Team Lost - Project2 Report

Members: Raman Sonkhla, Ashutosh Pandey, Shishir Garg

Highlights

- Algorithm used to calculate scores

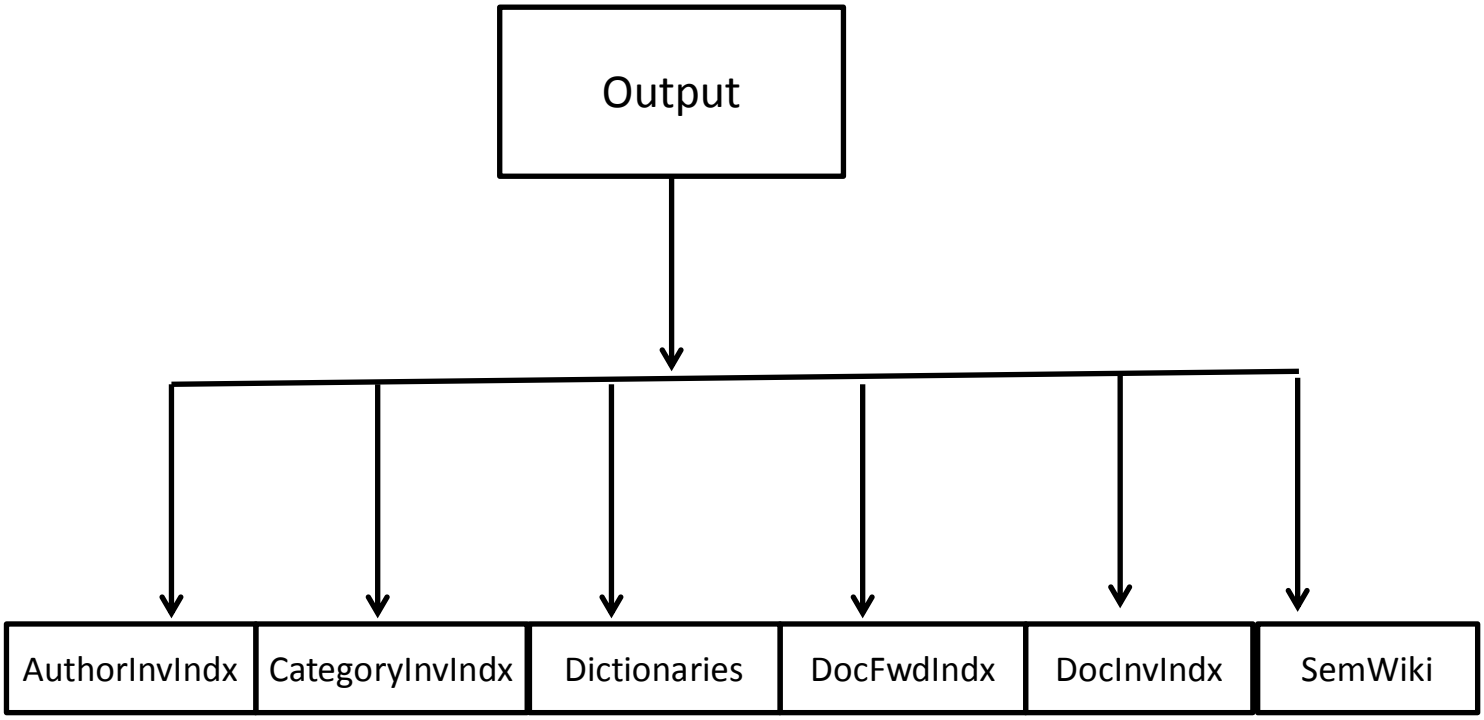
  For calculating the score of a document, we **do Document at a time traversal** through the postings lists of the query terms. We add the document scores in a heap with a fixed size equal to number of results desired by user.

  For each document, the document vector and query vector are sent to a generic function that assigns the correct weights depending on the scoring function(Boolean/ Cosine / Okapi).

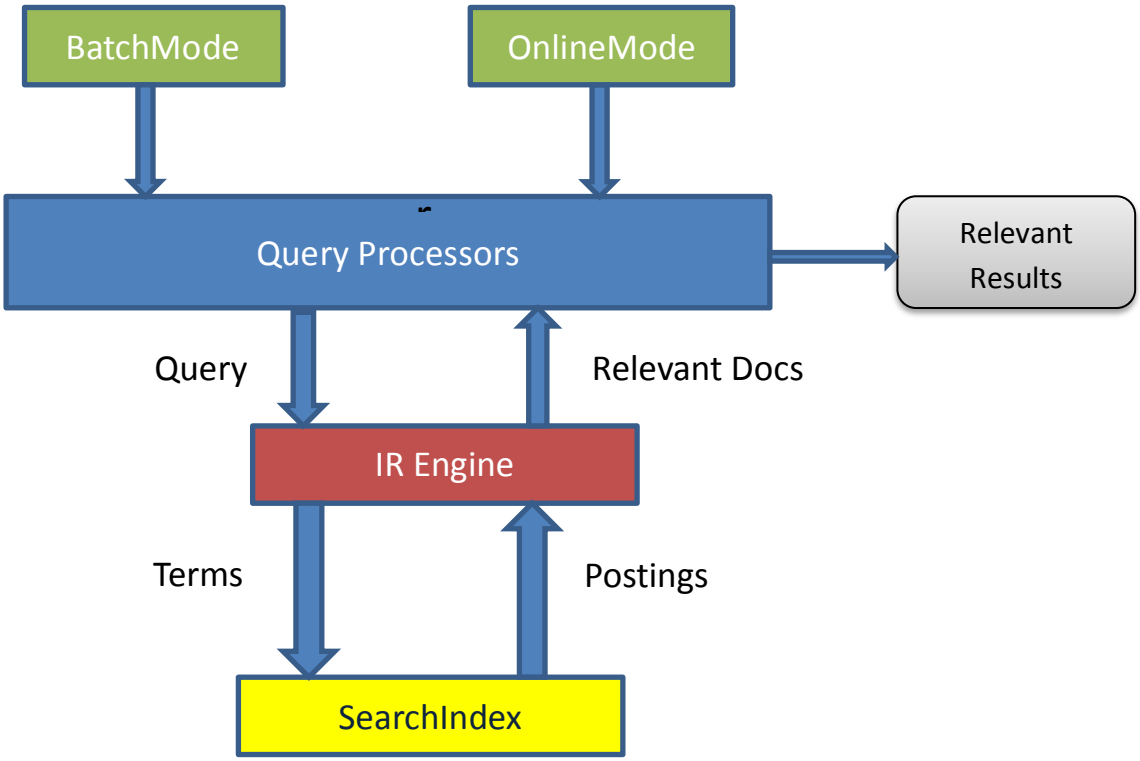
- Algorithm used to merge indexes
  The algorithm is an implementation of **SPIMI** (Single Pass In Memory Indexing). It opens all the temporary documents and writes the final index in one go. This has resulted in it being much faster than the previous indexer merge.

- Method to retrieve postings list
  The term-offset index is stored in memory as a hash map. Moreover, the offset comprises of byte offset and length of postings. This lets us fetch the postings of a term with minimum buffer.

