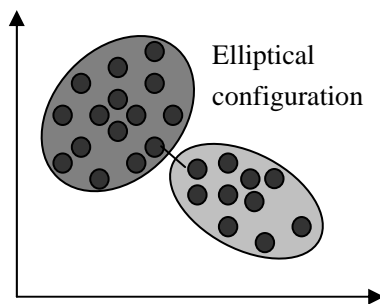


北京大学研究生课程期末考题，2003 年秋

课程名称：先进技术专题（网络信息体系结构）；时间：2003 年 1 月 7 日；地点：一教 105

1. 简述大规模通用 Web 搜索引擎的工作原理和设计其三大基本模块（收集、整理、服务）要考虑的主要问题（10 分）。
2. 一般地，描述词汇与文档关系的 term-doc 矩阵的元素可以是 term 的频率或布尔值，简要地解释他们的 SVD 效果是不同的（3 分）。
3. 发挥想象力，设计一个聚类策略解决 single linkage 难于处理的情形（3 分）。



4. 简要解释下列概念（8 分）
 - (1) “共有词汇概念” (shared bag of words)
 - (2) Vector Space Model
 - (3) TF*IDF
 - (4) Information Extraction
5. 试详细阐述 kNN 文本分类方法（6 分）
6. 按照下述描述和要求完成相关工作（10 分）
 - (1) 给定文档语料（网络实验室前 5 篇技术报告的标题）
 - d1. LilyTask 在共享内存下的设计和实现
 - d2. 动态 p2p 网络中的对象定位问题研究
 - d3. Sesame_backend 的设计与实现
 - d4. 燕川流媒体演示系统的设计
 - d5. 良师益友家教系统的设计与客户端的实现利用语言所在线切分软件我们分别得到：
LilyTask 在共享内存下的设计和实现
动态 p2p 网络中的对象定位问题研究
Sesame_backend 的设计与实现
燕川流媒体演示系统的设计

良师益友 家教 系统 的 设计 与 客 户 端 的 实 现

(2) 你的任务是设计一个针对这些文档的信息检索系统。具体要求是：

- A. 给出系统的有效词汇集和（说明取舍原因）
- B. 写出系统的 **term-doc** 矩阵（用布尔元素）
- C. 画出系统的倒排文件示意图
- D. 按照 **VSM**，给出针对查询“系统的设计”的结果输出