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 11 月 27 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

测试数据将于 2019 年 11 月 5 日（第二周）前发布。
测试数据格式与提交测评方式将同时公布。

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。