Per-Topic Scores: TREC 2009 Legal Track, Interactive Task

This Appendix reports scores, both message-based and document-based, and both pre-adjudicated and post-adjudicated, for the experimental runs submitted for the 7 test topics of the Interactive Task of the TREC 2009 Legal Track. (For more information on the task, please see the track overview paper, "Overview of the TREC 2009 Legal Track," in these proceedings.)

Also included in each table is a "fullset09i" reference run which consisted of the entire set of messages or documents in the collection.

The table headings are as follows:

- "Run" is the identifier of the experimental run.
- "Retrieved" is the number of items in the result set (an "item" is either a "message" or "document" depending on the unit of evaluation stated in the table caption). (From ":K:" in the l07_eval output.)
- "Recall" is the estimated recall of the result set (i.e., the estimated number of relevant items retrieved divided by the estimated total number of relevant items). (From ":est_K-Recall:" in the l07_eval output.)
- "Precision" is the estimated precision of the result set (i.e., the estimated number of relevant items retrieved divided by the sum of the estimated number of relevant and non-relevant items). (From ":est_K-Prec:" in the l07_eval output.)
- "F_1" is the estimated F_1 score of the result set (F_1 is 2*Precision*Recall/(Precision+Recall)). (From ":est_K-F1:" in the l07_eval output.)
- "Gray" is the estimated percentage of the result set that was gray (gray items are those for which a relevance judgment could not be determined; please see the track overview paper for more information). (From ":est_K-Gray:" in the l07_eval output.)
- "Num. Judged" is the actual number of judged items in the result set, followed in parentheses by the actual number of relevant (r), non-relevant (n) and gray (g) items. Note that because stratified sampling was used (i.e., different parts of the result sets were sampled with different draw probabilities) the estimated numbers of relevant and non-relevant items in a result set are not in general exactly proportional to the drawn numbers. (From ":K-jg_ret:", ":K-rel_ret:", ":K-nonrel_ret:" and ":K-gray_ret:" in the l07_eval output respectively.)

The table caption also reports the estimated number of relevant items in the collection for the test topic. (From ":est_rel:" in the l07_eval output.)

Version 2.6 of the l07_eval utility was used to produce the scores in this report. As an example, the command-line syntax to compute the message-based, post-adjudication scores of the fullset09i reference run was "`l07_eval run=fullset09i.msg q=qrels_msg_post_all.txt out=ignore1 out2=ignore2 out5=fullset09i.eval stringDisplay=100 M1000=7000000 probD=569034 estopt=1`". For message-based scoring, before running l07_eval, a separate utility was used to convert the input run, which was document-based, to a corresponding run of the associated messages. l07_eval reports the scores to 4 decimal places, but for this report the scores were subsequently rounded to 3 decimal places. Occasionally this sequence led to a minor roundoff error in this report, e.g., a score of 0.614454 would end up being rounded to 0.615 instead of 0.614.

As always for TREC experiments, it should be kept in mind that retrieval submissions typically are evaluating experimental techniques and do not necessarily reflect typical performance in practice. The test scenarios may have important differences from practical scenarios. Scores estimated from sampling inherently are subject to sampling error. Relevance assessments are also subject to error, even post-adjudication. Please consult the track overview paper and participant papers for more information on the research methodology and experiments conducted.

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
Clearwell09i	3542	0.061	0.399	**0.105**	0.020	472 (193r, 269n, 10g)
fullset09i	569034	**1.000**	0.042	0.080	0.037	2729 (328r, 2316n, 85g)
watlint	1330	0.036	**0.648**	0.069	0.033	211 (132r, 72n, 7g)
pittsis09	2204	0.016	0.168	0.029	0.013	283 (55r, 224n, 4g)
CGSHBCK	464	0.012	0.635	0.024	**0.042**	75 (45r, 27n, 3g)
CGSHBCK2	464	0.012	0.635	0.024	**0.042**	75 (45r, 27n, 3g)
CGSHBCK1	464	0.012	0.635	0.024	**0.042**	75 (45r, 27n, 3g)

