搜索引擎技术基础课程实验

2023/04/11

作业三：大作业（三选一，60分）

- 搜索引擎原型系统设计与实现（1-2人一组）
  - I）司法搜索引擎，II）知识搜索引擎，III）图片搜索引擎：（三选一）
  - 简介：实现一个中文/英文搜索引擎的原型系统（demo版本）。
  - 特点：需要实现搜索引擎中的常用算法功能（检索、查询推荐等）和用户界面，允许使用开源框架。
  - 评分：根据功能实现程度和主观使用体验打分（详见后文）。评分时会考虑三个选题的难度差异。
  - QA：维护了课程作业常见QA，位于https://cloud.tsinghua.edu.cn/d/e7148f4120e44bdd9aa7/，会同步大家私戳和群里发的问题。

I）司法搜索引擎

- 功能实现：
  - 核心功能（45分）：用户界面（20分），关键词检索（15分），类似案例检索（10分，要求能上传案例文件检索）。
  - 其他（15分）：精细化检索，高亮，标签显示，查询推荐与扩展，相似案例（文本相似/同法官/同法律等）推荐，等等，根据实现难度给分。
  - 主观体验（20分）：分展示、设计、技术、团队协作、QA共5个部分。
  - 实验报告（20分）：统一要求+测试关键词、测试案例的评测结果（参考第一次作业，要求有一个测试关键词和一个测试案例）。

- 可用数据集：https://cloud.tsinghua.edu.cn/d/e7148f4120e44bdd9aa7/目录下的Legal_data.zip
- 特别声明：本数据集为私有数据集，仅允许用于搜索引擎课程，请勿以任何方式挪用、公开数据集的部分或全部内容。

II）知识搜索引擎

- 功能实现：
  - 核心功能（45-55分）：用户界面（20分），实体检索（10分），知识推理（15-25分）。
  - 其他（5-15分）：结构化信息呈现，实体跳转，语音输入，查询纠错，基于Pagerank对检索结果排序等等，根据实现难度给分。
  - 主观体验（20分）：分展示、设计、技术、团队协作、QA共5个部分。
  - 实验报告（20分）：统一要求+测试实体检索、知识推理的评测结果，需要给出效果较好和效果较差的例子，并分析原因。

- 可用数据集：https://cloud.tsinghua.edu.cn/d/e7148f4120e44bdd9aa7/目录下的Xlore.zip
- 对应的论文：https://www.aminer.cn/pub/53e9a3cdb7602d9702ce01ab/xlore-a-large-scale-english-chinese-bilingual-knowledge-graph

III）图片搜索引擎

- 功能实现：
  - 核心功能（35分）：用户界面（25分），图片检索（10分）。
  - 其他（25分）：尺寸/颜色筛选（10分），以图搜图，查询纠错，跨语言搜索，跨数据集等等，根据实现难度给分。
  - 主观体验（20分）：分展示、设计、技术、团队协作、QA共5个部分。
  - 实验报告（20分）：统一要求+测试图片搜索功能。

- 可用数据集：
  - https://github.com/cvdfoundation/open-images-dataset;谷歌开放图片数据，共561G，可以选择其中任何一个packed file进行下载（约30G）；
  - https://github.com/fpingham/google-images-dataset；通过爬取谷歌图片数据创建个人的谷歌图片数据集，可以指定个人的查询list；
  - https://www.kaggle.com/competitions/landmark-retrieval-2019/overview：以图搜图数据集，选择该数据集的话只需要实现以图搜图作为图片检索功能

统一要求

- 代码编程语言：
  - 编程语言不限，后端可以使用Python、Java、PHP等，前端可以用JavaScript或View.js等框架，代码要加必要的注释，函数功能划分好；
  - 搜索功能的实现允许使用开源框架，如elastic,whoosh, solr, lucene, pyterrier, pylucene, pyserini等，也允许自己实现经典概率模型（**BM25**），向量空间模型，统计语言模型或更高级的检索算法
- 提交一份书面报告，格式为pdf，字数1000-3000字。内容必须包括：
  - 问题描述；
  - 实现模块；
  - 关键功能；
  - 测试结果和样例分析（详见三个任务要求）；
  - 使用了哪些开源资料；
  - 参考了他人提供的哪些代码（可能会查重）；