

## 1 布尔检索

- 1.1 信息检索概念
- 1.2 构建倒排索引 (倒排索引, 词项词典, 倒排记录表, 构建倒排记录流程)
- 1.3 布尔检索模型 (概念、词项关联矩阵, 布尔查询表示)
- 1.4 课后作业 1-1, 1-2, 1-3

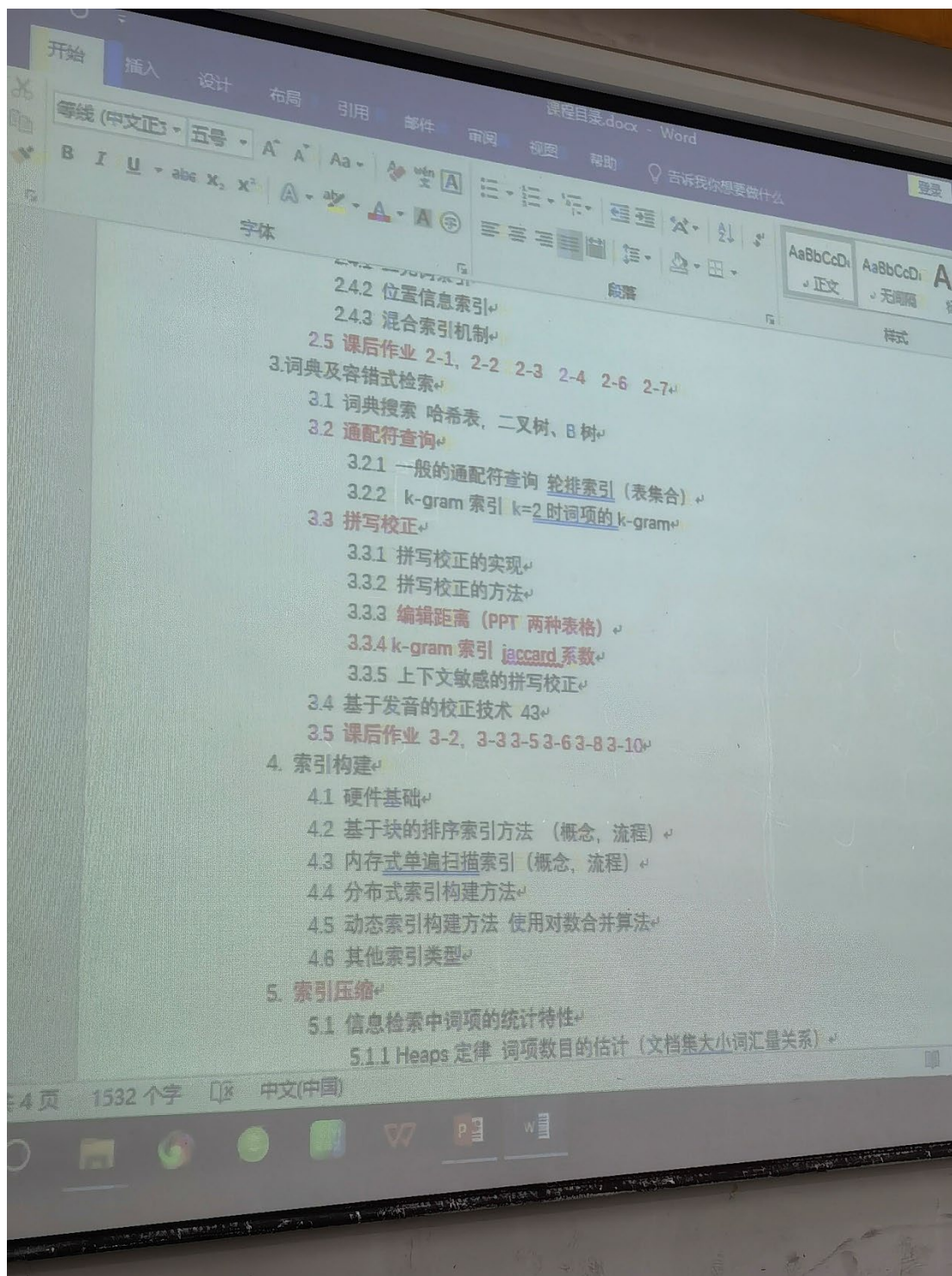
## 2 词项词典及倒排记录表

- 2.1 文档分析
  - 2.1.1 字符序列
  - 2.1.2 文档单位
- 2.2 词项集合确定
  - 2.2.1 词条化 (概念, 词条类、词项)
  - 2.2.2 去除停用词 (什么是停用词, 生成停用词的方法)
  - 2.2.3 词项归一化 (什么是此项归一化, 方法, 等价类)
  - 2.2.4 词干还原和词形归并 (概念 porter 最长匹配算法)
- 2.3 跳表合并算法 (上机练习代码思想)
- 2.4 含位置信息的倒排记录表
  - 2.4.1 二元词索引
  - 2.4.2 位置信息索引
  - 2.4.3 混合索引机制

