Overview of the TREC 2010 Legal Track

Gordon V. Cormack, gvcormac@plg.uwaterloo.ca
University of Waterloo, Ontario N2L 3G1, Canada

Maura R. Grossman, mrgrossman@wlrk.com
Wachtell, Lipton, Rosen & Katz
51 West 52nd Street, New York, NY 10019, USA

Bruce Hedin, bhedin@h5.com
H5, 71 Stevenson St., San Francisco, CA 94105, USA

Douglas W. Oard, oard@umd.edu
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA

Abstract

TREC 2010 was the fifth year of the Legal Track, which focuses on evaluation of search technology for discovery of electronically stored information in litigation and regulatory settings. The TREC 2010 Legal Track consisted of two distinct tasks: the *Learning task*, in which participants were required to estimate the probability of relevance for each document in a large collection, given a *seed set* of documents, each coded as responsive or non-responsive; and the *Interactive task*, in which participants were required to identify all relevant documents using a human-in-the-loop process.

1 Introduction

We are concerned with the document selection and review component of the *e-discovery* process, for which the objective is to identify as nearly as practicable all documents from a collection that are responsive to a *request for production* in civil litigation, while minimizing the number of non-responsive documents that are identified by the method. The Learning and Interactive tasks of the TREC 2010 Legal Track represent two different e-discovery scenarios:

- The Learning task represents the scenario in which preliminary search and assessment has yielded a set of documents that are coded as relevant or not; this *seed set* is then used as input to a process involving humans or technology to estimate the *probability* that each of the remaining documents in the collection is relevant.
- The Interactive task represents the process of using humans and technology, in consultation with a *Topic Authority*, to identify as well as possible *all relevant documents* in the collection, while simultaneously minimizing the number of false positives.

The Learning task derives from the TREC 2009 *Batch* task, while the Interactive task reprises the TREC 2009 Interactive task, with three new requests for production, and, in addition, a *privilege review* for which the objective is to identify documents that may be withheld from production because of attorney-client privilege or work-product protection.

For the document collection, both tasks used a newly processed variant of the Enron email dataset, containing about 1.3 million email messages captured by the Federal Energy Review Commission (FERC) from Enron, in the course of its investigation of Enron's collapse. The Learning task reused the seven requests for production (topics 201 through 207) from the TREC 2009 Interactive task (which had used a different variant of the Enron email dataset), and also one novel topic (topic 200). The Interactive task used three novel topics (topics 301, 302 and 303), in addition to a privilege review (topic 304).

Seventeen teams participated in the TREC 2010 Legal Track, as detailed in Table 1. Eight of the teams participated in the Learning task, while twelve participated in the Interactive task; three participated in both tasks. The detailed results given in the following sections identify each teams' results by a "run identifier" whose prefix is given in Table 1.

Run Identifier Prefix		
Learning Task	Interactive Task	Participating Group
	CS	Clearwell Systems Inc.
	SF	University of South Florida, IS/DS Department
	IS	Indian Statistical Institute, Kolkata
ITD	IT	IT.com, Inc.
	UW	University of Waterloo (Clarke)
	IN	Integreon Discovery Solutions
rmit	UM	RMIT University and University of Melbourne
	LA	Los Alamos National Laboratory
	MM	Waterford Technologies (MailMeter)
	EQ	Equivio
	UB	University at Buffalo, State University of New York
Bck	CB	Backstop LLP and Cleary Gottlieb Steen and Hamilton LLP
DUTH		Democritus University of Thrace, Greece
tcd		TCDI
xrce		XEROX
ot		Open Text Corporation
URSK		Ursinus College

Table 1: Run tags and group names for TREC 2010 Learning and Interactive task participants.

2 Document Collection

The document collection for both tasks was derived from the EDRM Enron Dataset, version 2, prepared by ZL Technologies in consultation with the Legal Track coordinators, and hosted by EDRM.[1] ZL acquired the full collection of 1.3 million Enron email messages from Lockheed Martin (formerly Aspen Systems) who captured and maintain the dataset on behalf of FERC. The EDRM dataset is available in two formats: EDRM XML and PST. The EDRM XML version contains a text rendering of each email message and attachment, as well as the original native format. The PST version contains the same messages, in a Microsoft proprietary format used by many commercial tools.

Both versions of the dataset approach 100GB in size, presenting an obstacle to participants. Furthermore, there are a large number of duplicate email messages in the dataset, that were captured more than once by Lockheed Martin. For TREC 2010, a list of 455,449 distinct messages were identified as canonical; all other messages duplicate one of the canonical messages. These messages contain about 230,143 attachment files; together these 455,449 messages plus 230,143 attachments form the 685,592 documents of the TREC 2010 Legal Track collection used for both the Learning and Interactive tasks. Text and native versions of these

[1] http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set-v2

documents were made available to participants, along with a mapping from the EDRM XML and PST files to their canonical counterparts in the TREC collection.

3 Relevance Assessments

In order to measure the efficacy of TREC participant efforts in the two tasks, it is necessary to compare their results to a *gold standard* indicating whether or not each document in the collection is relevant to a particular discovery request. The Learning task used eight distinct discovery requests, while the Interactive task used four. Ideally, a gold standard would indicate the relevance of each document to each topic, a total of eight million judgments.

It is impractical to use human assessors to render these eight million assessments. Instead, a sample of documents was identified for each topic, and assessors were asked to code only the documents in the sample as relevant or not. For the Learning task, 78,000 human assessments were used; for the Interactive task, 50,000 human assessments were used.

The Learning task assessments were rendered by individual volunteers, primarily, but not exclusively, law students. For each document and topic, three binary assessments were rendered, and the majority opinion was taken to be the gold standard. The Interactive task assessments were assessed by professional review companies. Ten percent of the documents for each topic were assigned to more than one reviewer; the agreement among these redundant assessments was used to estimate and correct for assessor error. In both cases, individual assessors were asked to review documents in batches of 500, and reviewed one or more batches.

The assessors used a new Web-based platform developed by the coordinators to view the documents and to record their relevance judgments. To avoid problems with local rendering software on each assessor's workstation, the assessors made their judgments based on pdf-formatted versions of the documents, as opposed to the original native format documents.

Assessors were provided with orientation and detailed guidelines created by a Topic Authority. For the Learning task, assessors were given 10 examples each of relevant and a non-relevant documents for their particular topic. The review platform included a "seek assistance" link which assessors were encouraged to use to request that the Topic Authority respond to questions to resolve any uncertainty that may have arisen as to particular documents.

In reviewing their bins, assessors were instructed to make a relevance judgment of relevant (R), not relevant (N), or broken (B) for every document in their bins. The latter code reflects the fact that a small percentage of documents (1.25%) were malformed and therefore could not be assessed.

4 Learning Task

For each of the eight topics, participants in the Learning Task were given a *seed set* – a list of documents within the collection, each coded as *relevant* or *not relevant* to the topic. For topic 200, which was new to TREC 2010, the seed set was constructed by TREC coordinators, using an interactive search and judging process. For topics 201 through 207, the seed sets were derived indirectly from the relevance assessments used to evaluate the TREC 2009 Interactive Task. It was not possible to used the relevance assessments directly, as TREC 2009 and TREC 2010 used different versions of the Enron dataset. The coordinators employed an approximate match strategy to find analogues to the TREC 2009 documents in the TREC 2009 dataset. Only documents with high similarity were used as seeds; the remainder were disregarded. Table 2 shows the number of relevant and non-relevant documents in each seed set.

The Learning task models the use of automated or semi-automated methods to guide review strategy for a multi-stage document review effort, organized as follows:

1. **Preliminary search and assessment.** The producing party analyzes the production request. Using ad hoc methods the team identifies a *seed set* of potentially responsive documents, and assesses each as responsive or not.

Topic	Relevant	Not Relevant	Total
200	230	621	851
201	168	523	691
202	1006	403	1409
203	67	892	959
204	59	1132	1191
205	333	1506	1839
206	19	336	355
207	80	511	591

Table 2: Seed set sizes for the TREC 2010 Learning Task.

2. **Learning by example.** A learning method is used to rank the documents in the collection from most to least likely to be responsive to the production request, and to estimate this likelihood for each document. The input to the learning method consists of the seed set, the assessments for the seed set, and the unranked collection; the output is a ranked list consisting of the document identifier and a probability of responsiveness for each document in the collection.

   The two components of learning by example – ranking and estimation – may be accomplished by the same method or by different methods. Either may be automated or manual. For example, ranking might be done using an information retrieval method or by human review using a five-point scale. Estimation might be done in the course of ranking or, for example, by sampling and reviewing documents at representative ranks.

3. **Review process.** A review process may be conducted, with strategy guided by the ranked list. One possible strategy is to review documents in order, thus discovering as many responsive documents as possible for a given amount of effort. Another possible strategy is triage: to review only mid-ranked documents, deeming, without further review, the top-ranked ones to be responsive, and the bottom-ranked ones to be non-responsive. Review strategy may be guided not only by the order of the ranked list, as outlined above, but also by the estimated effectiveness of various alternatives. Consider the strategy of reviewing the top-ranked documents. Where should a *cut* be made so that documents above the cut are reviewed and documents below are not? For triage, where should the two necessary cuts be made?

Using the seed set for each topic, participants were required to submit an estimate of the probability of relevance for every document in the collection. That is, each submitted run contained 5,484,736 probability estimates – 8 per document, and 685,592 per topic. Practically every review strategy decision boils down to the question,

For some particular set of documents, how many are responsive and how many are not?

To answer this question it suffices to answer the more detailed question:

What is the probability of each document in the set being relevant?

Given an answer to the second question, the answer to the first is simply the sum of the probabilities. For this reason, participants in the Learning task were required to provide an estimate of the probability of relevance for each document in the collection. The optimal relevance ranking follows from the probability estimate. If the probability estimate is accurate, documents with a higher probability are more likely to be relevant. At any given rank k, the expected number of relevant documents up to and including rank k is the sum of their probabilities, and this sum is maximized when the documents with the highest probabilities are given the highest ranks.

Stratum	Stratum Size	Sample Size	Sampling Rate
100	1063.0	1063.0	1.0
1000	7813.3	551.9	0.07
10000	73182.0	551.9	0.007
1000000	603533.6	553.2	0.0009

Table 3: Average stratum sizes, sample sizes, and sampling rates, over all topics, for Learning Task evaluation.

Furthermore, if the probability estimate is reasonable, the sum is itself an accurate estimate of the number of relevant documents among the top k. This estimate may be used to guide review strategy, as it allows the legal team to evaluate the tradeoff between review effort and the number of documents retrieved.

The number of relevant documents retrieved, as a fraction of the number of relevant documents in the collection, is known as *recall*. In document production for civil litigation, recall is typically more important than *precision,* the fraction of retrieved documents that are relevant. The major question to be answered in e-discovery is: if we examine the top-ranked k documents, what recall will be achieved? The answer to this question is encoded in the probability estimates submitted by TREC participants.

The TREC Legal Track evaluation process provides a post-hoc answer to that question, against which the participants' efforts may be compared. The *accuracy* of the estimate is defined to be

$$accuracy = 100\% \times \frac{\min(estimate,\, true\, value)}{\max(estimate,\, true\, value)}\,.$$

4.1 Relevance Assessment and Evaluation

For each submitted run, the Learning Task evaluation process uses stratified sampling and redundant assessment to count, for each possible $1 \leq k \leq 650000$, the number of responsive documents within the k highest-ranked within the run. From these counts are derived point estimates for recall at each possible cutoff value, as well as summary estimates of retrieval effectiveness over all cutoff values.

For each topic, each document in a stratified sample of 2,720 documents was assessed by three independent volunteer reviewers. The majority opinion of these three assessors was taken to be ground truth, and used as the gold standard against which the submitted runs were evaluated. Each reviewer had legal training; the majority were third-year law students who received pro bono credits from their academic institution.

The four strata whose sizes are detailed in Table 3, were defined as follows. The first stratum (100) consisted of any document that was ranked within the top 100 of any of the 20 runs. That is, stratum 100 was constructed by the TREC *pooling method*, with pool depth 100. The second stratum (1000) was also constructed using the pooling method, with pool dept 1000, and excluding all documents in the first stratum. The third stratum (10000) used a pool depth of 10000, excluding prior strata, while the last stratum (1000000) consisted of the entire corpus, excluding those documents in the first three strata.

4.2 Results

Eight participating groups submitted 20 runs. For each topic within each run, the number of relevant documents retrieved (and hence recall) was computed for all possible cutoff values k. That is, for each k between 1 and 650,000, the actual and estimated number of relevant documents was determined. Table 4 shows the estimated number of relevant documents for each topic, and Tables 5 through 8 show the resulting recall values for each run and each topic at four representative values of k: 20,000 (3% of the collection), 50,000 (7.5% of the collection), 100,000 (15% of the collection), and 200,000 (30% of the collection). In addition, the tables show the average recall over all topics, the estimated average recall over all topics, and the accuracy of the estimated average. From Table 5 we see that the best-performing system identifies 49.8% of all relevant documents within the top-ranked 3% of documents in the collection. Table 6 shows that the same system identifies 63.5% of all relevant documents within the top-ranked 7.5%. In terms of document review strategy, this indicates that a review team would have to examine two-and-a-half times as many

Topic	Estimate	95% C.I.
200	2,544	(479, 4608)
201	1,886	(1181, 2591)
202	6,312	(3793, 8832)
203	3,125	(2069, 4180)
204	6,362	(2786, 9937)
205	67,938	(53563, 82313)
206	866	(439, 1293)
207	20,929	(16256, 25603)

Table 4: Estimated total number of relevant documents (C.I.=Confidence Interval).

documents (but still only 7.5% of the entire collection) to find 30% more documents. Table 7 shows that the system identifies 74.3% of the relevant documents within the top-ranked 15% of the collection, while Table 8 shows that the system identifies 84.3% of the relevant documents within the top-ranked 30% of the collection.

It is apparent that, as expected, recall increases as the cutoff k increases, in a nonlinear fashion. The tradeoff between relevant documents retrieved and non-relevant documents retrieved at rank k may be expressed as a Receiver Operating Characteristic Curve, also known as a recall-fallout curve [4]. An ROC curve plots the fraction of relevant documents retrieved (recall) as a function of the fraction of non-relevant documents retrieved (fallout). While ROC curves are ubiquitous in signal detection and diagnostic test theory, recall-fallout curves have largely been supplanted by recall-precision curves in much of the work on information retrieval evaluation because of the emphasis of recall-precision curves on precision at early ranks. For e-discovery, ROC curves better illustrate system effectiveness at high recall levels.

Figures 1 through 4 show the ROC curves for each of the eight topics used in the Learning task. The top-performing run for each participant is plotted in the set of graphs for each topic. Figures 5 through 8 show the ROC curves for the top-performing run (according to AUC, see below) for each participant. Each graph in those figures shows the ROC curves for each topic. Note that the curves are plotted on a *logit* scale.[2]

A nearly perfect system generates a concave curve that rises steeply, rapidly approaching the top-left corner of the graph, then continues to the top-right corner; a random ranking yields a straight line along the main diagonal. In general, a superior curve represents superior effectiveness. A common summary measure of the height of the curve is the *area under the ROC curve (AUC)*. AUC is a number between 0 and 1, where 1 indicates perfection, and 0.5 indicates a random ranking. Table 9 shows the AUC results for every topic within every run and, in addition, the average AUC over all topics for each run.