Directory Structure

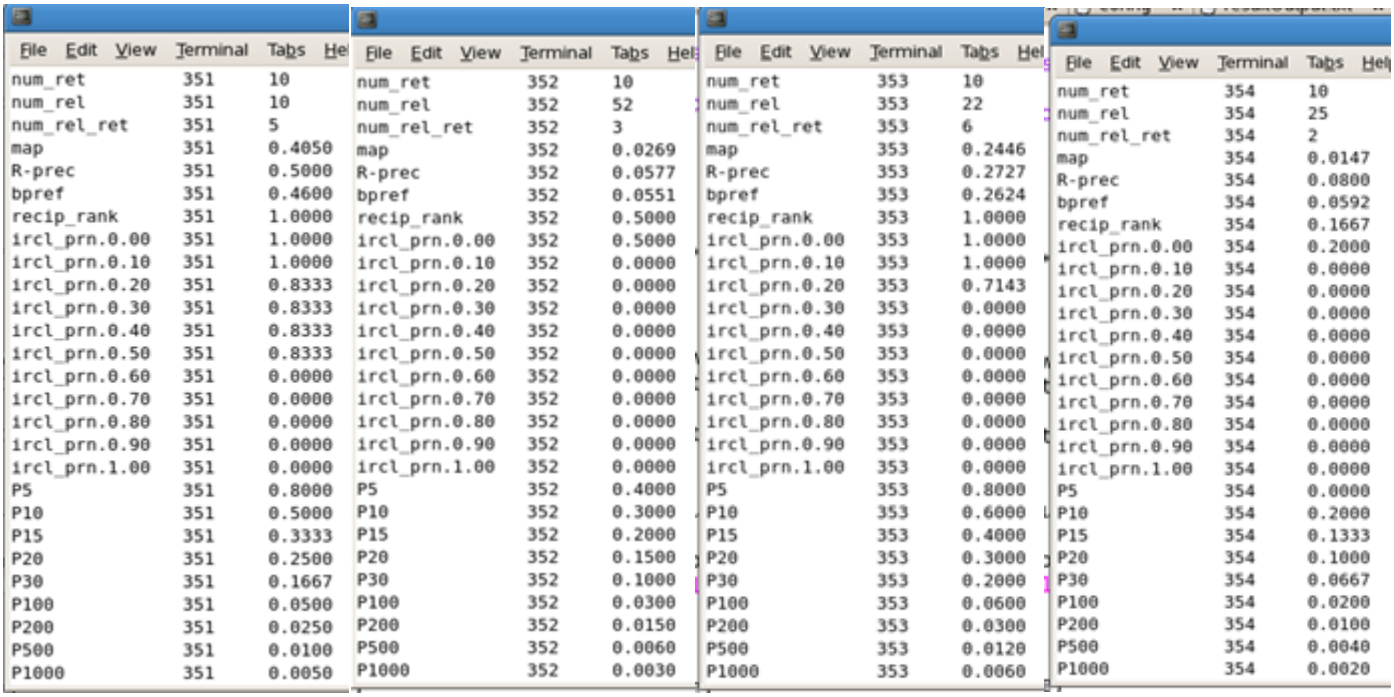


Flow Diagram



IREVAL results:

Okapi



num_ret	351	10
num_rel	351	10
num_rel_ret	351	5
map	351	0.4050
R-prec	351	0.5000
bpref	351	0.4600
recip_rank	351	1.0000
ircl_prn.0.00	351	1.0000
ircl_prn.0.10	351	1.0000
ircl_prn.0.20	351	0.8333
ircl_prn.0.30	351	0.8333
ircl_prn.0.40	351	0.8333
ircl_prn.0.50	351	0.8333
ircl_prn.0.60	351	0.0000
ircl_prn.0.70	351	0.0000
ircl_prn.0.80	351	0.0000
ircl_prn.0.90	351	0.0000
ircl_prn.1.00	351	0.0000
P5	351	0.8000
P10	351	0.5000
P15	351	0.3333
P20	351	0.2500
P30	351	0.1667
P100	351	0.0500
P200	351	0.0250
P500	351	0.0100
P1000	351	0.0050

num_ret	352	10
num_rel	352	52
num_rel_ret	352	3
map	352	0.0269
R-prec	352	0.0577
bpref	352	0.0551
recip_rank	352	0.5000
ircl_prn.0.00	352	0.5000
ircl_prn.0.10	352	0.0000
ircl_prn.0.20	352	0.0000
ircl_prn.0.30	352	0.0000
ircl_prn.0.40	352	0.0000
ircl_prn.0.50	352	0.0000
ircl_prn.0.60	352	0.0000
ircl_prn.0.70	352	0.0000
ircl_prn.0.80	352	0.0000
ircl_prn.0.90	352	0.0000
ircl_prn.1.00	352	0.0000
P5	352	0.4000
P10	352	0.3000
P15	352	0.2000
P20	352	0.1500
P30	352	0.1000
P100	352	0.0300
P200	352	0.0150
P500	352	0.0060
P1000	352	0.0030

num_ret	353	10
num_rel	353	22
num_rel_ret	353	6
map	353	0.2446
R-prec	353	0.2727
bpref	353	0.2624
recip_rank	353	1.0000
ircl_prn.0.00	353	1.0000
ircl_prn.0.10	353	1.0000
ircl_prn.0.20	353	0.7143
ircl_prn.0.30	353	0.0000
ircl_prn.0.40	353	0.0000
ircl_prn.0.50	353	0.0000
ircl_prn.0.60	353	0.0000
ircl_prn.0.70	353	0.0000
ircl_prn.0.80	353	0.0000
ircl_prn.0.90	353	0.0000
ircl_prn.1.00	353	0.0000
P5	353	0.8000
P10	353	0.6000
P15	353	0.4000
P20	353	0.3000
P30	353	0.2000
P100	353	0.0600
P200	353	0.0300
P500	353	0.0120
P1000	353	0.0060

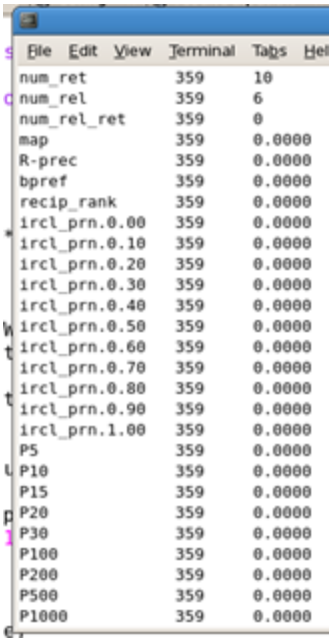
num_ret	354	10
num_rel	354	25
num_rel_ret	354	2
map	354	0.0147
R-prec	354	0.0800
bpref	354	0.0592
recip_rank	354	0.1667
ircl_prn.0.00	354	0.2000
ircl_prn.0.10	354	0.0000
ircl_prn.0.20	354	0.0000
ircl_prn.0.30	354	0.0000
ircl_prn.0.40	354	0.0000
ircl_prn.0.50	354	0.0000
ircl_prn.0.60	354	0.0000
ircl_prn.0.70	354	0.0000
ircl_prn.0.80	354	0.0000
ircl_prn.0.90	354	0.0000
ircl_prn.1.00	354	0.0000
P5	354	0.0000
P10	354	0.2000
P15	354	0.1333
P20	354	0.1000
P30	354	0.0667
P100	354	0.0200
P200	354	0.0100
P500	354	0.0040
P1000	354	0.0020

num_ret	355	10
num_rel	355	1
num_rel_ret	355	0
map	355	0.0000
R-prec	355	0.0000
bpref	355	0.0000
recip_rank	355	0.0000
ircl_prn.0.00	355	0.0000
ircl_prn.0.10	355	0.0000
ircl_prn.0.20	355	0.0000
ircl_prn.0.30	355	0.0000
ircl_prn.0.40	355	0.0000
ircl_prn.0.50	355	0.0000
ircl_prn.0.60	355	0.0000
ircl_prn.0.70	355	0.0000
ircl_prn.0.80	355	0.0000
ircl_prn.0.90	355	0.0000
ircl_prn.1.00	355	0.0000
P5	355	0.0000
P10	355	0.0000
P15	355	0.0000
P20	355	0.0000
P30	355	0.0000
P100	355	0.0000
P200	355	0.0000
P500	355	0.0000
P1000	355	0.0000

