TREC 2012 Crowdsourcing Track, Text Relevance Assessing Task (TRAT) results
Group: (HAC) ECS, University of Southampton
Run ID: OrcVBW16Conf

Run type: Secondary

Description of run:

Using topic analysis to select files to crowdsource, we obtained 2600 labels from Amazon Mechanical Turk workers. Independent Bayesian Classifier Combination was applied to crowdsourced labels, learning from Topic features extracted from the text. Reliability of workers is also learnt from the data and from test examples and used to weight crowdsourced labels. Confidence labels for the individual responses from the crowd are used to weight more confident responses more strongly. This version of the classifier uses middling priors that balance the prior belief that crowd members are accurate with the ability to spot the few that are not from patterns in the data. This allows us to disregard or treat as expert some of the responses from the crowd.

Results

Topic	#Docs	#Rel	TP	TN	FP	FN	TPR	TNR	FPR	FNR	LAM	AUC
411	2056	27	26	1631	398	1	0.946	0.804	0.196	0.054	0.105	0.923
416	1235	45	44	705	485	1	0.967	0.592	0.408	0.033	0.132	0.848
417	2992	75	53	2232	685	22	0.704	0.765	0.235	0.296	0.264	0.775
420	1136	37	30	712	387	7	0.803	0.648	0.352	0.197	0.268	0.762
427	1528	37	15	1219	272	22	0.408	0.817	0.183	0.592	0.363	0.667
432	2503	22	16	1724	757	6	0.717	0.695	0.305	0.283	0.294	0.763
438	1798	162	136	1016	620	26	0.837	0.621	0.379	0.163	0.256	0.775
445	1404	60	53	774	570	7	0.877	0.576	0.424	0.123	0.243	0.829
446	2020	156	135	1313	551	21	0.863	0.704	0.296	0.137	0.205	0.853
447	1588	16	6	1083	489	10	0.382	0.689	0.311	0.618	0.461	0.696
Average	1826.000	63.700	51.400	1240.900	521.400	12.300	0.751	0.691	0.309	0.249	0.259	0.789

Table 1: This table shows per-topic statistics and overall averages for the run OrcVBW16Conf. The topics are 10 randomly selected topics from the TREC 8 ad-hoc task. A relevant document is positive and a non-relevant document is negative. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) counts are based on an adjudicated set of relevance judgments that differs from the original TREC-8 ad-hoc qrels. The true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and the false negative rate (FNR) are all smoothed values. Details of the computation of the logistic average misclassification (LAM) rate and the area under the curve (AUC) are given in the track overview paper. Some runs did not report a probability of relevance and thus will have NA for their AUC score.

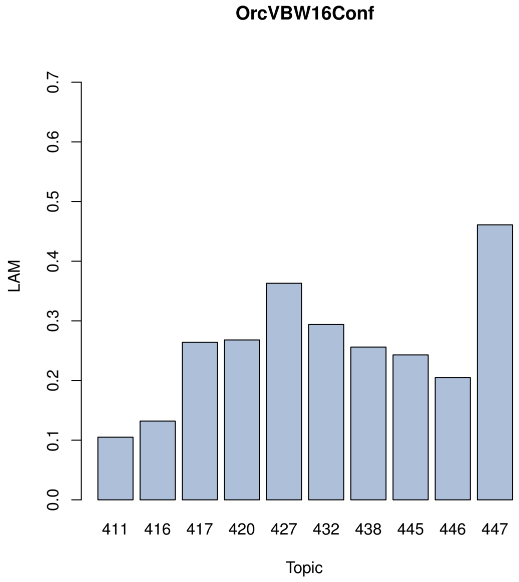

Figure 1: OrcVBW16Conf LAM

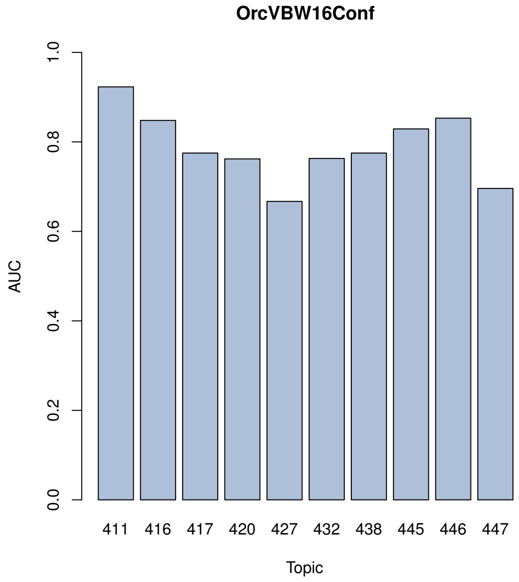

Figure 2: OrcVBW16Conf AUC