test 学习

ph835732abc

January 2024

1 Introduction

序：本说明包括从处理数据到输出 eval 结果的全过程

1.1 Step 1

process data：首先从原路径中读取 corpus 文件：
/share/peitian/Data/Datasets/llm-embedder/qa/msmarco/corpus.json
之后调用 process-corpus.py：

```
python scripts/process_corpus.py --collection-path
    /share/peitian/Data/Datasets/llm-embedder/qa/
    msmarco/corpus.json --output-folder /share/
    yutao/yifei/bm25_data/corpus
```

需要提前在主文件旁新建一个文件夹叫 “bm25-data” , 它会在主目录旁边新创建一个 corpus 文件夹，里面有 9 个.json 文件
之后从原路径中读取 queries 文件：
/share/peitian/Data/Datasets/llm-embedder/qa/msmarco/train.json
之后调用 process-queries.py：

```
python scripts/process_queries.py --collection-
    path /share/peitian/Data/Datasets/llm-embedder/
    qa/msmarco/train.json --output-folder /share/
    yutao/yifei/bm25_data/queries
```

它会在主目录旁边新创建一个 queries 文件夹，里面有 41 个.tsv 文件
这之后，corpus 的 json 文件中每一条数据格式为：
{”id”: id, ”contents”: contents}
queries 的 json 文件中每一条数据格式为：
{id query}

1.2 Step 2

在 bm25_data 里创建一个文件名为 index_msmarco，用于存储 index
之后调用命令：

```
target/appassembler/bin/IndexCollection
-collection JsonCollection
-input /share/yutao/yifei/bm25_data/corpus
-index /share/yutao/yifei/bm25_data/index_msmarco
-generator DefaultLuceneDocumentGenerator
-threads 9 -storePositions -storeDocvectors -
   storeRaw
```

1.3 Step 3

调用命令：

```
python scripts/mk_command.py
```

它会对之前保存的 41 个 query 的.tsv 文件分别创建 retrieval 的命令，将所有命令保存到 “commands.txt” 中
之后调用命令：

```
bash commands.txt
```

在 bm25_data 中构造了 41 个 retrieval 结果，分别保存为 f”run.msmarco-passage.num.tsv”
之后合并结果：
调用命令

```
python BM25-evaluation/scripts/summary_runs.py --
    input-folder /share/yutao/yifei/bm25_data/
    output_runs --output-file /share/yutao/yifei/
    bm25_data/output_runs/run.msmarco.tsv
```

它会生成最终的结果文件：

bm25_data/output_runs/run.msmarco.tsv 文件中每一行格式为：qid docid rank，中间用”隔开

之后 bm25_data/output_runs 中其他结果都没用了，可删

1.4 Step 4

首先将原来 train 的数据处理成 anserini 中 qrel 的格式：

调用指令

```
python BM25-evaluation/scripts/
    convert_train_to_trec_qrels.py --input /share/
    peitian/Data/Datasets/llm-embedder/qa/msmarco/
    train.json --output /share/yutao/yifei/
    bm25_data/qrels.tsv
```

之后调用 eval 函数：

```
python scripts/msmarco_passage_eval.py
/share/yutao/yifei/bm25_data/qrels.tsv /share/yutao/
yifei/bm25_data/output_runs/run.msmarco.tsv
```

会显示出 MRR @10 结果应该如下：

```
\#\#\# MRR @10: 0.23002808610498265 QueriesRanked:
    400775 \#\#\#
```

2 Conclusion

跋：在 scripts 中有一些代码是之前处理数据用的，后来优化了上述指令中用到的代码，一些冗余的代码就保留未删 冗余代码也先别删，没准后面有用 🤔