

Wahlpflichtmodul Information Retrieval und automatische Indexierung

Philipp Schaer – 2018-05-08 – philipp.schaer@th-koeln.de

1 Evaluation von Suchmaschinen mit trec_eval

Das Tool **trec_eval** ist ein Standard-Tool, das in der Evaluationskampagne TREC eingesetzt wird und das es erlaubt viele verschiedene Evaluationsmaße (P@n, MAP, NDCG, etc.) automatisch berechnen zu lassen.

1.1 Vorbereitungen

Für die Evaluation mit **trec_eval** brauchen Sie folgende Komponenten, die Sie allesamt in der folgenden ZIP-Datei finden: <https://ixtrieve.fh-koeln.de/lehre/mp-wp-biw-17s/trec-intro.zip>

- Das Programm **trec_eval** selbst. Das Archiv trec_eval.zip enthält eine Version für Windows (win32) und eine für MacOS. Bitte beachten Sie, dass die Dateien trec_eval.exe und cygwin1.dll immer zusammen im gleichen Verzeichnis liegen müssen; auf dem Mac existiert diese Einschränkung nicht.
- Die **Relevanzurteile** (qrels genannt) für die jeweiligen Topics und Dokumente. Die Relevanzurteile finden Sie in zwei Versionen:
 - Baseline-Kodierung (DE-all-base.txt)
 - Lingo-Kodierung (DE-all-lingo.txt – Nutzen wir später)
- Ein **Konvertierungsskript** (solr-trec.xsl), das aus den Solr-Ergebnissen, trec_eval-kompatible Run-Dateien erzeugt:

Dieses Skript kopieren Sie in den folgenden Solr-Unterverzeichnis (aus dem 1. Tutorial):

- **server/solr/girt_base/conf/xslt**

Den letzten Unterverzeichnis, also den Teil **xslt**, müssen Sie wahrscheinlich erst anlegen.

Das Konvertierungsskript enthält eine Zeile, die die Nummer des Topics enthält, für das Sie die Run-Datei erzeugen:

```
<xsl:text>126</xsl:text>
```

Bitte tragen Sie für jeden Run die richtige Topicnummer ein, ansonsten müssen Sie dies händisch z.B. per Texteditor nachtragen bzw. per Search/Replace ersetzen.

1.2 Erzeugen von Run-Dateien und Starten von trec_eval

Wenn Sie alle Bestandteile beisammenhaben, können Sie in Solr ein Ergebnis in das trec_eval-Format wie folgt konvertieren:

1. Suchen Sie ganz normal in der Web-Oberfläche, nach Ihrem Topic. In diesem Falle für Nummer 126, z.B. im Abstract: **abstract_txt:(Neue Kunst)**. Klicken Sie danach auf die graue Zeile in der Ausgabe.



Dashboard
Logging

Request-Handler (qt)

/select

common

q

neue kunst

http://localhost:8983/solr/girt_base/select?indent=on&q=neue kunst&wt=json

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
```

2. Die Anzeige wechselt dann in die JSON-Darstellung. In der Adresszeile des Browsers ändern Sie folgende Einträge, bzw. fügen Sie die Einträge hinzu:

- &fl=id,score → Die Ausgabe soll nur id und score enthalten
- &wt=xslt → Darstellung auf XSLT umstellen
- &tr=solr-trec.xsl → Welches XSLT soll verwendet werden
- &rows=1000 → 1000 Ergebnisse ausgeben

In der Summe sollte Ihre Adresszeile dann so aussehen:

[http://localhost:8983/solr/girt_base/select?q=abstract txt:\(neue%20kunst\)&fl=id,score&wt=xslt&tr=solr-trec.xsl&rows=1000](http://localhost:8983/solr/girt_base/select?q=abstract txt:(neue%20kunst)&fl=id,score&wt=xslt&tr=solr-trec.xsl&rows=1000)

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "q": "neue kunst",
      "indent": "on",
      "wt": "json"
    }
  },
  "response": { "numFound": 26223, "start": 0, "docs": [
    {
      ...
    }
  ]
}
```

```
126 Q0 GIRT-DE19950600372 0 10.655359 0
126 Q0 GIRT-DE19989185 1 10.206314 0
126 Q0 GIRT-DE19930113198 2 10.206314 0
126 Q0 GIRT-DE19940112579 3 10.147903 0
126 Q0 GIRT-DE19960106211 4 10.035259 0
126 Q0 GIRT-DE19940600886 5 9.840122 0
126 Q0 GIRT-DE19940600321 6 9.840122 0
126 Q0 GIRT-DE19950600749 7 9.840122 0
126 Q0 GIRT-DE19970600036 8 9.840122 0
126 Q0 GIRT-DE19950600260 9 9.823017 0
```