4.3 Evaluating F_1

The end goal of the discovery process is to produce a *set* of documents that are responsive to the request, not simply a prioritized list. To convert a a prioritized list to a set, it is necessary to choose a particular cutoff value k, and to include in the production set only the top-ranked k documents. The ideal set would, of course, include all the responsive documents and none of the non-responsive ones. In general, this ideal is impossible to realize: for any real ranking and for any k, there will always be some relevant documents that are not in the top-ranked k, or some non-relevant documents in the top k, or both. The challenge for evaluation, then, is to measure how close to ideal any particular set of produced documents is.

F_1 , used as the principal measure of effectiveness by the Interactive Task, is defined to be the harmonic mean of recall and precision:

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \ .$$

[2] $\text{logit}(x) = \log \frac{x}{1-x}$.

	Topic								Avg		
Run	200	201	202	203	204	205	206	207	Actual	Est	Acc
xrceLogA	16.3	47.2	82.7	60.5	29.8	24.7	49.4	87.5	49.8	51.7	96.2
xrceCalA	16.3	47.2	82.7	60.5	29.8	24.7	49.4	87.5	49.8	63.8	78.0
otL10bT	45.3	81.0	49.0	55.7	13.7	18.3	72.5	15.4	43.9	68.8	63.8
BckExtA	30.9	63.7	53.9	54.2	14.7	25.2	25.9	71.9	42.5	8.0	18.7
BckBigA	32.1	63.8	53.7	52.1	14.6	25.2	25.9	71.9	42.4	7.6	18.0
rmitindA	51.6	53.5	37.9	60.1	21.5	21.2	55.7	16.5	39.8	10.6	26.6
otL10FT	35.1	32.3	49.3	69.9	14.5	24.6	64.6	15.2	38.2	38.6	99.0
xrceNoRA	15.1	39.9	63.5	45.7	18.7	24.5	39.8	45.5	36.6	38.8	94.3
BckLitA	20.5	60.9	54.1	69.9	11.1	23.3	44.7	3.2	36.0	7.6	21.0
otL10rvlT	4.2	51.3	25.2	29.7	13.0	16.1	77.8	61.5	34.9	42.0	83.0
DUTHsdtA	17.7	55.3	39.4	55.4	6.7	16.0	36.1	12.5	29.9	86.5	34.6
DUTHsdeA	17.7	55.3	39.4	55.4	6.7	16.0	36.1	12.5	29.9	43.0	69.4
DUTHlrgA	17.7	55.3	39.4	55.4	6.7	16.0	36.1	12.5	29.9	69.5	43.0
rmitmlsT	2.3	16.4	40.1	32.2	16.3	18.6	48.5	15.1	23.7	14.5	61.2
URSK70T	15.2	15.4	8.8	33.2	26.0	6.0	56.8	12.3	21.7	37.8	57.4
URSLSIT	15.2	16.0	5.1	33.2	27.3	6.0	56.8	13.1	21.6	26.2	82.3
rmitmlfT	3.1	16.4	44.8	17.3	7.1	21.0	49.7	7.9	20.9	13.8	66.0
ITD	0.0	28.0	74.7	7.2	11.1	14.8	4.3	13.2	19.2	11.4	59.6
URSK35T	16.3	13.1	9.5	26.4	6.5	3.2	51.8	9.3	17.0	25.9	65.6
tcd1	2.1	1.3	0.2	0.0	19.7	0.0	11.5	3.8	4.8	9.2	52.2

Table 5: Recall (%) at *k=20,000 (3% cut).*

	Topic								Avg		
Run	200	201	202	203	204	205	206	207	Actual	Est	Acc
xrceLogA	27.9	65.0	88.2	73.3	44.2	39.3	78.2	92.2	63.5	68.3	93.0
xrceCalA	27.9	65.0	88.2	73.3	44.2	39.3	78.2	92.2	63.5	71.7	88.7
rmitindA	69.2	79.3	59.4	79.9	57.2	41.8	76.7	18.5	60.2	20.7	34.3
otL10bT	46.7	80.7	58.0	84.2	26.9	38.8	99.4	15.4	56.3	88.4	63.6
otL10rvlT	21.9	69.0	52.7	56.2	40.6	28.7	82.8	76.6	53.6	71.9	74.5
otL10FT	81.6	50.7	63.3	87.4	20.7	38.3	69.6	17.3	53.6	63.7	84.1
BckExtA	34.5	69.7	67.8	60.3	41.7	37.9	37.8	73.4	52.9	16.4	31.1
BckBigA	35.7	69.8	67.6	55.8	41.7	37.9	37.8	73.4	52.5	16.1	30.7
xrceNoRA	25.8	65.4	65.7	61.5	22.1	37.8	49.0	90.8	52.3	50.1	95.8
DUTHsdtA	29.7	80.4	58.1	74.1	22.8	40.2	67.9	13.8	48.4	87.1	55.5
DUTHsdeA	29.7	80.4	58.1	74.1	22.8	40.2	67.9	13.8	48.4	55.6	86.9
DUTHlrgA	29.7	80.4	58.1	74.1	22.8	40.2	67.9	13.8	48.4	87.0	55.6
BckLitA	32.1	68.8	63.8	76.2	13.8	36.5	49.0	4.2	43.0	16.7	38.8
rmitmlsT	3.9	23.9	55.8	52.9	21.5	29.3	55.7	15.4	32.3	29.5	91.3
rmitmlfT	9.6	24.0	53.4	20.5	18.7	28.1	56.0	11.9	27.8	28.0	99.0
URSK70T	19.6	17.2	10.3	39.4	32.8	12.2	73.6	15.6	27.6	60.1	45.9
URSLSIT	19.6	18.1	14.4	39.4	28.5	12.2	73.6	14.4	27.5	46.3	59.4
URSK35T	27.6	15.1	12.0	35.9	21.0	6.3	56.1	13.6	23.5	49.1	47.8
ITD	0.0	28.8	82.2	9.6	12.1	20.1	4.4	30.5	23.5	22.8	97.4
tcd1	15.7	15.9	0.4	12.0	23.1	5.3	38.5	10.8	15.2	20.0	76.2

Table 6: Recall (%) at *k=50,000 (7.5% cut).*

	Topic								Avg		
Run	200	201	202	203	204	205	206	207	Actual	Est	Acc
xrceLogA	33.7	94.5	91.0	82.5	69.9	51.4	78.5	93.0	74.3	79.0	94.1
xrceCalA	33.7	94.5	91.0	82.5	69.9	51.4	78.5	93.0	74.3	76.9	96.7
otL10rvlT	29.4	86.7	60.5	76.2	65.5	58.4	100.6	82.7	70.0	90.3	77.6
rmitindA	67.3	88.2	68.4	95.9	74.5	58.1	80.4	18.8	69.0	33.5	48.5
otL10FT	99.8	74.4	73.5	93.7	37.5	61.8	70.5	18.8	66.2	83.6	79.2
BckBigA	42.3	73.4	70.1	60.4	64.8	56.9	61.2	73.1	62.8	28.3	45.1
BckExtA	41.0	73.4	70.1	59.9	64.8	56.8	61.2	73.1	62.5	28.5	45.6
DUTHsdtA	41.1	85.6	70.0	85.5	65.3	58.0	72.9	16.3	61.8	88.1	70.2
DUTHsdeA	41.1	85.6	70.0	85.5	65.3	58.0	72.9	16.3	61.8	68.2	90.7
DUTHlrgA	41.1	85.6	70.0	85.5	65.3	58.0	72.9	16.3	61.8	93.1	66.4
otL10bT	45.5	81.0	63.7	83.2	36.0	55.5	97.7	24.8	60.9	97.3	62.6
xrceNoRA	25.8	64.8	70.0	66.8	31.7	49.4	72.7	85.7	58.4	60.2	96.9
BckLitA	40.9	72.5	72.8	88.4	30.6	56.1	60.3	8.9	53.8	29.2	54.3
rmitmlsT	15.9	42.4	67.1	57.4	37.0	36.9	59.2	15.3	41.4	47.3	87.6
tcd1	20.9	35.1	16.9	40.4	45.9	22.0	74.2	70.5	40.7	34.3	84.2
rmitmlfT	23.5	42.3	65.6	28.1	39.9	36.6	58.0	13.3	38.4	44.9	85.5
URSLSIT	31.0	19.8	19.4	42.1	43.6	13.3	74.7	20.9	33.1	65.4	50.6
URSK70T	31.0	19.0	10.7	42.1	40.8	13.3	74.7	17.1	31.1	78.3	39.7
URSK35T	36.2	16.2	13.3	37.0	33.9	13.2	76.5	16.0	30.3	72.8	41.6
ITD	0.0	34.2	86.7	11.2	18.2	29.2	8.3	53.9	30.2	36.3	83.2

Table 7: Recall (%) at *k=100,000 (15% cut).*

	Topic								Avg		
Run	200	201	202	203	204	205	206	207	Actual	Est	Acc
xrceLogA	77.9	97.1	97.8	91.3	73.8	66.7	77.8	92.0	84.3	88.9	94.8
xrceCalA	77.9	97.1	97.8	91.3	73.8	66.7	77.8	92.0	84.3	82.8	98.2
otL10FT	97.9	89.2	96.6	97.7	68.8	81.1	88.4	21.7	80.2	96.6	83.0
DUTHsdtA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6	90.1	88.3
DUTHsdeA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6	82.5	96.5
DUTHlrgA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6	96.3	82.6
otL10rvlT	39.8	88.3	64.5	83.4	85.2	82.9	99.6	86.6	78.8	98.9	79.7
BckExtA	78.9	74.0	75.4	71.4	67.5	75.4	85.0	80.9	76.1	49.7	65.4
BckBigA	80.7	74.1	75.4	66.4	67.4	75.4	85.0	80.9	75.7	49.6	65.5
rmitindA	72.9	92.3	72.5	98.0	79.2	85.0	80.9	19.8	75.1	53.4	71.1
tcd1	67.2	55.3	85.0	76.1	76.2	53.3	98.8	87.4	74.9	55.9	74.6
xrceNoRA	83.2	73.5	76.1	79.7	35.2	58.7	78.9	92.0	72.2	73.3	98.5
otL10bT	52.4	88.4	67.8	84.5	49.5	65.6	98.3	51.1	69.7	99.2	70.3
BckLitA	44.1	74.9	75.7	85.4	42.5	77.8	63.1	11.7	59.4	49.7	83.6
rmitmlsT	66.6	58.6	72.2	64.5	45.6	54.2	61.8	16.3	55.0	70.7	77.8
rmitmlfT	68.7	57.2	70.4	47.6	47.5	52.8	62.3	15.7	52.8	67.1	78.6
ITD	0.0	44.8	88.3	19.5	41.6	36.1	26.4	74.7	41.4	54.1	76.5
URSK70T	51.0	18.9	13.0	44.5	62.2	24.6	88.9	22.6	40.7	91.0	44.7
URSLSIT	51.0	21.0	21.1	44.5	50.6	24.6	88.9	22.5	40.5	83.5	48.5
URSK35T	51.3	25.1	15.2	45.1	40.5	27.7	91.8	18.3	39.4	93.3	42.2

Table 8: Recall (%) at *k=200,000 (30% cut).*

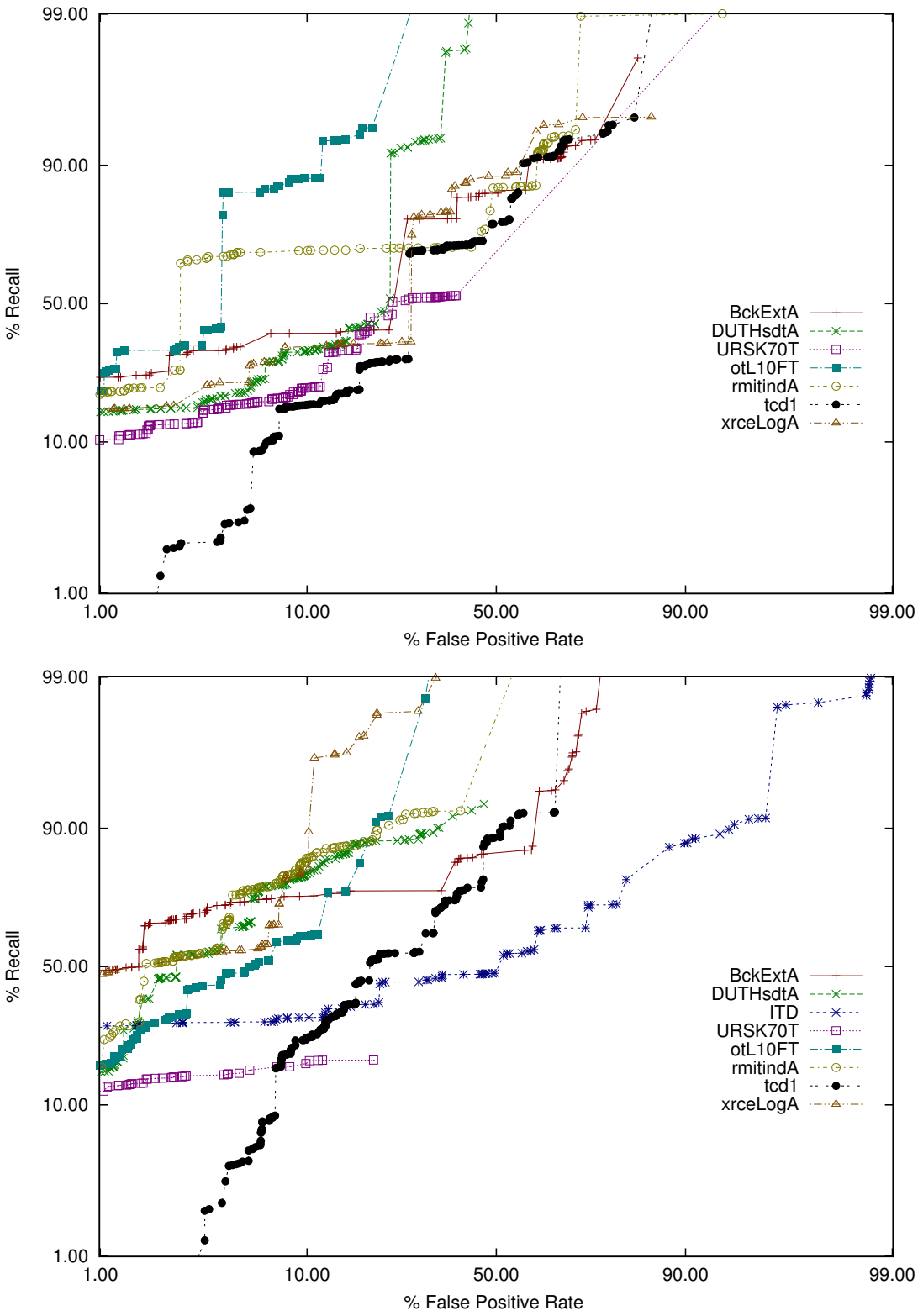

Figure 1: ROC curves for topics 200 (top) and 201 (bottom), best run per team.