Table 1: Message-based, Pre-Adjudication Scores for Topic 201 (22844.5 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	1330	0.778	**0.912**	**0.840**	0.024	211 (188r, 18n, 5g)
CGSHBCK	464	0.204	0.690	0.315	0.028	75 (49r, 24n, 2g)
CGSHBCK1	464	0.204	0.690	0.315	0.028	75 (49r, 24n, 2g)
CGSHBCK2	464	0.204	0.690	0.315	0.028	75 (49r, 24n, 2g)
Clearwell09i	3542	0.489	0.215	0.299	0.021	472 (116r, 346n, 10g)
pittsis09	2204	0.167	0.117	0.137	0.013	283 (42r, 237n, 4g)
fullset09i	569034	**1.000**	0.003	0.005	**0.037**	2729 (195r, 2448n, 86g)

Table 2: Message-based, Post-Adjudication Scores for Topic 201 (1523.9 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
Clearwell09i	4660	0.052	0.399	**0.091**	0.029	619 (253r, 347n, 19g)
watlint	2154	0.042	0.658	0.080	0.028	359 (229r, 120n, 10g)
fullset09i	847791	**1.000**	0.039	0.075	0.044	6455 (606r, 5605n, 244g)
CGSHBCK	875	0.018	**0.714**	0.036	**0.049**	147 (99r, 41n, 7g)
CGSHBCK1	875	0.018	**0.714**	0.036	**0.049**	147 (99r, 41n, 7g)
CGSHBCK2	875	0.018	**0.714**	0.036	**0.049**	147 (99r, 41n, 7g)
pittsis09	2204	0.010	0.164	0.020	0.013	283 (54r, 225n, 4g)

Table 3: Document-based, Pre-Adjudication Scores for Topic 201 (34243.1 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	2154	0.843	**0.937**	**0.888**	0.028	359 (327r, 22n, 10g)
CGSHBCK	875	0.290	0.806	0.426	**0.049**	147 (111r, 29n, 7g)
CGSHBCK1	875	0.290	0.806	0.426	**0.049**	147 (111r, 29n, 7g)
CGSHBCK2	875	0.290	0.806	0.426	**0.049**	147 (111r, 29n, 7g)
Clearwell09i	4660	0.498	0.276	0.355	0.029	619 (190r, 410n, 19g)
pittsis09	2204	0.104	0.117	0.110	0.013	283 (42r, 237n, 4g)
fullset09i	847791	**1.000**	0.003	0.005	0.044	6455 (341r, 5867n, 247g)

Table 4: Document-based, Post-Adjudication Scores for Topic 201 (2454.1 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
Clearwell09i	3423	0.176	0.602	**0.272**	0.030	803 (469r, 310n, 24g)
watlint	3002	0.158	**0.620**	0.251	**0.038**	714 (424r, 263n, 27g)
fullset09i	569034	**1.000**	0.021	0.040	0.030	3720 (625r, 2976n, 119g)

Table 5: Message-based, Pre-Adjudication Scores for Topic 202 (11372.9 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	3002	0.673	**0.884**	**0.764**	**0.036**	714 (607r, 81n, 26g)
Clearwell09i	3423	0.579	0.664	0.619	0.031	803 (517r, 261n, 25g)
fullset09i	569034	**1.000**	0.007	0.014	0.030	3720 (749r, 2851n, 120g)

Table 6: Message-based, Post-Adjudication Scores for Topic 202 (3801.3 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	8746	0.333	0.707	**0.453**	**0.033**	2105 (1433r, 602n, 70g)
Clearwell09i	6860	0.264	**0.728**	0.387	0.026	1595 (1131r, 423n, 41g)
fullset09i	847791	**1.000**	0.022	0.043	0.030	7435 (1743r, 5462n, 230g)

Table 7: Document-based, Pre-Adjudication Scores for Topic 202 (18250.5 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	8746	0.844	**0.934**	**0.887**	**0.033**	2105 (1899r, 136n, 70g)
Clearwell09i	6860	0.556	0.799	0.656	0.026	1595 (1242r, 312n, 41g)
fullset09i	847791	**1.000**	0.012	0.023	0.030	7435 (2097r, 5108n, 230g)

Table 8: Document-based, Post-Adjudication Scores for Topic 202 (9514.4 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	2222	0.158	0.255	**0.195**	**0.053**	272 (68r, 189n, 15g)
buffalo	9508	0.203	0.077	0.111	0.052	972 (82r, 839n, 51g)
CompEntrIT09	344	0.027	**0.283**	0.050	0.045	77 (19r, 55n, 3g)
fullset09i	569034	**1.000**	0.006	0.012	0.027	3320 (113r, 3093n, 114g)
CompCustIT09	84	0.003	0.142	0.006	0.048	25 (3r, 21n, 1g)

