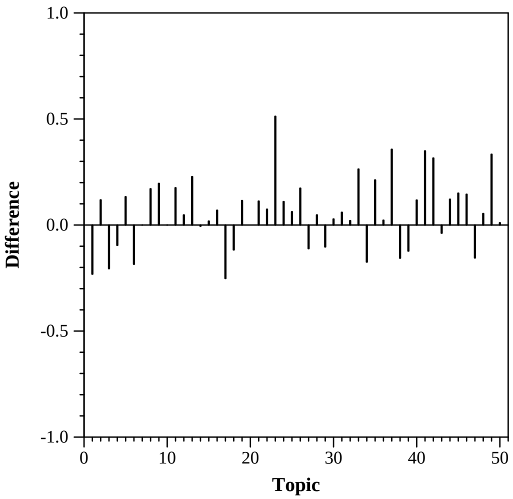
Phase 1: Each group submitted a set of 5 documents per topic to be used as relevance feedback input in Phase 2 by 3 to 5 groups. One or two sets submitted. Evaluation output includes number of relevant documents in set, and how well other groups did on this set compared to the other sets that that group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

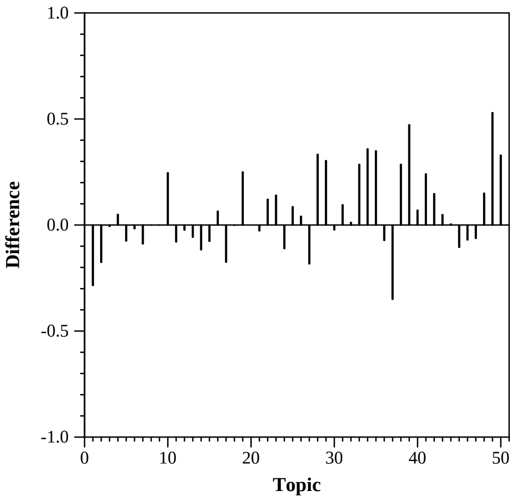
Phase 1 Summary Statistics			
RF Input Set		ugTr.1	
Total Num Rel in Set		80	
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	0	0
P(10)(all)	Full	0	0
statMAP (NEU) (all)	B	13	12
eMAP (UMass) (all)	B	10	15
Measure			Score
Score (all)			0.5400
Score (average over q)			0.5440



Score Per Topic Diff from Median, ugTr.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.

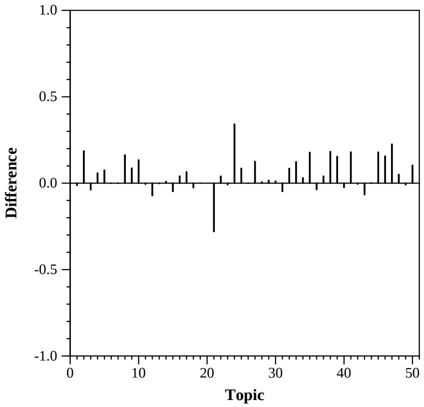
Phase 1 Summary Statistics			
RF Input Set		ugTr.2	
Total Num Rel in Set		126	
Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	0	0
P(10)(all)	Full	0	0
statMAP (NEU) (all)	B	6	15
eMAP (UMass) (all)	B	7	14
Measure			Score
Score (all)			0.6905
Score (average over q)			0.5411



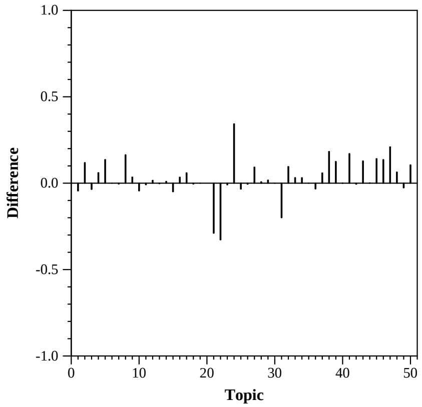
Score Per Topic Diff from Median, ugTr.2

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.

