IR Experimental Evaluation

Maria Maistro
mm@di.ku.dk
May 20, 2019

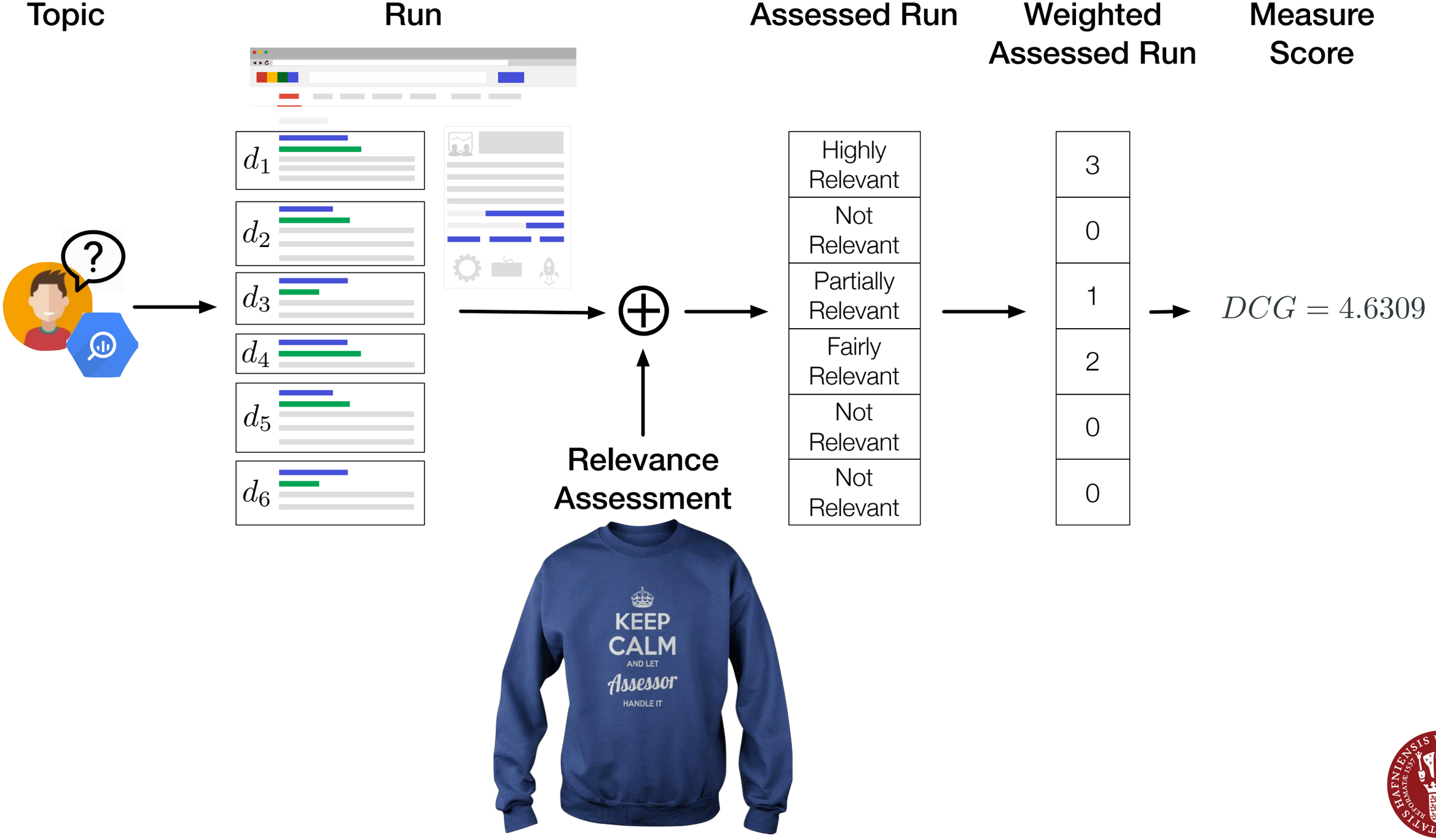
How Does Experimental Evaluation Work

- Cranfield Paradigm by Cyril W. Cleverdon
  - Dates back to mid 1960s
- Makes use of experimental collections:
  - documents (corpora)
  - topics
  - relevance judgments (binary or graded)
    also called relevance assessment
    or ground-truth (or qrels)
- Ensures comparability and repeatability of the experiments