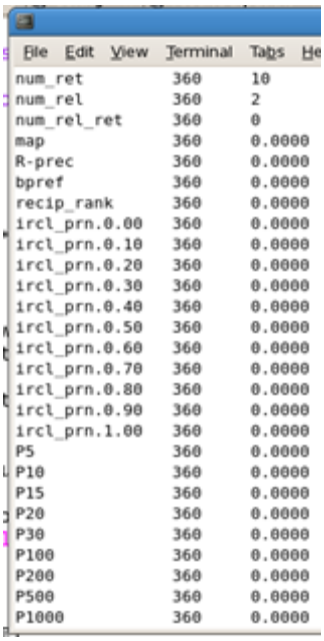
num_ret	356	10
num_rel	356	7
num_rel_ret	356	0
map	356	0.0000
R-prec	356	0.0000
bpref	356	0.0000
recip_rank	356	0.0000
ircl_prn.0.00	356	0.0000
ircl_prn.0.10	356	0.0000
ircl_prn.0.20	356	0.0000
ircl_prn.0.30	356	0.0000
ircl_prn.0.40	356	0.0000
ircl_prn.0.50	356	0.0000
ircl_prn.0.60	356	0.0000
ircl_prn.0.70	356	0.0000
ircl_prn.0.80	356	0.0000
ircl_prn.0.90	356	0.0000
ircl_prn.1.00	356	0.0000
P5	356	0.0000
P10	356	0.0000
P15	356	0.0000
P20	356	0.0000
P30	356	0.0000
P100	356	0.0000
P200	356	0.0000
P500	356	0.0000
P1000	356	0.0000

num_ret	357	10
num_rel	357	22
num_rel_ret	357	6
map	357	0.2727
R-prec	357	0.2727
bpref	357	0.2727
recip_rank	357	1.0000
ircl_prn.0.00	357	1.0000
ircl_prn.0.10	357	1.0000
ircl_prn.0.20	357	1.0000
ircl_prn.0.30	357	0.0000
ircl_prn.0.40	357	0.0000
ircl_prn.0.50	357	0.0000
ircl_prn.0.60	357	0.0000
ircl_prn.0.70	357	0.0000
ircl_prn.0.80	357	0.0000
ircl_prn.0.90	357	0.0000
ircl_prn.1.00	357	0.0000
P5	357	1.0000
P10	357	0.6000
P15	357	0.4000
P20	357	0.3000
P30	357	0.2000
P100	357	0.0600
P200	357	0.0300
P500	357	0.0120
P1000	357	0.0060

num_ret	358	10
num_rel	358	0
num_rel_ret	358	0
map	358	0.0000
R-prec	358	0.0000
bpref	358	0.0000
recip_rank	358	0.0000
ircl_prn.0.00	358	0.0000
ircl_prn.0.10	358	0.0000
ircl_prn.0.20	358	0.0000
ircl_prn.0.30	358	0.0000
ircl_prn.0.40	358	0.0000
ircl_prn.0.50	358	0.0000
ircl_prn.0.60	358	0.0000
ircl_prn.0.70	358	0.0000
ircl_prn.0.80	358	0.0000
ircl_prn.0.90	358	0.0000
ircl_prn.1.00	358	0.0000
P5	358	0.0000
P10	358	0.0000
P15	358	0.0000
P20	358	0.0000
P30	358	0.0000
P100	358	0.0000
P200	358	0.0000
P500	358	0.0000
P1000	358	0.0000



```
File  Edit  View  Terminal  Tabs  Hel
num_ret          359     10
num_rel          359     6
num_rel_ret      359     0
map              359     0.0000
R-prec           359     0.0000
bpref            359     0.0000
recip_rank       359     0.0000
ircl_prn.0.00    359     0.0000
ircl_prn.0.10    359     0.0000
ircl_prn.0.20    359     0.0000
ircl_prn.0.30    359     0.0000
ircl_prn.0.40    359     0.0000
ircl_prn.0.50    359     0.0000
ircl_prn.0.60    359     0.0000
ircl_prn.0.70    359     0.0000
ircl_prn.0.80    359     0.0000
ircl_prn.0.90    359     0.0000
ircl_prn.1.00    359     0.0000
P5               359     0.0000
P10              359     0.0000
P15              359     0.0000
P20              359     0.0000
P30              359     0.0000
P100             359     0.0000
P200             359     0.0000
P500             359     0.0000
P1000            359     0.0000
```

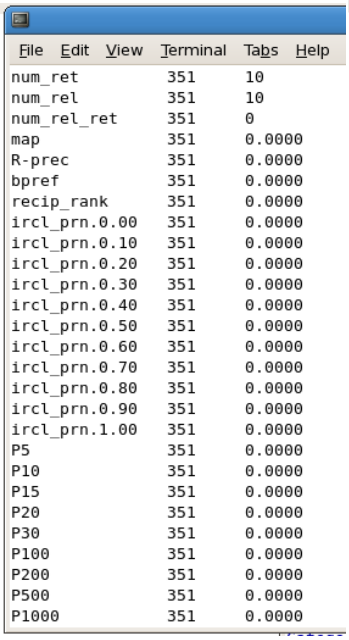


```
File  Edit  View  Terminal  Tabs  He
num_ret          360     10
num_rel          360     2
num_rel_ret      360     0
map              360     0.0000
R-prec           360     0.0000
bpref            360     0.0000
recip_rank       360     0.0000
ircl_prn.0.00    360     0.0000
ircl_prn.0.10    360     0.0000
ircl_prn.0.20    360     0.0000
ircl_prn.0.30    360     0.0000
ircl_prn.0.40    360     0.0000
ircl_prn.0.50    360     0.0000
ircl_prn.0.60    360     0.0000
ircl_prn.0.70    360     0.0000
ircl_prn.0.80    360     0.0000
ircl_prn.0.90    360     0.0000
ircl_prn.1.00    360     0.0000
P5               360     0.0000
P10              360     0.0000
P15              360     0.0000
P20              360     0.0000
P30              360     0.0000
P100             360     0.0000
P200             360     0.0000
P500             360     0.0000
P1000            360     0.0000
```

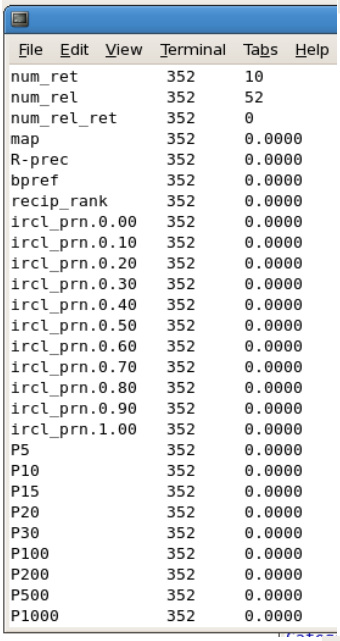
Okapi F-Measure

File Number	MAP	num_ret	num_rel	num_rel_ret	Recall	F-Measure
351	0.405	10	10	5	0.5	0.447514
352	0.0269	10	52	3	0.057692	0.036692
353	0.2446	10	22	6	0.272727	0.257899
354	0.0147	10	25	2	0.08	0.024836
355	0	10	1	0	0	0
356	0	10	7	0	0	0
357	0.2727	10	22	6	0.272727	0.272714
358	0	10	0	0	0	0
359	0	10	6	0	0	0
360	0	10	2	0	0	0

