Terrier Team | University of Glasgow

Dockerising Terrier for OSIRRC

Arthur Câmara — *TU Delft*

Craig Macdonald — *University of Glasgow*

Terrier.org is a Java IR platform. Based on over 20 years of experience in TREC participations, it supports many TREC test collections

One of the first platforms with integrated LTR support

- Can export results in SVMlight LTR format
- Jforests LambdaMART also included

Experimental Scala notebooks integration via Apache Spark (more later)

OSIIRC Terrier-Docker Image

Terrier Team | University of Glasgow

Our implementation used the following:

- `Dockerfile` – pre-requisites only
- `Init` – download Terrier
- `Index` – customisable for different TREC corpora
  - Supported corpora: Robust04, GOV2, Core18, CW09 & CW12
  - Configurable for positional information, and fields
- `Search` – runs Terrier's `batchretrieve` command
- `Train` – calls `Search` to generate training features and then runs Jforests LambdaMART
- `Interact` (more coming shortly)

Search Performances

Terrier Team | University of Glasgow

Method		GOV2		
		701-750	751-800	801-850
BM25	Vanilla	0.2461	0.3081	0.2629
	+QE	0.2621 (+6.50%)	0.3506 (+13.79%)	0.3118 (+18.60%)
	+Proximity	0.2537 (+3.09%)	0.3126 (+1.46%)	0.2724 (+3.61%)
	+QE +Proximity	0.2715 (+10.32%)	0.3507 (+13.83)	0.3085 (+17.34%)
PL2	Vanilla	0.2334	0.2884	0.2363
	+QE	0.2478 (+6.17%)	0.3160 (+9.57%)	0.2739 (+15.91%)
	+Proximity	0.2347 (+0.056%)	0.2835 (-1.70%)	0.2361 (-0.08%)
	+QE +Proximity	0.2455 (+5.18%)	0.3095 (+7.32%)	0.2628 (+11.21%)
DPH	Vanilla	0.2804	0.3311	0.2917
	+QE	**0.3120 (+11.27%)**	**0.3754 (+13.38%)**	**0.3439 (+17.90%)**
	+Proximity	0.2834 (+1.07%)	0.3255 (-1.69%)	0.2904 (+0.045%)
	+QE +Proximity	0.3064 (+9.27%)	0.3095 (-6.52%)	0.3288 (+12.72)

We chose a few weighting models, with/without query expansion and/or proximity

Interact – Using Notebooks for an IR Experiment

Terrier Team | University of Glasgow

In [1,2], we proposed Terrier-Spark, which allows Scala notebook for running Terrier experiments

In [17]:

```
//change this for your topics file
val topicsFile = "file:/path/to/topics.txt"
val qrelsFile = "file:/path/to/qrels.txt"

val topics = TopicSource.extractTRECTopics(topicsFile).toList.toDF("qid", "query").repartition(1)

val r1 = queryTransform.transform(topics)
//r1 is a dataframe with results for queries in topics
val qrelTransform = new QrelTransformer()
    .setQrelsFile(qrelsFile)

val r2 = qrelTransform.transform(r1)
//r2 is a dataframe as r1, but also includes a label column
val ndcg = new RankingEvaluator(Measure.NDCG, 20).evaluateByQuery(r2).toList

val newSchema = StructType(topics.schema.fields ++ Array(StructField("ndcg", DoubleType, false)))
val rtr = spark.createDataFrame(topics.rdd.zipWithIndex.map{ case (row, index) => Row.fromSeq(row.t
```

```
Querying concurrent:/work/indexes/robust04.properties for 250 queries
Got for 242108 results total
We have 311410 qrels
```

Out[17]:

In [18]:

```
%%dataframe
rtr
```

Out[18]:

qid	query	ndcg
301	international organized crime	0.0
302	poliomyelitis and post polio	0.17502679579397282
303	hubble telescope achievements	0.11854207483654515

Many experiments can be done in a notebook environment – I argue that, for replicability, we should aim similarly for IR: combining Docker & notebooks

[1] Combining Terrier with Apache Spark to create agile experimental information retrieval pipelines. Craig Macdonald. In *Proceedings of SIGIR 2018.*

[2] Agile Information Retrieval Experimentation with Terrier Notebooks. Craig Macdonald, Richard McCreadie and Iadh Ounis. In *Proceedings of DESIRES 2018.*

Other Lessons Learned

Terrier Team | University of Glasgow

Do you really have the original version of the corpus?

- Files change over time. It may have been [re+]compressed over time. From .z0 to .Z to .gz…

How much memory is in the container?

- It's not trivial to predict how much memory you need.
- We tried our best to give the JVM enough memory.

Can the classical indexer be more aggressive in using available memory?

- New Terrier 5.2 recognises available memory and optimises
- 10%+ Improvement of indexing time in some cases

Terrier Team | University of Glasgow

QUESTIONS?