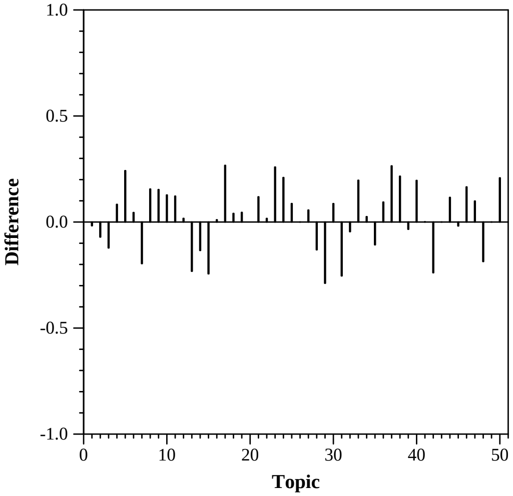
Phase 1: Each group submitted a set of 5 documents per topic to be used as relevance feedback input in Phase 2 by 3 to 5 groups. One or two sets submitted. Evaluation output includes number of relevant documents in set, and how well other groups did on this set compared to the other sets that that group 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	twen.1
Total Num Rel in Set	83

Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	7	7
P(10)(all)	Full	10	4
statMAP (NEU) (all)	B	9	10
eMAP (UMass) (all)	B	11	8

Measure	Score
Score (all)	0.4394
Score (average over q)	0.5119



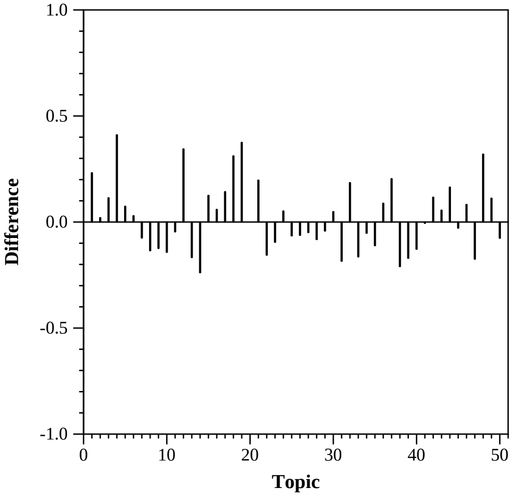
Score Per Topic Diff from Median, twen.1

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	twen.2
Total Num Rel in Set	71

Measure	Coll	Num Worse Than	Num Better Than
MAP(all)	Full	8	6
P(10)(all)	Full	7	7
statMAP (NEU) (all)	B	10	8
eMAP (UMass) (all)	B	12	6

Measure	Score
Score (all)	0.4219
Score (average over q)	0.5053



Score Per Topic Diff from Median, twen.2

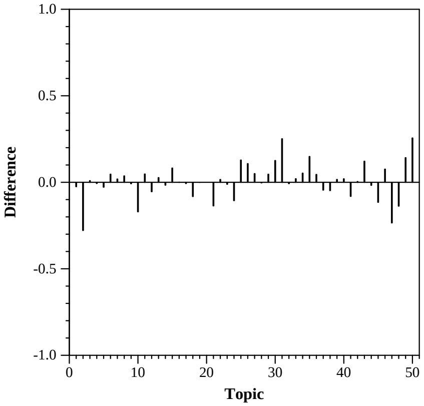
Phase 2: Each group ran with 7 to 8 different relevance feedback input documents, and ran a base case with no relevance feedback. Evaluated with two measures. If the group ran on the full collection, the measures were MAP and P(10). If the group ran on the B subset, the measures were statAP and eMAP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

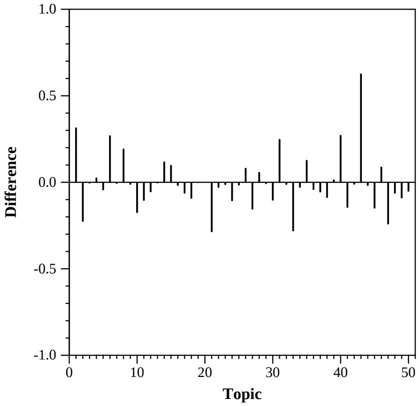
Phase 2 Run Summary Statistics

Document Collection : B (English1 Subset)