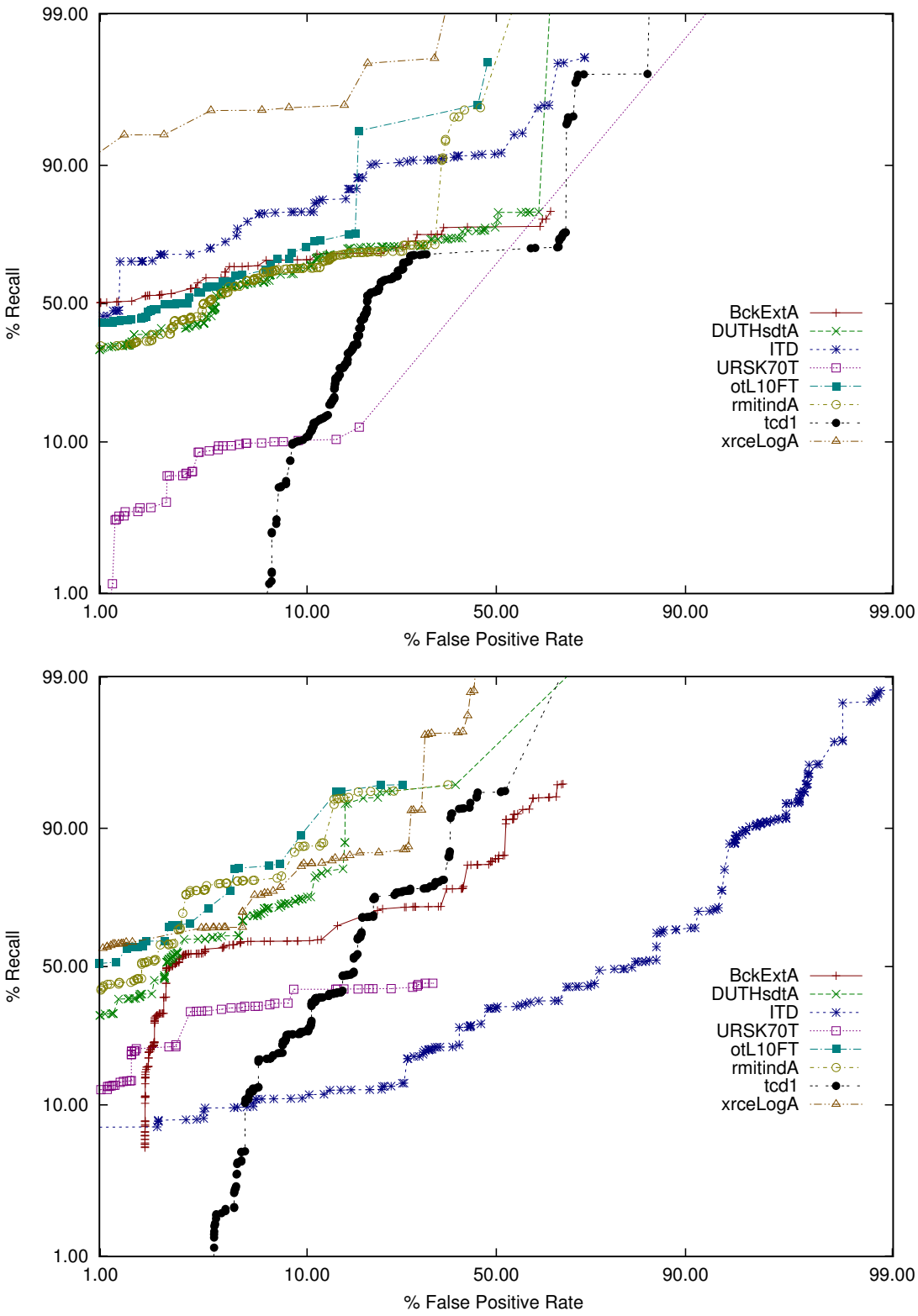

Figure 2: ROC curves for topics 202 (top) and 203 (bottom), best run per team.

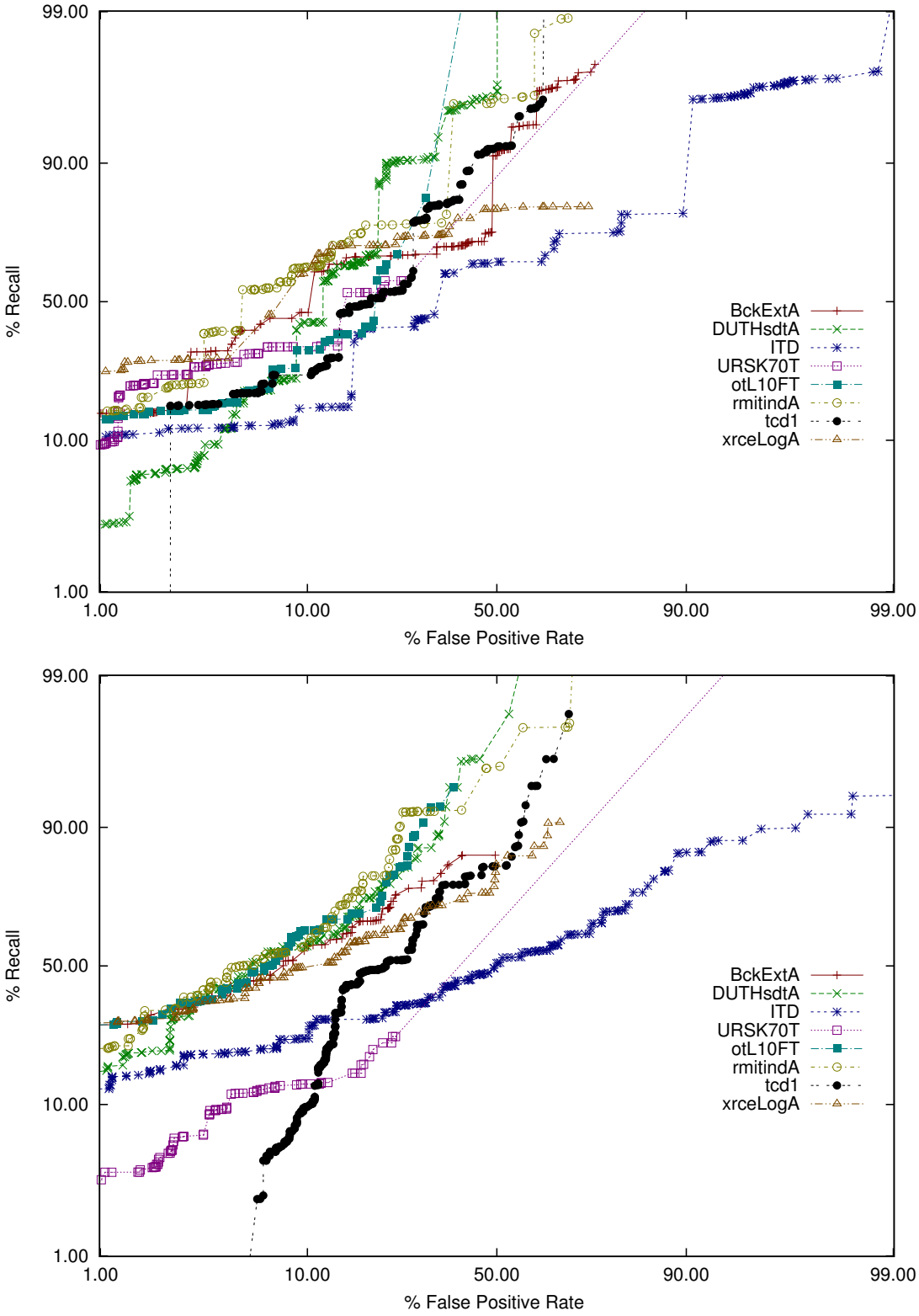

Figure 3: ROC curves for topics 204 (top) and 205 (bottom), best run per team.

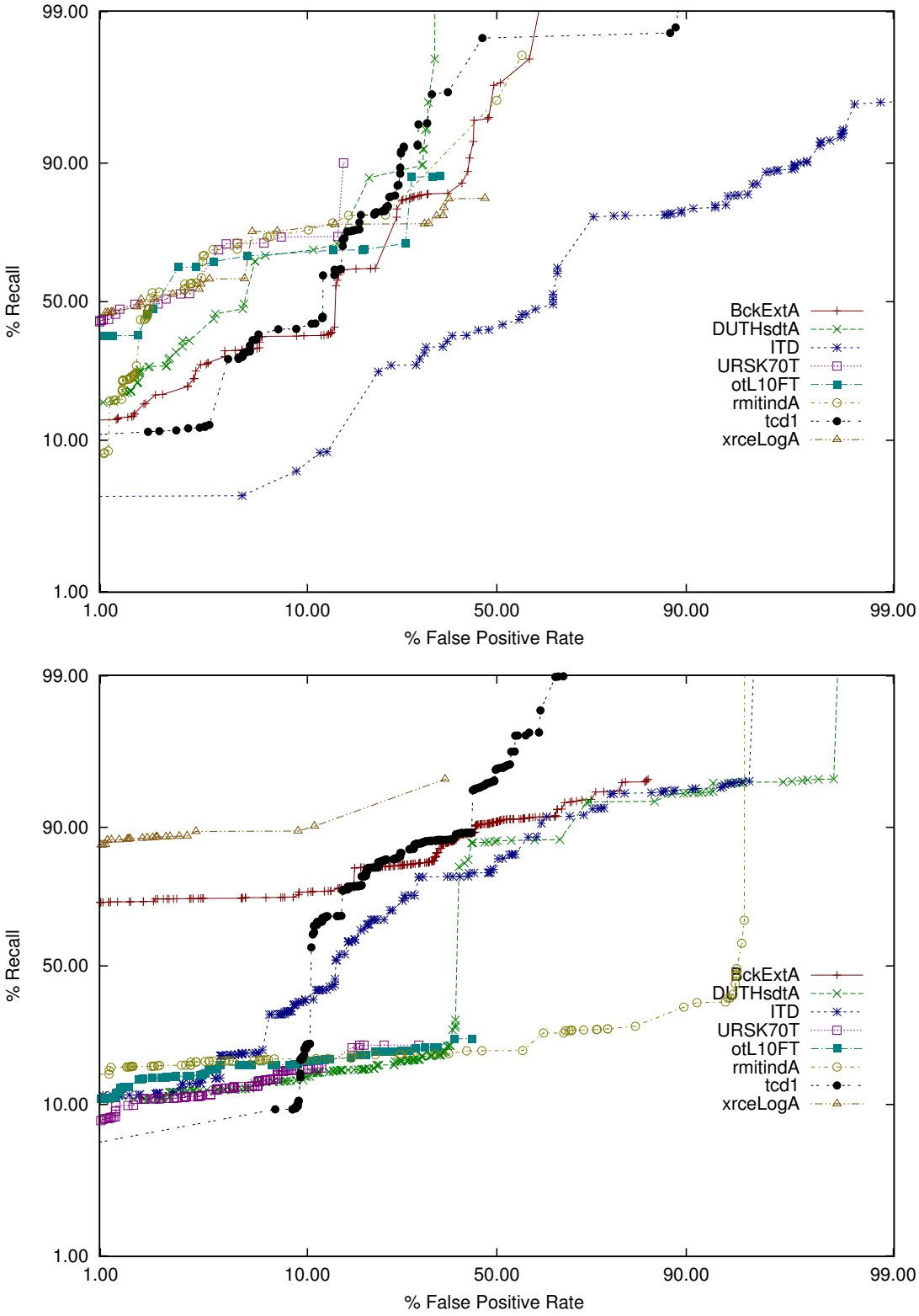

Figure 4: ROC curves for topics 206 (top) and 207 (bottom), best run per team.

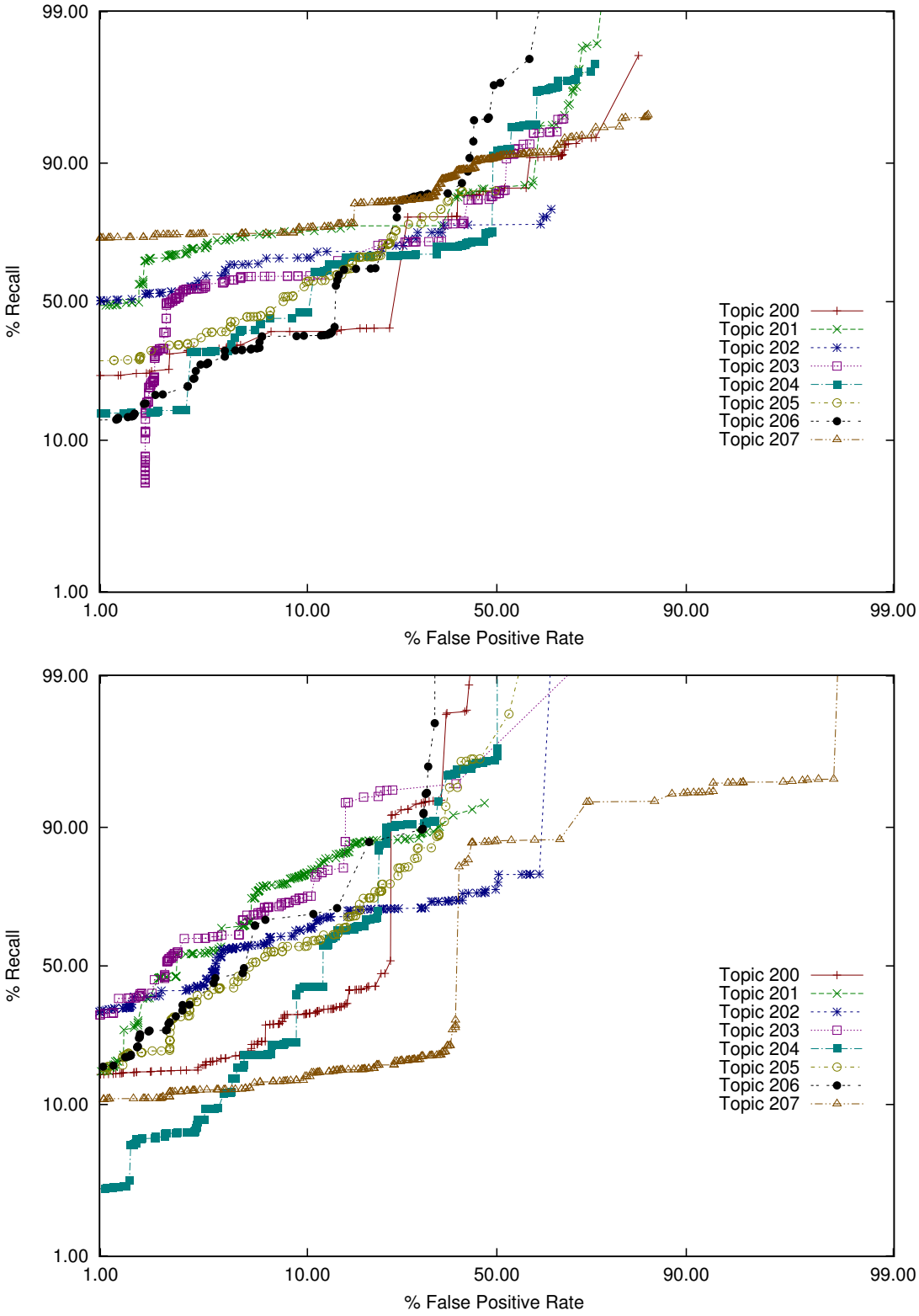

Figure 5: ROC curves for runs BckExtA (top) and DUTHsdtA (bottom).

