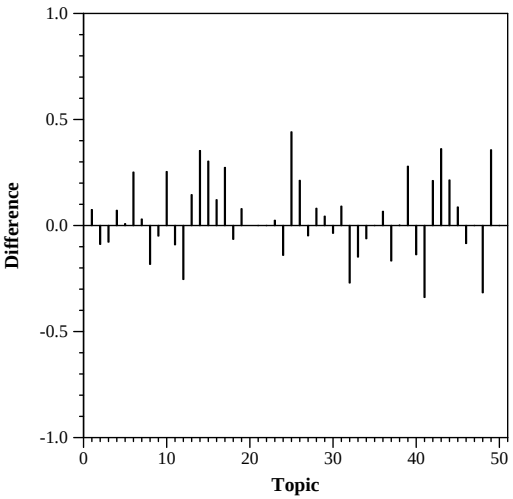
Phase 1: Each group submitted a set of 5 documents per topic to be used as relevance feedback input in Phase 2 by 3 to 5 groups. One or two sets submitted. Evaluation output includes number of relevant documents in set, and how well other groups did on this set compared to the other sets that that group ran (each group ran 7 to 8 Phase 1 sets). Comparison numbers totaled among the collection and evaluation measures used in Phase 2. Total score = B / (B + W) where B is the total number of runs/measures this set did better than, and W is the number this set did worse on.

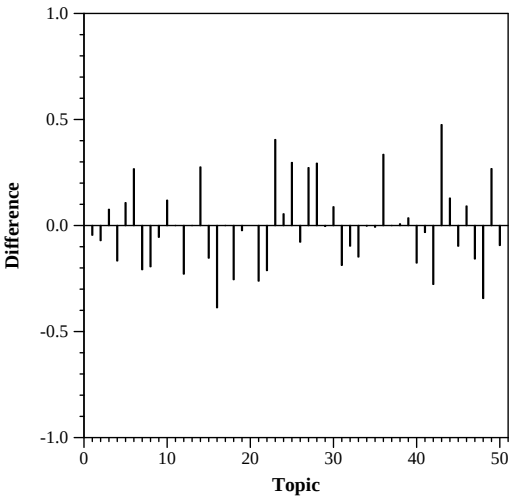
Phase 1 Summary Statistics			
RF Input Set	UCSC.1		
Total Num Rel in Set	126		
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	4	10
P(10)(all)	Full	4	10
statMAP (NEU) (all)	B	10	10
eMAP (UMass) (all)	B	8	12
Measure			Score
Score (all)			0.6176
Score (average over q)			0.5216



Score Per Topic Diff from Median, UCSC.1

Phase 1: Each group submitted a set of 5 documents per topic to be used as relevance feedback input in Phase 2 by 3 to 5 groups. One or two sets submitted. Evaluation output includes number of relevant documents in set, and how well other groups did on this set compared to the other sets that that group ran (each group ran 7 to 8 Phase 1 sets). Comparison numbers totaled among the collection and evaluation measures used in Phase 2. Total score = B / (B + W) where B is the total number of runs/measures this set did better than, and W is the number this set did worse on.

Phase 1 Summary Statistics			
RF Input Set	UCSC.2		
Total Num Rel in Set	116		
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	3	4
P(10)(all)	Full	3	4
statMAP (NEU) (all)	B	13	12
eMAP (UMass) (all)	B	12	13
Measure			Score
Score (all)			0.5156
Score (average over q)			0.4758

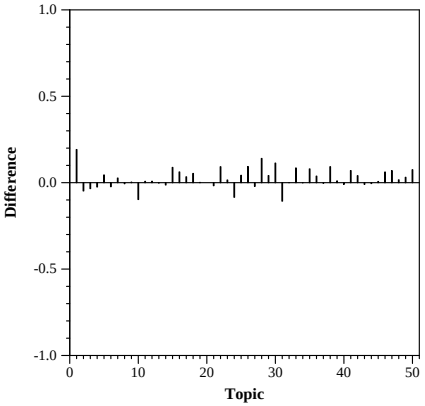


Score Per Topic Diff from Median, UCSC.2

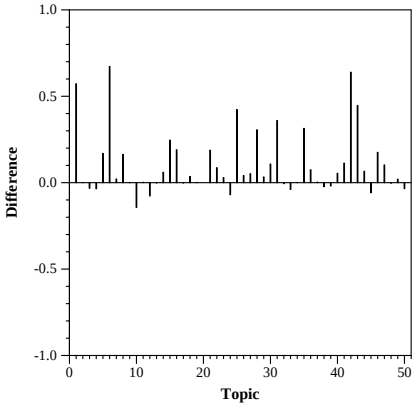
Phase 2: Each group ran with 7 to 8 different relevance feedback input documents, and ran a base case with no relevance feedback. Evaluated with two measures. If the group ran on the full collection, the measures were MAP and P(10). If the group ran on the B subset, the measures were statAP and eMAP (Million Query style evaluation).

In the Per topic Median Difference graphs, the median used is the global median measure (over all Phase 1 sets and base case) for each topic. THus it remains constant between graphs.