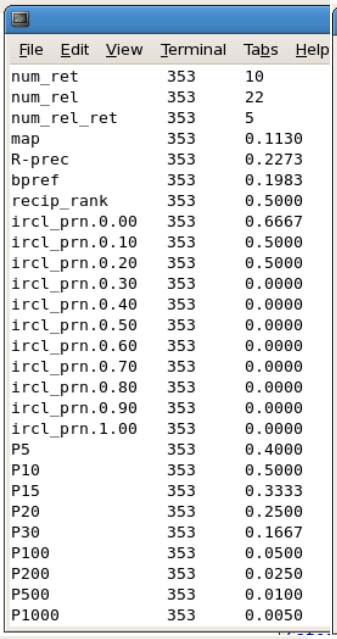
Cosine



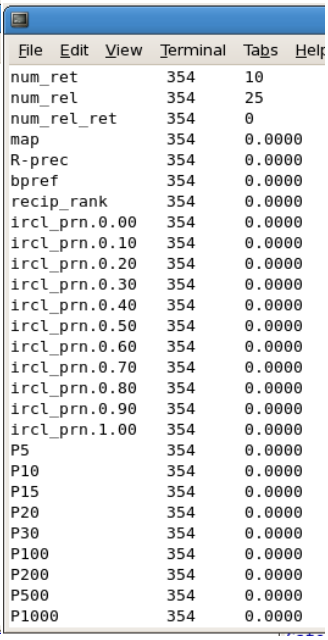
num_ret	351	10
num_rel	351	10
num_rel_ret	351	0
map	351	0.0000
R-prec	351	0.0000
bpref	351	0.0000
recip_rank	351	0.0000
ircl_prn.0.00	351	0.0000
ircl_prn.0.10	351	0.0000
ircl_prn.0.20	351	0.0000
ircl_prn.0.30	351	0.0000
ircl_prn.0.40	351	0.0000
ircl_prn.0.50	351	0.0000
ircl_prn.0.60	351	0.0000
ircl_prn.0.70	351	0.0000
ircl_prn.0.80	351	0.0000
ircl_prn.0.90	351	0.0000
ircl_prn.1.00	351	0.0000
P5	351	0.0000
P10	351	0.0000
P15	351	0.0000
P20	351	0.0000
P30	351	0.0000
P100	351	0.0000
P200	351	0.0000
P500	351	0.0000
P1000	351	0.0000



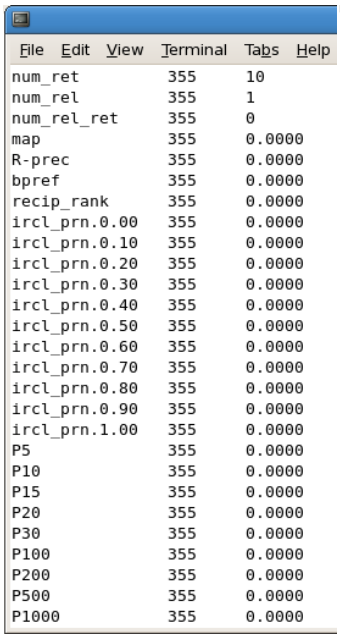
num_ret	352	10
num_rel	352	52
num_rel_ret	352	0
map	352	0.0000
R-prec	352	0.0000
bpref	352	0.0000
recip_rank	352	0.0000
ircl_prn.0.00	352	0.0000
ircl_prn.0.10	352	0.0000
ircl_prn.0.20	352	0.0000
ircl_prn.0.30	352	0.0000
ircl_prn.0.40	352	0.0000
ircl_prn.0.50	352	0.0000
ircl_prn.0.60	352	0.0000
ircl_prn.0.70	352	0.0000
ircl_prn.0.80	352	0.0000
ircl_prn.0.90	352	0.0000
ircl_prn.1.00	352	0.0000
P5	352	0.0000
P10	352	0.0000
P15	352	0.0000
P20	352	0.0000
P30	352	0.0000
P100	352	0.0000
P200	352	0.0000
P500	352	0.0000
P1000	352	0.0000



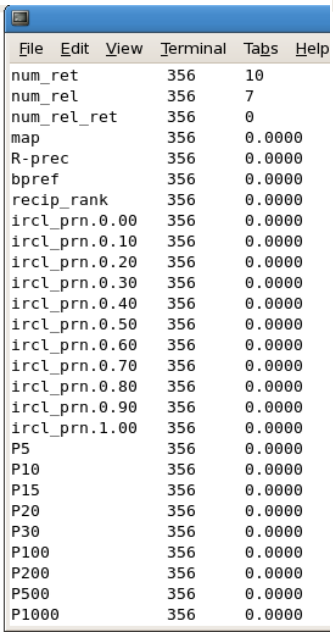
num_ret	353	10
num_rel	353	22
num_rel_ret	353	5
map	353	0.1130
R-prec	353	0.2273
bpref	353	0.1983
recip_rank	353	0.5000
ircl_prn.0.00	353	0.6667
ircl_prn.0.10	353	0.5000
ircl_prn.0.20	353	0.5000
ircl_prn.0.30	353	0.0000
ircl_prn.0.40	353	0.0000
ircl_prn.0.50	353	0.0000
ircl_prn.0.60	353	0.0000
ircl_prn.0.70	353	0.0000
ircl_prn.0.80	353	0.0000
ircl_prn.0.90	353	0.0000
ircl_prn.1.00	353	0.0000
P5	353	0.4000
P10	353	0.5000
P15	353	0.3333
P20	353	0.2500
P30	353	0.1667
P100	353	0.0500
P200	353	0.0250
P500	353	0.0100
P1000	353	0.0050



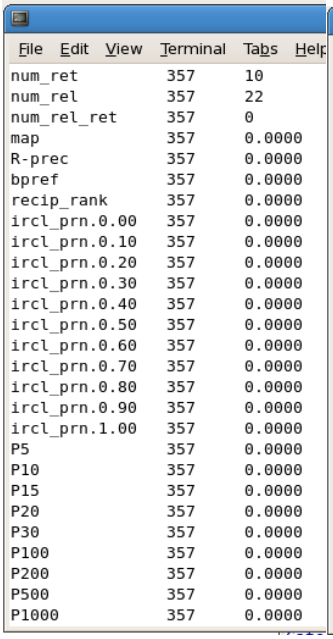
num_ret	354	10
num_rel	354	25
num_rel_ret	354	0
map	354	0.0000
R-prec	354	0.0000
bpref	354	0.0000
recip_rank	354	0.0000
ircl_prn.0.00	354	0.0000
ircl_prn.0.10	354	0.0000
ircl_prn.0.20	354	0.0000
ircl_prn.0.30	354	0.0000
ircl_prn.0.40	354	0.0000
ircl_prn.0.50	354	0.0000
ircl_prn.0.60	354	0.0000
ircl_prn.0.70	354	0.0000
ircl_prn.0.80	354	0.0000
ircl_prn.0.90	354	0.0000
ircl_prn.1.00	354	0.0000
P5	354	0.0000
P10	354	0.0000
P15	354	0.0000
P20	354	0.0000
P30	354	0.0000
P100	354	0.0000
P200	354	0.0000
P500	354	0.0000
P1000	354	0.0000



