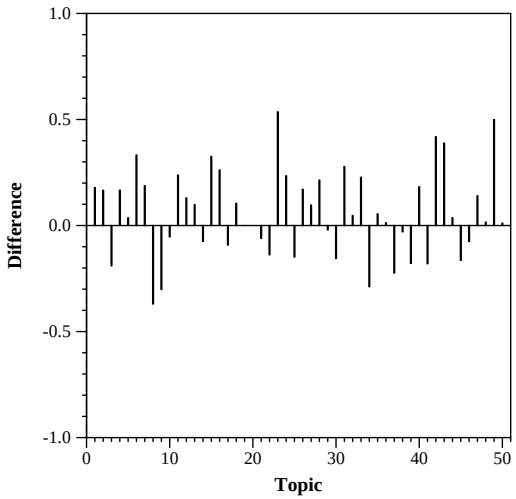
Phase 1: Each group submitted a set of 5 documents per topic to be used as relevance feedback input in Phase 2 by 3 to 5 groups. One or two sets submitted. Evaluation output includes number of relevant documents in set, and how well other groups did on this set compared to the other sets that that group 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.

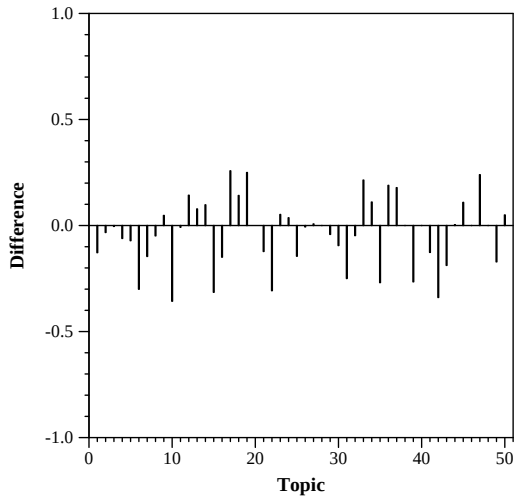
Phase 1 Summary Statistics			
RF Input Set	udel.1		
Total Num Rel in Set	99		
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	1	13
P(10)(all)	Full	1	13
statMAP (NEU) (all)	B	4	10
eMAP (UMass) (all)	B	4	10
Measure			Score
Score (all)			0.8214
Score (average over q)			0.5454



Score Per Topic Diff from Median, udel.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.

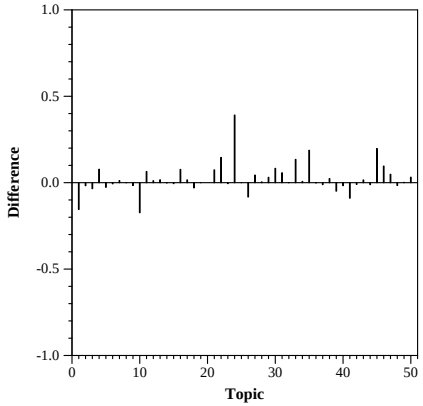
Phase 1 Summary Statistics			
RF Input Set	udel.2		
Total Num Rel in Set	79		
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	0	0
P(10)(all)	Full	0	0
statMAP (NEU) (all)	B	17	4
eMAP (UMass) (all)	B	17	4
Measure			Score
Score (all)			0.1905
Score (average over q)			0.4467



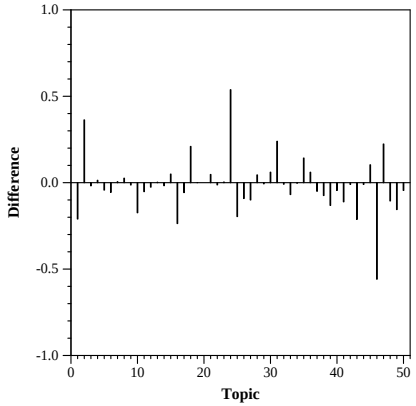
Score Per Topic Diff from Median, udel.2

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.