Table 9: Message-based, Pre-Adjudication Scores for Topic 203 (3406.1 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	2222	0.865	0.692	**0.769**	**0.053**	272 (183r, 74n, 15g)
CompEntrIT09	344	0.175	**0.895**	0.292	0.045	77 (65r, 9n, 3g)
buffalo	9508	0.592	0.111	0.186	0.052	972 (136r, 785n, 51g)
CompCustIT09	84	0.029	0.613	0.056	0.048	25 (14r, 10n, 1g)
fullset09i	569034	**1.000**	0.003	0.006	0.027	3320 (225r, 2981n, 114g)

Table 10: Message-based, Post-Adjudication Scores for Topic 203 (1684.9 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	2719	0.140	**0.232**	**0.175**	**0.054**	322 (72r, 232n, 18g)
buffalo	10016	0.164	0.070	0.098	0.050	1025 (80r, 893n, 52g)
CompEntrIT09	344	0.016	0.192	0.029	0.045	77 (13r, 61n, 3g)
fullset09i	847791	**1.000**	0.005	0.010	0.039	5710 (131r, 5296n, 283g)
CompCustIT09	84	0.000	0.000	0.000	0.048	25 (0r, 24n, 1g)

Table 11: Document-based, Pre-Adjudication Scores for Topic 203 (4058.8 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	2719	0.861	**0.645**	**0.737**	**0.054**	322 (202r, 102n, 18g)
CompEntrIT09	344	0.114	0.635	0.193	0.045	77 (46r, 28n, 3g)
buffalo	10016	0.540	0.104	0.174	0.050	1025 (136r, 837n, 52g)
CompCustIT09	84	0.011	0.244	0.020	0.048	25 (6r, 18n, 1g)
fullset09i	847791	**1.000**	0.002	0.004	0.039	5710 (248r, 5179n, 283g)

Table 12: Document-based, Post-Adjudication Scores for Topic 203 (1830.6 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
H52009	2919	0.137	**0.220**	**0.169**	**0.022**	245 (54r, 186n, 5g)
CGSHBCK	3741	0.072	0.089	0.080	0.010	306 (29r, 274n, 3g)
CGSHBCK1	3741	0.072	0.089	0.080	0.010	306 (29r, 274n, 3g)
CGSHBCK2	3741	0.072	0.089	0.080	0.010	306 (29r, 274n, 3g)
ADI2009Topic204	12748	0.122	0.045	0.065	0.018	910 (46r, 848n, 16g)
fullset09i	569034	**1.000**	0.008	0.016	0.019	3975 (92r, 3808n, 75g)

Table 13: Message-based, Pre-Adjudication Scores for Topic 204 (4569.6 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
H52009	2919	0.762	**0.844**	**0.801**	**0.022**	245 (202r, 38n, 5g)
CGSHBCK	3741	0.198	0.169	0.183	0.010	306 (59r, 244n, 3g)
CGSHBCK1	3741	0.198	0.169	0.183	0.010	306 (59r, 244n, 3g)
CGSHBCK2	3741	0.198	0.169	0.183	0.010	306 (59r, 244n, 3g)
ADI2009Topic204	12748	0.305	0.077	0.123	0.018	910 (84r, 810n, 16g)
fullset09i	569034	**1.000**	0.006	0.011	0.019	3975 (216r, 3684n, 75g)

Table 14: Message-based, Post-Adjudication Scores for Topic 204 (3162.7 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
H52009	2994	0.129	**0.225**	**0.164**	**0.021**	254 (57r, 192n, 5g)
CGSHBCK	4443	0.064	0.075	0.069	0.009	356 (29r, 324n, 3g)
CGSHBCK1	4443	0.064	0.075	0.069	0.009	356 (29r, 324n, 3g)
CGSHBCK2	4443	0.064	0.075	0.069	0.009	356 (29r, 324n, 3g)
ADI2009Topic204	21017	0.106	0.027	0.043	0.018	1466 (45r, 1395n, 26g)
fullset09i	847791	**1.000**	0.006	0.013	0.021	7289 (105r, 7024n, 160g)