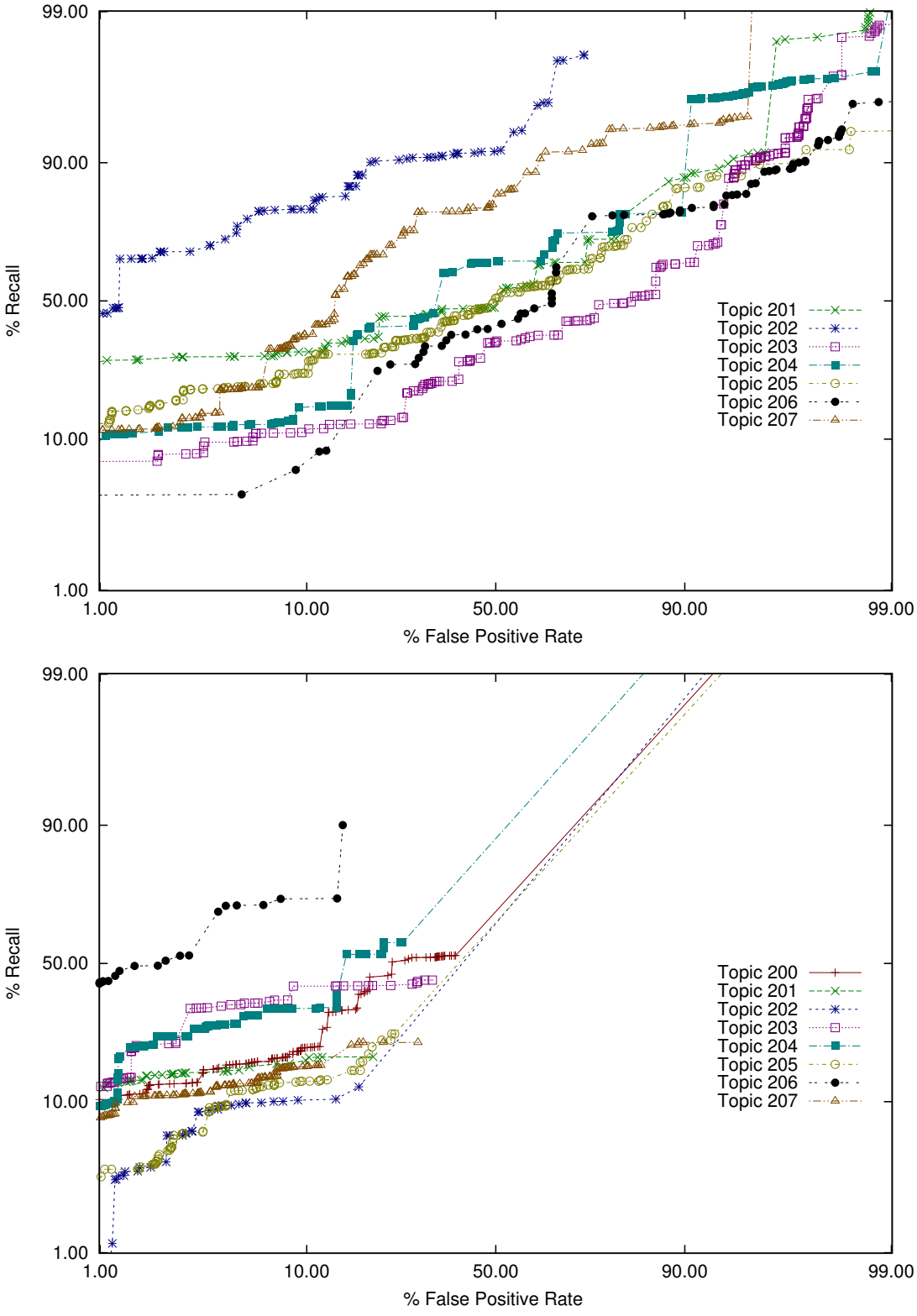

Figure 6: ROC curves for runs ITD (top) and URSK70T (bottom).

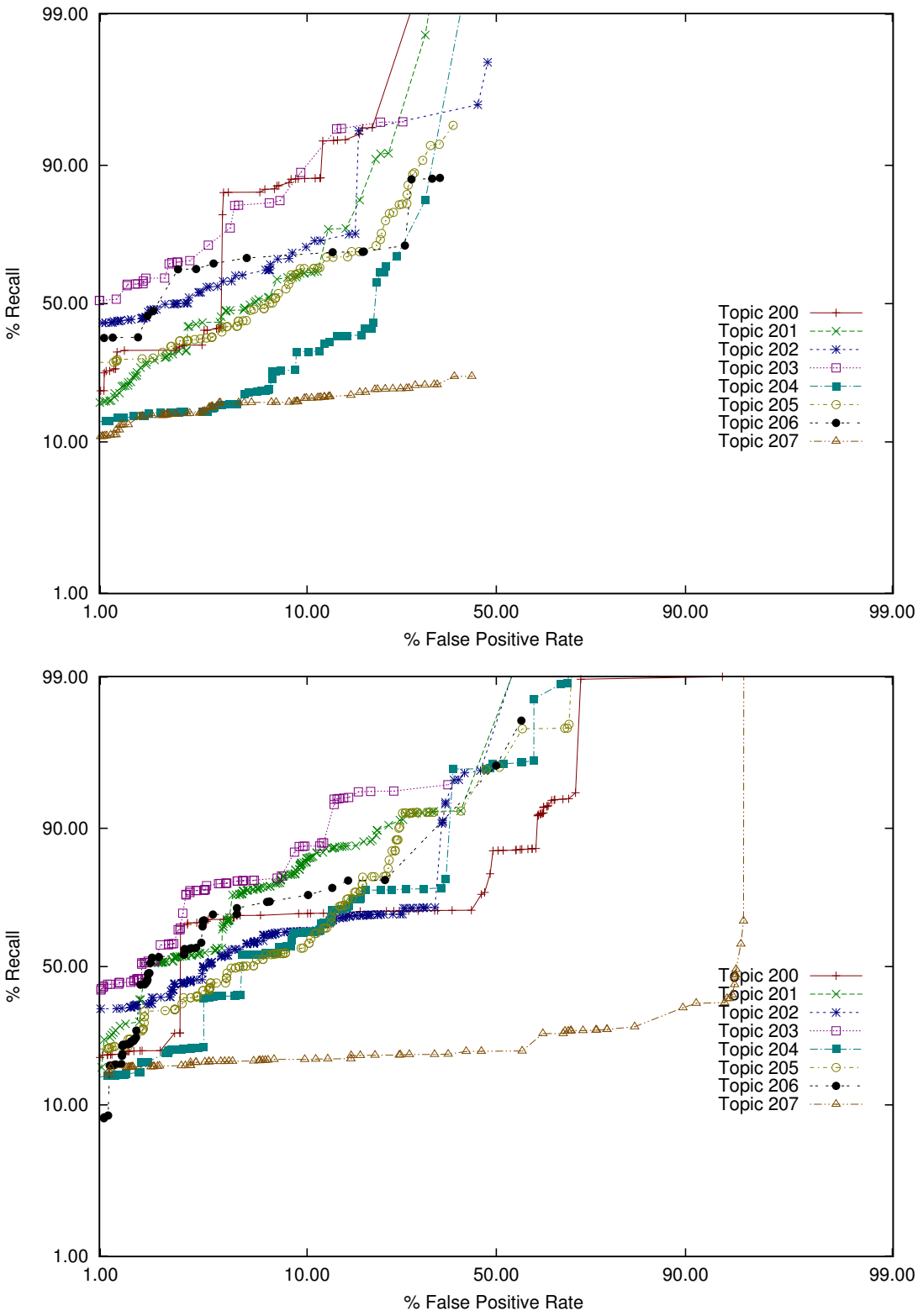

Figure 7: ROC curves for runs otL10FT (top) and rmitindA (bottom).

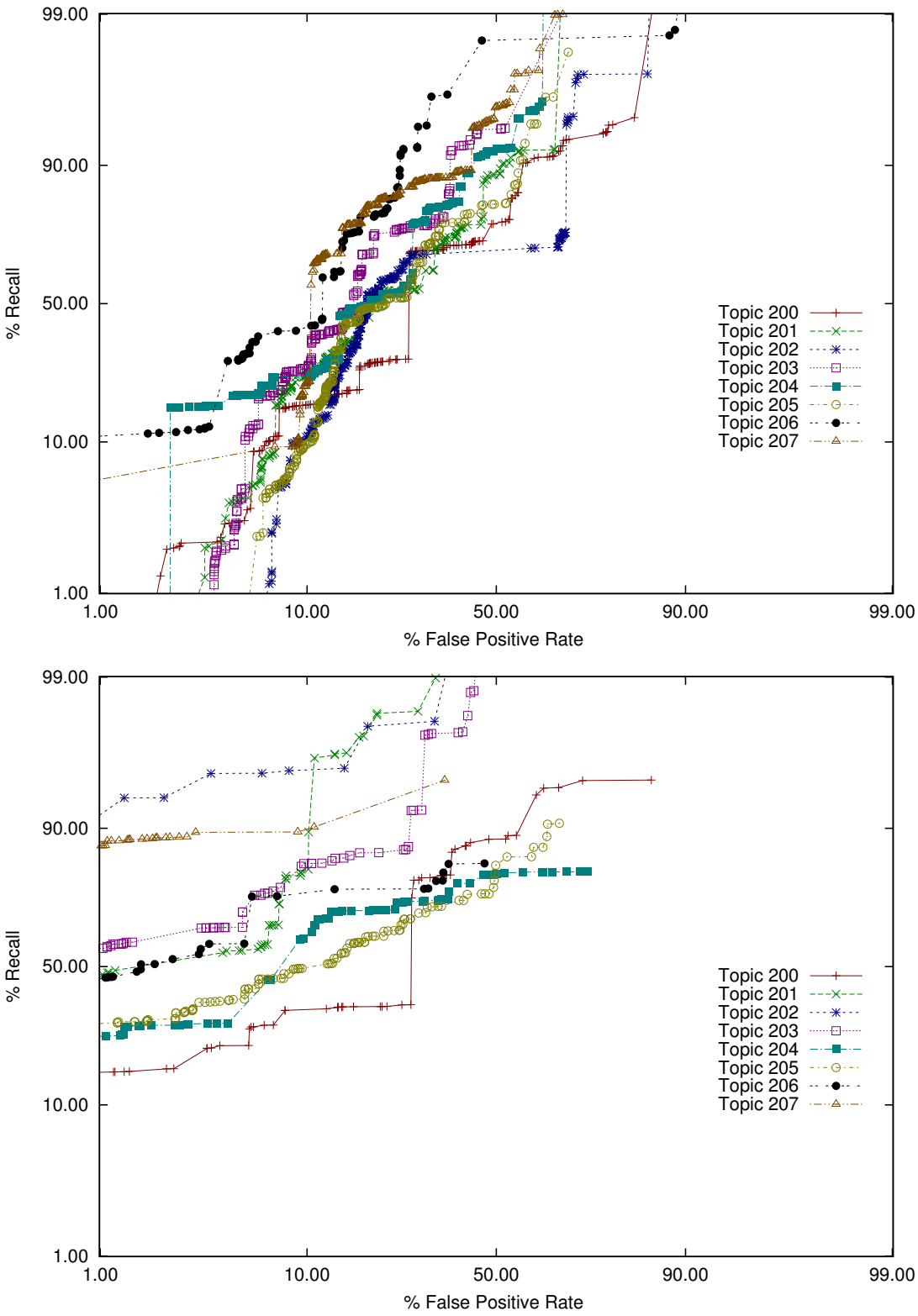

Figure 8: ROC curves for runs tcd1 (top) and xrceLogA (bottom).

	Topic								
Run	200	201	202	203	204	205	206	207	avg (95% C.I.)
xrceLogA	73.8	94.1	97.1	92.3	77.6	76.8	84.8	96.6	86.6 (80.8, 92.5)
xrceCalA	73.8	94.1	97.1	92.3	77.6	76.8	84.8	96.6	86.6 (80.8, 92.5)
otL10FT	94.8	89.8	92.0	95.9	80.3	86.7	89.7	47.0	84.5 (74.8, 94.3)
DUTHsdtA	82.7	90.4	81.0	92.9	86.8	87.3	91.3	60.3	84.1 (77.7, 90.5)
DUTHsdeA	82.7	90.4	81.0	92.9	86.8	87.3	91.3	60.3	84.1 (77.7, 90.5)
DUTHlrgA	82.7	90.4	81.0	92.9	86.8	87.3	91.3	60.3	84.1 (77.7, 90.5)
otL10rvlT	57.2	89.7	81.3	87.1	81.0	81.5	97.2	92.0	83.4 (76.0, 90.7)
BckExtA	76.3	84.8	82.7	81.5	79.9	81.3	83.7	87.4	82.2 (80.2, 84.2)
BckBigA	78.0	84.8	82.7	78.8	79.7	81.2	83.7	87.4	82.0 (80.1, 84.0)
rmitindA	81.6	91.2	86.8	94.3	87.0	87.9	87.6	27.0	80.4 (67.0, 93.8)
xrceNoRA	76.6	85.6	89.0	85.6	56.6	73.0	80.3	89.8	79.6 (72.8, 86.3)
otL10bT	67.6	91.1	84.1	87.9	52.2	71.0	97.0	72.8	78.0 (69.0, 87.0)
tcd1	68.3	73.0	77.9	79.4	78.2	72.7	89.7	84.2	77.9 (73.8, 82.1)
BckLitA	74.5	84.3	82.0	89.5	68.1	80.2	84.3	31.7	74.3 (63.1, 85.6)
rmitmlfT	72.8	74.2	85.2	64.8	71.6	72.9	75.4	28.4	68.2 (57.8, 78.6)
rmitmlsT	64.5	74.6	86.5	79.3	66.8	70.4	74.5	28.0	68.1 (57.3, 78.9)
URSK70T	61.5	30.2	35.2	48.0	73.6	33.7	87.5	39.6	51.2 (38.4, 64.0)
ITD	0.0	55.3	90.2	33.2	57.7	50.5	44.2	76.8	51.0 (34.3, 67.7)
URSK35T	65.4	31.9	36.6	48.0	62.9	34.2	86.3	38.7	50.5 (38.7, 62.3)
URSLSIT	61.5	30.9	32.1	48.0	69.9	33.7	87.5	39.4	50.4 (37.7, 63.0)

Table 9: AUC: Area under the receiver operating characteristic curve (C.I.=Confidence Interval).

While it is debatable whether or not F_1 aptly reflects the quality of a result set in any practical situation, optimizing F_1 is nonetheless a challenging proposition that requires good probability estimates for two purposes: to rank the documents, and to determine the optimal value of k.

Figure 10 shows the actual F_1 scores achieved on each topic by every run. The average over all topics ranges from 3.6% to 37.1%. Figure 11, in contrast, shows the *hypothetical* F_1 scores achieved on each topic by every run. This score represents the F_1 score that would have been achieved, had the probability estimates been accurate and therefore yielded the optimal value of k. That is, actual F_1 measures the quality of the ranking and also the accuracy of the probability estimates, while hypothetical F_1 measures only the quality of the ranking. A large discrepancy between actually and hypothetical F_1 indicates poor probability estimates. We see that the hypothetical F_1 scores range from 11.6% to 42.6%, indicating room for considerable improvement in the accuracy of probability estimation.

