Introduction to evaluation with trec_eval

IR System

How do we measure retrieval effectiveness?

Text Retrieval Conference (TREC)

- TREC is an annual conference that provides researchers with a platform upon which they can evaluate and collaborate on information retrieval techniques
- Run by NIST
- TREC has a number of different tracks and each task typically has a number of tasks
  - TREC provides the correct “answer” and a tool that allows participants to compare their own results to the answer
  - This comparison can be used to compare systems and/or for tweaking / further development

Precision

- Precision is the count of relevant documents in the answer set divided by the count of documents in the answer set indicating the proportion of retrieved documents which are relevant

$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}$$

Recall

- Recall is the count of relevant documents in the answer set divided by the count of relevant documents in the corpus showing the proportion of relevant documents which have been retrieved

$$\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}$$

Evaluation

- Evaluation can be undertaken with a program called trec_eval
- trec_eval
  - Reports average precision at various cut-off points
  - Single value summary measures
  - Precision and recall figures (interpolated)

RelDocs and DocRank

- 2 central files
  - RelDocs and DocRank
- RelDocs
  - Called qrels and contains the relevance judgements matching queries and documents
  - Forms ground truth
- DocRank
  - Generated by the system under investigation and is compared against RelDocs to ascertain retrieval effectiveness

Using trec_eval

RelDocs

This is your “ground-truth” with regards to relevance.

query_id	iter	**doc_id**	**rank**
301	0	FR940202-2-00150	0
301	0	CR93E-10505	0
301	0	CR93E-1282	1
302	0	CR93E-10071	0
302	0	CR93E-10276	0
302	0	CR93E-10279	0

qrels file

DocRank

This is the file the system under investigation creates

query_id	iter	**doc_id**	rank	**sim**	**run_id**
301	Q0	FBIS4-50478	1	3.340779	STANDARD
301	Q0	FR940202-2-00150	104	2.129133	STANDARD
301	Q0	FBIS4-45552	105	2.127882	STANDARD
301	Q0	FBIS4-49075	119	2.112576	STANDARD
301	Q0	FBIS3-27288	499	1.655729	STANDARD
302	Q0	FR940126-2-00106	1	3.903381	STANDARD
302	Q0	FBIS3-60449	200	1.374640	STANDARD
302	Q0	FBIS3-60572	499	1.099626	STANDARD

Required but ignored

result file

Using trec_eval

trec_eval output

./trec_eval test/qrels.test test/results.test

num_q	all	3
num_ret	all	1500
num_rel	all	561
num_rel_ret	all	131
map	all	0.1785
gm_ap	all	0.1051
R-prec	all	0.2174
bpref	all	0.1981
recip_rank	all	0.4064
ircl_prn.0.00	all	0.4665
ircl_prn.0.10	all	0.3884
ircl_prn.0.20	all	0.3186
ircl_prn.0.30	all	0.2732
ircl_prn.0.40	all	0.2666
ircl_prn.0.50	all	0.2184
ircl_prn.0.60	all	0.0822
ircl_prn.0.70	all	0.0348
ircl_prn.0.80	all	0.0312
ircl_prn.0.90	all	0.0312
ircl_prn.1.00	all	0.0312
P5	all	0.2667
P10	all	0.3000
P15	all	0.3111
P20	all	0.3667
P30	all	0.3333
P100	all	0.2467
P200	all	0.1600
P500	all	0.0873
P1000	all	0.0437