Table 15: Document-based, Pre-Adjudication Scores for Topic 204 (5114.0 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
H52009	2994	0.761	**0.838**	**0.797**	**0.021**	254 (208r, 41n, 5g)
CGSHBCK	4443	0.206	0.154	0.176	0.009	356 (63r, 290n, 3g)
CGSHBCK1	4443	0.206	0.154	0.176	0.009	356 (63r, 290n, 3g)
CGSHBCK2	4443	0.206	0.154	0.176	0.009	356 (63r, 290n, 3g)
ADI2009Topic204	21017	0.301	0.048	0.083	0.018	1466 (85r, 1355n, 26g)
fullset09i	847791	**1.000**	0.004	0.008	0.021	7289 (224r, 6905n, 160g)

Table 16: Document-based, Post-Adjudication Scores for Topic 204 (3242.1 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
clearwell01	59225	0.406	0.517	**0.455**	0.037	1602 (798r, 744n, 60g)
IntegreonB	33237	0.183	0.418	0.255	**0.043**	916 (369r, 508n, 39g)
Equivio205R1	13736	0.151	**0.805**	0.254	0.012	384 (304r, 75n, 5g)
fullset09i	569034	**1.000**	0.132	0.233	0.033	3270 (1066r, 2087n, 117g)

Table 17: Message-based, Pre-Adjudication Scores for Topic 205 (72525.2 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
Equivio205R1	13736	0.463	**0.915**	**0.615**	0.012	384 (345r, 34n, 5g)
clearwell01	59225	0.673	0.321	0.434	0.049	1602 (490r, 1034n, 78g)
IntegreonB	33237	0.292	0.251	0.270	**0.061**	916 (220r, 640n, 56g)
fullset09i	569034	**1.000**	0.049	0.093	0.034	3270 (614r, 2515n, 141g)

Table 18: Message-based, Post-Adjudication Scores for Topic 205 (26839.3 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
clearwell01	79262	0.348	0.459	**0.396**	0.038	2109 (933r, 1096n, 80g)
IntegreonB	69959	0.228	0.324	0.268	**0.087**	2135 (622r, 1287n, 226g)
Equivio205R1	16989	0.129	**0.789**	0.222	0.007	455 (355r, 97n, 3g)
fullset09i	847791	**1.000**	0.125	0.221	0.042	6367 (1631r, 4289n, 447g)

Table 19: Document-based, Pre-Adjudication Scores for Topic 205 (99003.9 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
Equivio205R1	16989	0.435	**0.901**	**0.586**	0.007	455 (405r, 47n, 3g)
clearwell01	79262	0.607	0.272	0.376	0.038	2109 (554r, 1475n, 80g)
IntegreonB	69959	0.292	0.141	0.191	**0.087**	2135 (275r, 1634n, 226g)
fullset09i	847791	**1.000**	0.042	0.081	0.042	6367 (799r, 5121n, 447g)

Table 20: Document-based, Post-Adjudication Scores for Topic 205 (33614.0 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
fullset09i	569034	**1.000**	0.043	**0.082**	0.045	3397 (146r, 3074n, 177g)
CGSHBCK2	33501	0.057	0.042	0.049	0.064	782 (44r, 687n, 51g)
LogikIT09t	26675	0.037	0.034	0.036	0.065	664 (24r, 597n, 43g)
CGSHBCK1	305	0.007	0.608	0.015	**0.077**	32 (18r, 11n, 3g)
CGSHBCK	241	0.006	**0.612**	0.013	0.000	23 (15r, 8n, 0g)

Table 21: Message-based, Pre-Adjudication Scores for Topic 206 (23064.8 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
fullset09i	569034	**1.000**	0.029	**0.056**	0.046	3397 (111r, 3108n, 178g)
CGSHBCK2	33501	0.076	0.038	0.051	0.065	782 (41r, 689n, 52g)
LogikIT09t	26675	0.042	0.026	0.032	0.065	664 (19r, 602n, 43g)
CGSHBCK1	305	0.011	0.608	0.021	**0.077**	32 (18r, 11n, 3g)
CGSHBCK	241	0.009	**0.612**	0.018	0.000	23 (15r, 8n, 0g)

Table 22: Message-based, Post-Adjudication Scores for Topic 206 (15695.1 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
fullset09i	847791	**1.000**	0.043	**0.083**	0.048	7371 (235r, 6860n, 276g)
CGSHBCK2	39967	0.036	0.033	0.034	**0.052**	969 (43r, 875n, 51g)
LogikIT09t	87627	0.034	0.013	0.018	0.018	2385 (30r, 2312n, 43g)
CGSHBCK1	317	0.005	0.568	0.010	0.050	33 (18r, 13n, 2g)
CGSHBCK	252	0.004	**0.612**	0.008	0.000	23 (15r, 8n, 0g)

