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

大作业内容

- 第一部分：在 TREC Precision Medicine (PM) 2017 数据上进行检索竞赛（20 分）
- 第二部分：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的结果检查。建议 bash 纯命令行界面）
- 实验报告（10 分）

第一部分：检索竞赛

- PM 任务介绍
  - 主页： http://www.trec-cds.org/2017.html
  - 要求：匹配病人信息，检索相关文档，主要有两个子任务。**大作业只做第二个任务，即 Clinical trials**
    - Scientific Abstract：是医疗文献的摘要部分， 目标是为医生提供学术研究上相关的治疗信息 （不需参与此任务）
    - **Clinical trials：是病人病历数据库， 目标是为医生提供与此病人相关的电子病历**
- 数据集
  - 是 ClinicalTrials.gov 上的一个 snapshot， 包含了 24 万多的电子病历，可从 PM 任务主页下载

PM 查询

- 简要的病人信息， 包括疾病， 基因， 年龄，性别以及其它信息， e.g.

<topic number="1">
<disease>Acute lymphoblastic leukemia</disease>
<gene>ABL1, PTPN11</gene>
<demographic>12-year-old male</demographic>
<other>No relevant factors</other>
</topic>

- 通常做法是将disease 字段作为查询，其它字段作为辅助信息
  - 如何利用辅助信息?

提交结果文件格式

标准 TREC 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：

1 Q0 NCT02571829 0 23.3981874263057 I_LIKE_IR

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。

评价指标

- 评价指标
  - P@5, P@10, P@15
  - Ground truth 在 IR 里面通常是一个叫做 qrels 的文件

    https://trec.nist.gov/data/precmed2017.html
  - 如何计算
    - 用 trec_eval 脚本计算
      - https://trec.nist.gov/trec_eval/
      - 运行示范： ./trec_eval qrels res
  - 训练时候可以采用 P@10 作为主要观测指标来选择模型

一些注意点

- 明确建立的索引的域
  - Brief (Official?) Title， Description， MeSH Terms， Inclusion (Exclusion) Criteria, …
- 话题的处理
  - 是否所有的域都是同等重要的?
  - 是否所有信息都是有用的？（PM2018 已经去掉了 other 这个 field）
- 是否后过滤
  - 如年龄的过滤， 检索到的病历与话题是否匹配
  - 性别
  - …

使用的系统

- 建议使用开源工具
- 利用开源工具 API 实现自己的功能
- 参赛队伍报告中一般会提及使用的系统

竞赛规则参考资料

- PM2017 总结报告
  - https://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf

- 可参考其它竞赛队伍的报告
  - https://trec.nist.gov/pubs/trec26/xref.html#med

检索竞赛评分规则

- 参与形式：小组完成
- 检索效果（20分）：
  - 训练数据较少， 故需要选择5折交叉验证， 统一采用如下的话题（即查询）划分形式
    - [28, 29, 25, 22, 6, 7], [26, 11, 1, 18, 21, 4], [19, 24, 27, 30, 12, 23], [13, 14, 3, 16, 8, 9], [15, 20, 5, 10, 17, 2]
    - 每一折中，使用3部分训练，1部分验证，1部分测试（即test set）
  - 报告模型的结果请对每一折的test set取平均
  - 不得在测试查询上进行训练！违者视为作弊
- 实验报告（10分）：
  - 对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

- 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？如有必要给出公式
- 描述详细实验步骤
  - 训练
  - 测试，汇报最终得到的 5 个 test set 上的平均 P@10
  - 要求能看出没有在测试查询集上进行训练
- 汇报最终在 TREC PM 2017 上测试结果

结果提交

- 将所有材料做成一个压缩包， Email 至 benhe@ucas.ac.cn
  - 源代码
  - 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)