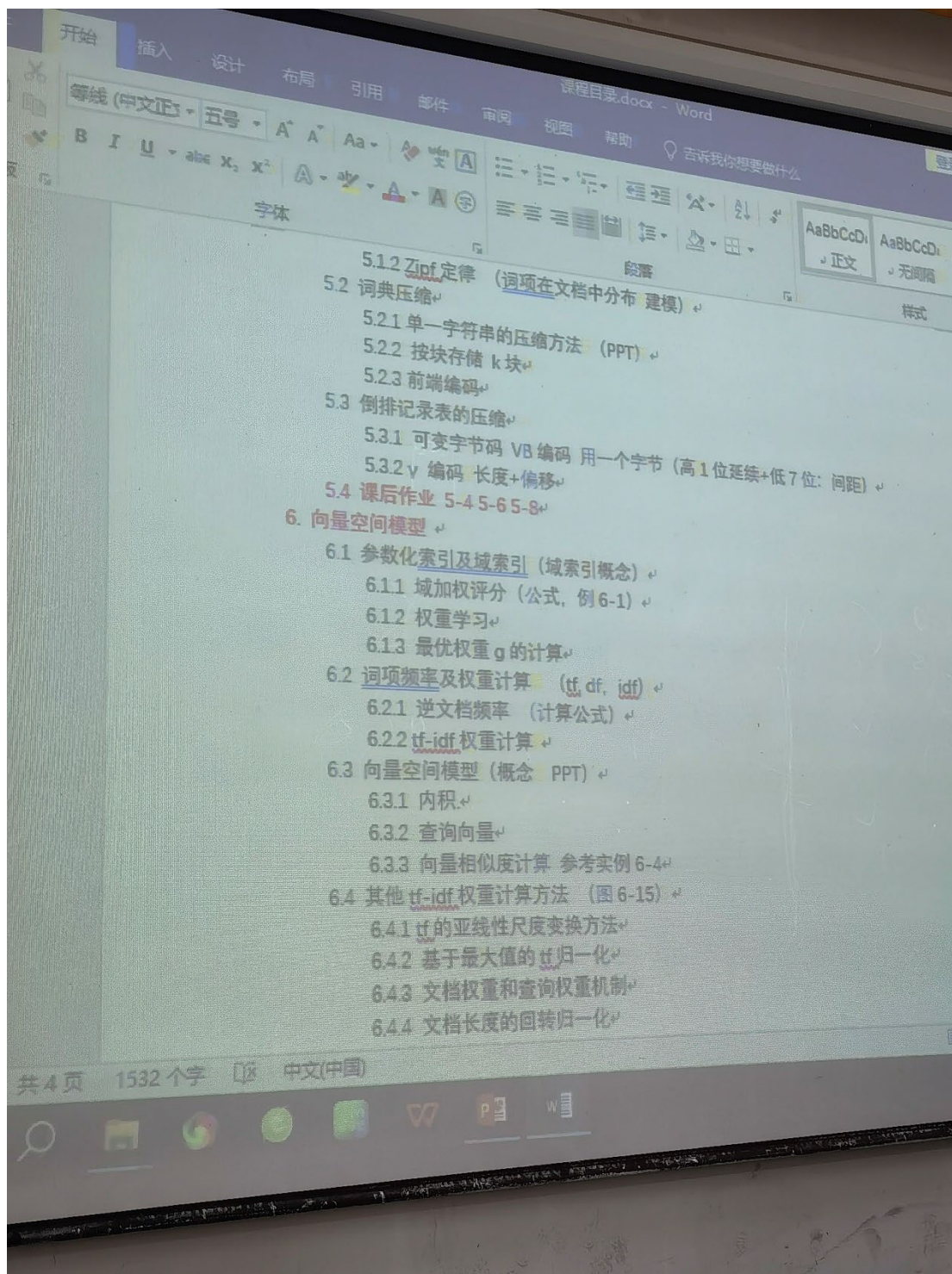
## 2.5 课后作业 2-1, 2-2 2-3 2-4 2-6 2-7

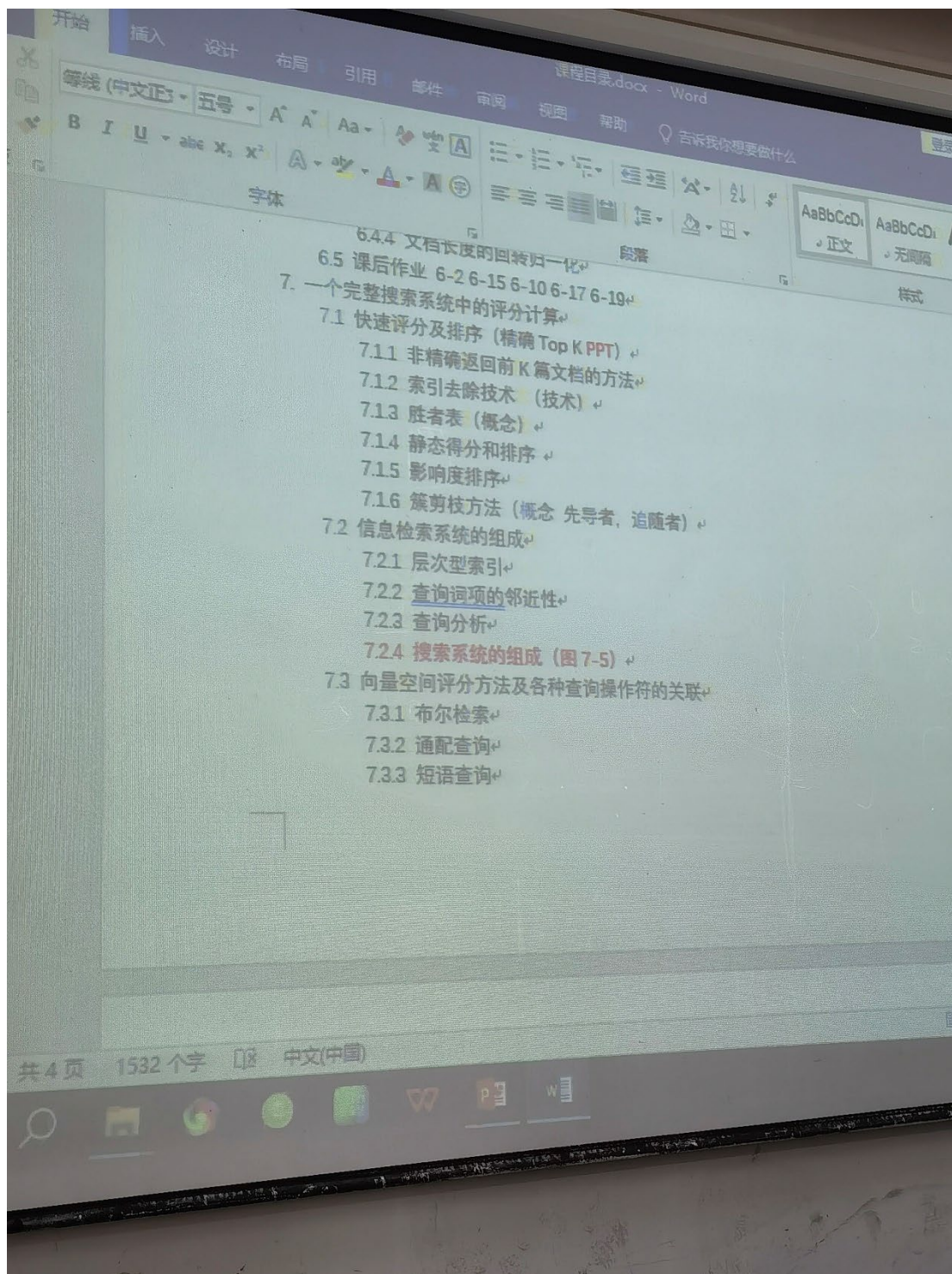
## 3 词典及容错式检索

- 3.1 词典搜索 哈希表, 二叉树、B 树
- 3.2 通配符查询
  - 3.2.1 一般的通配符查询 轮排索引 (表集合)

