《互联网数据挖掘》本科生课程

期末复习

万小军

北京大学王选计算机研究所

https://wanxiaojun.github.io/

2019年12月17日

考试安排

- **闭卷考试**：侧重基础内容、注重发散思维
- 考试时间
  - 2019年12月31日周二上午(待确认)
- 考试地点
  - 待确认

考试题目类型

- 计算题
  - 写出主要计算步骤与最终结果；
- 简答题
  - 简要解释基本概念、回答基本问题；
- 论述题
  - 结合自己的互联网体验与思考进行论述；

考试复习知识点

- 词袋模型
- 文档余弦相似度计算
- 倒排索引构建与优点
- 布尔检索模型及其优缺点
- Web搜索架构
- PageRank算法
- HITS算法
- 信息检索评价指标MAP的计算
- 关联规则挖掘过程与Apriori算法；
- 朴素贝叶斯分类算法
- K近邻分类算法
- 分类与回归的联系与区别
- K均值聚类算法
- 凝聚式聚类算法
- 半监督聚类之COP K-means算法
- 自然语言处理领域的歧义现象
- 正向最大匹配分词与逆向最大匹配分词
- 无向图度数中心性、中介中心性与亲近中心性的计算(未规范化与规范化)
- 基于图排序(PageRank)的文档摘要方法
- 基于句子分类的文档摘要方法
- 基于PMI的情感词汇获取方法及文本情感分类方法
- 观点抽取的目的和主要步骤
- 基于用户/物品的协同推荐算法
- 基于矩阵分解的协同推荐算法
- 智能问答系统架构

考试复习提纲

- 思考：
  - 对当前主流中文互联网搜索系统有所了解、比较与思考。
  - 思考社交媒体对搜索技术的挑战以及解决办法。
  - 思考移动搜索的挑战与关键技术。
  - 面向特定领域设计一个智能问答系统。
  - 面向特定领域的知识库构建及其应用。

课程答疑

- 电子邮件
- 微信群
- 办公室
  - 北大东南门对面计算机研究所大楼2层