大作业组织形式

- 以小组为单位， 10 月 13 日前将分组情况发至助教邮箱 ranyanhuaemail@163.com，并抄送教师 benhe@ucas.ac.cn

  邮件标题： IR 大作业分组 _[组长姓名]

  邮件内容：

  组长 Email 、手机号

  小组成员姓名、学号

- 每组不超过 5 人

  若自己一个人一组，也请发邮件告知

大作业组队群

2017-秋-现代信息检索

该二维码7天内(10月1日前)有效，重新进入将更新

大作业内容

- 第一部分：编写索引器，构建Shakespear-Merchant语料索引 （15分）
- 第二部分：在TREC CDS 2014&2015数据上进行检索竞赛（20分）
- 第三部分：编写界面程序（无具体要求，但应具备最基本的功能满足任务一、二的结果检查。建议bash纯命令行界面）
- 实验报告（15分）
- 课程结束前会组织一两次课堂报告，各组自愿报名，有加分
  - 汇报已有进展
  - 将采取什么方法进一步提高结果

第一部分：编写索引器

- 内容：编写索引器，构建给定语料的词典与倒排索引
- Shakespear-Merchant 语料
  - 《威尼斯商人》剧本，语料规模极小，用于测试索引器
  - 下载地址：http://gucasir.org/ModernIR/shakespeare-merchant.trec.tgz
  - 解压命令： tar zxvf shakespear-merchant.trec.tgz

语料格式

- 按剧本场景分为 22 个文档，每个文档有如下格式：

  <DOC>

  <DOCNO>StringID</DOCID>

  <title>The Title</title>

  Content goes here

  <speaker> Name </speaker> Speech

  </DOC>

- 其中， DOC 标识文档起止位置， DOCNO 为文档字符串 ID ， title 为标题。

功能要求

- 基本要求：构建词典和倒排索引
  - 实现 Single-pass In-memory Indexing
  - 实现倒排索引的 Gamma 或 VB 编码压缩 / 解压
  - 实现词典的单一字符串形式压缩 / 解压，任意数据结构（如哈希表、 B 树等）
  - 实现关键字的查找，命令行中 Print 给定关键字的倒排记录表
  - 给出以下语料统计量：词项数量，文档数量，词条数量，文档平均长度（词条数量）
  - 编程语言不限，但必须提交代码和说明文档
- 对停用词去除、词干还原等无要求，但应实现最基本的词条化功能
  - 例如：将所有非字母和非数字字符转换为空格，不考虑纯数字词项

第二部分：检索竞赛

- 采用类似 TREC 竞赛的形式
  - 以小组形式在给定数据上进行实验
  - 鼓励创新思维
    - 评分：综合考虑实验结果和使用的新方法、提出的新思路
- Collection: TREC CDS 2014 & 2015
  - 一个医疗文献数据集，是 TREC Clinical Decision Support(CDS) 2014&2015 任务使用的文档集。
  - 可从以下 CDS 任务主页下载，包括查询 (topics) 和相关性标记 (qrels)

    - 数据集 : http://www.trec-cds.org/2014.html

    - 查询 (topics) 和相关判定 (qrels): http://www.trec-cds.org/2016.html

使用的系统

- 不排斥使用开源工具
- 也可以基于第一部分的索引器进一步开发实现检索功能，并用于第二部分的大作业
  - 鉴于此项工作的难度，有加分

查询：只使用 summary 域

<topic number="1" type="diagnosis">

<description>

A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.

</description>

<summary>

58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.

</summary>

</topic>

竞赛规则参考资料

Simpson M S, Voorhees E M, Hersh W. Overview of the trec 2014 clinical decision support track[R]. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD, 2014.

https://pdfs.semanticscholar.org/fcf8/1b7641c0cd7be089051018a53fabfa685da0.pdf

Roberts K, Simpson M S, Voorhees E M, et al. Overview of the TREC 2015 Clinical Decision Support Track[C]//TREC. 2015.

http://trec.nist.gov/pubs/trec24/papers/Overview-CL.pdf

提交结果文件格式

标准 trec_eval 格式，具体如下：

< 查询 ID> Q0 < 文档 ID> < 文档排序 > < 文档评分 > < 系统 ID>

例如：

501 Q0 WTX046-B13-199 1 16.827150770678543 InL2c7.0

其中 Q0 没有具体意义，仅起到分隔作用，方便结果文件的脚本处理。

检索竞赛评分规则

- 检索效果（ 20分）：
  - 训练：
    - TREC CDS 2014 任务的30个查询用于训练所有模型参数
  - 测试：
    - TREC CDS 2015（ A ）任务的30个查询用于最终评测，只提交这一部分查询上的结果
    - 不得在测试查询上进行训练！违者视为作弊
  - 评价指标： infNDCG
    - 评价指标计算工具下载： www-nlpir.nist.gov/projects/t01v/trecvid.tools/sample_eval/sample_eval.pl
  - 可以用开源工具，使用自己实现的系统有加分
- 实验报告（ 15分）：
  - 对索引器代码和运行方法进行说明
  - 详细描述实验中采用的技术
  - 对于提出的新方法、新技术有得分奖励
    - 新检索模型、新相关反馈方法等，或对现有模型、方法的提高和修正

可能需要采用的技术

- 检索模型
- 相关反馈 / 查询扩展
- 词嵌入
- 深度神经网络
- 其它方法

实验报告

- 索引器系统结构、实现方案、主要代码类以及运行方法的说明
- 使用了什么技术？基于什么原理？分别给出公式
- 描述详细实验步骤
  - 训练
  - 测试
  - 要求能看出没有在测试查询集上进行训练
- 汇报最终在 TREC CDS 2015 (A) 上的 30 个测试查询上获得的 infNDCG

结果提交

- 将所有材料做成一个压缩包， Email 至 benhe@ucas.ac.cn
  - 第一、二部分的源代码
  - 第一、二部分的可执行程序
  - 符合 trec_eval 格式的结果文件
  - 实验报告
  - 但不提交中间文件，避免附件过大
  - 提交时限： 1 月 1 日之前

提交材料的要求

- 代码清晰明确
- 建议使用 Linux ，推荐 Ubuntu 环境
- 实验报告中应明确说明如何运行程序
  - 要求“一键式”运行得到报告中的结果
    - 报告中明确给出需运行的脚本命令
    - 运行一个脚本命令（如 bash 或 python ），完成建立索引、模型训练、模型测试这三个步骤
    - 说明最终产生的 TREC 结果文件存放的位置（要求和打包提交的结果文件一致）

如需安装额外的软件包，应明确给出安装命令 (例如 sudo apt-get install xxxx ， conda install pytorch torchvision -c soumith)