Web 信息处理与应用：实验 1
信息检索部分
实验于 2019 年 10 月 30 日开始，为期四周，两人一组进行分组实验。
请于 2019 年 12 月 4 日前将实验报告发送至课程邮箱：ustcweb2019@163.com

**实验总体要求：**
给定若干数量的文档和查询条件，请为每个查询条件返回前 20 条最相关的文档。
其中，每条文档与查询条件的相关性评级取值为{0, 1, 2, 3}，3 为最相关，0 为不相关。
返回结果将通过 F1 值与 NDCG@20 进行评价。

**必备考核内容：文档排序（无监督与有监督均可，方法自选）**
可选考核内容：建立索引（是否进行分词自定，索引方法自选）
主要评价搜索结果，对搜索界面（前端）不做要求，如有兴趣可进行尝试。
严禁抄袭代码，一经查实本实验作 0 分处理。

**数据文件格式说明：**
**训练数据**包含“文档数据集.csv”和“查询-文档相关性标签.csv”两个文件。
其中：
“文档数据集.csv”的第一行为格式说明，此后每行对应一个文档。
- 每行的内容包括“文档 ID,文档 URL,文档标题,文档内容”，以“,”进行分隔。

| doc_id | doc_url | doc_title | content |
|---|---|---|---|
| d7831761 | http://wenku.baidu.com/view/1aed6b325f0e7cd185... | 公司员工考勤表格式范本_百度文库 | 百度文库;实用文档;表格/模板;表格类模板暂无评价 \| 0人阅读 \| 0次下载 \|举报文档公... |
| d8389921 | http://www.doc88.com/p-6945939027688.html | 2013福师《中国古代小说研究》在线作业一答案 - 道客巴巴 | 浏览次数:108 内容提示: 福师《中国古代小说研究》 在线作业一满分答案 试卷总分: 10... |
| d418777 | http://www.ht88.com/downinfo/571956.html | 过零丁洋ppt26 人教版 | 网站首页;下载首页;初中课件;八年级下册课件资源类别:人教版 / 初中课件 / 八年级下册... |
| d863006 | http://www.jianli-moban.com/n2955c8.aspx | 人事经理英文简历模板 | 网站首页;英文简历;人事经理英文简历模板□[日期:2013-08-04] □□来源: 作者:... |
| d9136077 | http://www.docin.com/p-50727035.html&endPro=true | 香港和澳门的回归教学叙事 - 豆丁网 | 中学教育;初中教育《香港和澳门的回归》是八年级下册第四单元民族团结与祖国统一中,关于香港和澳... |
| ... | ... | ... | ... |

“查询-文档相关性标签.csv”的第一行为格式说明，此后每行对应一个查询-文档对。
- 每行的内容包括“查询,文档标题,文档 ID,相关性标签”，以“,”进行分隔。
- 相关性标签取值为{0, 1, 2, 3}，3 为最相关，0 为不相关。
- 每个查询对应的文档数量不等。

| query | doc_title | doc_id | label |
|---|---|---|---|
| 药品养护汇总分析 | 10月份药品养护汇总分析 - 豆丁网 | d6893042 | 3 |
| 药品养护汇总分析 | 药品养护质量信息汇总分析报告_文档资料库 | d5709647 | 3 |
| 药品养护汇总分析 | 药品养护质量信息汇总分析报告_完整版 - 道客巴巴 | d919596 | 3 |
| 药品养护汇总分析 | 药品养护质量信息汇总分析报告_百度文库 | d6893040 | 3 |
| 药品养护汇总分析 | 月份药品养护汇总分析_百度文库 | d919590 | 3 |

训练数据获取方式；
百度网盘：https://pan.baidu.com/s/1TzP4OSa4AorenE5vp-WQGA 提取码: 6tkk
睿客（校内网盘）：http://rec.ustc.edu.cn/share/e0a74d50-fa4f-11e9-a970-4396d9d7eb68

**测试数据**共包含 13213 篇文档和 470 个 query。其中，test_querys.csv 文件为查询条件列表，包括 Query 与 Query ID 两项内容。

| query | query_id |
|---|---|
| 年均增长率怎么算 | q170910 |
| 他得的红圈圈最多课件 | q28400 |
| 撒哈拉以南的非洲ppt | q197250 |
| 哈兰研究勋章 | q115920 |
| 红头文件格式模板下载 | q410940 |
| ... | ... |
| 将相和ppt | q110759 |
| 创业构思范文 | q70519 |
| 3000多年前+就有了分数记号 | q517399 |
| 年级下册数学月考试卷及答案 | q257519 |
| 上海打工一般工资多少 | q517159 |

test_docs.csv 文件与训练数据中“文档数据集.csv”文件格式一致，包含了每篇文档的编号，链接，标题以及内容。

测试数据获取方式；
百度网盘：链接: https://pan.baidu.com/s/11c8T28zv4JPVtquK7qVyFA 提取码: mjap

**提交数据格式说明：**
最终提交文档如下图所示，对于 test_querys 文件中的每个 query，返回最相关的 20 篇文档编号（返回编号即可，无需返回内容）。最终提交形式如 submission.csv 所示，其中 query_id 为查询编号，doc_id 为对应的相关文档编号。每个 query_id 对应于 20 个 doc_id。

| | query_id | doc_id |
|---|---|---|
| 0 | q350667 | d8888888 |
| 1 | q350667 | d8888888 |
| 2 | q350667 | d8888888 |
| 3 | q350667 | d8888888 |
| 4 | q350667 | d8888888 |
| ... | ... | ... |
| 9395 | q145053 | d8888888 |
| 9396 | q145053 | d8888888 |
| 9397 | q145053 | d8888888 |
| 9398 | q145053 | d8888888 |
| 9399 | q145053 | d8888888 |

9400 rows × 2 columns

最后提交的 submission.csv 应为 9400（即 470*20）行 2 列的 csv 文件，其中第一列为 query_id，第二列为 doc_id。提交前请确认文件格式正确，每个 query_id 对应 20 个 doc_id，不要缺少或增加。否则可能将影响你的结果评估。

在最终结果提交前，将安排若干次测试提交并反馈结果（只反馈指标，不反馈正确答案），第一次提交将于 11 月 12 日前进行，提交方式将于课程主页、课程 QQ 群等渠道公布。

本说明文档将根据实验进行不断更新。更新时将通过课程主页、课程 QQ 群及课上等渠道进行通知，敬请关注。