      1. 查询分析 NLP
      2. 相关度计算
   5. 排序
   6. 用户反馈
      1. refine query , relaxing query
9. 搜索引擎的工作原理
   1. **爬行和抓取**
      1. 爬虫程序（Spider，robot ）
      2. 搜索引擎从已知的数据库出发，就像正常用户的[浏览器](http://baike.baidu.com/view/7718.htm)一样访问这些网页并抓取文件。
      3. 搜索引擎通过这些爬虫去爬互联网上的外链，从这个网站爬到另一个网站，去跟踪网页中的链接，访问更多的网页
      4. 这些新的网址会被存入数据库等待搜索。
   2. 建立索引
      1. 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).
      2. 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**
      1. 用户在搜索[引擎](http://baike.baidu.com/view/53607.htm)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](http://baike.baidu.com/view/58654.htm)进行处理，
      2. 如中文特有的分词处理，去除[停止词](http://baike.baidu.com/view/2860648.htm)，判断是否需要启动整合搜索，
      3. 判断是否有拼写错误或错别字等情况。
   4. 排序
      1. 对[搜索词](http://baike.baidu.com/view/58654.htm)处理后，搜索引擎程序便开始工作，从索引数据库中找出所有包含搜索词的[网页](http://baike.baidu.com/view/828.htm)，并且根据排名算法计算出哪些网页应该排在前面，然后按照一定格式返回到“搜索”页面。
      2. 再好的搜索引擎也无法与人相比，这就是为什么网站要进行[搜索引擎优化](http://baike.baidu.com/view/7147.htm)。没有[SEO](http://baike.baidu.com/view/1047.htm)的帮助，[搜索](http://baike.baidu.com/view/8638.htm)引擎常常并不能正确的返回最相关、最权威、最有用的[信息](http://baike.baidu.com/view/1527.htm)。
10. 搜索引擎的评价
    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
11. 网络爬虫技术
    * + 1. 爬虫的定义
           1. 一种自动获取网页内容的程序，是搜索引擎的重要组成部分, 通俗的讲，也就是通过HTML源码解析来获得想要的内容
        2. 爬取过程的描述
           1. 从一个或若干初始网页的[URL](http://baike.baidu.com/view/1496.htm)开始，
           2. Fetch and parse them
           3. Extract URLs they point to
           4. Place the extracted URLs on a queue
           5. Fetch each URL on the queue and repeat
           6. 直到满足系统的一定停止条件。
        3. 爬虫的功能
           1. 鲁棒性
           2. 礼貌性
           3. 分布式
           4. 可扩展性
           5. 功能可扩展性
           6. 新鲜性
        4. 爬取策略
           1. 理论上，两者能够在大致的时间里完成所有的整个静态网页的爬取工作。
           2. 工程上，网络爬虫更应该定义为“如何在有限的时间里最多的爬下那些重要的网页”
           3. DFS 要限定爬取的深度
           4. 在爬取时为了防止有些错误链接导致的无穷递归爬取，需要限定爬取的深度。
           5. 此外层次越深的网页对用户来说可用的信息越少，理论上呈对数的倒数关系
        5. URL查重：哈希法
        6. 文档查重：文档指纹
        7. 常见的开源软件
           1. Nutch
           2. Heritrix
12. 网页分析技术
13. 什么是正则表达式？
    1. 是由一组普通字符和一组元字符组成的字符串，
    2. 用来表示符合一定模式的一组字符串，
    3. 常用于字符串处理，表单验证等场合，
    4. 表示能力与正规文法相同。
14. 正则表达式用来做什么？
    1. 正则表通常被用来检索、替换那些符合某个模式(规则)的文本。
15. 基于正则表达式获取内容的步骤
    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
16. 正则表达式的特点
    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
17. 什么是DOM
    1. 文档对象模型
    2. 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。
18. 开源HTML解析器
    1. **JAVA：HTMLParser,jsoup等**
    2. **jsoup 是一款Java 的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据。**
    3. **C/C++：htmlcxx等**
    4. **Python：Beautiful Soup等**
    5. **C#：Winista.Htmlparser.Net等**
19. **正则表达式和DOM的比较**
    1. 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
    2. 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于HTML DOM树的爬取程序。
20. 什么是元搜索
    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
21. 网站防爬博弈
    1. 网站防爬取措施
       1. Robot 协议
       2. IP屏蔽
       3. 登录
       4. JavaScript渲染
    2. 方案
       1. 模拟浏览器工作
22. 词项词典
23. 如何建立词项词典？
    1. 文档解析
    2. 词条化
    3. 词项归一化
    4. 词干还原
    5. 词形归并
    6. 停用词
24. 什么是词条化？
    1. 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
25. 什么是词项归一化？
    1. 将文档和查询中的词条“归一化”成一致的形式
26. 归一化的结果
    1. 在IR系统的词项词典中，形成多个近似词项的一个等价类
27. 词项归一化的**策略**：建立同义词扩展表
    1. 为每个查询维护一张包含多个词的查询扩展词表
    2. 在建立索引建构时就对词进行扩展
28. 西方同音词算法
    1. SoundEx算法
29. 什么是词干还原？
    1. 通常指去除单词两端词缀的启发式过程
    2. 词干还原能够提高召回率，但是会降低准确率
30. 英文处理中最常用的**词干还原**算法
    1. Porter算法
31. 什么是词形归并?
    1. 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。
    2. 词形归并可以减少词项词典中的词项数量
32. 词干还原和词形归并的区别
    1. 代表意义不同。
       1. Stemming通常指很粗略的去除单词两端词缀的启发式过程。
       2. 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. 两个过程的区别还在于：
       1. 词干还原在一般情况下会将多个派生相关词合并在一起，
       2. 而词形归并通常只将同一词元的不同屈折形式进行合并。
    3. 词干还原和词形归并，都体现了不同语言之间的差异性
33. 什么是停用词？
    1. 应用太广泛，区分度太低
    2. 对这样的词搜索引擎无法保证能够给出真正相关的搜索结果，难以帮助缩小搜索范围，同时还会降低搜索的效率
34. 消除停用词的优缺点
    1. 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
    2. 缺点：
       1. 有时消除的停用词对检索是有意义的
35. 如何确定停用词
    1. 查表法
    2. 基于文档频率
36. 中文分词
    * + 1. 什么是分词？
           1. 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. 分词方法
           1. 基于理解的分词方法
           2. 基于字符串匹配的分词方法
           3. 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
           1. 优点：