Cyril W. Cleverdon

Cleverdon, C. W. (1962). Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.
Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

Evaluation with Test Collections in a Nutshell



Large-scale Evaluation Initiatives: TREC

TREC (Text REtrieval Conference), USA, since 1992

https://trec.nist.gov/

Donna Harman

Ellen M. Voorhees

Large-scale Evaluation Initiatives: NTCIR

NTCIR (NII Testbeds and Community for Information access Research), Japan, since 1999

http://research.nii.ac.jp/ntcir/index-en.html

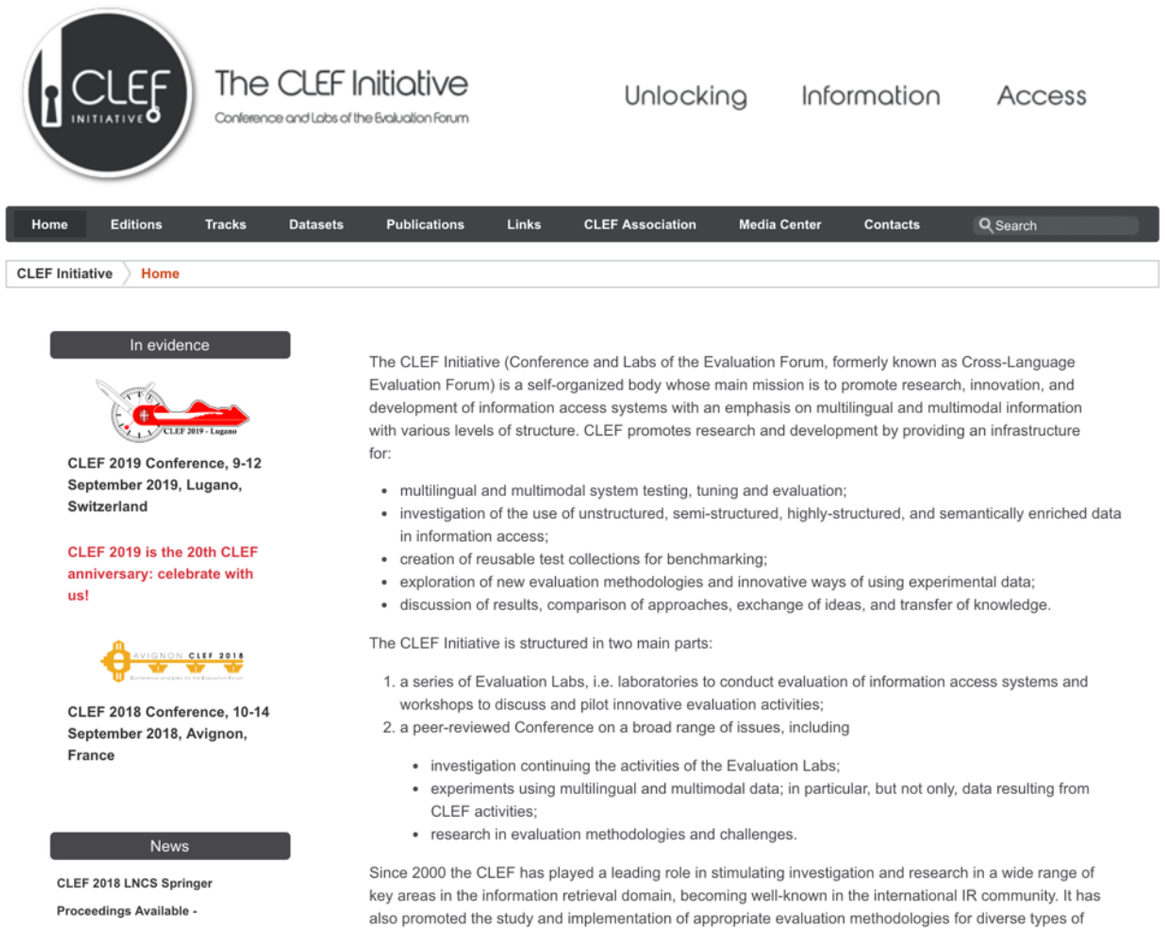
Noriko Kando

Large-scale Evaluation Initiatives: CLEF

CLEF (Conference and Labs of the Evaluation Forum), Europe, since 2000

http://www.clef-initiative.eu/

Carol Ann Peters



Nicola Ferro

Large-scale Evaluation Initiatives: FIRE

FIRE (Forum for Information Retrieval Evaluation), India, since 2008

http://fire.irsi.res.in/

Mandar Mitra

Prasenjit Majumder

Example of Documents

```
<DOC>
<DOCNO> LA010189-0003 </DOCNO>
<DOCID> 3 </DOCID>
<DATE>
<P>
January 1, 1989, Sunday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Book Review; Page 2; Book Review Desk
</P>
</SECTION>
<LENGTH>
<P>
1194 words
</P>
</LENGTH>
<HEADLINE>
<P>
PERUVIAN MEMORIES AND THE 'SHINING PATH';
</P>
<P>
TUNGSTEN A NOVEL BY CESAR VALLEJO; TRANSLATED BY ROBERT MEZEY; FOREWORD BY
KEVIN J. O'CONNOR (SYRACUSE UNIVERSITY PRESS: $19.95; 168 PP.; 0-8156-0226-X)
</P>
</HEADLINE>
<BYLINE>
<P>
By Edith Grossman, Grossman is a critic and translator of Latin American
literature. She teaches at Dominican College in New York State, is the author
of The Antipoetry of Nicanor Parra and recently translated Gabriel Garcia
Marquez's Love in the Time of Cholera.
</P>
</BYLINE>
<TEXT>
<P>
Why read "Tungsten," a partisan novel written almost 60 years ago in which evil