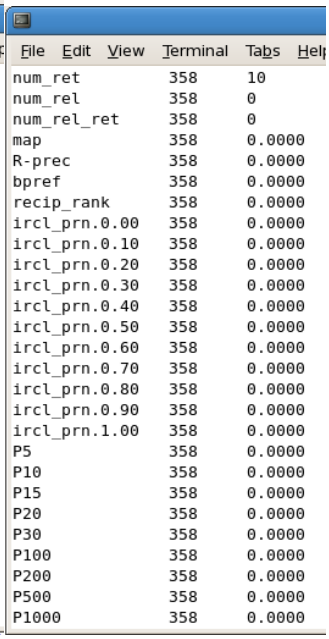
num_ret	355	10
num_rel	355	1
num_rel_ret	355	0
map	355	0.0000
R-prec	355	0.0000
bpref	355	0.0000
recip_rank	355	0.0000
ircl_prn.0.00	355	0.0000
ircl_prn.0.10	355	0.0000
ircl_prn.0.20	355	0.0000
ircl_prn.0.30	355	0.0000
ircl_prn.0.40	355	0.0000
ircl_prn.0.50	355	0.0000
ircl_prn.0.60	355	0.0000
ircl_prn.0.70	355	0.0000
ircl_prn.0.80	355	0.0000
ircl_prn.0.90	355	0.0000
ircl_prn.1.00	355	0.0000
P5	355	0.0000
P10	355	0.0000
P15	355	0.0000
P20	355	0.0000
P30	355	0.0000
P100	355	0.0000
P200	355	0.0000
P500	355	0.0000
P1000	355	0.0000



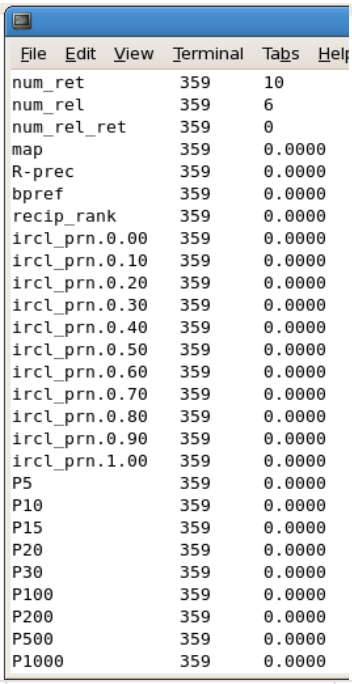
num_ret	356	10
num_rel	356	7
num_rel_ret	356	0
map	356	0.0000
R-prec	356	0.0000
bpref	356	0.0000
recip_rank	356	0.0000
ircl_prn.0.00	356	0.0000
ircl_prn.0.10	356	0.0000
ircl_prn.0.20	356	0.0000
ircl_prn.0.30	356	0.0000
ircl_prn.0.40	356	0.0000
ircl_prn.0.50	356	0.0000
ircl_prn.0.60	356	0.0000
ircl_prn.0.70	356	0.0000
ircl_prn.0.80	356	0.0000
ircl_prn.0.90	356	0.0000
ircl_prn.1.00	356	0.0000
P5	356	0.0000
P10	356	0.0000
P15	356	0.0000
P20	356	0.0000
P30	356	0.0000
P100	356	0.0000
P200	356	0.0000
P500	356	0.0000
P1000	356	0.0000



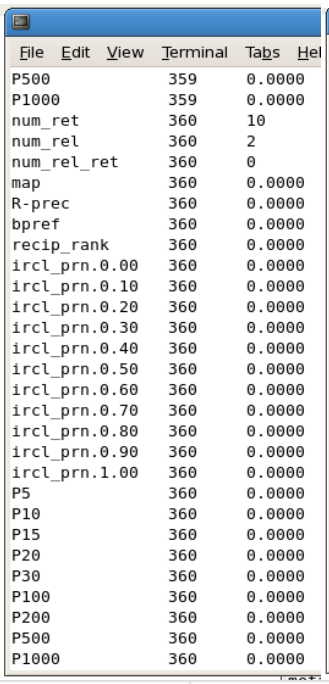
num_ret	357	10
num_rel	357	22
num_rel_ret	357	0
map	357	0.0000
R-prec	357	0.0000
bpref	357	0.0000
recip_rank	357	0.0000
ircl_prn.0.00	357	0.0000
ircl_prn.0.10	357	0.0000
ircl_prn.0.20	357	0.0000
ircl_prn.0.30	357	0.0000
ircl_prn.0.40	357	0.0000
ircl_prn.0.50	357	0.0000
ircl_prn.0.60	357	0.0000
ircl_prn.0.70	357	0.0000
ircl_prn.0.80	357	0.0000
ircl_prn.0.90	357	0.0000
ircl_prn.1.00	357	0.0000
P5	357	0.0000
P10	357	0.0000
P15	357	0.0000
P20	357	0.0000
P30	357	0.0000
P100	357	0.0000
P200	357	0.0000
P500	357	0.0000
P1000	357	0.0000



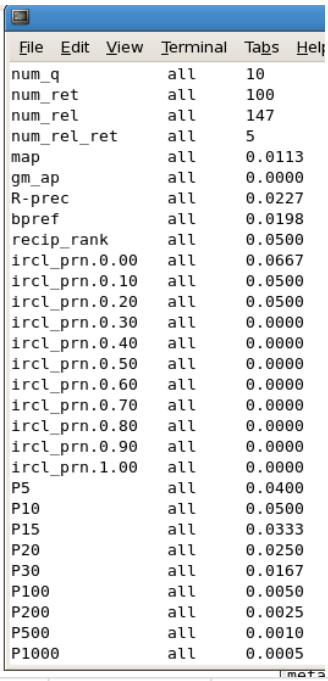
num_ret	358	10
num_rel	358	0
num_rel_ret	358	0
map	358	0.0000
R-prec	358	0.0000
bpref	358	0.0000
recip_rank	358	0.0000
ircl_prn.0.00	358	0.0000
ircl_prn.0.10	358	0.0000
ircl_prn.0.20	358	0.0000
ircl_prn.0.30	358	0.0000
ircl_prn.0.40	358	0.0000
ircl_prn.0.50	358	0.0000
ircl_prn.0.60	358	0.0000
ircl_prn.0.70	358	0.0000
ircl_prn.0.80	358	0.0000
ircl_prn.0.90	358	0.0000
ircl_prn.1.00	358	0.0000
P5	358	0.0000
P10	358	0.0000
P15	358	0.0000
P20	358	0.0000
P30	358	0.0000
P100	358	0.0000
P200	358	0.0000
P500	358	0.0000
P1000	358	0.0000



num_ret	359	10
num_rel	359	6
num_rel_ret	359	0
map	359	0.0000
R-prec	359	0.0000
bpref	359	0.0000
recip_rank	359	0.0000
ircl_prn.0.00	359	0.0000
ircl_prn.0.10	359	0.0000
ircl_prn.0.20	359	0.0000
ircl_prn.0.30	359	0.0000
ircl_prn.0.40	359	0.0000
ircl_prn.0.50	359	0.0000
ircl_prn.0.60	359	0.0000
ircl_prn.0.70	359	0.0000
ircl_prn.0.80	359	0.0000
ircl_prn.0.90	359	0.0000
ircl_prn.1.00	359	0.0000
P5	359	0.0000
P10	359	0.0000
P15	359	0.0000
P20	359	0.0000
P30	359	0.0000
P100	359	0.0000
P200	359	0.0000
P500	359	0.0000
P1000	359	0.0000