程序简单易行，开发周期短；

仅需很少的语言资源（词表），

不需要任何词法、句法、语义资源。

可以自定义词库，增加新词

* + - * 1. 缺点

Out of Vocabulary

歧义消解能力差；

切分正确率不高，一般在95%左右。

* + - 1. 基于统计的分词方法有哪些？基本思想是什么？
         1. HMM、CRF、SVM、深度学习等算法
         2. 用字与字相邻出现的频率来反应成词的可靠度，统计语料中相邻出现的各个字的组合的频度，当组合频度高于某一个临界值时，我们便可认为此字组可能构成一个词语。
      2. 基于统计的主要统计模型有？
         1. N元文法模型（N-gram）
         2. 隐马尔可夫模型（Hidden Markov Model ，HMM）
         3. 最大熵马尔可夫模型（MEMM）
         4. 条件随机场模型（Conditional Random Fields，CRF）
         5. 深度学习模型
      3. 基于统计的分词方法的优缺点：
         1. 优点：

分词准确度高；

能够平衡地看待词表词和未登录词的识别问题。

* + - * 1. 缺点：

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

对常用词的识别精度差，时空开销大

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

* + - 1. 什么是马尔科夫过程？
         1. 是指数学中具有马尔可夫性质的离散事件[随机过程](http://baike.baidu.com/view/18964.htm)
      2. HMM模型是一个五元组
         1. 状态值集合
         2. ObservedSet: 观察值集合
         3. TransProbMatrix: 转移概率矩阵
         4. EmitProbMatrix: 发射概率矩阵
         5. InitStatus: 初始状态分布
      3. HMM模型可以用来解决三种问题
         1. 模型参数学习问题
         2. 预测问题
         3. 评估观察序列概率
      4. HMM 模型应用
         1. 股票预测
         2. 模式识别
         3. 语音识别
         4. 分词
      5. Viterbi 算法是什么
         1. 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。
      6. 常见开源分词软件有？
         1. [庖丁解牛分词](http://code.google.com/p/paoding/)
         2. IKAnalyzer
         3. 结巴分词等；