is unmitigated, virtue unblemished, and the characters seem to lack dimension
or complexity as they play out their predetermined roles on the stage of the
class struggle? ...
</P>
<P>
...
</P>
<P>
Kevin J. O'Connor has written a thoughtful and informative foreword, Robert
Mezey has done a fine job of translating Vallejo's often quirky Spanish, and
both of them deserve our gratitude for their sensitivity and skill in bringing
the work to the attention of an English-language audience. ...
</P>
</TEXT>
<TYPE>
<P>
Book Review
</P>
</TYPE>
</DOC>
```

Example of Topics

```
<top>

<num> Number: 303
<title> Hubble Telescope Achievements

<desc> Description:
Identify positive accomplishments of the Hubble telescope since it
was launched in 1991.

<narr> Narrative:
Documents are relevant that show the Hubble telescope has produced
new data, better quality data than previously available, data that
has increased human knowledge of the universe, or data that has led
to disproving previously existing theories or hypotheses.  Documents
limited to the shortcomings of the telescope would be irrelevant.
Details of repairs or modifications to the telescope without
reference to positive achievements would not be relevant.

</top>
```

- Topics consists of:
  - title: a brief statement expressing the information need. It resembles the typical search engine query
  - description: more detailed formulation of the information need
  - narrative: instructions for assessors on when to consider a document relevant
- Typical experimental collections make use of 50 topics

Example of Run (trec_eval format)

- Runs are textual files whose field are separated by tab or space
- Typically, there are 50 topics and 1,000 documents are retrieved for each topic