P500	359	0.0000
P1000	359	0.0000
num_ret	360	10
num_rel	360	2
num_rel_ret	360	0
map	360	0.0000
R-prec	360	0.0000
bpref	360	0.0000
recip_rank	360	0.0000
ircl_prn.0.00	360	0.0000
ircl_prn.0.10	360	0.0000
ircl_prn.0.20	360	0.0000
ircl_prn.0.30	360	0.0000
ircl_prn.0.40	360	0.0000
ircl_prn.0.50	360	0.0000
ircl_prn.0.60	360	0.0000
ircl_prn.0.70	360	0.0000
ircl_prn.0.80	360	0.0000
ircl_prn.0.90	360	0.0000
ircl_prn.1.00	360	0.0000
P5	360	0.0000
P10	360	0.0000
P15	360	0.0000
P20	360	0.0000
P30	360	0.0000
P100	360	0.0000
P200	360	0.0000
P500	360	0.0000
P1000	360	0.0000

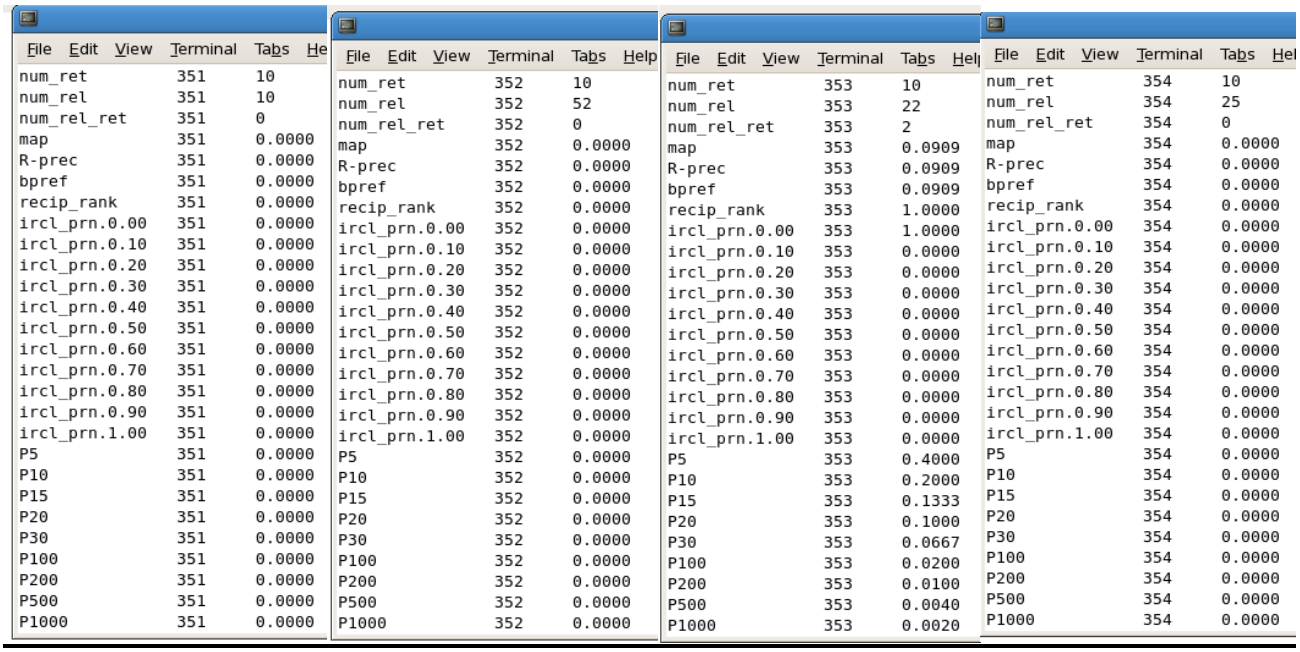


num_q	all	10
num_ret	all	100
num_rel	all	147
num_rel_ret	all	5
map	all	0.0113
gm_ap	all	0.0000
R-prec	all	0.0227
bpref	all	0.0198
recip_rank	all	0.0500
ircl_prn.0.00	all	0.0667
ircl_prn.0.10	all	0.0500
ircl_prn.0.20	all	0.0500
ircl_prn.0.30	all	0.0000
ircl_prn.0.40	all	0.0000
ircl_prn.0.50	all	0.0000
ircl_prn.0.60	all	0.0000
ircl_prn.0.70	all	0.0000
ircl_prn.0.80	all	0.0000
ircl_prn.0.90	all	0.0000
ircl_prn.1.00	all	0.0000
P5	all	0.0400
P10	all	0.0500
P15	all	0.0333
P20	all	0.0250
P30	all	0.0167
P100	all	0.0050
P200	all	0.0025
P500	all	0.0010
P1000	all	0.0005

Cosine F-Measure

File Number	MAP	num_ret	num_rel	num_rel_ret	Recall	F-Measure
351	0	10	10	0	0	0
352	0	10	52	0	0	0
353	0.113	10	22	5	0.227272727	0.045454545
354	0	10	25	0	0	0
355	0	10	1	0	0	0
356	0	10	7	0	0	0
357	0	10	22	0	0	0
358	0	10	0	0	0	0
359	0	10	6	0	0	0
360	0	10	2	0	0	0