Die Ausgabe wechselt dann in das trec_eval-Format.

Je nach Konfiguration Ihres Browsers sehen Sie das Ergebnis direkt, oder es wird Ihnen als Datei-Download angeboten.

3. Speichern Sie die erzeugten Ergebnisse in eine einfache Textdatei ab, z.B. in die Datei **girt_base_run126.txt**. Nehmen Sie nun die passende Datei mit den Relevanzurteilen zur Hand und starten Sie trec_eval wie folgt:

trec_eval DE-all-base.txt girt_base_run126.txt

trec_eval liefert Ihnen dann eine Analyse des Runs und ermittelt eine ganze Reihe an Evaluationsmaßen.

```
[1108][schaer@p033068:~/sciebo/BIW-Wahlpflicht/trec]$ ./trec_eval DE-all-base.txt girt_base_run126.txt
runid          all      0
num_q          all      1
num_ret        all     1000
num_rel        all     192
num_rel_ret    all     127
map            all     0.1715
gm_map         all     0.1715
Rprec          all     0.2917
bpref          all     0.4244
recip_rank     all     0.2500
iprec_at_recall_0.00  all     0.4048
iprec_at_recall_0.10  all     0.4048
iprec_at_recall_0.20  all     0.3611
iprec_at_recall_0.30  all     0.3080
iprec_at_recall_0.40  all     0.2646
iprec_at_recall_0.50  all     0.1818
iprec_at_recall_0.60  all     0.1492
iprec_at_recall_0.70  all     0.0000
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     0.2000
P_10          all     0.3000
P_15          all     0.2667
P_20          all     0.2500
P_30          all     0.3000
P_100         all     0.3600
P_200         all     0.2900
P_500         all     0.1880
P_1000        all     0.1270
```

2 Aufgabe

Evaluieren Sie fünf Topics für die verschiedenen Indexierungsverfahren aus Tutorial 1.

Wir verteilen in der Veranstaltung je fünf Topics aus der GIRT-Kollektion, die Sie jeweils bearbeiten. Sie dürfen die Anfrage händisch anpassen, z.B. indem Sie mit Phrasen suchen. Dies ist Ihnen vollkommen freigestellt. Es sollte nur für jede Indexierungsvariante die gleiche Anfrage verwendet werden – Ansonsten wäre es ein unfairer Vergleich.

Wir haben uns auf folgendes Setting in der Veranstaltung geeinigt:

- Jeder bearbeitet 5 Topics:
 - n.a.
- Für ein Topic sollen Sie allerdings zusätzlich auch alle Tricks und Kniffe anwenden um zu schauen, was am besten aus dem Indexat herauszuholen ist.

Vergleichen Sie die Ergebnisse Ihrer Varianten für Ihre fünf Topics mit trec_eval und ermitteln Sie die Werte für:

- P@5
- P@10
- Recall (num_rel / num_rel_ret)
- Rprec
- bpref

Stellen Sie eine Tabelle mit einer Übersicht zusammen. Diese Werte werden wir in der nächsten Sitzung als Grundlage für einen Vergleich der Systeme verwenden.

Bei Fragen können Sie sich gerne per Forum im Moodle an mich wenden. Viel Erfolg.