Topic ID	Fixed	Document ID	Rank	Score	Run ID
101	Q0	LA129013-951	1	0.4315	updrun
101	Q0	AP880219-0139	2	0.4278	updrun
101	Q0	LA551208-001	3	0.4278	updrun
101	Q0	AP880223-0104	4	0.3197	updrun
101	Q0	LA149197-059	5	0.3005	updrun
		…			
102	Q0	AP880273-4504	1	0.7687	updrun
102	Q0	LA149197-045	2	0.7011	updrun
102	Q0	AP820113-2304	3	0.6950	updrun

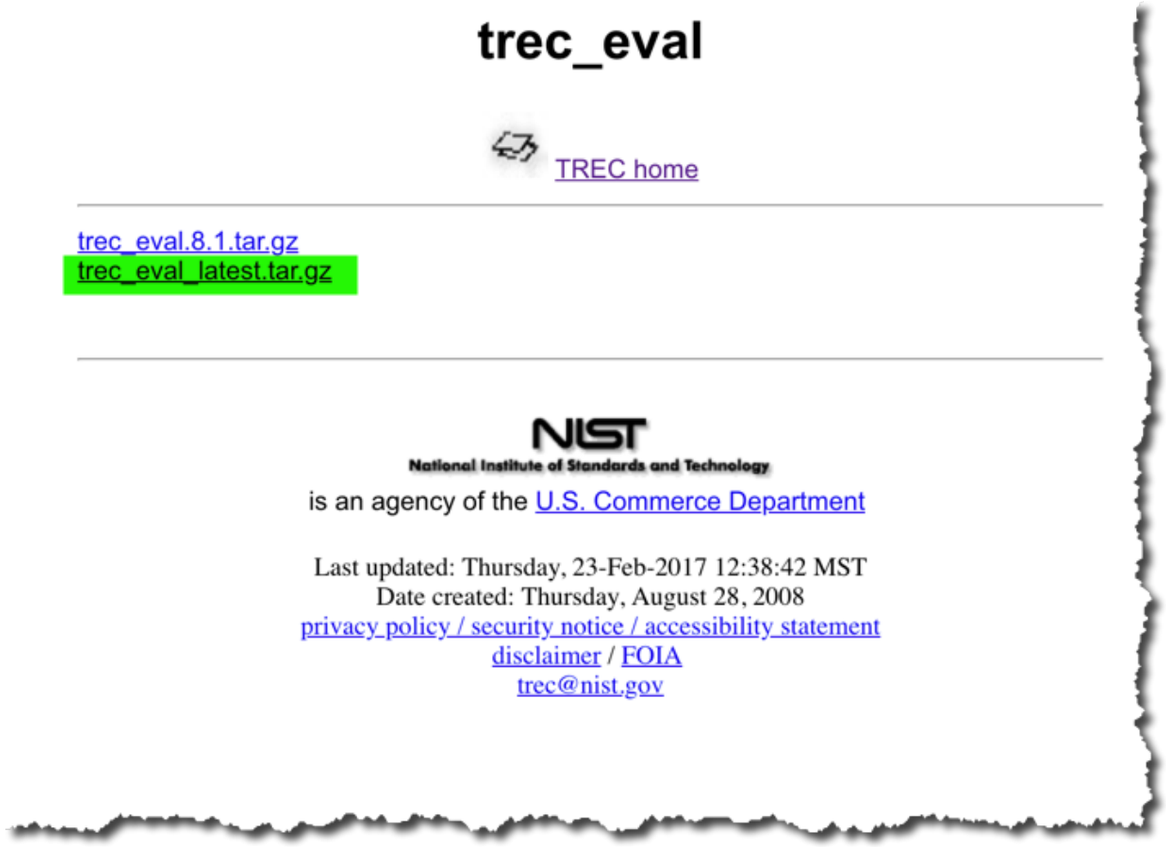
Example of Ground-truth (trec_eval format)

- Relevance judgements (qrels) are textual files whose field are separated by tab or space
- Typically, for each topic there are 300-700 judgement documents and the number of judged document vary from topic to topic