Boolean

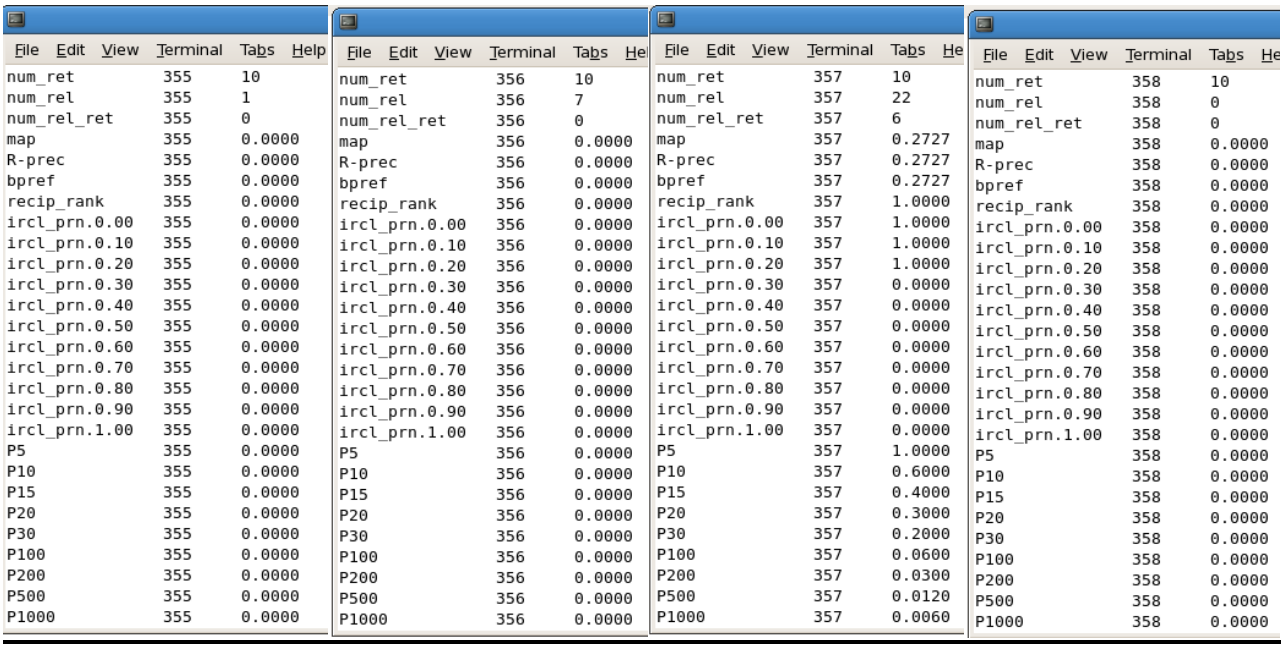


num_ret	351	10
num_rel	351	10
num_rel_ret	351	0
map	351	0.0000
R-prec	351	0.0000
bpref	351	0.0000
recip_rank	351	0.0000
ircl_prn.0.00	351	0.0000
ircl_prn.0.10	351	0.0000
ircl_prn.0.20	351	0.0000
ircl_prn.0.30	351	0.0000
ircl_prn.0.40	351	0.0000
ircl_prn.0.50	351	0.0000
ircl_prn.0.60	351	0.0000
ircl_prn.0.70	351	0.0000
ircl_prn.0.80	351	0.0000
ircl_prn.0.90	351	0.0000
ircl_prn.1.00	351	0.0000
P5	351	0.0000
P10	351	0.0000
P15	351	0.0000
P20	351	0.0000
P30	351	0.0000
P100	351	0.0000
P200	351	0.0000
P500	351	0.0000
P1000	351	0.0000

num_ret	352	10
num_rel	352	52
num_rel_ret	352	0
map	352	0.0000
R-prec	352	0.0000
bpref	352	0.0000
recip_rank	352	0.0000
ircl_prn.0.00	352	0.0000
ircl_prn.0.10	352	0.0000
ircl_prn.0.20	352	0.0000
ircl_prn.0.30	352	0.0000
ircl_prn.0.40	352	0.0000
ircl_prn.0.50	352	0.0000
ircl_prn.0.60	352	0.0000
ircl_prn.0.70	352	0.0000
ircl_prn.0.80	352	0.0000
ircl_prn.0.90	352	0.0000
ircl_prn.1.00	352	0.0000
P5	352	0.0000
P10	352	0.0000
P15	352	0.0000
P20	352	0.0000
P30	352	0.0000
P100	352	0.0000
P200	352	0.0000
P500	352	0.0000
P1000	352	0.0000

num_ret	353	10
num_rel	353	22
num_rel_ret	353	2
map	353	0.0909
R-prec	353	0.0909
bpref	353	0.0909
recip_rank	353	1.0000
ircl_prn.0.00	353	1.0000
ircl_prn.0.10	353	0.0000
ircl_prn.0.20	353	0.0000
ircl_prn.0.30	353	0.0000
ircl_prn.0.40	353	0.0000
ircl_prn.0.50	353	0.0000
ircl_prn.0.60	353	0.0000
ircl_prn.0.70	353	0.0000
ircl_prn.0.80	353	0.0000
ircl_prn.0.90	353	0.0000
ircl_prn.1.00	353	0.0000
P5	353	0.4000
P10	353	0.2000
P15	353	0.1333
P20	353	0.1000
P30	353	0.0667
P100	353	0.0200
P200	353	0.0100
P500	353	0.0040
P1000	353	0.0020

num_ret	354	10
num_rel	354	25
num_rel_ret	354	0
map	354	0.0000
R-prec	354	0.0000
bpref	354	0.0000
recip_rank	354	0.0000
ircl_prn.0.00	354	0.0000
ircl_prn.0.10	354	0.0000
ircl_prn.0.20	354	0.0000
ircl_prn.0.30	354	0.0000
ircl_prn.0.40	354	0.0000
ircl_prn.0.50	354	0.0000
ircl_prn.0.60	354	0.0000
ircl_prn.0.70	354	0.0000
ircl_prn.0.80	354	0.0000
ircl_prn.0.90	354	0.0000
ircl_prn.1.00	354	0.0000
P5	354	0.0000
P10	354	0.0000
P15	354	0.0000
P20	354	0.0000
P30	354	0.0000
P100	354	0.0000
P200	354	0.0000
P500	354	0.0000
P1000	354	0.0000

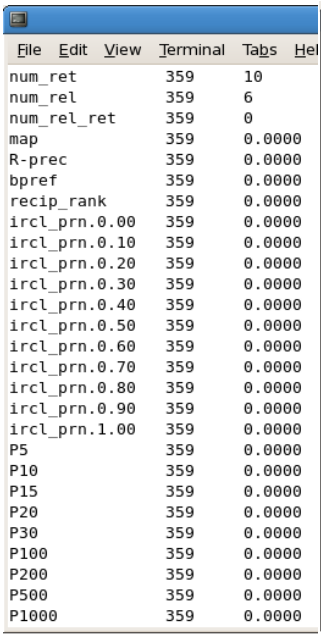


num_ret	355	10
num_rel	355	1
num_rel_ret	355	0
map	355	0.0000
R-prec	355	0.0000
bpref	355	0.0000
recip_rank	355	0.0000
ircl_prn.0.00	355	0.0000
ircl_prn.0.10	355	0.0000
ircl_prn.0.20	355	0.0000
ircl_prn.0.30	355	0.0000
ircl_prn.0.40	355	0.0000
ircl_prn.0.50	355	0.0000
ircl_prn.0.60	355	0.0000
ircl_prn.0.70	355	0.0000
ircl_prn.0.80	355	0.0000
ircl_prn.0.90	355	0.0000
ircl_prn.1.00	355	0.0000
P5	355	0.0000
P10	355	0.0000
P15	355	0.0000
P20	355	0.0000
P30	355	0.0000
P100	355	0.0000
P200	355	0.0000
P500	355	0.0000
P1000	355	0.0000

num_ret	356	10
num_rel	356	7
num_rel_ret	356	0
map	356	0.0000
R-prec	356	0.0000
bpref	356	0.0000
recip_rank	356	0.0000
ircl_prn.0.00	356	0.0000
ircl_prn.0.10	356	0.0000
ircl_prn.0.20	356	0.0000
ircl_prn.0.30	356	0.0000
ircl_prn.0.40	356	0.0000
ircl_prn.0.50	356	0.0000
ircl_prn.0.60	356	0.0000
ircl_prn.0.70	356	0.0000
ircl_prn.0.80	356	0.0000
ircl_prn.0.90	356	0.0000
ircl_prn.1.00	356	0.0000
P5	356	0.0000
P10	356	0.0000
P15	356	0.0000
P20	356	0.0000
P30	356	0.0000
P100	356	0.0000
P200	356	0.0000
P500	356	0.0000
P1000	356	0.0000