4.4 Discussion

Four broad themes stand out in the Learning track results. First, to the extent that we can use the random-selection main diagonal as a surrogate for manual review, most teams convincingly beat manual review on most topics. From this we can conclude that learning methods have a place in the design of cost-effective document review processes. Second, no team found more than half of the relevant documents that we estimate to exist for more than half of the topics when the result set was limited to 20,000 documents. From this we can conclude that further research on the application of learning techniques to this task is called for. Third, if we interpret the average accuracy of system estimates to be a measure of the degree to which systems can help their users to determine a cost-effective point at which to cease further learning, present systems seem particularly poor at that important task. Indeed, some systems do well at that "stopping-guidance" task at some points in the ranked list but not at others, while others do consistently less well. Xerox Research Centre Europe (xrce) is among the very few that do fairly well at this "stopping-guidance"

Run	Topic								
	200	201	202	203	204	205	206	207	Avg
xrceCalA	17.7	31.6	64.9	30.4	20.0	38.6	4.7	88.8	37.1
xrceLogA	3.8	27.1	50.7	33.4	26.0	43.7	4.8	89.9	34.9
xrceNoRA	15.1	27.0	57.6	22.0	13.4	35.2	1.8	21.4	24.2
otL10FT	8.2	6.8	33.2	25.8	9.5	40.6	6.2	16.9	18.4
otL10bT	20.0	15.1	24.9	16.8	6.4	34.2	4.1	16.1	17.2
DUTHsdtA	5.7	6.0	25.6	16.6	4.4	20.9	4.8	12.3	12.0
otL10rvlT	0.9	6.0	14.4	7.3	6.7	28.5	4.7	20.6	11.1
DUTHlrgA	4.2	6.4	18.4	17.6	3.3	24.8	3.5	10.5	11.1
rmitmlsT	0.4	3.6	17.1	12.1	7.1	31.0	2.5	10.6	10.6
rmitmlfT	0.9	3.8	22.0	6.8	3.9	32.6	2.2	7.4	10.0
rmitindA	1.0	4.1	12.0	4.6	3.1	47.5	0.9	2.3	9.4
BckExtA	1.0	0.8	2.4	1.4	3.1	45.0	0.3	7.9	7.7
BckBigA	1.0	0.8	2.4	1.3	3.1	44.5	0.3	7.9	7.7
BckLitA	1.0	0.8	2.4	1.5	2.9	44.4	0.3	2.1	6.9
tcd1	1.2	0.8	5.5	1.9	3.0	28.5	1.0	11.0	6.6
DUTHsdeA	1.3	4.8	25.6	0.7	0.6	1.1	2.7	12.3	6.1
URSK70T	2.8	5.3	0.6	7.4	8.5	6.9	4.8	11.4	6.0
ITD	0.0	1.1	2.3	1.0	2.4	16.5	0.2	17.5	5.1
URSLSIT	2.8	3.4	2.8	7.4	2.3	6.9	4.8	3.8	4.3
URSK35T	2.1	1.4	3.2	4.7	3.2	3.5	2.9	8.1	3.6

Table 10: F_1 scores achieved by submitted runs, using submitted probability estimates to estimate optimal cutoff k.

task consistently, suggesting that their paper would be well worth reading.

5 Interactive Task

The Legal Track's Interactive task more fully models the conditions and objectives of a search for documents that are responsive to a production request served during the discovery phase of a civil lawsuit. The 2010 exercise represented the third year that the Interactive task, in its current design, was featured in the Legal Track. Results from the first two years can be found in the track overviews for 2008 [7] and 2009 [6]. In this year's overview, we briefly review the task design (Section 5.1); describe the specific features that defined the 2010 exercise (Section 5.2); summarize the results obtained for each of the 2010 topics (Section 5.3); and provide additional analysis on certain points of interest (Section 5.4).

5.1 Task Design

The most complete discussion of the design of the Interactive task, and of the reasoning behind it, can be found in the 2008 task guidelines [2]. While the core features of the task have not changed since 2008, we have, in each of the subsequent years, introduced various modifications to the design in an effort to improve the effectiveness and efficiency of the exercise. In this section, we briefly review the task's core features (Section 5.1.1) and summarize the modifications introduced for the 2010 running of the exercise (Section 5.1.2).

5.1.1 Core Features

The real-world circumstance modeled by the Interactive task is that of a search for documents that are responsive to a discovery request served in civil litigation. The task is designed to gauge the effectiveness

	Topic								
Run	200	201	202	203	204	205	206	207	Avg
xrceLogA	19.1	34.5	70.6	39.4	26.6	45.9	14.3	90.3	42.6
xrceCalA	19.1	34.5	70.6	39.4	26.6	45.9	14.3	90.3	42.6
BckBigA	18.9	40.7	56.6	36.4	15.2	47.3	3.4	81.8	37.5
xrceNoRA	16.9	28.5	60.8	34.8	16.9	48.3	12.9	65.0	35.5
BckExtA	18.9	40.7	56.8	15.7	15.4	47.2	3.4	81.8	35.0
BckLitA	24.8	43.0	53.3	39.4	7.0	46.1	11.6	6.7	29.0
otL10bT	25.8	23.1	32.8	32.9	8.1	47.3	37.0	22.9	28.7
otL10FT	21.6	14.1	40.2	33.5	13.3	51.2	27.3	18.6	27.5
otL10rvlT	3.4	19.5	15.0	11.1	10.2	48.8	26.9	72.6	25.9
rmitindA	15.3	13.0	37.8	32.2	17.1	52.1	5.9	24.3	24.7
DUTHsdtA	8.9	11.6	32.1	24.2	8.2	51.4	7.6	16.7	20.1
DUTHsdeA	8.9	11.6	32.1	24.2	8.2	51.4	7.6	16.7	20.1
DUTHlrgA	8.9	11.6	32.1	24.2	8.2	51.4	7.6	16.7	20.1
ITD	0.0	24.1	53.5	8.6	10.1	26.6	7.0	19.1	18.6
rmitmlsT	1.8	7.1	38.2	19.6	8.7	35.1	7.1	16.4	16.8
rmitmlfT	2.3	6.8	41.7	11.3	5.3	34.5	18.7	11.2	16.5
URSLSIT	7.4	10.1	4.6	9.1	20.9	18.0	20.1	16.3	13.3
URSK70T	7.4	9.6	4.5	9.1	15.6	18.0	20.1	14.1	12.3
URSK35T	8.7	11.9	4.9	12.0	5.3	18.0	21.0	11.5	11.7
tcd1	1.9	1.5	5.6	3.2	10.3	31.8	9.0	25.6	11.1

Table 11: Hypothetical F_1 scores achieved by submitted runs, using optimal cutoff k. These numbers represent the best possible F_1 that could have been achieved with the same relevance ranking and a better choice of k.

of various approaches to document retrieval (whether they be fully automated, fully manual, or something in between) at meeting the requirements of this sort of search. The core features of the task design are as follows.

Complaint and topics. Context for the Interactive task is provided by a mock complaint that sets forth the legal and factual basis for the hypothetical lawsuit that motivates the discovery requests at the heart of the exercise. Associated with the complaint are document requests that specify the categories of documents which must be located and produced. For purposes of the Interactive task, each of these document requests serves as a separate topic. The goal of a team participating in a given topic is to retrieve all, and only, documents relevant to that topic (as defined by the "Topic Authority;" see below).

Most of the topics featured in the Interactive task have been modeled on typical subject-matter requests for production (i.e., they seek documents pertinent to the legal and factual issues that are the focus of the litigation). The 2010 exercise, for the first time, also featured a "privilege" topic, a topic requiring the identification of any and all documents that could be withheld from production on grounds of privilege or work product protection.

The Topic Authority as the source of the operative standard of relevance. A key role in the Interactive task is played by the "Topic Authority." The Topic Authority plays the role of a senior attorney who is charged with overseeing a client's response to a request for production and who, in that capacity, must certify to the court that the client's response to the request is complete and correct (commensurate with a reasonable and good-faith effort). In keeping with that role, it is this attorney who, weighing considerations of genuine subject-matter relevance as well as pragmatic considerations of legal strategy and tactics, holds ultimate responsibility for deciding what is and is not to be considered responsive for purposes of the document production (or, in the terms of the Interactive task, what is and is not to be considered relevant to a target topic). The Topic Authority's role, then, is to be the source for the authoritative conception of relevance that each participating team, in the role of a hired cohort of manual reviewers or of a vendor of document-retrieval services, will be asked to replicate across the full document collection.

Now, needless to say, different lawyers may well take different approaches to discovery, and the interpretation which one lawyer would give to a request will likely not be identical to the interpretation another lawyer would give to the same request. That is nothing more than to state that responsiveness, like relevance, is, to a certain extent, subjective. In the real-world circumstance modeled by the Interactive task, however, there is, from the perspective of a vendor hired to assist in the search for responsive documents, only one lawyer (or litigation team) that matters, the client and its counsel, and one interpretation of responsiveness that matters, that of the client and its counsel, whose retrieval requirements the vendor has been hired to fulfill. Once fully situated in a real-world context, therefore, in which there is a client and its counsel who set the requirements and a vendor who is hired to implement those requirements, the subjectivity of relevance gives way to a single operative standard of relevance, that of the client and its counsel (or, in the terms of the Interactive task, that of the Topic Authority). In the Interactive task, it is this single standard of relevance, set by the Topic Authority, that defines the target set of documents for each topic.

Each topic has one, and only one, Topic Authority, and each Topic Authority has responsibility for one, and only one, topic.

Allowance for interaction with the Topic Authority. If it is the Topic Authority who defines the target (i.e., who determines what should and should not be considered relevant to a topic), it is essential that provision be made for teams to be able to interact with the Topic Authority in order to gain a better understanding of the Topic Authority's conception of relevance. In the Interactive task, this provision takes the following form. Each team can ask, for each topic for which it plans to submit results, for up to 10 hours of a Topic Authority's time for purposes of clarifying a topic. A team can call upon a Topic Authority at any point in the exercise, from the kickoff of the task to the deadline for the submission of results. How a team makes use of the Topic Authority's time is largely unrestricted: a team can ask the Topic Authority to pass judgment on exemplar documents; a team can submit questions to the Topic Authority by email; a team can arrange for conference calls to discuss aspects of the topic. One constraint that is placed on communication between the teams and their designated Topic Authorities is introduced in order to minimize the sharing of information developed by one team with another; while the Topic Authorities are instructed to generally be

free in sharing the understanding they have of their topic, they are asked to avoid volunteering to one team specific information that was developed only in the course of interaction with another team.

Participant submissions. Each participant's final deliverable is a binary classification (relevant / not relevant) of the full test collection for relevance to each target topic in which it has chosen to participate.

Composition of evaluation samples. Once participants have completed their submissions, we are in a position to draw the evaluation samples for each topic. The evaluation samples are composed using a stratified sampling design that allows for disproportionate allocation among strata. More specifically, the samples are composed as follows.

With regard to stratification, strata are defined on the basis of participant submissions. The sets of documents deemed relevant by the participants in a topic allow for a straightforward submission-based stratification of the collection: one stratum contains the documents all participants submitted as relevant, another stratum contains the documents no participant submitted as relevant, and other strata are defined for each of the other possible submission combinations. If, for example, there are five teams that submitted results for a topic, the collection will be partitioned into $2^5 = 32$ strata.

With regard to allocation among strata in composing the samples, strata are represented largely in keeping with their full-collection proportions. In order to ensure that a sufficient number of documents are drawn from all strata, however, some small strata may be over-represented, and some large strata under-represented, relative to their full-collection proportions. In particular, the "All-N" stratum, the stratum containing the documents all participants have deemed not relevant, tends to be very large, often making up 90% or more of the collection. While we do sample extensively from this stratum, as we want to include in the analysis an ample number of documents that no participant deemed relevant, we under-represent it in the sample, relative to its full-collection proportion, in order to be able also to represent adequately the strata defined for documents that one or more participants have deemed relevant.

Sampling within a stratum is simple random selection without replacement.

Two-stage assessment protocol. Once evaluation samples are drawn, the documents in each sample are assessed for relevance to the topic for which the sample was drawn. In the Interactive task, a two-stage process is followed in order to arrive at final sample assessments. In the first stage, a team of assessors is assigned to each sample and asked to complete, under the guidance of the Topic Authority, a first-pass assessment of the sample. In the second stage, a subset of those first-pass assessments are escalated to the Topic Authority for final adjudication. These subsets consist primarily of assessments that a participant, after reviewing the results of the first-pass assessment, has chosen to appeal to the Topic Authority (because the participant believes the first-pass assessment to be out of keeping with the Topic Authority's guidance). In the 2010 Interactive task, we also, for the first time, escalated a number of non-appealed assessments to the Topic Authority for final adjudication (additional details in Section 5.1.2 below).

Effectiveness measures. Once all escalated assessments have been adjudicated by the Topic Authority, we are in a position to obtain final estimates of the level of effectiveness achieved by each submission. Given the binary nature of the submissions, we look to set-based metrics to gauge effectiveness. In the Interactive task, the metrics used are recall, precision, and, as a summary measure of effectiveness, F_1. For further detail on the estimation procedures followed in the Interactive task, see the appendix to the Overview of the TREC 2008 Legal Track [7].

5.1.2 New to the 2010 Exercise

While keeping constant the core features of the design of the Interactive task, we did introduce a small number of new features to the 2010 task with an eye to improving the efficiency of the task and to laying the groundwork for further analysis of task results. Chief among these were modifications to the assessment and adjudication process. The modifications are as follows.

Professional assessors. In 2008, the first-pass assessment was carried out entirely by individual volunteers: law students and legal professionals who volunteered to assess one or two 500-document batches of documents. In 2009, for the first time, we were able to engage, on a pro bono basis, the services of firms offering professional document-review services to carry out the first-pass assessment for some, but not all, of that year's evaluation samples; the samples for other topics again were assessed by individual volunteers. In

2010, we were again able to engage, on a pro bono basis, the services of firms offering professional document-review services; for 2010, the document-review professionals were able to carry out the first-pass assessment for all of the evaluation samples.

Dual assessment. In 2008 and 2009, the first-pass assessment consisted of our gathering, from the team of assessors, a single relevance assessment on each document in the evaluation sample. In 2010, we also gathered, on each document in a subsample of the evaluation sample, a second independent assessment. We did so (i) in order to have data to support further analysis of rates of assessor error and of rates of inter-assessor agreement and (ii) in order to identify a set of documents that, even if not appealed, would be good candidates for escalation to the Topic Authority for adjudication (e.g., cases of conflicting assessment).

The dual-assessment subsample was selected as follows. The full evaluation sample was selected (in accordance with the stratified design described above) and randomly distributed into 25 assessment batches of approximately equal size. Then, for purposes of obtaining dual assessments, a supplemental set of messages was added to each batch; the supplement consisted of messages drawn, via simple random selection without replacement, from messages already included in the sample but not assigned to the batch for which the supplement was intended (and not already added as a dual-assessment supplement to another batch). In all, for each topic, about 10% of the messages in the evaluation sample were selected for dual assessment in this way.

Adjudication of non-appealed assessments. As noted above, the Interactive task follows a two-stage procedure in arriving at final sample assessments: first-pass assessment followed by adjudication. The purpose of the adjudication stage is to identify and correct any errors made in the first-pass assessment ("errors" being relevance assessments that are not in keeping with the Topic Authority's definition of relevance). In 2008 and 2009, adjudication was an entirely appeals-driven process; that is to say, the set of documents escalated to the Topic Authority for final adjudication consisted of all, and only, the documents the first-pass assessments of which participants, after reviewing the results of the first-pass review, had appealed. We have found that an entirely appeals-driven adjudication process is an effective mechanism for correcting first-pass errors, as long as participants make diligent use of the appeals process; we have also found, however, that, in cases in which participants in a topic elect to make little use of appeals (e.g., Topic 206 from the 2009 exercise), the adjudication process will likely leave many first-pass errors uncorrected [6] [9]. For 2010, therefore, we decided to supplement the adjudication set with a certain number of non-appealed documents. Further details on the selection of non-appealed messages for adjudication are provided in Section 5.3.4 below.

Adjudication materials. In both 2008 and 2009, we asked participants to prepare, in support of their appeals, documents detailing the specific grounds for each of their appeals. The purpose of these "grounds for appeals" documents was to enable participants to direct the Topic Authority to specific features of a document that they believed, given the Topic Authority's prior relevance guidance, would be decisive as to the document's being deemed relevant or not.