1. 布尔模型与倒排索引
2. 什么是信息检索模型?
   1. 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) >
3. 基于内容的信息检索模型
   1. 集合论模型：布尔模型、模糊集合模型、扩展布尔模型
   2. 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
   3. 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
   4. 深度学习模型
4. 什么是布尔检索模型
   1. 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，每篇文档：索引词（0或1）的集合。
5. 什么是“词袋”模型
   1. 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，将其仅仅看作一个词集合
      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
6. 倒排索引是什么，有哪两部分组成？
   1. 词项 + 倒排记录
   2. 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。
   3. 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
   1. 二元词索引
   2. 位置信息索引
8. 布尔检索模型的特点
   1. 优点：
      1. 查询简单，因此容易理解
      2. 通过使用复杂的布尔表达式，可方便地控制查询结果
   2. 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
9. 向量空间模型
10. 什么是Jaccard系数，公式是什么？
    1. 一种常用的衡量两个**集合**A,B**重叠度**的方法
    2. Jaccard(A,B) = |A ∩ B| / |A ∪ B|
11. 什么是tf？
    1. 词项频率：词项*t*在文档*d*中出现的次数，记为*tft,d*
    2. 一种替代原始tf的方法: 对数词频
12. 什么是idf，公式是什么？
    1. df文档频率，出现词项的文档数目
    2. *idft* = log10(*N*/*dft*)
    3. *idft* 是反映词项*t*的信息量的一个指标
13. 什么是tf-idf？
    1. tf-idf 是信息检索中最著名的权重计算方法
    2. tf-idf值随着词项在单个文档中出现次数(tf)增加而增大
    3. tf-idf值随着词项在文档集中数目(df)增加而减小
    4. 词项*t*的tf-idf 由它的tf和idf组合而成
    5. *wt,d*=(1+log *tft,d*) × log10(*N*/*dft*)
14. tf-idf的作用
    1. TF-IDF是一种[统计方法](http://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%96%B9%E6%B3%95)，
    2. 用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。
    3. 字词的重要性随着它在文件中出现的次数成正比增加，
    4. 但同时会随着它在[语料库](http://baike.baidu.com/item/%E8%AF%AD%E6%96%99%E5%BA%93)中出现的频率成反比下降
15. 什么是向量空间模型？
    1. 每篇文档表示成一个基于tf-idf权重的实值向量∈ *R*|*V*|
16. 向量相似度的计算
    1. Jaccard相似度:比较文本相似度，用于文本查重与去重
    2. 欧式距离
    3. 余玄相似度
17. 向量空间模型的特点
    1. 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
    2. 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
18. 相似度计算
19. 不同范数的距离：
    1. L1 范数：曼哈顿距离
    2. L2 范数：欧几里得距离
    3. Lp范数：明科夫斯基距离
    4. 切比雪夫距离
20. 标准化欧氏距离改进的地方：
    1. 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
21. 什么是加权欧氏距离？
    1. 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
22. 什么是编辑距离？
    1. 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
23. 海明距离
    1. 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。
24. 什么是谷本系数？
    1. 广义Jaccard相似度，元素的取值可以是实数。又叫作谷本系数
25. 去重算法
    1. shingle算法
    2. 局部敏感哈希
       1. MinHash算法
       2. SimHash算法
26. shingle算法的基本思想？
    1. Shingle算法的核心思想是将文件相似性问题转换为集合的相似性问题
27. 局部敏感哈希可以用来降维
28. MinHash的用处
    1. 可以用来快速估算两个集合的相似度。
    2. 用于在搜索引擎中检测重复网页。
    3. 它也可以应用于大规模聚类问题
29. SimHash的步骤
    1. 分词、hash、加权、合并、降维
30. 检索排序
31. 什么是精确TOP K 排序？
    1. 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
32. 精确TOP K 排序的三种加速方法：
    1. 快速计算余弦
    2. 堆排序法*N*中选*K*
    3. 提前终止计算
33. 什么是非精确TOP k 检索？
    1. 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
34. 非精确 TOP K检索的五种策略
    1. 索引去除
    2. 胜者表
    3. 静态得分
    4. 影响度排序
    5. 簇剪枝方法—预处理
35. 链接分析算法？
    1. Pagerank 算法
       1. 强调链接数量与质量整体关系
    2. HITS算法
       1. 强调权威页与枢纽页的 相互增强关系
    3. Hilltop算法
       1. 强调链接与链接之间相关性与质量度
       2. 由Pagerank 算法和HITS算法融合
    4. LALSA算法
       1. 强调优质链接可进行间接性权重传递
36. 描述PageRank算法
    1. 假设一个浏览者在网络上随机行走，在稳定状态下，每个页面都有一个访问概率 – 用这个概率作为页面的分数
37. PageRank算法的核心思想
    1. 投票
    2. 强调链接数量和链接质量的整体关系
    3. 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”
    4. 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
38. PageRank算法的公式
    1. 
39. HITS算法的基本思想
    1. 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)
40. PageRank 和 HITS 算法的区别：
    1. 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。
    2. 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；
    3. 而PageRank算法是独立于检索主题，因此也常被称为Query-independent算法
41. Direct Hit 算法
    1. 注重信息的质量和用户反馈的排序方法
42. 机器学习L2R的三种方法
    1. 单文档方法
       1. 损失函数评估单个 doc 的预测得分和真实得分之间差异
    2. 文档对方法
       1. 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
    3. 文档列表方法
       1. 搜索结果列表整体作为一个训练实例