Table 23: Document-based, Pre-Adjudication Scores for Topic 206 (35316.5 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
fullset09i	847791	**1.000**	0.032	**0.063**	0.048	7371 (185r, 6910n, 276g)
CGSHBCK2	39967	0.045	0.030	0.036	**0.052**	969 (41r, 877n, 51g)
LogikIT09t	87627	0.040	0.011	0.017	0.018	2385 (26r, 2316n, 43g)
CGSHBCK1	317	0.006	0.568	0.013	0.050	33 (18r, 13n, 2g)
CGSHBCK	252	0.006	**0.612**	0.011	0.000	23 (15r, 8n, 0g)

Table 24: Document-based, Post-Adjudication Scores for Topic 206 (26343.7 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	7116	0.709	**0.749**	**0.728**	0.000	284 (212r, 72n, 0g)
CGSHBCK	7911	0.719	0.695	0.706	0.016	316 (216r, 95n, 5g)
CGSHBCK1	7911	0.719	0.695	0.706	0.016	316 (216r, 95n, 5g)
CGSHBCK2	7911	0.719	0.695	0.706	0.016	316 (216r, 95n, 5g)
Equivio207R1	5706	0.461	0.616	0.527	0.014	224 (137r, 84n, 3g)
LogikIT09t	25404	0.514	0.156	0.239	**0.023**	1009 (153r, 833n, 23g)
fullset09i	569034	**1.000**	0.014	0.027	0.022	3795 (278r, 3432n, 85g)

Table 25: Message-based, Pre-Adjudication Scores for Topic 207 (7526.2 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	7116	0.761	**0.907**	**0.828**	0.003	284 (256r, 27n, 1g)
CGSHBCK	7911	0.768	0.834	0.799	0.016	316 (259r, 52n, 5g)
CGSHBCK1	7911	0.768	0.834	0.799	0.016	316 (259r, 52n, 5g)
CGSHBCK2	7911	0.768	0.834	0.799	0.016	316 (259r, 52n, 5g)
Equivio207R1	5706	0.483	0.725	0.580	0.014	224 (161r, 60n, 3g)
LogikIT09t	25404	0.538	0.183	0.273	**0.023**	1009 (180r, 806n, 23g)
fullset09i	569034	**1.000**	0.015	0.030	0.022	3795 (330r, 3379n, 86g)

Table 26: Message-based, Post-Adjudication Scores for Topic 207 (8454.0 Est. Relevant Messages)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	23252	0.885	0.881	**0.883**	0.003	970 (851r, 116n, 3g)
CGSHBCK	22254	0.857	**0.889**	0.872	0.005	934 (825r, 104n, 5g)
CGSHBCK1	22254	0.857	**0.889**	0.872	0.005	934 (825r, 104n, 5g)
CGSHBCK2	22254	0.857	**0.889**	0.872	0.005	934 (825r, 104n, 5g)
Equivio207R1	20852	0.800	0.865	0.831	0.004	890 (768r, 119n, 3g)
LogikIT09t	86740	0.734	0.224	0.343	0.008	3184 (704r, 2455n, 25g)
fullset09i	847791	**1.000**	0.030	0.059	**0.037**	8658 (941r, 7377n, 340g)

Table 27: Document-based, Pre-Adjudication Scores for Topic 207 (24299.4 Est. Relevant Documents)

Run	Retrieved	Recall	Precision	F_1	Gray	Num. Judged
watlint	23252	0.896	**0.972**	**0.932**	0.005	970 (937r, 28n, 5g)
CGSHBCK	22254	0.834	0.941	0.884	0.005	934 (873r, 56n, 5g)
CGSHBCK1	22254	0.834	0.941	0.884	0.005	934 (873r, 56n, 5g)
CGSHBCK2	22254	0.834	0.941	0.884	0.005	934 (873r, 56n, 5g)
Equivio207R1	20852	0.784	0.921	0.847	0.004	890 (818r, 69n, 3g)
LogikIT09t	86740	0.723	0.240	0.360	0.008	3184 (754r, 2405n, 25g)
fullset09i	847791	**1.000**	0.033	0.064	**0.037**	8658 (1040r, 7275n, 343g)

Table 28: Document-based, Post-Adjudication Scores for Topic 207 (26419.9 Est. Relevant Documents)