上机报告，作业要求

徐悦甦(Yueshen Xu)

ysxu@xidian.edu.cn / xuyueshen@163.com

知识与数据工程研究中心

西安电子科技大学
XIDIAN UNIVERSITY

两次实验上机说明

西安电子科技大学
XIDIAN UNIVERSITY

- **内容：实现小型信息检索引擎**
  - **两部分（爬虫/数据 + 检索/展示）**
    - 第一次（爬虫/数据）：针对静态网页与普通动态网页
      - 使用Java/Python等语言实现对于网页的Http请求
      - 使用beautifulsoup等工具实现对HTML标签内容的提取
    - 第二次（检索/展示）：使用Lucene
      - 使用Lucene提供的接口/API实现从索引建立到关键字搜索的任务
      - 如果有时间，可以使用简单的HTML，将检索出的结果以网页的形式展示出来
    - 提交报告
      - 提交代码 + 实验报告（电子版 + 纸质版 ➔ 必须）

两次实验上机说明

西安电子科技大学 XIDIAN UNIVERSITY

- **要求**
  - 出勤；姓名/学号
  - 单独完成，没有组队
  - 电子版报告提交至： xdseirclass@163.com
    - 命名：学号+姓名+信息检索上机报告(一/二).doc/pdf （命名专业）
    - 发之前检查，不要损坏
    - 纸质版直接交给我
    - 每一次上机一份，共两份，每次时间充足，两周
- **时间与地点**
  - 4月12日晚（结束）；4月26日晚；7:00 ~ 9:30
  - G楼314

随堂作业

西安电子科技大学 XIDIAN UNIVERSITY

聚类方法实现

- **作业：针对**K-Means或K-Medoid方法，完成报告，并在报告中附代码（一个方法即可，不再要求跑代码）
  - 建议一种报告格式(可根据自己的想法调整)：

    *题目 → 姓名/学号 → K-Means(或K-Medoid)方法步骤 → 学习代码过程中遇到的问题(或对代码实现的理解) → 心得与总结 → 附代码
- 不再要求一定跑代码，但要附代码，做出对代码的分析；仍然鼓励跑代码，可参照已有代码
- 时间充足，五周（5.25之前），鼓励提前交
  - 邮箱：xdseirclass@163.com ；报告+附代码
  - 命名：学号+姓名+信息检索课堂作业.doc/pdf

课程大作业

西安电子科技大学 XIDIAN UNIVERSITY

网页分类方法概述

- **作业：**完成针对**网页**分类方法的综述性报告（.doc/.pdf），并最好在报名中写明你对**网页**分类方法的思考
  - 注意，不是泛泛的分类方法，而是针对**网页**的分类方法
  - 资料来源：Google，百度，搜狗
  - **报告要正式；**一种报告内容安排的建议：

    * 题目→ 姓名/学号 → （ 摘要 ）→ 概述 → 选取两到三种方法（一种太少了）详述 → 自己见解与思考 → 总结 → 参考文献/资源

    *参考文献/资源：可论文，可PPT，可网页（博客），可教材

    *<可根据自己的想法进行调整>

课程大作业

西安电子科技大学 XIDIAN UNIVERSITY

- **数据集与代码**：大作业不再要求实现，只要求报告
- **不要抄袭**，没必要抄袭
- 交pdf或word均可，建议大家在里面配图（不一定要自己画，但要给出来源）
- 主要看内容，格式也不要太乱
- 从现在开始就可以做，课程结束后一周以内deadline（越早越好）
  - 邮箱：xdseirclass@163.com；报告+代码
  - 命名：学号+姓名+信息检索大作业.doc/pdf

课件与作业说明地址


徐悦甦的留言板
西安电子科技大学
XIDIAN UNIVERSITY
徐悦甦的主页：
个人信息
科学研究
论文成果
荣誉奖励
科研团队
课程教学
课程教学
1. 承担课程与竞赛：
主讲：《信息检索》
《信息检索》课件：
§ 4.1 文本聚类：/ysxu/files/20170427_104024.pdf
§ 4.2 文本分类： /ysxu/files/20170506_135325.pdf
§ 5 推荐系统：/ysxu/files/20170516_183350.pdf
§ 6 语义网： /ysxu/files/20170514_153659.pdf

各位同学，请按时上交报告与作业