十一、

1. 什么是查准率和查全率？
   1. **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**
   2. **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**
2. **什么是F值？**
   1. *F*值(*F*-measure)：召回率*R*和查准率*P*的加权调和平均值
   2. 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
3. 什么是R-查准率？
   1. 计算序列中第R个位置文献的查准率
   2. P@10
      1. 目前Web信息检索中最常用的测度
      2. 检查网页排序算法的好坏
4. 查准率查全率曲线的特点？
   1. 原始的曲线常常呈现锯齿状
   2. 需要去掉锯齿，进行平滑。采用插值查准率,记为pinterp
   3. 在查全率为*r*的位置的插值查准率，定义为查全率不小于*r*的位置上的查准率的最大值
5. **什么是MAP**？
   1. 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**
   2. 看好MAP的计算方式！！！
6. 什么是NDCG？
   1. 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

十二、 相关反馈及查询扩展

1. 什么是K-Gram索引？
   1. K-GRAM的定义：k个连续的**字母**的组合
   2. 索引结构：
      1. k-gram index的dictionary是所有词的k-gram的集合
      2. k-gram index的posting是匹配k-gram的单词序列
2. 拼写校正的方法？
   1. 词项独立校正
      1. 编辑距离方法
      2. k-gram重复度法
   2. 上下文敏感的校正
3. 拼写校正的步骤？
   1. 利用k-gram计算Jaccard系数
      1. 给定一个查询Q，计算出Q所对应的k-gram，
      2. k-gram索引遍历，
      3. 对每个单词和Q计算Jaccard系数
      4. 取得Jaccard系数高出阈值的单词
   2. 总结：
      1. 可以先进行k-gram索引，然后再进行编辑距离的计算
      2. 计算query与集合S的编辑距离，找最小的。作为纠错词项
4. 提高召回率的方法？
   1. **局部(local)方法**: 对用户查询进行局部的即时的分析
      1. **交互式相关反馈**
   2. **全局(Global)方法**: 进行一次性的全局分析(比如分析整个文档集)来产生同/近义词词典
      1. **查询扩展**