Topic ID	Fixed	Document ID	Judgement
101	0	AP880212-0047	1
101	0	AP880219-0139	0
101	0	AP880219-0166	0
101	0	AP880222-0172	0
101	0	AP880223-0104	0
…			
102	0	LA120763-901	0
102	0	LA190863-113	1
102	0	AP880273-4504	1

trec_eval

https://trec.nist.gov/trec_eval/



trec_eval

TREC home

trec_eval.8.1.tar.gz
trec_eval_latest.tar.gz

NIST
National Institute of Standards and Technology
is an agency of the U.S. Commerce Department

Last updated: Thursday, 23-Feb-2017 12:38:42 MST
Date created: Thursday, August 28, 2008
privacy policy / security notice / accessibility statement
disclaimer / FOIA
trec@nist.gov

- trec_eval is a software developed by TREC to compute most of the commonly used evaluation measures

- Download the latest version 9.0, i.e. trec_eval_latest.tar.gz

Compiling trec_eval

```
CN15037:~ tjz514$ cd /Users/tjz514/Desktop/trec_eval.9.0
CN15037:trec_eval.9.0 tjz514$ make
gcc -g -I.  -Wall -DVERSIONID=\"9.0\"  -o trec_eval trec_eval.c formats.c meas_init.c meas_acc.c meas_avg.c meas_print_single.c meas_print_final.c get_qre
ls.c get_trec_results.c get_prefs.c get_qrels_prefs.c get_qrels_jg.c form_res_rels.c form_res_rels_jg.c form_prefs_counts.c utility_pool.c get_zscores.c c
onvert_zscores.c measures.c  m_map.c m_P.c m_num_q.c m_num_ret.c m_num_rel.c m_num_rel_ret.c m_gm_map.c m_Rprec.c m_recip_rank.c m_bpref.c m_iprec_at_reca
ll.c m_recall.c m_Rprec_mult.c m_utility.c m_11pt_avg.c m_ndcg.c m_ndcg_cut.c m_Rndcg.c m_ndcg_rel.c m_binG.c m_G.c m_rel_P.c m_success.c m_infap.c m_map_
cut.c m_gm_bpref.c m_runid.c m_relstring.c m_set_P.c m_set_recall.c m_set_rel_P.c m_set_map.c m_set_F.c m_num_nonrel_judged_ret.c m_prefs_num_prefs_poss.c
 m_prefs_num_prefs_ful.c m_prefs_num_prefs_ful_ret.c m_prefs_simp.c m_prefs_pair.c m_prefs_avgjg.c m_prefs_avgjg_Rnonrel.c m_prefs_simp_ret.c m_prefs_pair
_ret.c m_prefs_avgjg_ret.c m_prefs_avgjg_Rnonrel_ret.c m_prefs_simp_imp.c m_prefs_pair_imp.c m_prefs_avgjg_imp.c m_map_avgjg.c m_Rprec_mult_avgjg.c m_P_av
gjg.c -lm
In file included from trec_eval.c:110:
./sysfunc.h:31:9: warning: 'bzero' macro redefined [-Wmacro-redefined]
#define bzero(dest,len)      memset(dest,'\0',len)
        ^
```

- Go into the source directory and compile `trec_eval` by running **`make`** (warnings may arise)
- Once done, you should see an executable named **`trec_eval`**

```
-rw-r--r--@  1 tjz514  SCIENCE\Domain Users    1299 Mar  2  2008 sysfunc.h
drwxr-xr-x@ 24 tjz514  SCIENCE\Domain Users     768 Apr 29  2008 test
-rwxr-xr-x   1 tjz514  SCIENCE\Domain Users  197896 May 20 14:22 trec_eval
-rw-r--r--@  1 tjz514  SCIENCE\Domain Users   19842 Dec 31  2008 trec_eval.c
drwxr-xr-x   3 tjz514  SCIENCE\Domain Users      96 May 20 14:22 trec_eval.dSYM
-rw-r--r--@  1 tjz514  SCIENCE\Domain Users    9493 Apr 11  2008 trec_eval.h
-rw-r--r--@  1 tjz514  SCIENCE\Domain Users   11479 Mar  2  2008 trec_format.h
-rw-r--r--@  1 tjz514  SCIENCE\Domain Users    1652 Mar  2  2008 utility_pool.c
```

