

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



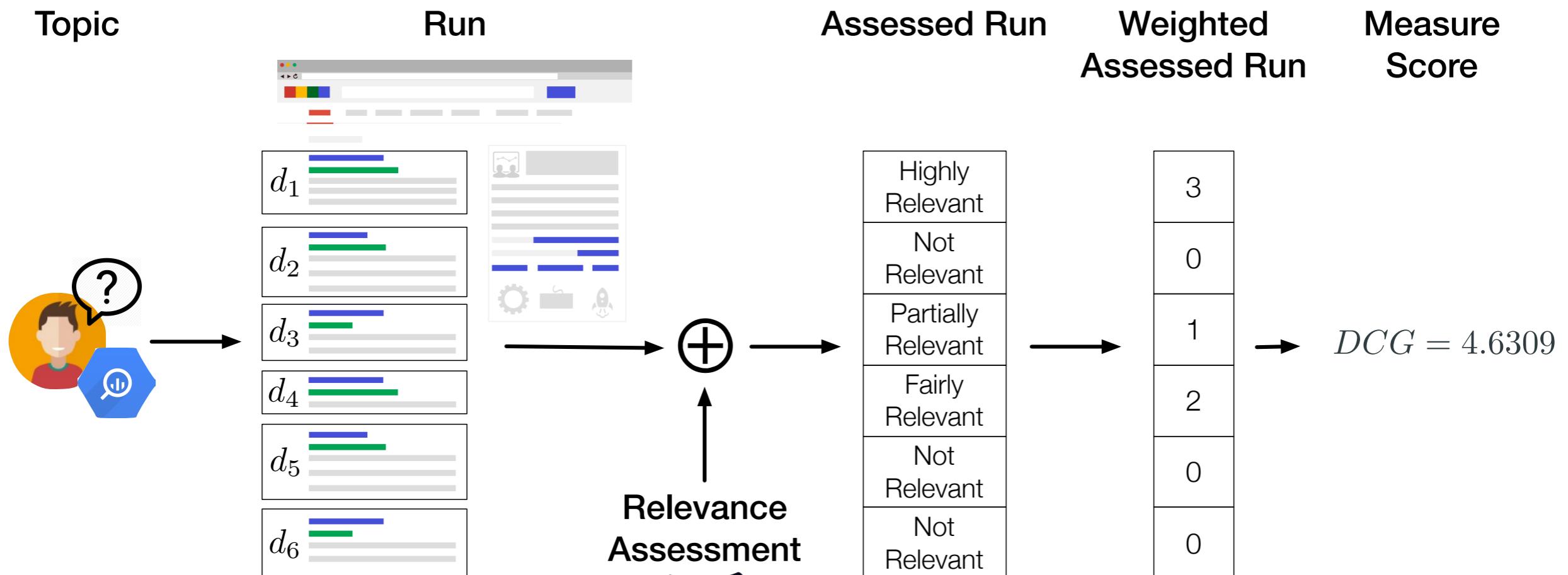
Cyril W. Cleverdon

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/>

Text REtrieval Conference (TREC)
*...to encourage research in information retrieval
from large text collections.*

Overview Other Evaluations

Publications Information for Active Participants

Tracks Frequently Asked Questions

Past TREC Results Contact Information

Data

Application deadline to participate in TREC 2018 is now past. [Celebration of the 25th TREC: November 15, 2016](#)

[TREC Economic Impact Study](#)

[TREC Statement on Product Testing and Advertising](#)

The TREC Conference series is co-sponsored by the NIST [Information Technology Laboratory's \(ITL\) Retrieval Group](#) of the [Information Access Division \(IAD\)](#)

Contact us at: trec (at) 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>

The screenshot shows the NTCIR website homepage. At the top, there's a banner with green leaves and orange circular icons for 'About NTCIR' and 'FAQ'. Below the banner is a search bar. The main navigation menu includes 'Publications/Online Proceedings', 'Data/Tools', 'NTCIR CMS Site', 'Related URL's', and 'Contact us'. On the left, a sidebar for 'NTCIR 14' lists various conference-related links. The central content area features a blue header for 'NTCIR-14' and a sub-header 'The 14th NTCIR (2018 - 2019) Evaluation of Information Access Technologies January 2018 - June 2019 Conference: June 10-13, 2019, NII, Tokyo, Japan'. Below this, a 'What's New' section lists recent updates, including task registrations and user agreement form releases.