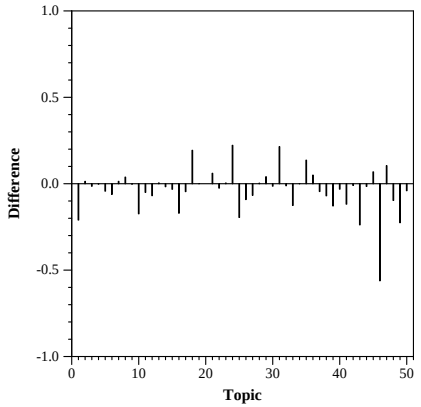
Phase 2 Run Summary Statistics		
Document Collection : B (English1 Subset)		
Run ID	statAP	eMAP
udel2.base	0.1689	0.0421
udel2.SIEL.1	0.1311	0.0355
udel2.Sab.1	0.1092	0.0328
udel2.UCSC.1	0.1720	0.0382
udel2.WatS.1	0.1387	0.0367
udel2.fub.1	0.1702	0.0377
udel2.twen.1	0.1443	0.0383
udel2.udel.1	0.1762	0.0393
udel2.udel.2	0.1480	0.0350



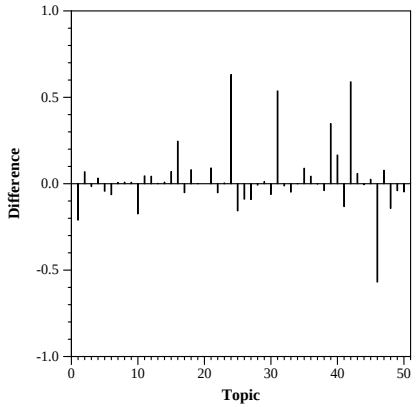
statAP Per Topic Median Diff, udel2.base



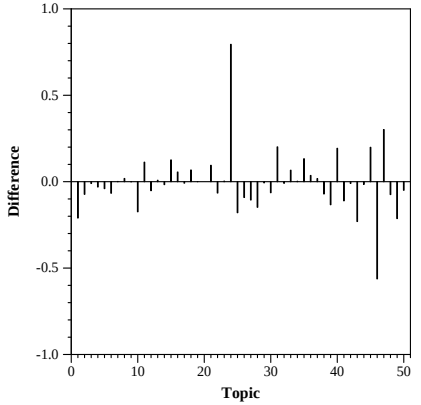
statAP Per Topic Median Diff, udel2.SIEL.1



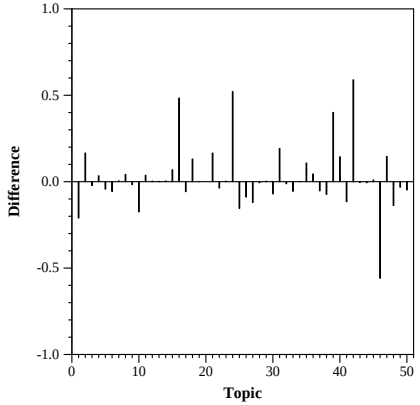
statAP Per Topic Median Diff, udel2.Sab.1



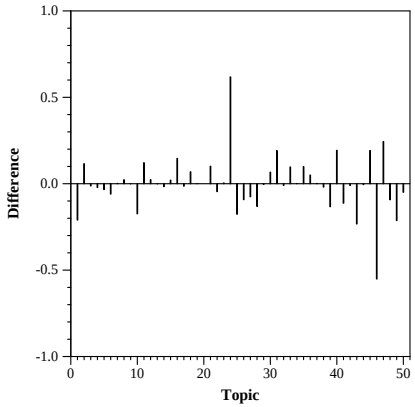
statAP Per Topic Median Diff, udel2.UCSC.1



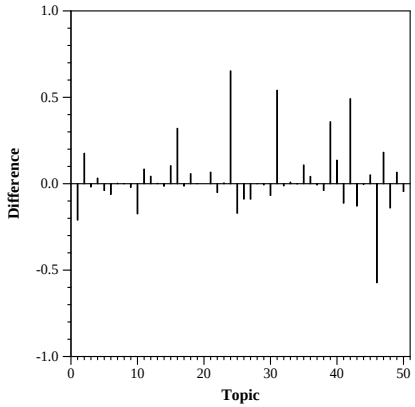
statAP Per Topic Median Diff, udel2.WatS.1



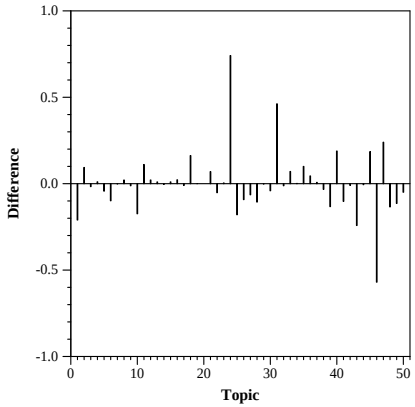
statAP Per Topic Median Diff, udel2.fub.1



statAP Per Topic Median Diff, udel2.twen.1



statAP Per Topic Median Diff, udel2.udel.1



statAP Per Topic Median Diff, udel2.udel.2