十三、 隐语义空间

1. 什么是LSA？
   1. 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
2. LSA的缺点是什么？
   1. 无法解决多义词的问题
   2. 特征向量的方向没有对应的物理解释
   3. SVD的计算复杂度很高，而且当有新的文档来到时，若要更新模型需重新训练
   4. 维数的选择是ad-hoc的
   5. LSA具有词袋模型的缺点，即在一篇文章，或者一个句子中忽略词语的先后顺序
   6. 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
3. 什么是pLSA
   1. PLSA是以统计学的角度来看待[LSA](http://baike.baidu.com/item/LSA)，是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
4. 什么是主题模型？
   1. 一篇文档(Document) 可以由多个主题(Topic) 混合而成
5. **“文档-词项”的生成模型的训练？**
   1. 按照概率选择一篇文档d
   2. 选定文档后，从主题分布中按照概率选择一个隐含的主题类别p(z|d)
   3. 选定后，从词分布中按照概率p(w|z)选择一个词
6. PLSA生成文档的过程?
   1. 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。
   2. 自动地发现文档集中的主题（分布）
      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)
7. PLSA有哪些应用？
   1. 文本聚类
   2. 文本分类
8. PLSA的优势？
   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Multi-nomial分布假设更符合文本特性
   3. pLSA的优化目标是是KL-divergence最小，而不是依赖于最小均方误差等准则
   4. 可以利用各种model selection和complexity control准则来确定topic
9. 生成过程中的公式？
   1. p(w|d) = p(w|z) \* p(z|d)
10. 什么是LDA模型？
    1. 一个隐含狄利克雷分布的主题模型
11. Gensim是一个用于从文档中自动提取语义主题的Python库

十四、 词嵌入

1. 什么是表征学习？
   1. 在[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0)中，**表征学习**是学习一个特征的技术的集合
   2. 将原始数据转换成为能够被机器学习来有效开发的一种形式
2. 什么是嵌入？
   1. 是一种可用于将离散变量表示成连续向量的方法
3. 词的表示方法？
   1. One-hot
   2. 分布式表示
4. ONE-HOT 和 分布式表示的区别？
   1. ONE-HOT 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
   2. ONE-HOT 语义鸿沟，维度灾难，稀疏, 无法表示unseen words
   3. 分布式表示的一种低维实数向量，每一维可以看成词的语义或者主题信息，维度压缩解决语义鸿沟，
   4. 基于学习模型，快速对于unseen words 进行表示
5. 什么是语言模型和统计语言模型？
   1. 语言模型根据语言客观事实而进行的语言抽象数学建模
   2. 统计语言模型为上下文相关的特性建立数学模型
6. 语言模型的公式
   1. S 的概率 P(S)等于每一个词出现的概率相乘
   2. P(S) =***P***(ω1)•***P***(ω2|ω1)•***P***(ω3|ω1,ω2)•••***P***(ωn|ω1，ω2，...，ωn-1)
7. 什么是n-gram语言模型？
   1. N-1阶马尔可夫假设:
   2. 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
   3. 对应的语言模型称为N元模型(N-Gram Model)
8. n-gram语言模型的应用
   1. **拼写纠错**
   2. **语音识别**
   3. **音字转换**
9. **n-gram语言模型的缺点**
   1. 简单有效
   2. 只考虑了词的位置关系，
   3. 没有考虑词之间的相似度，词语法和词语义，
   4. 还存在数据稀疏的问题
10. 神经网络模型？
    1. 每个模型包含基层输入，基层将获得输出
11. word2vec
    1. 用一个简单模型训练出连续的词向量
    2. 基于词向量的表达，训练一个连续的N-gram神经网络模型
12. CBoW模型
    1. CBoW模型等价于一个词袋模型的向量乘以一个embedding矩阵，从而得到一个连续的embedding向量
13. **word2vec 应用**
    1. google的word2vec工具
    2. **word2vec**模型训练 Python语言平台 **Gensim**