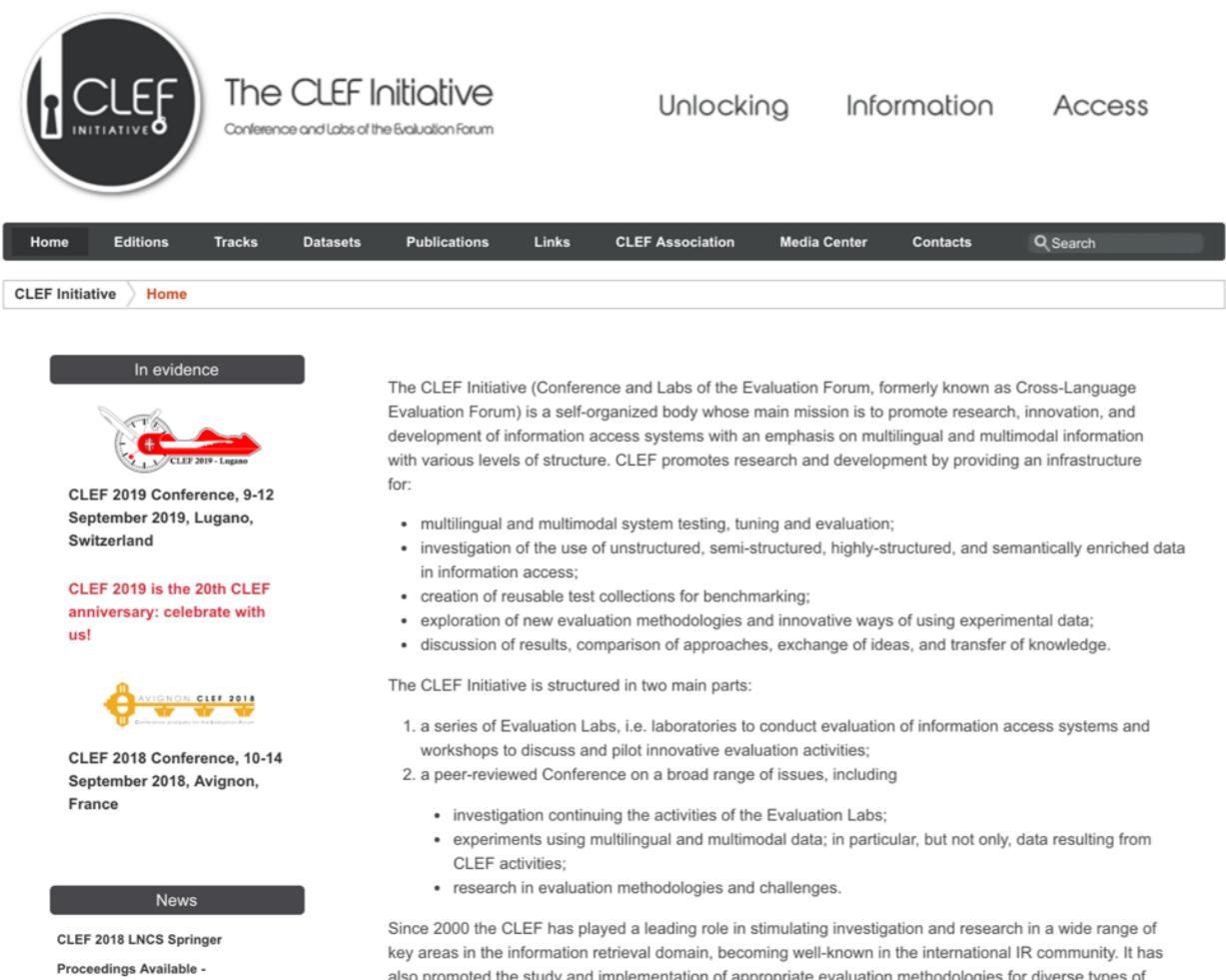
Noriko Kando



Large-scale Evaluation Initiatives: CLEF

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

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



The screenshot shows the CLEF Initiative website. At the top left is the CLEF logo. To its right is the text "The CLEF Initiative" and "Conference and Labs of the Evaluation Forum". Below this is a navigation bar with links: Home, Editions, Tracks, Datasets, Publications, Links, CLEF Association, Media Center, Contacts, and a search bar. A sub-navigation bar below it shows "CLEF Initiative > Home". On the left, there's a section titled "In evidence" featuring logos for CLEF 2019 (Lugano) and CLEF 2018 (Avignon). Below this is a "News" section with a link to "CLEF 2018 LNCS Springer Proceedings Available -". The main content area contains text about the CLEF Initiative's mission and structure, followed by a list of activities and a detailed description of its two main parts: Evaluation Labs and a Conference.

The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. CLEF promotes research and development by providing an infrastructure for:

- multilingual and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies and innovative ways of using experimental data;
- discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

The CLEF Initiative is structured in two main parts:

1. a series of Evaluation Labs, i.e. laboratories to conduct evaluation of information access systems and workshops to discuss and pilot innovative evaluation activities;
2. a peer-reviewed Conference on a broad range of issues, including
 - investigation continuing the activities of the Evaluation Labs;
 - experiments using multilingual and multimodal data; in particular, but not only, data resulting from CLEF activities;
 - research in evaluation methodologies and challenges.

Since 2000 the CLEF has played a leading role in stimulating investigation and research in a wide range of key areas in the information retrieval domain, becoming well-known in the international IR community. It has also promoted the study and implementation of appropriate evaluation methodologies for diverse types of



Carol Ann Peters



Nicola Ferro



Large-scale Evaluation Initiatives: FIRE