The approach was taken both for the sake of efficiency (enabling the Topic Authority to skip to the salient parts of often long documents) and for the sake of accuracy (enabling participants to draw the Topic Authority's attention to the often subtle features of a document). The approach did raise some concerns, however, as some participants found the preparation of the "grounds for appeals" documents rather burdensome and others argued that information about what the first-pass assessment (and the proposed alternative) was might, in some way, influence the Topic Authority's final assessment.

In 2010, therefore, on a trial basis, we decided to implement a blind adjudication protocol. Participants were not asked to document the grounds for their appeals; they were asked simply to submit a list of the documents they wished to have adjudicated. The Topic Authorities, when given their adjudication sets, were given no information as to the first-pass assessment, the alternative assessment being proposed by the appealing team, or whether or not the document was included in the set as a result of an appeal. We analyze the results of this year's protocol below.

5.2 Task Specifics

The following features defined the specific landscape of the 2010 Interactive task.

5.2.1 Test Collection

The document collection used for the 2010 Interactive task was that derived from the EDRM Enron Dataset, v. 2; see Section 2 above for details on this collection.

5.2.2 Topics & Topic Authorities

For 2010, the exercise included an entirely new mock complaint and three associated document requests, each of which served as a separate topic [1]. In addition to these three subject-matter topics, the task also featured a "privilege" topic, a topic designed to model a search for material that could be withheld from a production on grounds of privilege or work-product protection. For each of the four topics, a separate Topic Authority was designated. The topic statements and Topic Authorities were as follows .

- **Topic 301.** All documents or communications that describe, discuss, refer to, report on, or relate to onshore or offshore oil and gas drilling or extraction activities, whether past, present or future, actual, anticipated, possible or potential, including, but not limited to, all business and other plans relating thereto, all anticipated revenues therefrom, and all risk calculations or risk management analyses in connection therewith.
  - **Topic Authority:** Mira Edelman (Hughes Hubbard & Reed).
- **Topic 302.** All documents or communications that describe, discuss, refer to, report on, or relate to actual, anticipated, possible or potential responses to oil and gas spills, blowouts or releases, or pipeline eruptions, whether past, present or future, including, but not limited to, any assessment, evaluation, remediation or repair activities, contingency plans and/or environmental disaster, recovery or clean-up efforts.
  - **Topic Authority:** John F. Curran (Stroz Friedberg).
- **Topic 303.** All documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto.
  - **Topic Authority:** Robert E. Singleton (Squire, Sanders, & Dempsey).
- **Topic 304.** Should Defendants choose to withhold from production any documents or communications in the TREC Legal Track Enron Collection on the basis of a claim of privilege, attorney work-product, or any other applicable protection, they should identify all such documents or communications.

  Important procedural note specific to Topic 304. Solely for the purpose of the TREC 2010 Legal Track, participants who choose to submit results for Topic 304 should identify any and all documents or communications in the TREC Legal Track Enron Collection that are subject to a claim of privilege, attorney work-product, or other applicable protection, regardless of whether they are responsive to any of the Requests for Production specified above.
  - **Topic Authority:** Michael Roman Geske (Aphelion Legal Solutions).

5.2.3 Participating Teams

The Interactive task received submissions from twelve teams, who, collectively, submitted a total of 22 single-topic runs. The twelve teams that submitted results for evaluation are shown in Table 1.

It should be noted that Douglas Oard, a track coordinator who was on sabbatical at the University of Melbourne and RMIT during a part of this period, participated in that team's research and that Gordon Cormack, of the University of Waterloo and also a track coordinator, offered some advice to the University of Waterloo team.

A team could ask to participate in as many, or as few, topics as it chose. Given constraints on the number of teams for which a Topic Authority could take responsibility (typically, a maximum of eight teams), we indicated that we might not be able to give all teams all of their choices and asked teams to rank their topic selections in order of preference. Topics were assigned largely on a first-come-first-serve basis. For the 2010 task, it turned out that we were able to give all teams their preferred topics. Table 12 shows the number of runs submitted by each team for each topic (in the table, an empty cell represents no submissions for the given team-topic combination).

	Topics				**Total**
Team	**301**	**302**	**303**	**304**	**Runs**
CB			2	4	**6**
CS	1				**1**
EQ			1		**1**
IN		1		1	**2**
IS	1	1			**2**
IT	1		1		**2**
LA		1			**1**
MM		1			**1**
SF	1				**1**
UB			1		**1**
UM		1			**1**
UW	1	1	1		**3**
Total Runs	**5**	**6**	**6**	**5**	**22**

Table 12: Runs submitted for each topic.

As can be seen from the table, in most cases, each team submitted, in accordance with the task guidelines, just one run for each topic in which it participated. In one case, however, a team asked for, and was given, permission to submit multiple runs for a single topic.

The Cleary-Backstop team (CB) wished, for both Topic 303 and Topic 304, to have two submissions evaluated, one taking into account "family" associations between documents and the other not; in the following, these runs are designated CB1 (with family associations) and CB2 (without family associations). The team also wished, for Topic 304, to have two additional runs evaluated. Both of these runs targeted a broader notion of "potentially privileged;" again one taking into account family associations and the other not. These two "broader" runs are designated CB3 (with family associations) and CB4 (without family associations).

5.2.4 Assessors

As noted above, for the 2010 Interactive task, the first-pass assessment of the evaluation samples for all four topics was carried out by firms that provide professional contract or managed review services, using their typical review processes and procedures.

5.2.5 Unit of Assessment

In evaluating the effectiveness of approaches to assessing the relevance of email messages, one must decide whether one wants to assess effectiveness at the *message* level (i.e., treat the parent email together with all of its attachments as the unit of assessment) or to assess effectiveness at the *document* level (i.e., treat each of the components of an email message (the parent email and each child attachment) as a distinct unit of assessment).

For the 2010 exercise, in an effort to gather data on both levels, we asked participants to submit their results at the document level (in order to enable document-level analysis) from which we would then derive message-level values (which would serve as the primary basis for evaluation). The specific rules governing the assignment of assessments were as follows.

- A parent email should be deemed relevant either if, in itself, it has content that meets the definition of relevance or if any of its attachments meet that definition; contextual information contained in all components of the email message should be taken into account in determining relevance.
- An email attachment should be deemed relevant if it has content that meets the Topic Authority's definition of relevance; in making this determination, contextual information contained in associated documents (parent email or sibling attachments) should be taken into account.
- A message will count as relevant if at least one of its component documents (parent email or attachments) has been found relevant.
- For purposes of scoring, the primary level is the message-level; document-level analysis is on between documents reviewed and supplementary. By contrast, the Learning task reports only document-level analysis.

5.3 Task Results

The 2010 Interactive task got underway, with the release of the final task guidelines [5] and of the mock complaint and associated topics [1], on July 6, 2010. In this section, we summarize the results of the exercise.

5.3.1 Team-TA Interaction

As noted above, the Interactive task permits teams to call on up to 10 hours (600 minutes) of a Topic Authority's time for purposes of clarifying the scope and intent of a topic. Of course, teams are not required to use their full allotment and, in previous years, we have seen considerable variation in the amount of Topic Authority time that teams choose to use.

Figure 9 summarizes, for the 2010 exercise, the participants' use of the Topic Authorities' time for each topic. In the diagram, each bar represents the total time allowed for team-TA interaction (600 minutes); the gray portion of the bar represents the amount of the permitted time that was actually used by a team (with the number of minutes used indicated just above the gray portion).

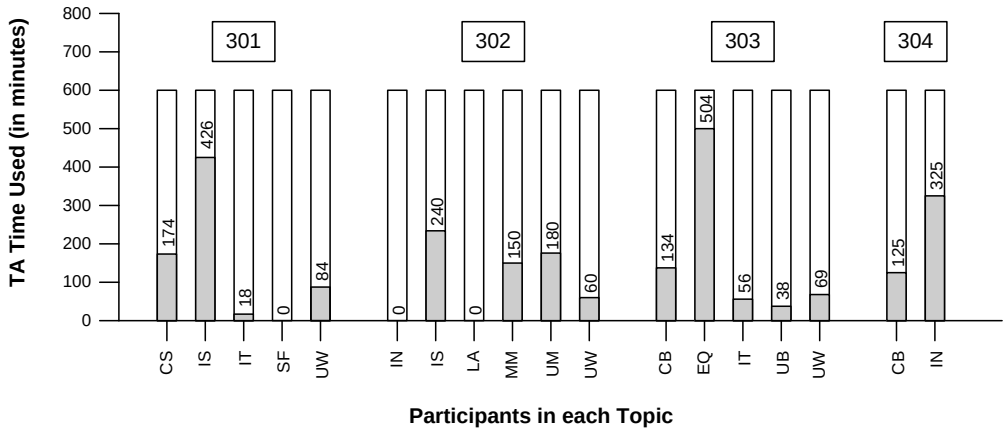

Figure 9: Team–TA interaction time.

As can be seen from the diagram, there is, once again, considerable variation in the extent to which teams utilized their allotted time for interacting with the Topic Authority: some teams used less than an hour of their available time, while others used seven hours or more. On the whole, however, teams tended to use considerably less than the maximum amount of time that they were allowed. In 15 of the 18 participant-TA interactions, the participant utilized less than 50% of the time available for interacting with the Topic Authority; in only three instances (301-IS; 303-EQ; 304-IN) did a team utilize more than 50% of the time permitted. We consider below (Section 5.4.1) whether there is any correlation between the amount of time spent interacting with the Topic Authority in the preparation of a run and the effectiveness of the run that results.

5.3.2 Submissions

Participants submitted their results on or before September 16, 2010. Table 13 summarizes, at the message level, the submissions received for each topic. The table shows (for the complement of the union of all submissions; for the union of all submissions; for each submission; and for the intersection of all submissions): (i) the number of messages that belong to each designated subset, (ii) the proportion, out of all messages in the full collection, that each subset represents, and (iii) the proportion, out of the union of all submissions, that each subset represents. (Recall that the full collection consists of 455,449 messages.)

The submission data alone, of course, tell us little about the effectiveness of the runs; that can be gauged only after review of the sampling and assessment data. The submission data do, however, permit a couple of initial observations.

First, with regard to the yield of the topics, we see that, for three of the four topics, the union of all submissions represents a relatively sizeable proportion of the collection: for each of Topics 301, 303, and 304, the union of all submissions represents 8% or more of the full collection. For Topic 302, on the other hand, the proportion of the collection submitted as relevant by at least one team is quite small: for this topic, the union of all submissions represents approximately 1% of the collection (and most of that can be attributed to a single submission). The submission data suggest, then, that three of the four topics are moderate to high yielding and that one of the four is low yielding. Of course, actual yields can be determined only after reviewing the sampling and assessment data (see Section 5.3.5).

Second, with regard to differences among submissions, we see that there is considerable variation in the number of messages that participants found relevant. Looking simply at the ratio of the largest to the smallest submission for each topic, we see that, for Topic 301, nearly 40 (39.7) messages were submitted as relevant by the largest submission (SF) for every one message submitted as relevant by the smallest (IS); for Topic 302, the ratio of largest to smallest submission is 50.9; for Topic 303, the ratio is 14.4; and, for Topic 304, the ratio is 18.6. It is clear that, across the full set of participants in any given topic, there was not a common understanding of (or at least a consistent implementation of) the scope that the Topic Authority intended for the topic; that is not to say, of course, that some subset of participants in a topic did not arrive at something approaching a shared understanding of the topic's intended scope.

Of course, what matters, in the end, is how closely each of the various submissions overlaps with the subset of messages that actually meet the Topic Authority's definition of relevance. In order to gauge that, we turn to sampling and assessment.

5.3.3 Stratification & Sampling

Once the submissions were received, the collection was stratified for each topic and evaluation samples were drawn. Stratification followed the submission-based design noted above (Section 5.1.1), whereby one stratum was defined for messages all participants found relevant (the "All-R" stratum), another for messages no participant found relevant (the "All-N" stratum), and others for the various possible cases of conflicting assessment among participants. The operative unit for stratification was the message, and messages were assigned intact (parent email together with all attachments) to strata.

Samples were composed following the allocation plan sketched above (Section 5.1.1), whereby strata are represented in the sample largely in accordance with their full-collection proportions. An exception to

Topic	Subset	Msg Count	Proportion of Full Collection	Proportion of Submitted as R
301	Submitted as R in no run	417,092	0.916	n.a.
	Submitted as R in at least one run	38,357	0.084	1.000
	Submitted as R in CS	5,428	0.012	0.142
	Submitted as R in IS	593	0.001	0.015
	Submitted as R in IT	13,170	0.029	0.343
	Submitted as R in SF	23,514	0.052	0.613
	Submitted as R in UW	619	0.001	0.016
	Submitted as R in all runs	18	< 0.001	< 0.001
302	Submitted as R in no run	450,575	0.989	n.a.
	Submitted as R in at least one run	4,874	0.011	1.000
	Submitted as R in IN	4,483	0.010	0.920
	Submitted as R in IS	88	< 0.001	0.018
	Submitted as R in LA	121	< 0.001	0.025
	Submitted as R in MM	164	< 0.001	0.034
	Submitted as R in UM	264	0.001	0.054
	Submitted as R in UW	135	< 0.001	0.028
	Submitted as R in all runs	2	< 0.001	< 0.001
303	Submitted as R in no run	411,961	0.905	n.a.
	Submitted as R in at least one run	43,488	0.095	1.000
	Submitted as R in CB1	7,536	0.017	0.173
	Submitted as R in CB2	10,025	0.022	0.231
	Submitted as R in EQ	17,245	0.038	0.397
	Submitted as R in IT	11,664	0.026	0.268
	Submitted as R in UB	30,185	0.066	0.694
	Submitted as R in UW	2,101	0.005	0.048
	Submitted as R in all runs	321	0.001	0.007
304	Submitted as R in no run	398,233	0.874	n.a.
	Submitted as R in at least one run	57,216	0.126	1.000
	Submitted as R in CB1	12,554	0.028	0.219
	Submitted as R in CB2	13,737	0.030	0.240
	Submitted as R in CB3	42,560	0.093	0.744
	Submitted as R in CB4	54,927	0.121	0.960
	Submitted as R in IN	2,961	0.007	0.052
	Submitted as R in all runs	568	0.001	0.010

Table 13: Submission Data (message-level).

proportionate representation is made in the case of the very large All-N stratum, which is under-represented in the sample relative to its full-collection proportions, thereby allowing each of the R strata to be somewhat over-represented relative to their full-collection sizes. Selection within a stratum was made using simple random selection without replacement. The operative unit for selection into a sample was the message, and any message selected was included intact (parent email together with all attachments) in the sample.

Tables showing, for each topic, the stratum-by-stratum partitioning of the collection, the samples drawn from each stratum, and the pre- and post-adjudication assessments attached to those samples are provided in an appendix to this document (Appendix A). For purposes of this section, we present, in Table 14, a high-level view of the outcome of the stratification and sample selection process. In the table, we aggregate, for each topic, the totals for each of the individual R strata into a single row (labeled "R Strata," with the number of non-empty individual strata so aggregated noted in parentheses) and present the view of collection and sample composition that results.

