文件解释

数据集

- `IRwork/corpus/wiki_webq_corpus.tsv` ：wiki语料库
- `IRwork/data/webq-train.json`，抽样后的WebQ训练集
- `IRwork/data/webq-dev.json`，抽样后的WebQ测试集
- `IRwork/data/webq-text.txt`，测试集，每一行为一个Question。
- `IRwork/data/webq-text.csv`，测试集，每一行为一个QA对。

初始文件树

```
IRwork
├── cal_hit_multi.py # 计算hit@k的文件
├── corpus
|   └── wiki_webq_corpus.tsv # 原始wiki语料库
├── data # 数据集
|   ├── webq-dev.json
|   ├── webq-test.csv
|   ├── webq-test.txt
|   └── webq-train.json
├── DPR
|   ├── build
|   ├── CHANGELOG.md
|   ├── CODE_OF_CONDUCT.md
|   ├── conf
|   ├── CONTRIBUTING.md
|   ├── dpr
|   ├── dpr.egg-info
|   ├── LICENSE
|   ├── outputs
|   ├── README.md
|   ├── requirements.txt
|   └── setup.py
├── Retriever # 流程脚本
|   ├── bash_gen.sh
|   ├── bash_retr.sh
|   ├── gen_embedding.py
|   └── retriever.py
└── utils # 所需组件
    ├── __init__.py
    ├── __pycache__
    └── retriever_utils.py
```

DPR示例

我们以DPR为例，完成了大作业的流程供大家参考。

环境搭建

1. `conda create -n dpr python=3.8`

2. `cd IRwork/DPR`,`pip install -r requirements.txt`

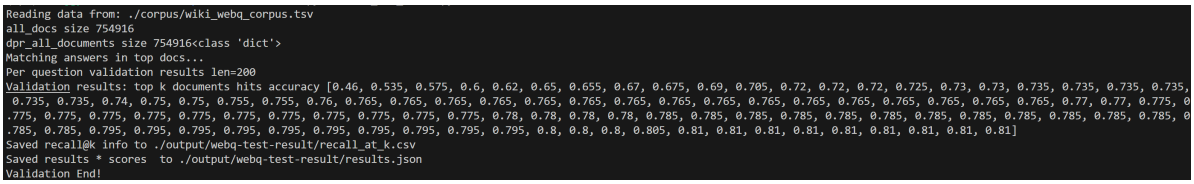
3. `python -m spacy download en_core_web_sm`

4. `cd IRwork`，将DPR加入路径`export PYTHONPATH=$(pwd)/DPR`

检索流程

1. `cd IRwork/Retriever`
2. `bash bash_gen.sh`，用`facebook/dpr-ctx_encoder-multiset-base`模型来生成语料库的embedding，保存在`IRwork/corpus/ctx_embeddings.pkl`中。
3. `bash bash_repr.sh`，用模型生成问题的embedding和语料库的faiss索引，进行检索。
4. `python IRwork/cal_hit_multi.py`，计算hit@k，k∈[1,100]。索引结果保存在`IRwork/output/webq-test-result/results.json`，hit@k结果保存在`IRwork/output/webq-test-result/recall_at_k.csv`。

输出的hit@k如下



输出的csv保存了该结果，如图所示，每一行的<k,hit>代表测试集的前k个检索结果中包含正确答案的比例：

最终文件树如下

```
├── IRwork
|   ├── cal_hit_multi.py # 根据检索结果计算hit@k的脚本
|   ├── corpus
|   |   ├── ctx_embeddings.pkl # wiki语料库的embedding
|   |   └── wiki_webq_corpus.tsv # 原始wiki语料库 格式为(id,text,title)
|   ├── data
|   |   ├── webq-dev.json # 验证集
|   |   ├── webq-test.csv # 测试集的<问题,答案>对
|   |   ├── webq-test.txt # 测试集的问题
|   |   └── webq-train.json # 训练集
|   ├── DPR
```

```
|   |   ├── build
|   |   ├── CHANGELOG.md
|   |   ├── CODE_OF_CONDUCT.md
|   |   ├── conf
|   |   ├── CONTRIBUTING.md
|   |   ├── dpr
|   |   ├── dpr.egg-info
|   |   ├── LICENSE
|   |   ├── outputs
|   |   ├── README.md
|   |   ├── requirements.txt
|   |   └── setup.py
|   ├── index
|   |   └── webq_index # 语料库的faiss索引
|   ├── output
|   |   ├── result.pkl # 检索的序列化结果
|   |   └── webq-test-result # 保存最终检索json和hit@k的文件夹
|   ├── Retriever
|   |   ├── bash_gen.sh # 生成语料库embedding的脚本
|   |   ├── bash_retr.sh # 生成检索结果与语料库faiss索引的脚本
|   |   ├── gen_embedding.py
|   |   └── retriever.py
|   └── utils
|       ├── __init__.py
|       ├── __pycache__
|       └── retriever_utils.py # 检索所需组件
```