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

<http://fire.irs.i.res.in/>



Mandar Mitra



Prasenjit Majumder

The 10th meeting of *Forum for Information Retrieval Evaluation* 2018 will be held in Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India. Started in 2008 with the aim of building a South Asian counterpart for TREC, CLEF and NTCIR, FIRE has since evolved continuously to meet the new challenges in multilingual information access. It has expanded to include new domains like plagiarism detection, legal information access, mixed script information retrieval and spoken document retrieval to name a few.

Continuing the trend started in 2015, the FIRE will consist of a peer-reviewed conference track along with evaluation tasks. We invite full and short papers from information retrieval, natural language processing, and related domains. Please refer to the call for papers or submission guidelines for more information.



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 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:~ tgz514$ cd /Users/tgz514/Desktop/trec_eval.9.0
CN15037:trec_eval.9.0 tgz514$ 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_qrels.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 convert_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_recall.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_rnid.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_avgjg.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 tgz514 SCIENCE\Domain Users 1299 Mar  2 2008 sysfunc.h
drwxr-xr-x@ 24 tgz514 SCIENCE\Domain Users 768 Apr 29 2008 test
-rwxr-xr-x  1 tgz514 SCIENCE\Domain Users 197896 May 20 14:22 trec_eval
-rw-r--r--@ 1 tgz514 SCIENCE\Domain Users 19842 Dec 31 2008 trec_eval.c
drwxr-xr-x  3 tgz514 SCIENCE\Domain Users  96 May 20 14:22 trec_eval.dSYM
-rw-r--r--@ 1 tgz514 SCIENCE\Domain Users 9493 Apr 11 2008 trec_eval.h
-rw-r--r--@ 1 tgz514 SCIENCE\Domain Users 11479 Mar  2 2008 trec_format.h
-rw-r--r--@ 1 tgz514 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 tjjz514$ ./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
    preceding 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 tjjz514$ ./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 tjj514$ ./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?