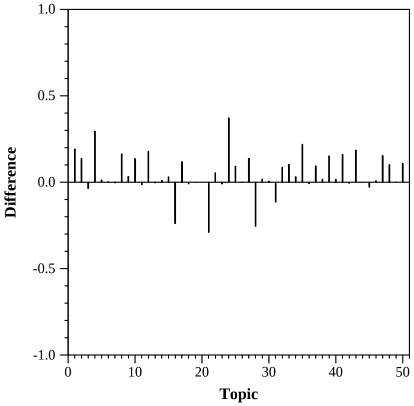
Phase 2 Run Summary Statistics		
Document Collection : B (English1 Subset)		
Run ID	statAP	eMAP
ugTr.base	0.1970	0.0461
ugTr.CMU.1	0.1764	0.0409
ugTr.UMas.1	0.1958	0.0421
ugTr.UPD.1	0.1715	0.0399
ugTr.YUIR.1	0.1900	0.0460
ugTr.hit2.1	0.1667	0.0414
ugTr.ilps.2	0.1756	0.0406
ugTr.ugTr.1	0.2081	0.0464
ugTr.ugTr.2	0.1810	0.0409



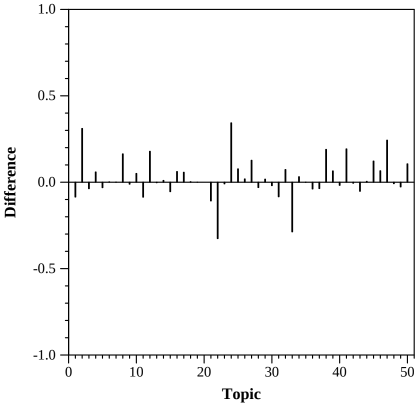
statAP Per Topic Median Diff, ugTr.base



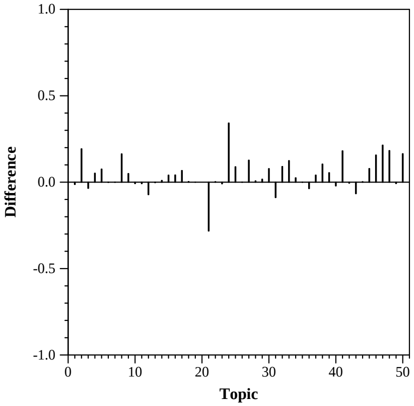
statAP Per Topic Median Diff, ugTr.CMU.1



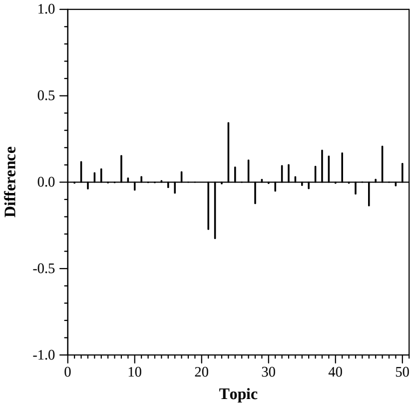
statAP Per Topic Median Diff, ugTr.UMas.1



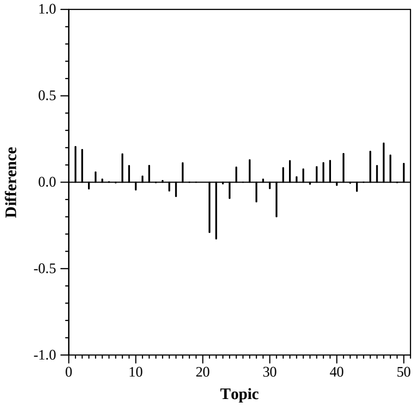
statAP Per Topic Median Diff, ugTr.UPD.1



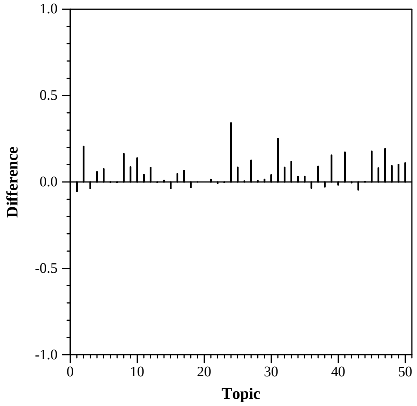
statAP Per Topic Median Diff, ugTr.YUIR.1



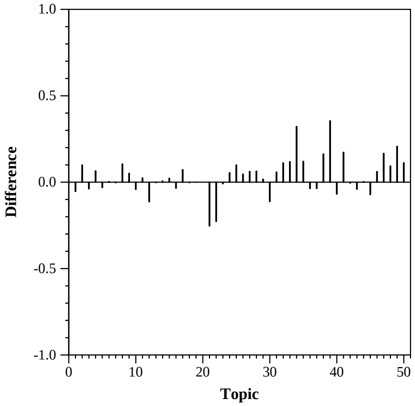
statAP Per Topic Median Diff, ugTr.hit2.1



statAP Per Topic Median Diff, ugTr.ilps.2



statAP Per Topic Median Diff, ugTr.ugTr.1



statAP Per Topic Median Diff, ugTr.ugTr.2