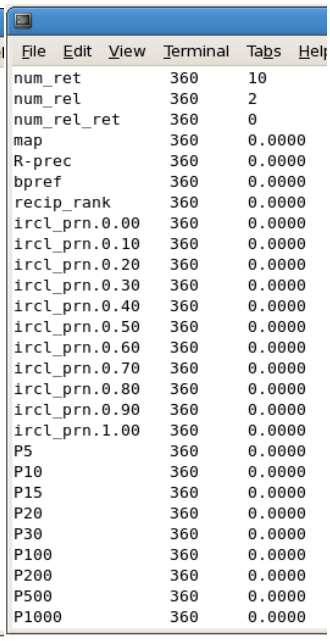
num_ret	357	10
num_rel	357	22
num_rel_ret	357	6
map	357	0.2727
R-prec	357	0.2727
bpref	357	0.2727
recip_rank	357	1.0000
ircl_prn.0.00	357	1.0000
ircl_prn.0.10	357	1.0000
ircl_prn.0.20	357	1.0000
ircl_prn.0.30	357	0.0000
ircl_prn.0.40	357	0.0000
ircl_prn.0.50	357	0.0000
ircl_prn.0.60	357	0.0000
ircl_prn.0.70	357	0.0000
ircl_prn.0.80	357	0.0000
ircl_prn.0.90	357	0.0000
ircl_prn.1.00	357	0.0000
P5	357	1.0000
P10	357	0.6000
P15	357	0.4000
P20	357	0.3000
P30	357	0.2000
P100	357	0.0600
P200	357	0.0300
P500	357	0.0120
P1000	357	0.0060

num_ret	358	10
num_rel	358	0
num_rel_ret	358	0
map	358	0.0000
R-prec	358	0.0000
bpref	358	0.0000
recip_rank	358	0.0000
ircl_prn.0.00	358	0.0000
ircl_prn.0.10	358	0.0000
ircl_prn.0.20	358	0.0000
ircl_prn.0.30	358	0.0000
ircl_prn.0.40	358	0.0000
ircl_prn.0.50	358	0.0000
ircl_prn.0.60	358	0.0000
ircl_prn.0.70	358	0.0000
ircl_prn.0.80	358	0.0000
ircl_prn.0.90	358	0.0000
ircl_prn.1.00	358	0.0000
P5	358	0.0000
P10	358	0.0000
P15	358	0.0000
P20	358	0.0000
P30	358	0.0000
P100	358	0.0000
P200	358	0.0000
P500	358	0.0000
P1000	358	0.0000



File Edit View Terminal Tabs Hel

```
num_ret           359     10
num_rel           359     6
num_rel_ret       359     0
map               359     0.0000
R-prec            359     0.0000
bpref             359     0.0000
recip_rank        359     0.0000
ircl_prn.0.00     359     0.0000
ircl_prn.0.10     359     0.0000
ircl_prn.0.20     359     0.0000
ircl_prn.0.30     359     0.0000
ircl_prn.0.40     359     0.0000
ircl_prn.0.50     359     0.0000
ircl_prn.0.60     359     0.0000
ircl_prn.0.70     359     0.0000
ircl_prn.0.80     359     0.0000
ircl_prn.0.90     359     0.0000
ircl_prn.1.00     359     0.0000
P5                359     0.0000
P10               359     0.0000
P15               359     0.0000
P20               359     0.0000
P30               359     0.0000
P100              359     0.0000
P200              359     0.0000
P500              359     0.0000
P1000             359     0.0000
```



File Edit View Terminal Tabs Help

```
num_ret           360     10
num_rel           360     2
num_rel_ret       360     0
map               360     0.0000
R-prec            360     0.0000
bpref             360     0.0000
recip_rank        360     0.0000
ircl_prn.0.00     360     0.0000
ircl_prn.0.10     360     0.0000
ircl_prn.0.20     360     0.0000
ircl_prn.0.30     360     0.0000
ircl_prn.0.40     360     0.0000
ircl_prn.0.50     360     0.0000
ircl_prn.0.60     360     0.0000
ircl_prn.0.70     360     0.0000
ircl_prn.0.80     360     0.0000
ircl_prn.0.90     360     0.0000
ircl_prn.1.00     360     0.0000
P5                360     0.0000
P10               360     0.0000
P15               360     0.0000
P20               360     0.0000
P30               360     0.0000
P100              360     0.0000
P200              360     0.0000
P500              360     0.0000
P1000             360     0.0000
```

Boolean F-Measure

File Number	MAP	num_ret	num_rel	num_rel_ret	Recall	F-Measure
351	0	10	10	0	0	0
352	0	10	52	0	0	0
353	0.09	10	22	2	0.090909	0.090452
354	0	10	25	0	0	0
355	0	10	1	0	0	0
356	0	10	7	0	0	0
357	0.27	10	33	6	0.181818	0.217304
358	0	10	0	0	0	0
359	0	10	2	6	3	0
360	0	10	2	0	0	0

Contributions of Team Members

Ashutosh Pandey

Phase1:

- Plain Text Parser
- Term, Term Count Dictionaries
- Single Pass In Memory Merge

Phase 2:

- Implementation of Single Pass In Memory Merge
- Interface to All Indices: Efficient Implementation of Posting List Retrieval using Input and Output Buffering
- Generation of statistics of program performance on TREC data using different scoring methods.

Shishir Garg

Phase 1:

Indexer.h and Indexer.cpp

- Class to hold inverted index structure and associated variables.
- Function UpdateIndex to update index as each file is processed and write to the disk if memory is exceeded.
- Parameters.cpp - Class to hold all parameters provided in configuration file
- Function to read and parse parameters given file path.
- Stopword.cpp - Function to filter tokens based on a file of stopwords. Not used in final implementation.

Phase 2:

- Stopword Removal.
- Boolean.cpp, Okapi.cpp and Cosine.cpp. Not used in final implementation.
- Project2_report.pdf

Raman Sonkhla

Work done Phase-1

1. Wiki files processing
    - Metadata parsing and SemiWiki file generation.
    - Generation of dictionaries and repositories related to wiki files.
2. Program design
    - Algorithm flow (participation of all team members)
    - Generic classes for Data collection (crawler), Dictionaries generation and dumping.

Work done Phase-2

1. Inverted index generation
    - Re-did the Inverted index from scratch. Index being generated Phase-1 did not have any information regarding: Term frequency in a document, Document frequency, Document length, etc. This data is required for score / compute documents similarity. Also temp-inverted index has data like length of postings etc, for efficient merging operation.
    - Implemented Hash map based inverted index generation to reduce indexing time.
    - Generation of Inverted indexes for Author and categories, and Champions list implementation. Champions list is not being used right now. But the frame work is complete. We can easy add this feature in phase 3, if we have some extra time.
    - Some important source files
        i. InvertedIndexGenerator(.h/.cpp)
        ii. InvertedIndexHashMap(.h/.cpp)
        iii. PostingsValue(.h/.cpp)
2. Query processor
    - Added support for both Batch mode and online mode of queries.
    - Added logic for robustness against unexpected query and fault tolerance.
    - Added support for generation of output in the expected format, including static snippets generation using SemiWiki files. Currently we are not using fwd index. Using it, simple k-mean clustering and a window size we implement a form of key-word-in-context (KWIC) snippet.
    - Some important source files
        i. QuerryProcessor(.h/.cpp)
        ii. Parameters(.h/.cpp)
3. Information retrieval engine
    - Added support for processing query vector given by Query processor.
    - Have implemented 3 ranking methods (Boolean, Cosine and Okapi).
    - Implemented a priority queue (of size as requested by Query processor) to store and update scores of most relevant documents.
    - Some important source files
        i. IREngine(.h/.cpp)
        ii. HeapDocScore(.h/.cpp)