Getting Help From trec_eval

To get help about the usage of `trec_eval`, run

`trec_eval -h`

```
CN15037:trec_eval.9.0 tjz514$ ./trec_eval -h
trec_eval [-h] [-q] [-m measure[.params] [-c] [-n] [-l <num>]
   [-D debug_level] [-N <num>] [-M <num>] [-R rel_format] [-T results_format]
   rel_info_file  results_file

Calculate and print various evaluation measures, evaluating the results
in results_file against the relevance info in rel_info_file.

There are a fair number of options, of which only the lower case options are
normally ever used.
 --help:
 -h: Print full help message and exit. Full help message will include
     descriptions for any measures designated by a '-m' parameter, and
     input file format descriptions for any rel_info_format given by '-R'
     and any top results_format given by '-T.'
     Thus to see all info about preference measures use
          trec_eval -h -m all_prefs -R prefs -T trec_results
 --version:
 -v: Print version of trec_eval and exit.
 --query_eval_wanted:
 -q: In addition to summary evaluation, give evaluation for each query/topic
 --measure measure_name[.measure_params]:
 -m measure: Add 'measure' to the lists of measures to calculate and print.
    If 'measure' contains a '.', then the name of the measure is everything
    preceeding the period, and everything to the right of the period is
    assumed to be a list of parameters for the measure, separated by ','.
```

Using trec_eval to Compute Set-based Measures

To compute set-based evaluation measures (Precision, Recall, F-measure) run:

`trec_eval -q -m set qrels.txt run.txt`

- `-q` prints topic-by-topic results
- `-m` selects which measures to compute, use set for set-based evaluation measures

```
CN15037:trec_eval.9.0 tjz514$ ./trec_eval -q -m set ../data/qrels_TREC_08.txt ../data/apl8n.txt
num_ret                 401     1000
num_rel                 401     300
num_rel_ret             401     104
utility                 401     -792.0000
set_P                   401     0.1040
set_relative_P          401     0.3467
set_recall              401     0.3467
set_map                 401     0.0361
set_F                   401     0.1600
num_ret                 402     1000
num_rel                 402     80
num_rel_ret             402     77
utility                 402     -846.0000
set_P                   402     0.0770
set_relative_P          402     0.9625
set_recall              402     0.9625
set_map                 402     0.0741
set_F                   402     0.1426
```

Using trec_eval to Compute Rank-based Measures

To compute all the evaluation measures run

`trec_eval -q -m all_trec qrels.txt run.txt`

```
CN15037:trec_eval.9.0 tjz514$ ./trec_eval -q -m all_trec ../data/qrels_TREC_08.txt ../data/apl8n.txt
num_ret                 401     1000
num_rel                 401     300
num_rel_ret             401     104
map                     401     0.1382
Rprec                   401     0.2267
bpref                   401     0.2409
recip_rank              401     1.0000
iprec_at_recall_0.00    401     1.0000
iprec_at_recall_0.10    401     0.6458
iprec_at_recall_0.20    401     0.3261
iprec_at_recall_0.30    401     0.1327
iprec_at_recall_0.40    401     0.0000
iprec_at_recall_0.50    401     0.0000
iprec_at_recall_0.60    401     0.0000
iprec_at_recall_0.70    401     0.0000
iprec_at_recall_0.80    401     0.0000
iprec_at_recall_0.90    401     0.0000
iprec_at_recall_1.00    401     0.0000
P_5                     401     0.6000
P_10                    401     0.6000
P_15                    401     0.6667
P_20                    401     0.6500
P_30                    401     0.6333
P_100                   401     0.4400
P_200                   401     0.3050
P_500                   401     0.1660
P_1000                  401     0.1040
relstring               401     '1100101110'
recall_5                401     0.0100
recall_10               401     0.0200
recall_15               401     0.0333
```

Questions?