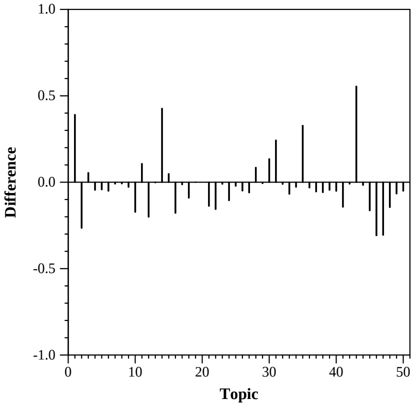
Run ID	statAP	eMAP
twen.base	0.1523	0.0382
twen.FDU.1	0.1415	0.0311
twen.UCSC.2	0.1294	0.0343
twen.YUIR.2	0.1384	0.0398
twen.ilps.1	0.1481	0.0369
twen.twen.1	0.1528	0.0317
twen.twen.2	0.1344	0.0306
twen.ugTr.1	0.1297	0.0327



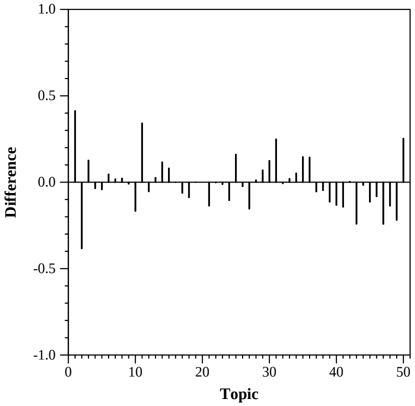
statAP Per Topic Median Diff, twen.base



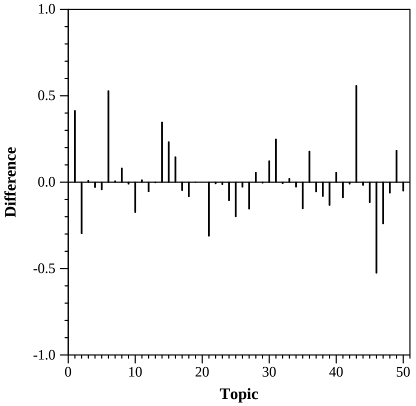
statAP Per Topic Median Diff, twen.FDU.1



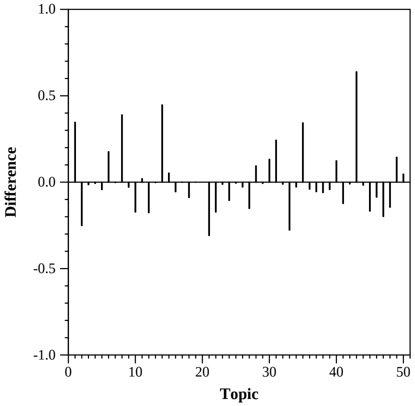
statAP Per Topic Median Diff, twen.UCSC.2



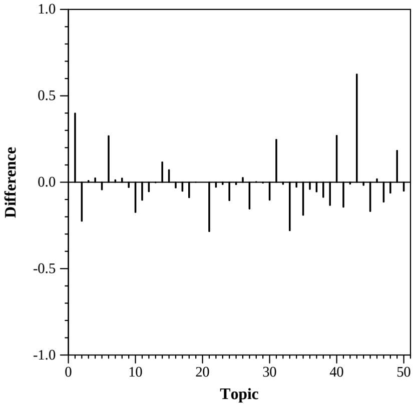
statAP Per Topic Median Diff, twen.YUIR.2



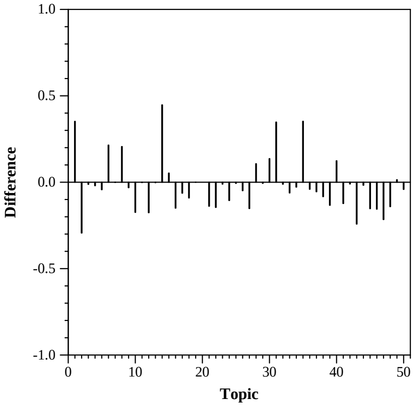
statAP Per Topic Median Diff, twen.ilps.1



statAP Per Topic Median Diff, twen.twen.1



statAP Per Topic Median Diff, twen.twen.2



statAP Per Topic Median Diff, twen.ugTr.1