P = 131/561

Mean Average Precision

Interpolated precision at different values of recall.
Use to draw P/R graphs

Average precision at various counts of retrieved documents

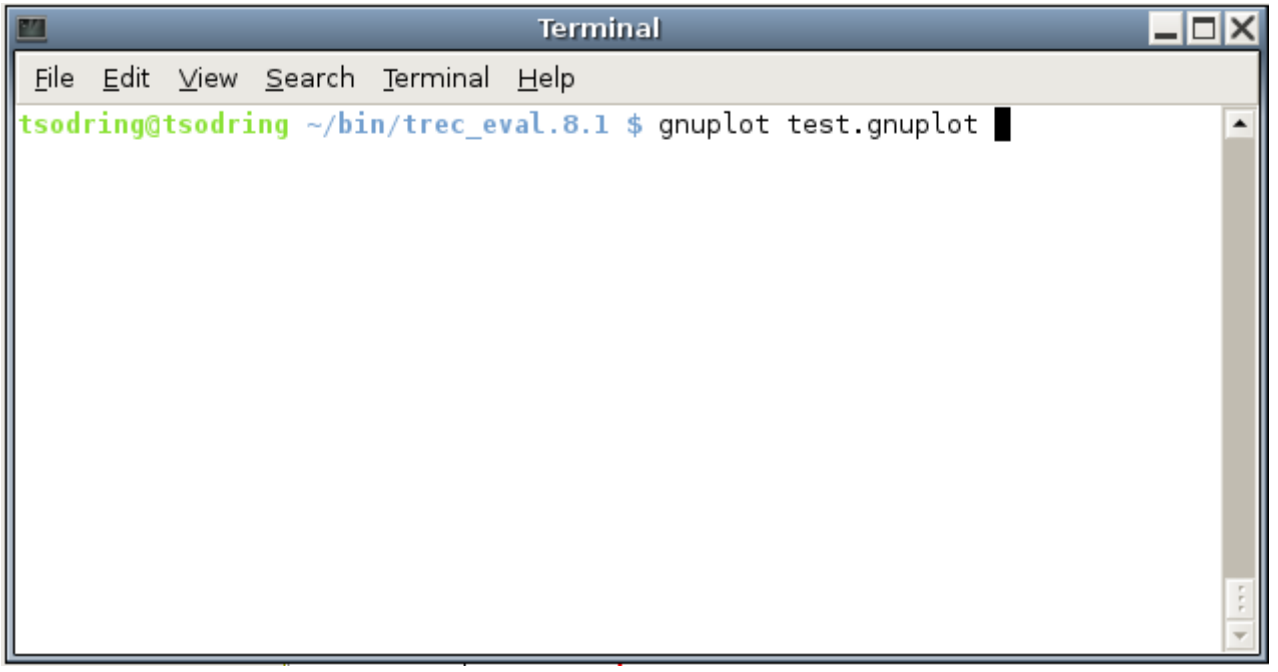
Graphing P/R

test.gnuplot

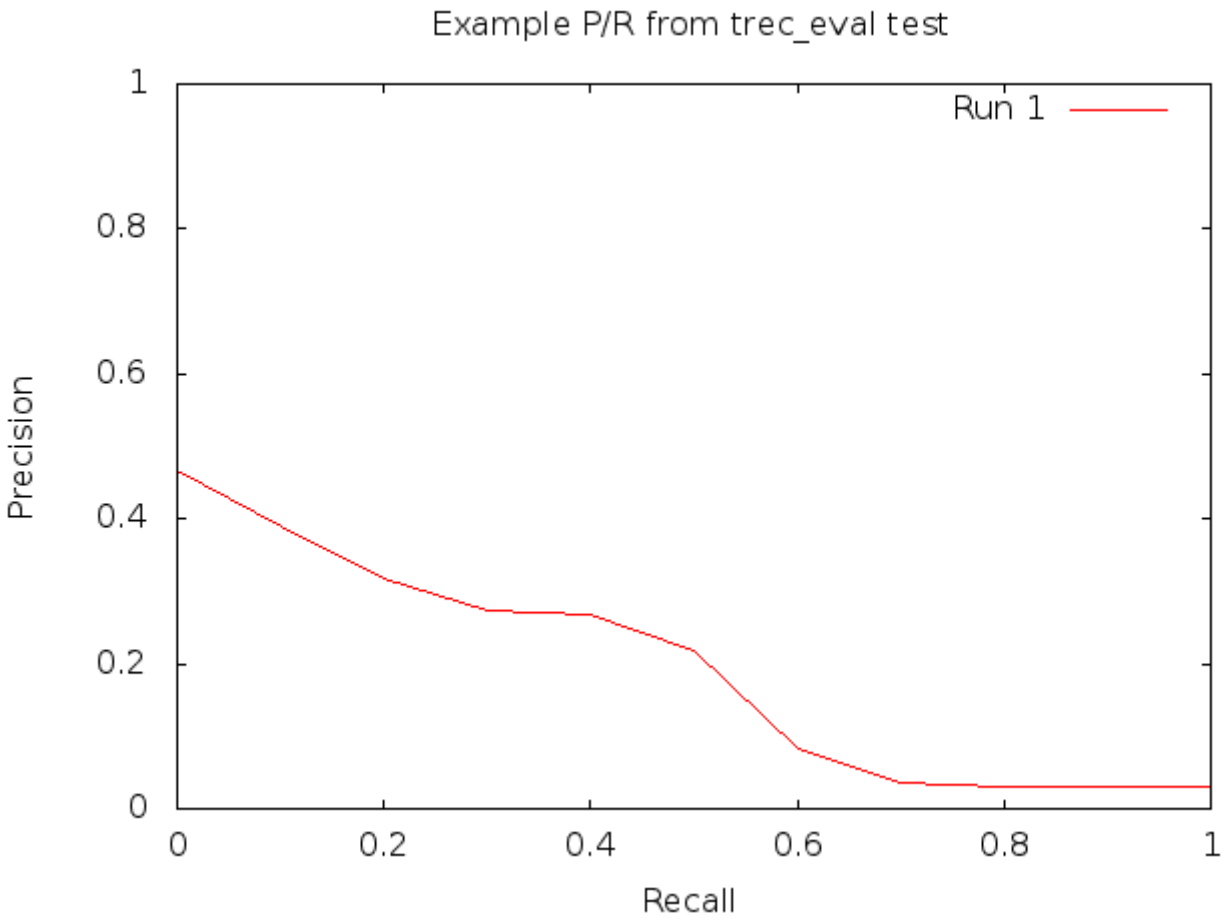
```
set term png
set output "pr_graph.png"
set title "Example P/R from trec_eval test"
set ylabel "Precision"
set xlabel "Recall"
set xrange [0:1]
set yrange [0:1]
set xtics 0,.2,1
set ytics 0,.2,1

plot 'data.dat' title "Run 1" with lines
```

data.dat

```
0.00 0.4665
0.10 0.3884
0.20 0.3186
0.30 0.2732
0.40 0.2666
0.50 0.2184
0.60 0.0822
0.70 0.0348
0.80 0.0312
0.90 0.0312
1.00 0.0312
```



Output from gnuplot



How to use trec_eval

- Compare P/R or MAP
  - of various systems
  - tweaking parameters
  - of retrieval models
- Average Precision for say retrieval at 5 docs
  - Compare systems or retrieval models

- Learn to script gnuplot with php or perl
  - Really see increased Turn around times

Further reading / references

- For textual explanation see also:
  - http://infoscience.epfl.ch/record/115460/files/Free_sofware_for_IR.pdf
  - http://ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm
  - http://trec.nist.gov/pubs/trec10/appendices/measures.pdf