		Messages				**Documents**			
		Full Collection		**Sample**		**Full Collection**		**Sample**	
Topic	**Stratum**	**Count**	**Prp**	**Count**	**Prp**	**Count**	**Prp**	**Count**	**Prp**
301	R Strata (29)	38,357	0.084	2,767	0.474	97,039	0.142	6,678	0.598
	All-N Stratum (1)	417,092	0.916	3,075	0.526	588,553	0.858	4,488	0.402
	Total (30)	455,449	1.000	5,842	1.000	685,592	1.000	11,166	1.000
302	R Strata (52)	4,874	0.011	1,979	0.342	17,368	0.025	6,746	0.549
	All-N Stratum (1)	450,575	0.989	3,800	0.658	668,224	0.975	5,534	0.451
	Total (53)	455,449	1.000	5,779	1.000	685,592	1.000	12,280	1.000
303	R Strata (53)	43,488	0.095	4,370	0.614	72,545	0.106	7,266	0.639
	All-N Stratum (1)	411,961	0.905	2,750	0.386	613,047	0.894	4,110	0.361
	Total (54)	455,449	1.000	7,120	1.000	685,592	1.000	11,376	1.000
304	R Strata (25)	57,216	0.126	3,491	0.516	114,869	0.168	6,915	0.601
	All-N Stratum (1)	398,233	0.874	3,275	0.484	570,723	0.832	4,594	0.399
	Total (26)	455,449	1.000	6,766	1.000	685,592	1.000	11,509	1.000

Table 14: Stratification & sampling — high-level view.

The table enables us to make a few observations. First, with regard to the size of samples, we see that the samples are all fairly large. We set out, taking into account the capacity of our review resources, to construct samples that were in the neighborhood of 11,500 documents for each topic. Given that our unit of selection was the message, however, and not the document, the number of documents included in each sample could not be precisely specified in advance. The results of the selection process can be seen in the table. In terms of messages, the samples ranged from 5,779 messages (for Topic 302) to 7,120 messages (for Topic 303), with the average size of a sample being 6,377 messages. In terms of documents, we see that the samples ranged in size from 11,166 documents (Topic 301), to 12,280 documents (Topic 302), with the average size of a sample coming to 11,583 documents.

Second, comparing the size of the set formed by aggregating the R strata to the size of the All-N stratum, we see, as we saw in the previous section (5.3.2), that, in the full collection, the R strata, collectively, represent a relatively small proportion of the population, representing, depending on topic, between 1% (Topic 302) and 13% (Topic 304) of the messages in the collection. Looking at representation in the sample, on the other hand, we see that, in accordance with our sampling design, the R strata are represented in higher proportions, and the All-N strata in lower proportions, than their full-collection proportions would dictate: the R strata represent between 34% (Topic 302) and 61% (Topic 303) of the messages in the evaluation samples.

Third, comparing, for each topic, the document-to-message ratio found in the subset formed by the R strata (full collection) to the document-to-message ratio found in the N stratum (also full collection), we see that, for all topics, the document-to-message ratio is higher in the R strata than it is in the N stratum. For Topic 301, the ratio of the ratios is 1.8 (i.e., the document-to message ratio in the aggregated R strata is 1.8 times that in the N stratum); for Topic 302, the ratio of the ratios is 2.4; for Topic 303, 1.1; and, for Topic 304, 1.4. The same trend was observed in the 2009 exercise [6]. The explanation for the trend could lie in the distribution of relevant documents across messages, in the nature of the retrieval systems evaluated, or in some combination of both. Determining which explanation is correct will require further analysis.

5.3.4 Assessment & Adjudication

As noted above (Section 5.1.1), the Interactive task follows a two-stage assessment protocol, whereby an initial relevance assessment is made of each document in each evaluation sample and then a selection of those first-pass assessments are escalated to the pertinent Topic Authority for final adjudication. In this section, we summarize the results of assessment and adjudication for the 2010 exercise.

First-Pass Assessment. Once the evaluation samples were drawn, they were made available to review teams for first-pass assessment. The review teams, for the 2010 exercise, were all staffed by commercial providers of document-review services. Four firms volunteered their services, with each firm taking responsibility for the review of the sample for one of the four 2010 topics.

In order to conduct their reviews, the review teams were provided with detailed assessment guidelines (compiled largely from the relevance guidance that the Topic Authority had provided the participants in the course of the exercise). In addition, at the outset of each review team's work, an orientation call was held with the Topic Authority for the team's topic; on the call, the Topic Authority outlined his or her approach to the topic, and the review team had the opportunity to ask any initial questions it had regarding the relevance criteria to be applied in assessing documents. Finally, once the review got under way, an email channel was opened, whereby the review team could ask the Topic Authority any questions that arose, whether regarding specific documents or regarding the relevance criteria in general, in the course of their assessment of the evaluation sample.

In assessing their samples, the review teams were instructed to make a relevance judgment (relevant (R) or not relevant (N)) for each document in their samples. A small number of documents in each sample were such as not to permit a relevance judgment by the review team (due, e.g., to errors in the processing of the data or due to non-English content); in these cases, the review teams were instructed to code the document as "broken." Out of the 46,331 documents reviewed across all four topics, 2,608 (5.6% of the total) were found to be "broken" and so counted as non-assessable.

It should be noted that, although, as described above (Section 5.1.2), each of the samples was organized into 25 batches, with each batch consisting of approximately 500 documents, it is not necessarily the case that the review firms that provided the sample assessments observed the batch organization in assigning documents to individual assessors. Some firms found it to be more in keeping with their usual review procedures simply to take the sample as a whole and then to allocate documents among assessors in the way that they believed would be most efficient. As a result, one should not assume that there exists a one-to-one relation between a given batch and any individual assessor.

Dual Assessment. As noted above (Section 5.1.2), for the 2010 exercise, we gathered a second independent assessment on a subset of the messages included in each evaluation sample. The dual-assessment subset was chosen by random selection from messages already included in the sample. In order to gather the second assessments, a second instance of each message selected for dual assessment was included in the full set passed to the review team (with the second instance randomly ordered with regard to the first). Both assessments were thus supplied by the same review team; indeed, it is not impossible that, in some cases, the same individual supplied both assessments. What we can say about the two assessments is that they represent distinct assessments of the same message on two different occasions.

Table 15 summarizes the overall rates of agreement achieved on the dual-assessed messages. The table shows, for each topic and for the aggregate of all four topics, (i) the total number of messages included in the dual-assessment subset, (ii) the number of those on which the two assessments were in agreement (both as a

count of messages and as a proportion of total), and (iii) the number of those on which the two assessments were in conflict (both as a count of messages and as a proportion of total).

Topic	Dual Asmnt Messages	Assessments Agree		Assessments Conflict	
		Msgs	Of Total	Msgs	Of Total
301	642	582	0.907	60	0.093
302	662	637	0.962	25	0.038
303	762	686	0.900	76	0.100
304	703	612	0.871	91	0.129
Total	2,769	2,517	0.909	252	0.091

Table 15: Dual Assessment — Overall Rates of Agreement.

As can be seen from the table, the overall rates of agreement on twice-assessed messages are fairly high, ranging from 87% (Topic 304) to 96% (Topic 302), with an aggregate rate of agreement of 91%. Overall rates of agreement can be misleading, however, in that, in cases in which the vast majority of documents are not relevant (as is typical in e-discovery), a high overall rate of agreement will result simply from the assessors' agreement on the large number of documents that are not remotely relevant. Looking at Topic 302, for example, we see that the assessors for this topic achieved the highest rate of agreement; we also know, however, judging from participant submissions (Table 13), that Topic 302 is likely to be significantly lower-yielding than the others (and this hypothesis is borne out by analysis of the post-adjudication assessment data, Section 5.3.5). We cannot say, therefore, whether the higher rate of agreement observed for Topic 302 is simply a function of the lower yield of the topic or genuinely reflects a higher level of consistency, on the part of the Topic 302 assessors, in applying the relevance criteria for that topic.

What we would like is a metric that is not sensitive to the yield of a topic, and for that we turn to the overlap metric. Overlap is a gauge of interassessor consistency when making positive assessments. More specifically, it tells us the proportion, out of all documents judged relevant by either of the assessors, that are judged relevant by both of the assessors (or, in other words, the proportion that the intersection of the two sets of R assessments represents out of the union of the two sets of R assessments).

Table 16 summarizes the overlap data on the dual-assessed messages. The table shows, for each topic and for the aggregate of all four topics, (i) the total number of dual-assessed messages that were judged relevant on at least one occasion, (ii) the number of dual-assessed messages that were judged relevant on both occasions (both as a count of messages and as a proportion of total), and (iii) the number of dual-assessed messages that were judged relevant on one occasion and non-relevant on the other occasion (both as a count of messages and as a proportion of total).

As can be seen from the table, when we confine our attention to just the dual-assessed messages that received a positive assessment on at least one occasion, we find rates of interassessor consistency that are much lower than the overall rates of agreement: overlap ranges from 31% (Topic 302) to 61% (Topic 303), coming to 50% on the aggregated data. Indeed, the dual-assessment set that showed the highest rate of overall agreement (Topic 302) also showed the lowest rate of overlap, suggesting that the high overall rate was in fact primarily a result of the low yield of the topic. These data are evidence that, while, overall, the assessors almost always agreed on the assessment to assign to a document, once in the neighborhood of potentially relevant documents, the assessors showed a lower rate of agreement as to how the relevance criteria should be applied.

The conflicts in assessment among the dual-assessed messages are obvious candidates for escalation to the Topic Authority for final adjudication. We next consider the data on the appeal and adjudication of the first-pass assessments.

Adjudication. As noted above (Section 5.1.2), for the 2010 exercise, the set of first-pass assessments

Topic	Union of R Asmnts	Assessments Agree		Assessments Conflict	
		Msgs	Of Total	Msgs	Of Total
301	132	72	0.545	60	0.455
302	35	11	0.314	24	0.686
303	189	115	0.608	74	0.392
304	146	55	0.377	91	0.623
Total	502	253	0.504	249	0.496

Table 16: Dual Assessment — Overlap.

escalated to the Topic Authority for final adjudication derived from two sources: (i) first-pass assessments appealed by one or more of the participants and (ii) non-appealed first-pass assessments.

With regard to the appeals, once first-pass assessment was complete, participants were provided (i) with the first-pass assessments made on all documents in the evaluation sample, (ii) with the (message-level) probability of selection associated with each document in the sample, and (iii) with preliminary (i.e., pre-adjudication) estimates of the recall, precision, and F_1 scores achieved in their submitted runs. Participants were then invited to appeal any first-pass assessments that they believed were incorrect (i.e., out of keeping with the Topic Authority's relevance guidance). Participants were asked to submit their appeals at the document level (i.e., on the assessment assigned to a specific parent or attachment). In a departure from previous implementations of the Interactive task, participants were not asked to prepare documentation of the grounds for their appeals; participants simply submitted lists of the IDs of the documents the assessments of which they wished to challenge. There was no limit on the number of assessments that a participant could appeal, and all appealed assessments were included in the set escalated to the Topic Authority for adjudication. Although appeals were made at the document level, messages were included intact in the adjudication set (i.e., if the assessment of any one component of a message was appealed, all components of that message were included in the set sent to the Topic Authority for adjudication).

With regard to the non-appealed messages included in the adjudication set, these derived from the following sources (all after the exclusion of messages already included in the adjudication set via the appeals process):

- dual-assessed messages on which the two first-pass assessments were in conflict;
- dual-assessed messages on which the two first-pass assessments were in agreement;
- single-assessed messages from the All-N stratum on which the first-pass assessment was Relevant;
- single-assessed messages from the All-N stratum on which the first-pass assessment was Not Relevant;
- single-assessed messages from the R-strata.

In composing the non-appealed subset to be included in the adjudication set, priority was given to cases of (non-appealed) dual-assessment conflict and to cases of R assessment in All-N stratum (i.e., the first and the third of the sources listed above). Selection of messages within each of the subsets was made via simple random selection without replacement, and the unit of selection was the message (not document).

Table 17 summarizes the composition of the adjudication set for each of the 2010 topics. Shown are the number of messages (and documents) included in the set both via the appeals process and via the sampling of non-appealed messages (both as counts and as proportions of the full sample).

As can be seen from the table, there was a fair amount of topic-to-topic variation in the number of messages appealed. Expressed as a percentage of the messages in the full evaluation sample, appeals ranged from 3.5% of the sample (Topic 304) to 11.3% of the sample (Topic 303), with the average across the four topics coming to 6.6% of the sample. Participants in Topic 303 clearly made the most extensive use of the

Topic	Source	Messages		Documents	
		Msgs	Of Sample	Docs	Of Sample
301	Appeal	397	0.068	811	0.073
	Non-Appeal	200	0.034	357	0.032
	Total	597	0.102	1,168	0.105
302	Appeal	282	0.049	898	0.073
	Non-Appeal	198	0.034	606	0.049
	Total	480	0.083	1,504	0.122
303	Appeal	802	0.113	1,773	0.156
	Non-Appeal	187	0.026	270	0.024
	Total	989	0.139	2,043	0.180
304	Appeal	237	0.035	564	0.049
	Non-Appeal	299	0.044	543	0.047
	Total	536	0.079	1,107	0.096
All Topics	Appeal	1,718	0.067	4,046	0.087
	Non-Appeal	884	0.035	1,776	0.038
	Total	2,602	0.102	5,822	0.126

Table 17: Adjudication Set — sources.

appeals mechanism, collectively appealing more than twice the number of messages appealed by participants in any of the other topics.

As for non-appealed messages included in the adjudication set, the budget for these was largely a matter of the adjudication capacity of the Topic Authorities, and so the numbers are fairly consistent across topics: as a percentage of messages in the full evaluation sample, non-appealed messages included in the adjudication set ranged from 2.6% of the sample (Topic 303) to 4.4% of the sample (Topic 304), with the average across the four topics coming to 3.5% of the sample.

The full adjudication sets (both appealed and non-appealed) represent fairly sizeable proportions of the evaluation samples. In terms of messages, between 7.9% (Topic 304) and 13.9% (Topic 303) of the evaluation samples were escalated to the Topic Authority for final adjudication.

Once selected, the adjudication sets were made available to the Topic Authorities for final assessment. In making their assessments, the Topic Authorities had access to the assessment guidelines they had prepared for the first-pass assessors, as well as any other materials they had compiled in the course of their interactions with the participants. The Topic Authorities did not, as noted above, have access to any documentation of the grounds on which a participant was appealing a given first-pass assessment. The Topic Authorities were not, in fact, made aware of the first-pass assessments that initially had been rendered on the documents in their adjudication sets nor were they made aware of which documents had been included in the set via an appeal and which had been otherwise included. Topic Authorities were asked to make their assessments at the document level (taking appropriate account of the context provided by message body and associated attachments, as in the initial assessment stage).

The results of the adjudication process are summarized in Table 18. For each topic, the table breaks down the results by source (appealed, non-appealed, total), and shows (i) the total number of messages adjudicated from the source, (ii) the total number of messages for which adjudication resulted in no change in the (message-level) assessment (both as a count and as a proportion of all messages adjudicated from the given source), and the total number of messages for which adjudication did result in a change in the assessment (both as a count and as a proportion of all messages adjudicated from the given source).

Topic	Source	Total Msgs	No Change Msgs	No Change Of Source	Change Msgs	Change Of Source
301	Appeal	397	250	0.630	147	0.370
	Non-Appeal	200	138	0.690	62	0.310
	Total	597	388	0.650	209	0.350
302	Appeal	282	164	0.582	118	0.418
	Non-Appeal	198	176	0.889	22	0.111
	Total	480	340	0.708	140	0.292
303	Appeal	802	498	0.621	304	0.379
	Non-Appeal	187	145	0.775	42	0.225
	Total	989	643	0.650	346	0.350
304	Appeal	237	150	0.633	87	0.367
	Non-Appeal	299	211	0.706	88	0.294
	Total	536	361	0.674	175	0.326
All Topics	Appeal	1,718	1,062	0.618	656	0.382
	Non-Appeal	884	670	0.758	214	0.242
	Total	2,602	1,732	0.666	870	0.334

Table 18: Adjudication Set — summary of results.

