大作业组织形式

- 以小组为单位，每组不超过5人
- 以小组为单位进行评分，所有小组成员大作业成绩一样
- 实验报告首页应包含所有小组成员的个人信息

大作业内容

- 在TREC 2020 Deep Learning Passage Ranking数据上进行检索竞赛
- 提交以下材料

  a)结果文件：标准TREC格式

  b)程序代码：应具备对提交结果的检查功能

  c)实验报告：模型训练、验证、测试的过程

第一部分：检索竞赛

- Deep Learning Track介绍
  - 主页：https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020
  - 要求：匹配查询信息，从大规模语料中检索出相关文档，分为两个任务，Document Ranking和Passage Ranking。
  - **大作业只做第二个任务的Re-ranking子任务，即Passage Re-ranking Subtask。**
- 数据集
  - 由于Passage Ranking任务的原始语料较大，本次大作业只需在抽样处理后的数据上进行实验。
  - 可从课程网站上下载：IR_2022_Project.zip。

实验数据

- 本次大作业的实验数据IR_2022_Project.zip包含：

Description	Filename	Num Records	Format
Train Passages	collection.train.sampled.tsv	39,820	tsv: pid, passage
Train Queries	queries.train.sampled.tsv	20,000	tsv: qid, query
Train Triples (ID)	qidpidtriples.train.sampled.tsv	20,000	tsv: qid, pos_pid, neg_pid
Validation TopFile	msmarco-passagetest2019-43-top1000.tsv	41,042	tsv: qid, pid, query, passage
Validation Qrels	2019qrels-pass.txt	9,260	txt: qid, “Q0”, pid, rating
Test TopFile	msmarco-passagetest2020-54-top1000.tsv	50,024	tsv: qid, pid, query, passage
Test Qrels	2020qrels-pass.txt	11,386	txt: qid, “0”, pid, rating

- 共有20K个训练查询，40K个正负训练样本，43个验证查询，54个测试查询；只对官方提供的测试查询的初始top1000进行重排。

提交结果文件格式

- 提交的结果文件要求是**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	2571829	1	23.3981	I_LIKE_IR
1	Q0	1037798	2	22.8745	I_LIKE_IR
1	Q0	2948430	3	20.9023	I_LIKE_IR
1	Q0	5038329	4	16.1211	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
  - NDCG@10 （ndcg_cut_10）
  - Ground truth,在IR里面通常是一个叫做qrels的文件
  - 如何计算
    - 用trec_eval脚本计算
      - https://trec.nist.gov/trec_eval/
      - 运行示范：./trec_eval -m ndcg_cut qrels res
  - 模型训练时可以将验证集上的NDCG@10作为主要观测指标来选择模型

使用的系统

- 可以使用开源工具
- 利用开源工具API实现自己的功能
- 实验报告中应提及所使用的系统

竞赛相关参考资料

- TREC 2020 DL Track总结报告

  https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf

- 可参考其它竞赛队伍的报告

  https://trec.nist.gov/pubs/trec29/xref.html#deep

检索竞赛评分规则

- 参与形式：小组完成
- 检索效果（20分）：
  - 在给定的训练样本上进行模型训练，在给定的验证查询上进行模型验证选择
  - 汇报在**54个测试查询**上结果
  - 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
    - 对现有模型、方法的提高和修正

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
  - 数据处理
  - 模型训练
  - 验证测试
  - 要求能看出没有在测试查询集上进行训练
- 汇报在TREC 2020 DL的54个查询上的测试结果
- 明确给出最终提交的在测试集上得到的NDCG@10

结果提交

- 将所有材料做成一个压缩包，Email至 chenxuanang19@mails.ucas.ac.cn
- 邮件标题：IR大作业_[组长姓名]
  - 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：2022年12月31日24点

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），完成模型训练、模型选择、模型测试等步骤，得到报告中的测试结果
    - 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 xxx，conda install xxx，pip install xxx)