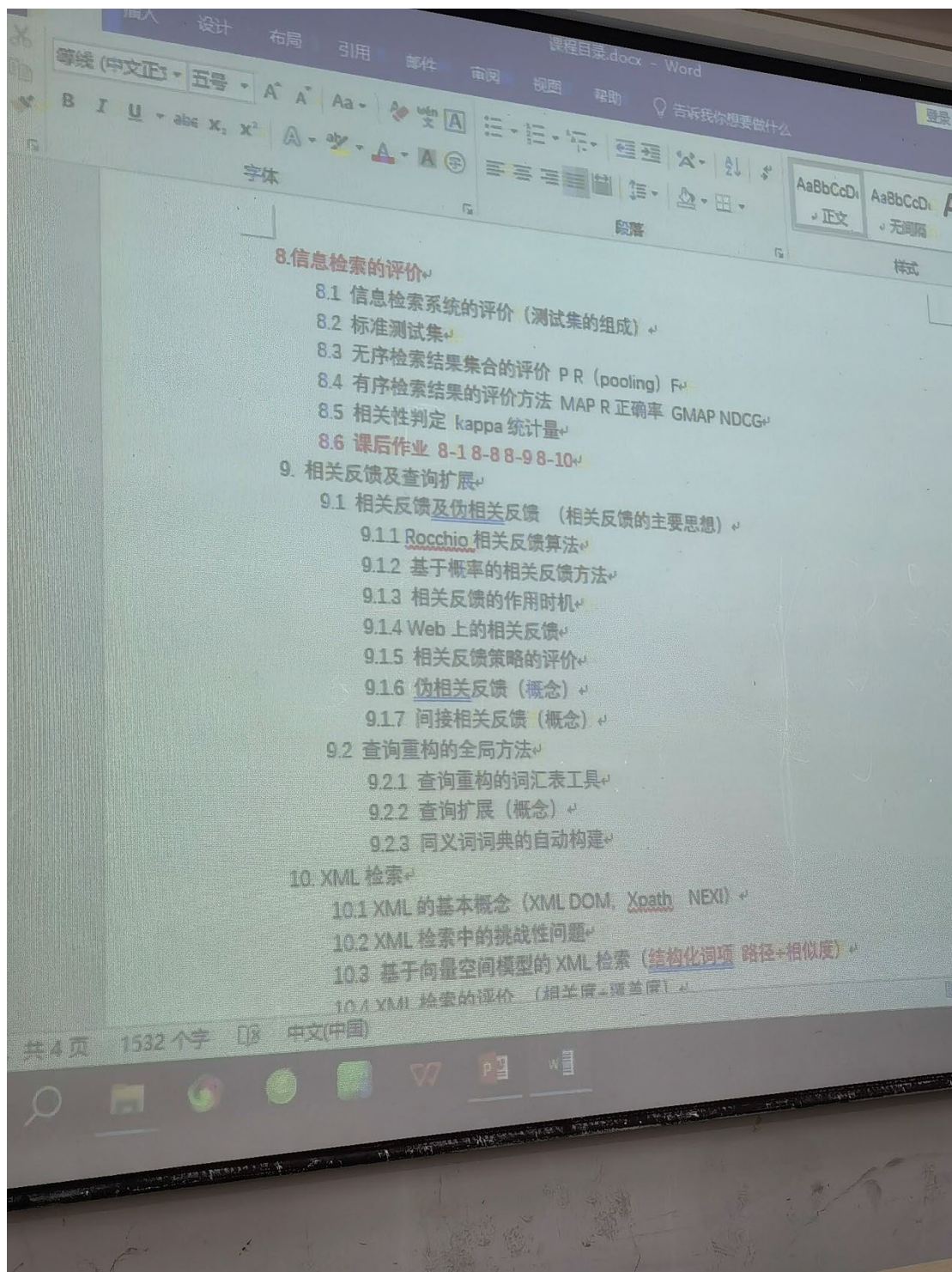












开始 (中文正)

五号

字体

段落

AaBbCcDd  
正文

AaBbCcDd  
无间隔

样式

## 8. 信息检索的评价

8.1 信息检索系统的评价 (测试集的组成)

8.2 标准测试集

8.3 无序检索结果集合的评价 PR (pooling) F<sub>1</sub>

8.4 有序检索结果的评价方法 MAP R 正确率 GMAP NDCG

8.5 相关性判定 kappa 统计量

8.6 课后作业 8-1 8-8 8-9 8-10

## 9. 相关反馈及查询扩展

9.1 相关反馈及伪相关反馈 (相关反馈的主要思想)

9.1.1 Rocchio 相关反馈算法

9.1.2 基于概率的相关反馈方法

9.1.3 相关反馈的作用时机

9.1.4 Web 上的相关反馈

9.1.5 相关反馈策略的评价

9.1.6 伪相关反馈 (概念)

9.1.7 间接相关反馈 (概念)

9.2 查询重构的全局方法

9.2.1 查询重构的词汇表工具

9.2.2 查询扩展 (概念)

9.2.3 同义词词典的自动构建

## 10. XML 检索

10.1 XML 的基本概念 (XML DOM, XPath, XSLT)

10.2 XML 检索中的挑战性问题

10.3 基于向量空间模型的 XML 检索 (结构化词项 路径+相似度)

10.4 XML 检索的评价 (相关性+覆盖率)

