《信息检索与数据挖掘》复习指南2016版

# [期中考试前]

1. 给定微型文档集，画出文档集对应的词项—文档矩阵；画出该文档集的倒排索引。
2. 布尔检索中查询的优化。
3. 文档词条化、词项归一化、词干还原和词形归并。
4. 倒排索引的构成与构建的步骤。
5. 倒排记录表的合并。
6. 基于跳表的倒排记录表。
7. 内存受限给倒排索引构建和维护带来了什么问题，如何应对？
8. 倒排索引字典和倒排记录部分各自可用哪些方式实现压缩存储？
9. 何为词项频率tf（term frequencey）、逆文档频率idf（inverse document frequency）？
10. 向量空间模型下为何常用余弦相似度而非欧式距离来度量相关性？
11. 假定*C*为*M*×*N* 的词项-文档计数(tf)矩阵。矩阵*CCT* 的物理意义？矩阵*CTC*的物理意义？如果*C*为*M*×*N* 的词项-文档权重(tf-idf)矩阵呢？
12. 举例说明词项-文档计数(tf)矩阵的SVD分解在文档分析中有何应用。
13. 用向量模型表示文档时一义多词（synonymy）会带来什么问题，有何应对措施？
14. 用向量模型表示文档时一词多义（polysemy）会带来什么问题，有何应对措施？
15. 简述LSA(Latent Semantic Analysis)的基本原理。
16. 推荐系统中效用矩阵（user-item matrix）UV分解后得到的U矩阵和V矩阵的物理意义。
17. 什么是推荐系统中的协同过滤（Collaborative Filtering）？
18. 推荐系统中，为何不直接把user-item矩阵空白位置填0，然后用SVD分解获取隐藏在user-item矩阵中的users和items的关联信息？

# [期中考试后，期末考试的主要内容]

1. 正确率和召回率的计算。
2. 无序检索结果的评价和有序检索结果评价有何区别？
3. 一般的P-R曲线为何出现锯齿的形状？
4. 请解释MAP指标 $MAP\left(Q\right)=\frac{1}{\left|Q\right|}\sum\_{j=1}^{\left|Q\right|}\frac{1}{m\_{j}}\sum\_{k=1}^{m\_{j}}Precision(R\_{jk})$ 的含义。
5. 简述Rocchio 相关反馈算法的基本思想。
6. 同义词词典自动构建有哪些可能的方法？
7. BIM（binary independence model）中估计排序函数的时候“平滑”本质是什么？
8. 1-gram模型需要估计的参数有多少个？N-gram模型需要估计的参数个数又是多少？
9. $请解释P\left(M\_{d}\right)=K\_{q}\prod\_{t\in q}^{}P(t|M\_{d})^{tf\_{t,d}},K\_{q}=L\_{d}!/(tf\_{t1,d}!tf\_{t2,d}!\cdots tf\_{tM,d}!)$式对应的文档生成模型的假设条件。
10. 比较两个文档生成模型可以用什么样的度量？
11. 翻译模型：$P\left(M\_{d}\right)=\prod\_{t\in q}^{}\sum\_{v\in V}^{}P(t|M\_{d})T(t|v)$，公式的含义？
12. 硬分类和软分类的区别？
13. 朴素贝叶斯分类中在计算MAP的时候引入了对数，意义何在？
14. NB分类中，Multinomial model和Bernoulli model有何不同？
15. NB分类中，Multinomial model和Bernoulli model各自需要估计哪些参数？
16. 举例说明何为过学习（overfitting）?
17. 文本分类中特征选择起什么作用？
18. 文本分类的评价与IR系统无序检索评价有何差异？
19. 举例说明什么是偏差-方差折衷准则？
20. Rocchio文本分类方法应用的局限性是什么？
21. 列举课程中学习过的线性文本分类器、非线性文本分类器。
22. 说明线性分类器如何克服或缓解噪音文档的问题。
23. 何为SVM中的核技巧（Kernel Trick）？
24. 什么是概率图模型？
25. 概率图模型表示的要素有哪些？【结构：G(V,E)；参数：CPTs】
26. 概率图有关研究的主要内容是什么？
27. 什么是文档的主题模型，常见主题模型有哪些？
28. pLSA文档模型有哪些缺点？
29. LDA文档模型的基本原理。
30. CNN的大致结构和应用现状。
31. RNN的大致结构和应用现状。
32. K-mean聚类的时候如何确定K？
33. 聚类结果的评价和分类方法的评价有何不同？
34. 何为基于模型的聚类？
35. 采用采用单连接和全连接方法进行聚类各有何优缺点？
36. 列举在课程中学习过的非监督聚类方法。
37. 请举例说明哈希函数在IR系统中的应用。
38. 简述哈希表的优缺点。
39. 简述布隆过滤器(Bloom Filter)的基本原理，调研除电子邮件黑名单外的其他应用。
40. 简述基于Shingles的近似重复文档检测的原理。
41. 如何在计算机中存储Web图？
42. Pagerank算法原理及计算过程。
43. HITS算法原理及计算过程。
44. Web检索系统的整体框架和各模块的主要技术。
45. 举例阐明多媒体信息分析与检索中的语义鸿沟(Sematic Gap)问题。
46. 图像分类算法不同发展阶段的思路有何差异？
47. 举例说明在图像局部特征提取时如何利用人的视觉感知特性。
48. 应用Sparse Coding时，Coding和Pooling各起什么作用？
49. 举例阐述深度学习（Deep Learning）方法在多媒体信息分析与检索中的应用。
50. 请阐述个人关于词向量（Word Embedding）应用的想法。