信息检索大作业：精准医疗信息检索

项目组：7

组员：

查雨捷 2018104119

谢华伦 2018104129

刘伟 2018104118

目录

一、 实验目的 3

二、 实验描述 3

三、 检索数据集、查寻集 3

四、 实验环境 3

五、 实验过程 3

1. 提取数据 3

2. 创建索引 4

3. 检索 5

4. 结果评估 6

六、 实验结果 7

七、 结果分析 8

八、 讨论与展望 9

1. 实验心得 9

2. 实验展望 9

一、实验目的

学会使用一个信息检索系统完成给定的信息检索任务，包括创建索引、选择检索模型并设置参数、评价检索结果等等。

二、实验描述

使用 Terrier 信息检索系统，完成 TREC 2017 Precision Medicine Track 的检索任务，至少给出 10 组不同参数配置或检索方案的结果，并给出每组检索结果的各评价指标的得分，进行分析比较。

三、检索数据集、查寻集

REC 2017 Precision Medicine Track：
http://www.trec-cds.org/2017.html

四、实验环境

1.机器参数：8 核 32G 内存 Ubuntu16.04

2.软件版本：Terrier5.0， JRE1.8

五、实验过程

1. 提取数据

(1）处理 medline_xml 部分的数据集
提取所有 XML 中每篇文献的 PMID 标签内容作为 DocID，提取 ArticleTitle 和 AbstractText 标签中的内容作为文档内容（以空格分隔）。处理后的格式如下：

```
<DOC>
<DOCNO>26749840</DOCNO>
A New 3'-Prenyloxypsoralen from the Raw Fruits of Aegle marmelos and its
 Cytotoxic Activity.
</DOC>
<DOC>
```

（2）处理 extra_abstracts 部分的数据集

将所有 TXT 文件的文件名（去除后缀部分）作为 DocID，提取文本中的 Title 和 Background 部分作为文档内容（以空格分隔）。处理后的格式如下

```
<DOC>
<DOCNO>AACR_2012-751</DOCNO>
Title: Mass balance, excretion and metabolism of [14C] ASA404 in cancer p
atients in a phase I trial   Purpose: To determine the mass balance, excr
etion and metabolism of the small molecule flavonoid tumor vascular disru
pting agent ASA404 in patients with advanced cancer. Methods: Seven cance
r patients were given a single dose of 3000 mg [14C] ASA404 by intravenou
s infusion over 20 minutes prior to collection of samples of plasma, urin
e and faeces.
</DOC>
```

2. 创建索引

（1）文件路径

Terrier 工具存放在主目录 (/home/centos) 的 terrier-project-5.0 下。处理过后的数据集、查询集和相关文档集存放在 ~/centos/info 目录下，目录结构如下：

```
info/
├── data/          # 由所有xml和txt处理得到的TREC数据集
├── topic.xml      # 处理后的查询集
└── eval.qrels     # 相关文档集
```

(2) 初始化

```
cd terrier-project-5.0
./bin/trec_setup.sh /home/centos/info/data
```

在 terrier-project-5.0 目录 下执行命令对 TREC 数据集初始化 Terrier

初始化成功后可以在 ./etc/collection.spec 文件中查看到所有将要用于建立索引的文件。

(3) 创建索引

```
cd terrier-project-5.0
./bin/trec_setup.sh /home/centos/info/data
```

执行命令创建索引，其中 -j 参数表示实用更快的单通道索引，-p 表示使用多线程执行任务

```
TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=num
TrecQueryTags.process=TOP,num,disease,gene,demographi,other
```

索引数据会保存在 ./var/index 目录中

3. 检索

```
<TOP>
    <num>1<num>
    <disease>Liposarcoma
    <gene>CDK4 Amplification
    <demographic>38-year-old male
    <other>GERD
  </TOP>
```

首先需要将官网获取到的查询集处理为 TREC 格式，保存为 ~/info/topic.xml：

```
TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=num
TrecQueryTags.process=TOP,num,disease,gene,demographi,other
```

并在 ./etc/terrier.properties 配置文件中指明查询集中需要处理的标签：

```
./bin/trec_terrier.sh -r -Dtrec.model=PL2 -Dtrec.topics=/home/centos/info/topic.xml
```

同样在 terrier-project-5.0 中执行：

用-Dtrec.model 参数指定加权模型，以-Dtrec.topics 参数指定查询集 topic 文件的路径。

```
1 Q0 23852861 0 27.650466126956033 PL2
1 Q0 14694526 1 26.725727557350652 PL2
1 Q0 21910158 2 25.930928587323233 PL2
1 Q0 26336885 3 25.26427458798432 PL2
1 Q0 24487315 4 25.011965009320974 PL2
1 Q0 25679065 5 25.011965009320974 PL2
1 Q0 16938516 6 24.89633649447346 PL2
1 Q0 25121597 7 24.769926337996235 PL2
1 Q0 19737942 8 24.362111976231834 PL2
1 Q0 15221942 9 23.928335452147458 PL2
1 Q0 19574885 10 23.810128390463664 PL2
1 Q0 2013455 11 23.12209117032631 PL2
1 Q0 11475677 12 22.860751890956287 PL2
1 Q0 11505267 13 22.856427509921623 PL2
```