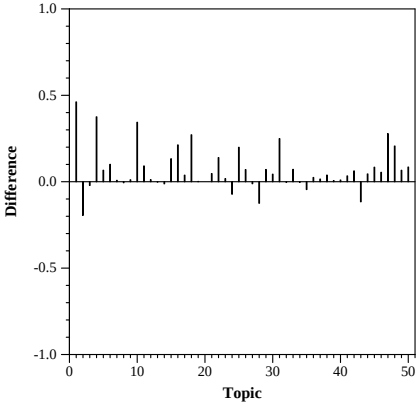
Phase 2 Run Summary Statistics		
Document Collection : B (English1 Subset)		
Run ID	statAP	eMAP
UCSC.base	0.1718	0.0455
UCSC.CMIC.1	0.2535	0.0536
UCSC.FDU.1	0.2167	0.0473
UCSC.MSRC.1	0.2119	0.0449
UCSC.UCSC.1	0.2402	0.0520
UCSC.UCSC.2	0.2478	0.0527
UCSC.UMas.2	0.2351	0.0499
UCSC.udel.2	0.2020	0.0456
UCSC.ugTr.2	0.2467	0.0522



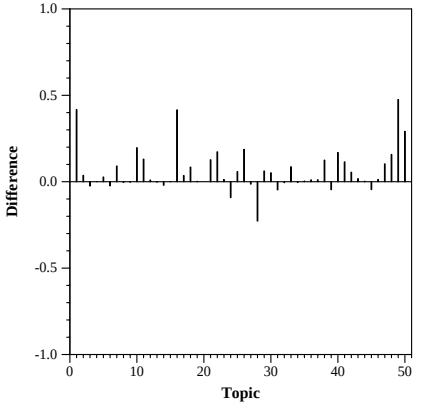
statAP Per Topic Median Diff, UCSC.base



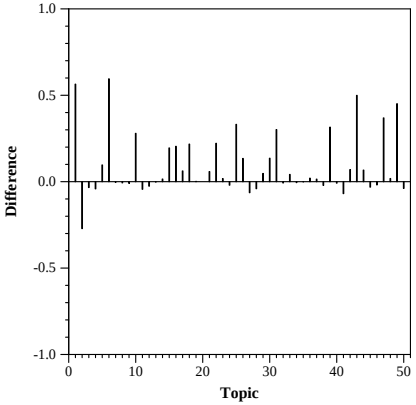
statAP Per Topic Median Diff, UCSC.CMIC.1



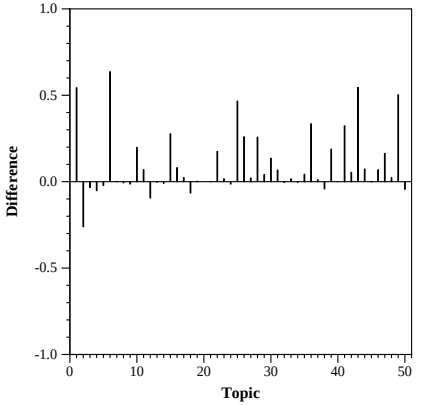
statAP Per Topic Median Diff, UCSC.FDU.1



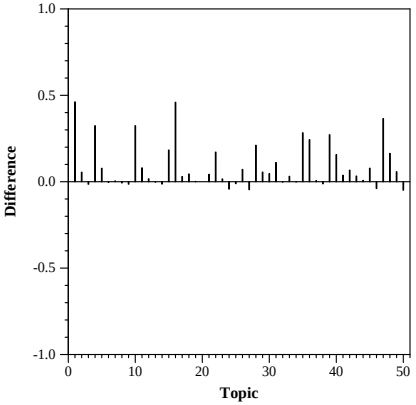
statAP Per Topic Median Diff, UCSC.MSRC.1



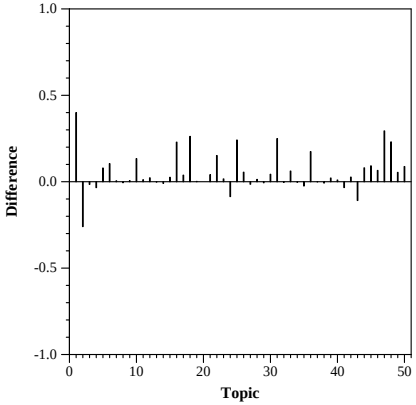
statAP Per Topic Median Diff, UCSC.UCSC.1



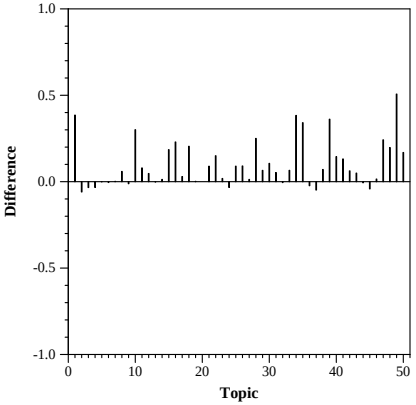
statAP Per Topic Median Diff, UCSC.UCSC.2



statAP Per Topic Median Diff, UCSC.UMas.2



statAP Per Topic Median Diff, UCSC.udel.2



statAP Per Topic Median Diff, UCSC.ugTr.2