十五、 图片检索

1. 什么是基于内容的图像检索？
   1. CBIR: 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:图像特征提取和匹配
2. 颜色特征的表示
   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**
3. **什么是感知哈希算法？**
   1. 对每张图片生成一个"指纹"（fingerprint）字符串
   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
4. **基于颜色特征的快速图片检索的方法？**
   1. 根据图片特征进行相似度计算
   2. 降维
   3. 感知哈希算法
5. 基于统计特征的纹理特征提取方法？
   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
   4. 基于邻域灰度差别矩阵的纹理特征
6. 灰度共生矩阵分析——特征
   1. 能量（角二阶矩）ASM，是灰度共生矩阵元素值的平方和，反映了图像灰度分布均匀程度和纹理粗细度
   2. 对比度CON，反映了图像的清晰度和纹理沟纹深浅的程度
   3. **自相关，反应了图像纹理的一致性**
   4. 熵，它表示了图像中纹理的非均匀程度或复杂程度
   5. 逆差矩，反映图像纹理的同质性
7. Tamura纹理特征和灰度共生矩阵表示的主要区别是Tamura纹理特征中所有纹理特征都在视觉上有意义
8. Tamura的六种属性
   1. 对比度(contrast)、粗糙度(coarseness)、方向性(directionality)对于图像检索尤为重要
   2. 线像度(1ine likeness)、规整度(regularity)和粗略度(roughness)
9. 什么是LBP？
   1. 局部二值模式，一种有效的纹理描述[算子](http://baike.baidu.com/view/53313.htm)，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
10. 基于信号处理方法描述纹理特征
    1. 傅里叶频谱法
    2. Gabor纹理
    3. **小波**
11. LBP的特点？
    1. 对光照具有不变性。
    2. 具有旋转不变性
    3. 灰度不变性等
12. 如何构造LBP？
    1. LBP算子定义为在3\*3的窗口内，
    2. 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\*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
13. LBP的应用中，如纹理分类、人脸分析等
14. 形状的描述符大体可以分为两大类
    1. 基于轮廓的形状描述符
       1. 描述形状目标区域边界轮廓的像素集合，称为基于轮廓的形状描述符
       2. 链码(用一串数字表示图像中目标的边界)，傅里叶描述子，小波变换
    2. 基于区域的形状描述符
       1. 对形状目标区域内所有像素集合的描述
       2. 区域面积，几何矩
15. 大津法Otsu's method
    1. 证明了 "类内差异最小"与"类间差异最大"是同一件事
    2. 计算方法：
       1. 灰度值小于阈值的像素为 n1 个，
       2. 大于等于阈值的像素为 n2 个
       3. w1 和 w2 表示这两种像素各自的比重
       4. w1 = n1 / n
       5. 类内差异 = w1(σ1的平方) + w2(σ2的平方)
       6. 类间差异 = w1w2(μ1-μ2)^2
16. 常见的图像局部特征
    1. **HOG特征描述**
    2. **SIFT特征描述算子**
    3. 其他特征
       1. **LBP特征描述**
       2. Harr特征描述
       3. SURF特征描述算子
       4. ORB特征描述算子
17. HOG是什么？
    1. 方向梯度直方图，Histogram of Oriented Gradient, HOG
    2. 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
    3. 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
18. Hog特征结合 SVM分类器已经被广泛应用于图像识别中，尤其在行人检测中获得了极大的成功
19. HOG特征如何提取？
    1. 灰度化
    2. 采用Gamma校正法对输入图像进行颜色空间的标准化（归一化）
    3. 计算图像每个像素的梯度
    4. 将图像划分成小cells
    5. 统计每个cell的梯度直方图
    6. 将每几个cell组成一个block
    7. 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的HOG特征descriptor了
20. HOG算法的优缺点？
    1. 优点
       1. 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。
       2. 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。
       3. 因此HOG特征是特别适合于做图像中的人体检测的
21. 什么是SIFT？
    1. 尺度不变特征转换，Scale-invariant feature transform
    2. 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
22. SIFT算法大致分为四个步骤
    1. 步骤一：建立尺度空间
    2. 步骤二：在尺度空间中检测极值点，并进行精确定位和筛选
    3. 步骤三：特征点方向赋值，
    4. 步骤四：计算特征描述子
23. 基于特征点的图像相似度计算 基于SIFT描述子
    1. 通过找到匹配点的个数来判断两幅图像是否一致，
       1. 暴力匹配
       2. 计算向量的欧氏距离, 找距离最小的样本向量
24. 图像检索算法
    1. 图像检索领域：将局部特征表示成全局特征的编码
    2. 通常继承了局部特征的部分不变性，如对平移、旋转、缩放、光照和遮挡等与语义相关不大的因素保持不变
25. 三种经典的编码
    1. [BoW](http://yongyuan.name/blog/Bag%20of%20visual%20words%20model:%20recognizing%20object%20categories)
    2. [VLAD](https://hal.inria.fr/inria-00633013/document)**局部聚合向量**
    3. [FV](https://hal.inria.fr/hal-00830491/document)
26. BOF算法流程
    1. 1.用surf算法生成图像库中每幅图的特征点及描述符。
    2. 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. 3.生成每幅图像的BOF，
    4. 4.通过tf-idf对频数表加上权重，生成最终的bof。
    5. 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. 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