查询结果保存在./var/results 下 PL2_1.res 文件中

4. 结果评估

```
./bin/trec_terrier.sh -e -Dtrec.qrels=/home/centos/info/eval.qrels
```

执行命令对查询结果进行评分，以-Dtrec.qrels 指定相关文档集作为评估标准文件

```
runid   all     PL2
num_q   all     30
num_ret all     30000
num_rel all     3875
num_rel_ret     all     1586
map     all     0.0804
gm_map  all     0.0334
Rprec   all     0.1385
bpref   all     0.1721
recip_rank      all     0.5497
iprec_at_recall_0.00    all     0.5980
iprec_at_recall_0.10    all     0.2295
iprec_at_recall_0.20    all     0.1351
iprec_at_recall_0.30    all     0.1006
iprec_at_recall_0.40    all     0.0721
iprec_at_recall_0.50    all     0.0482
iprec_at_recall_0.60    all     0.0266
iprec_at_recall_0.70    all     0.0063
iprec_at_recall_0.80    all     0.0000
iprec_at_recall_0.90    all     0.0000
iprec_at_recall_1.00    all     0.0000
P_5     all     0.3333
P_10    all     0.2933
P_15    all     0.2600
P_20    all     0.2500
P_30    all     0.2289
P_100   all     0.1650
P_200   all     0.1268
```

评价结果会保存在 ./var/results/PL2_0.eval 文件中

六、实验结果

Terrier 提供了许多加权模型的实现，选择不同的权重模型依次执行检索，对所有结果进行分析，如下表

1. PL2（DFR）：随机性的泊松估计，第一归一化的拉普拉斯连续，以及术语频率归一化的归一化 2；
2. BM25：BM25 概率模型；

3. BB2（DFR）：随机性的 Bose-Einstein 模型，第一次归一化的两个伯努利过程的比率，以及术语频率归一化的归一化 2；
4. DPH（DFR）：使用 Popper 归一化（无参数）的不同超几何 DFR 模型；
5. LGD（DFR）：对数逻辑 DFR 模型；
6. IFB2（DFR）：随机性的反向项频率模型，第一次归一化的两个伯努利过程的比率，以及术语频率归一化的归一化 2；
7. DLH13（DFR）：DLH 的改进版本（无参数）；
8. InL2（DFR）：随机性的逆文档频率模型，第一次归一化的拉普拉斯序列，以及术语频率归一化的归一化 2；
9. DFRee（DFR）：另一种超几何模型，平均需要两个信息度量；
10. Hiemstra_LM：Hiemstra 的语言模型.

检索模型	MAP	P@10
PL2	0.0884	0.2933
BM25	0.0908	0.2967
BB2	0.1396	0.3967
DPH	0.0666	0.2500
LGD	0.0604	0.2033
IFB2	0.1405	0.3967
DLH13	0.1280	0.4800
InL2	0.0898	0.2867
DFRee	0.0492	0.2267
Hiemstra_LM	0.0802	0.2533

七、结果分析

1. MAP 是反映系统在全部相关文档上性能的单值指标。系统检索出来的相关文档越靠前(rank 越高)，MAP 就可能越高。如果系统没有返回相关文档，则准确率默认为 0；
2. P@10 是系统对于查询返回的前十个结果的准确率，由于尝试了不同的加强模型，在处理数据集和查询集方便做得比较完善，因此本小组 P@10 结果比较好；
3. MRR是把标准答案在被评价系统给出结果中的排序取倒数作为它的准确度，再对所有的问题取平均；
4. 在 NDCG 中，文档的相关度可以分为多个等级进行打分。由于每个查询语句所能检索到

的结果文档集合长度不一，p 值的不同会对 DCG 的计算有较大的影响。所以不能对不同查询语句的 DCG 进行求平均，需要进行归一化处理。NDCG 就是用 IDCG 进行归一化处理，表示当前 DCG 比 IDCG 还差多大的距离。

八、讨论与展望

1. 实验心得

本小组查询性能达到预期，利用了服务器平台处理数据源、创建索引和查询，实验过程并不一帆风顺。

在开始的时候，我们选取 Galago 作为检索工具，编写了 python 程序对数据源 xml 文档进行了单文件测试，在成功后对所有文档解压并准备检索，但全部解压为 xml 和 txt 文档过程漫长且占储存空间太大，在此过程中我们找到了可以使用 python 的 lib 库 gzip 打开 gz 文件免去解压过程，并获得了实验室服务器的使用权限，于是我们将数据源上传至服务器，采用多线程对数据源压缩文件处理，处理用时大约 2 小时。

之后，我们对比了三种检索工具，从使用便捷性、文档齐全性和创新性三个方面综合考虑，最终选择了 Terrier 作为本小组的检索工具。

确认了检索工具之后我们开始创建索引，利用官网文件，选择了单通道创建，用时约 50 分钟。在选择检索模型的过程中，我们比较了官网给出的各种加权模型，依据实验要求选择了十个相对不同的加权检索模型进行实验的检索评估，得到的检索结果较好。

2. 实验展望

实验的不足之处主要在与没有很好的理解官网的检索模型，由检索工具 Terrier 没有得到 MRR 和 NDCG 两个评价指标；实验的完善之处在于建立了完整的数据源索引模型，使用了十种不同的加权检索模型进行检索，正确使用官网给出的命令参数，得到的检索评价指标较好。