The table enables a few observations. First, we see that, for appealed messages, the rate at which first-pass assessments were overturned is fairly consistent across topics. Overturn rates range from 36.7% (Topic 304) to 41.8% (Topic 302); on average, about 38% of the messages included in the adjudication set via the appeals process saw a change in assessment.

Second, we see that some non-appealed messages also saw a change in assessment as a result of adjudication by the Topic Authority. Overturn rates for non-appealed messages range from 11.1% (Topic 302) to 31.0% (Topic 301), with an average across topics of 23.5%.

Third, while, for all topics, the overturn rate was greater for the appealed messages than it was for the non-appealed messages, the difference in rates was, at least for some of the topics, not as great as we might have expected. Comparing the odds of an overturn for an appealed message to the odds of an overturn for a non-appealed message, we see that, for Topic 301, the odds ratio is 1.31 (the odds of an appealed message having its first-pass assessment overturned are just 1.3 times those of a non-appealed message having its first-pass assessment overturned). For Topic 302, on the other hand, we see a much greater difference in overturn rates: the odds ratio for Topic 302 is 5.76. For Topic 303, the odds ratio is 2.11; for Topic 304, the ratio is 1.39.

We take a closer look at some of the implications of these adjudication data below (Section 5.4). For now, we turn to participants' final scores.

5.3.5 Final Results

Once the Topic Authorities had completed their reviews of their adjudication sets and the sample assessments had been finalized, we were in a position to calculate final estimates of the overall yield for each topic and of the recall, precision, and F_1 achieved in each run submitted by participants. Before turning to those estimates, we add two further notes by way of background to our calculations.

The first note concerns the derivation of message-level values in cases of partially-assessed messages (i.e., cases in which a message has one or more components that have been assessed as "broken"). In most of these cases, we simply ignore the unjudged components and derive the message-level value on the basis of

the judged components: if any judged component is assessed as Relevant, the message counts as Relevant; if all of the judged components are assessed as Not Relevant, the message counts as Not Relevant. There are, however, a small number of cases in which none of the judged components have been assessed as Relevant, but there is at least one unjudged component that has been submitted by a participant as Relevant; in this case, the message counts as Unjudged (rather than Not Relevant), because the component(s) that the participant found Relevant did not receive a definitive assessment.

The second note concerns the use of the adjudicated assessments. As noted above, some of the (non-appealed) messages included in the adjudication set were chosen via random selection from pertinent subsets of the evaluation sample. For purposes of calculating the estimates reported below, any changes in assessment on such messages that occurred as a result of adjudication affect only the message actually adjudicated: we do not project from the messages selected for adjudication to the larger subsets from which they were drawn. Put another way, the basis for calculating participant scores was simply the set of post-adjudication assessments associated with the full evaluation sample; the procedures used to arrive at those scores are those detailed in the Overview to the 2008 Legal Track [7].

We now turn to the estimates themselves. Table 19 reports the estimated full-collection yield of relevant messages for each of the four Interactive topics; yield is reported both as a count of messages and as a proportion of the full collection. (Recall that the full collection consisted of 455,449 messages.)

	Relevant Messages		**Of Full Collection**	
Topic	**Est.**	**95% C.I.**	**Est.**	**95% C.I.**
301	18,973	(16,688, 21,258)	0.042	(0.037, 0.047)
302	575	(174, 976)	0.001	(0.0004, 0.002)
303	12,124	(11,261, 12,987)	0.027	(0.025, 0.029)
304	20,176	(18,427, 21,925)	0.044	(0.040, 0.048)

Table 19: Estimated yields (C.I.=Confidence Interval).

As can be seen from the table, our hypothesis, based on the submission data (see Section 5.3.2), that Topic 302 was low yielding has been borne out by the assessment results: we estimate that just 0.1% of the messages in the collection are relevant to this topic (which was on the subject of responses to oil and gas spills). The other three topics, on the other hand, while not extremely high-yielding, do find representation in substantial numbers of emails: 4.2% of the collection is relevant to Topic 301, 2.7% to Topic 303, and 4.4% to Topic 304.

Table 20 reports measures of how effective the participants were at retrieving the messages relevant to each topic. More specifically, the table reports, for each run submitted, estimates of the message-level recall, precision, and F_1 achieved in the run.

The data presented in the table permit a few observations on the results observed for each topic. With regard to Topic 301, we see that, while some of the submitted runs achieved relatively high levels of precision (with three of the five runs scoring over 50% on the point estimate for this metric), all of the runs found recall a challenge (with no run scoring above 25% on recall). Looking at the relatively high yield of this topic (4.2%), we see that the Topic Authority took a rather broad view of what was relevant to Topic 301; it appears that none of the participants succeeded in capturing what the Topic Authority viewed as the full scope of the topic.

With regard to Topic 302, the lowest yielding of the topics, we see that participants again scored better on precision than they did on recall: five of the six runs score better than 40% on precision, but none of the six exceed 25% on recall.

The strongest scores were turned in for Topic 303 (on the subject of lobbying). Five of the six runs for this topic scored above 70% on either precision or recall, and two of the six scored above 50% on both

Topic	Run	Recall		Precision		F_1	
		Est.	**95% C.I.**	**Est.**	**95% C.I.**	**Est.**	**95% C.I.**
301	CS	**0.165**	(0.142, 0.187)	**0.579**	(0.541, 0.616)	**0.256**	(0.229, 0.284)
	IT	**0.205**	(0.174, 0.236)	**0.295**	(0.268, 0.322)	**0.242**	(0.219, 0.265)
	SF	**0.239**	(0.204, 0.274)	**0.193**	(0.177, 0.210)	**0.214**	(0.197, 0.231)
	IS	**0.027**	(0.023, 0.031)	**0.867**	(0.781, 0.952)	**0.052**	(0.045, 0.060)
	UW	**0.019**	(0.014, 0.023)	**0.578**	(0.465, 0.691)	**0.036**	(0.028, 0.045)
302	UM	**0.200**	(0.060, 0.340)	**0.450**	(0.426, 0.475)	**0.277**	(0.143, 0.411)
	UW	**0.169**	(0.051, 0.288)	**0.732**	(0.691, 0.773)	**0.275**	(0.119, 0.431)
	MM	**0.115**	(0.035, 0.195)	**0.410**	(0.395, 0.426)	**0.180**	(0.082, 0.277)
	LA	**0.096**	(0.029, 0.163)	**0.481**	(0.445, 0.517)	**0.160**	(0.066, 0.253)
	IS	**0.090**	(0.027, 0.153)	**0.693**	(0.693, 0.693)	**0.160**	(0.061, 0.259)
	IN	**0.135**	(0.039, 0.232)	**0.017**	(0.015, 0.020)	**0.031**	(0.026, 0.035)
303	EQ	**0.801**	(0.738, 0.865)	**0.577**	(0.557, 0.597)	**0.671**	(0.645, 0.697)
	CB2	**0.572**	(0.526, 0.617)	**0.705**	(0.680, 0.730)	**0.631**	(0.602, 0.661)
	CB1	**0.452**	(0.415, 0.488)	**0.734**	(0.706, 0.762)	**0.559**	(0.530, 0.588)
	UB	**0.723**	(0.665, 0.781)	**0.300**	(0.289, 0.311)	**0.424**	(0.409, 0.439)
	IT	**0.248**	(0.226, 0.271)	**0.259**	(0.245, 0.273)	**0.254**	(0.240, 0.267)
	UW	**0.134**	(0.121, 0.147)	**0.773**	(0.722, 0.824)	**0.228**	(0.209, 0.247)
304	CB3	**0.633**	(0.568, 0.698)	**0.302**	(0.285, 0.318)	**0.408**	(0.388, 0.429)
	CB4	**0.715**	(0.643, 0.788)	**0.264**	(0.250, 0.278)	**0.385**	(0.367, 0.404)
	CB2	**0.271**	(0.239, 0.303)	**0.402**	(0.370, 0.435)	**0.324**	(0.298, 0.349)
	CB1	**0.201**	(0.175, 0.228)	**0.327**	(0.295, 0.360)	**0.249**	(0.227, 0.272)
	IN	**0.072**	(0.059, 0.086)	**0.494**	(0.416, 0.572)	**0.126**	(0.106, 0.146)

Table 20: Post-adjudication estimates of recall, precision, and F_1.

precision and recall. It appears that participants in this topic were generally more successful in capturing the Topic Authority's understanding of the intent and scope of the topic.

For Topic 304, the "privilege" topic, the two runs that scored highest on F_1 did so by achieving relatively high scores on recall (greater than 60%) while scoring lower on precision (less than 35%). Interestingly, these were the two runs that deliberately took a broad view of the topic (see Section 5.2.3); it may be that retrieval efforts that focus on "potentially privileged" (rather than genuinely privileged) are more successful at capturing the genuinely privileged material, even if that comes at the cost of some loss of precision.

The results for all four topics are summarized in Figure 10. The figure plots the post-adjudication results for each of the 22 submitted runs on a precision-recall diagram. In the figure, topics are distinguished by shape as per the legend. That the majority of the points lie on the left-hand side of the diagram underlines the fact that the submissions in the 2010 Interactive task generally performed better on the precision metric than they did on recall.

5.4 Further Analysis

The results of the 2010 Interactive task raise a number of questions that merit further study. In this section, we confine ourselves to a brief look at a few points of interest.

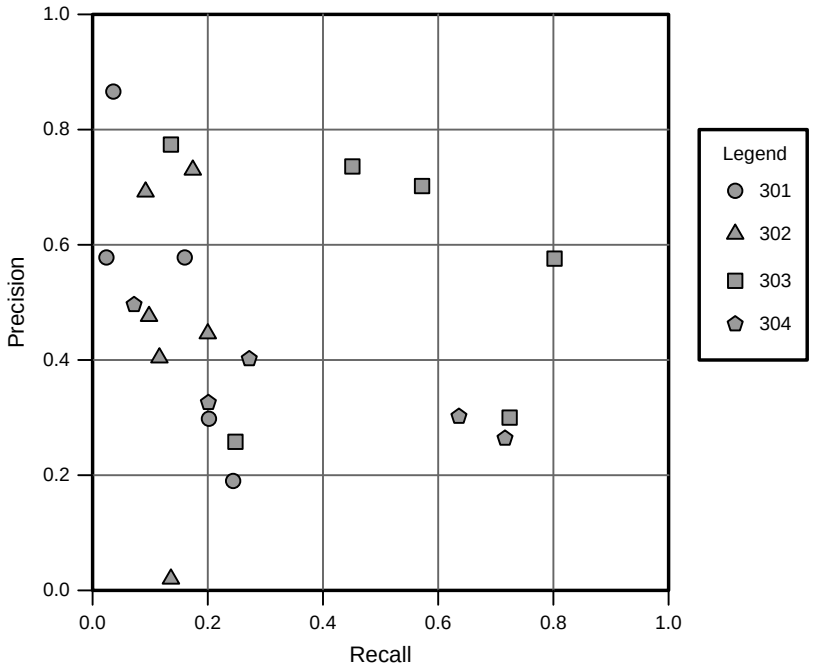

Figure 10: Interactive runs — recall and precision.

5.4.1 Team-TA Interaction

Earlier (Section 5.3.1), we saw that there was considerable variation in the amount of time teams chose to spend with the Topic Authorities for the purpose of clarifying the intent and scope of the target topics; times ranged from zero minutes in three instances to 504 minutes in another instance. Such variation prompts the question of whether there is a correlation between the amount of time spent with a Topic Authority and retrieval effectiveness.

Figure 11 plots, for each run in the 2010 exercise, the run's retrieval effectiveness (as measured by post-adjudication F_1 scores) against the time spent with the Topic Authority in preparing the run.

As can be seen from the chart, it is true that the run that scored highest, in terms of F_1, of any of the 2010 runs (303-EQ), is also the run that utilized the most TA time (504 minutes), suggesting that perhaps there is some correlation between effectiveness and time spent with the Topic Authority. When we look at the other runs, however, it is hard to discern a pattern: some of the runs that scored low on F_1 used a lot of TA time, and some of the runs that scored relatively high on F_1 made limited use of the Topic Authority's time. When we test for the significance of the correlation, moreover, by calculating an estimate and 95% confidence interval for the Pearson product-moment correlation coefficient, we find that our doubts about a correlation are borne out; while the point estimate for the coefficient is positive (0.200), the 95% confidence interval (-0.242, 0.573) includes zero: the data are not evidence of a positive correlation between effectiveness and time spent with the Topic Authority.

These results, on the question of Team-TA interaction, for the 2010 exercise are similar to those obtained in the 2009 exercise. Evidently, in looking for factors that drive effectiveness, we have to look not merely at quantitative measures of Team-TA interaction but also at qualitative aspects of that interaction.

5.4.2 Resource Utilization

Apart from the question of how well the amount of time spent with the Topic Authority correlates with retrieval effectiveness, we are also interested in how well other measures of resource utilization correlate

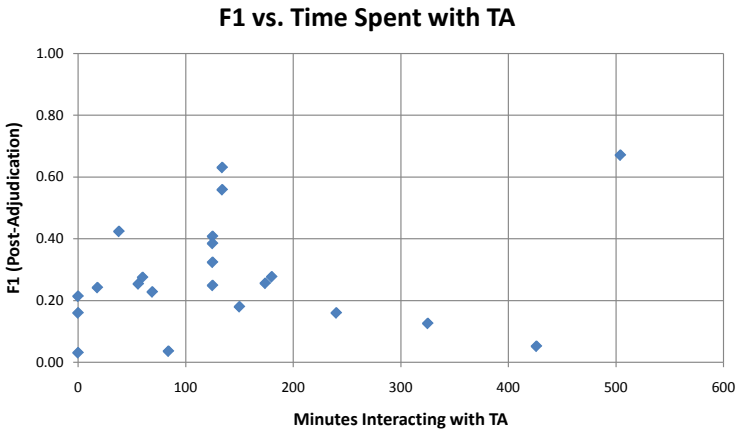

Figure 11: Interactive runs — F_1 vs. TA-time.

with effectiveness. For the 2010 Interactive task, we looked at two such measures: (i) overall time spent in preparing a submission and (ii) the number of documents reviewed in the course of preparing a submission.

With regard to the first measure (overall preparation time), we asked that participants, upon submission of their results, report an estimate of the total person-hours they had spent in preparing their submissions for each topic. In estimating their total hours, participants were asked to include any hours spent either on work specific to a particular topic or on work specific to the loading and analysis of the data set; participants were told not to include any hours spent on the development of tools and methods that, though used for the Interactive task, had general application beyond the exercise. With regard to the second measure (documents reviewed), participants were asked to report the number of documents that their team had manually reviewed in the course of preparing their submission.

Figure 12 looks at how well each of these measures of resource utilization correlates with effectiveness. The figure shows two charts. In the left-hand chart, we plot, for each run in the 2010 exercise, the run's retrieval effectiveness (as measured by post-adjudication F_1 scores) against the total time spent (in person-hours) in preparing the run. In the right-hand chart, we plot, for each run in the 2010 exercise, the run's retrieval effectiveness (again, as measured by post-adjudication F_1 scores) against the total number of documents manually reviewed in the course of preparing the run,

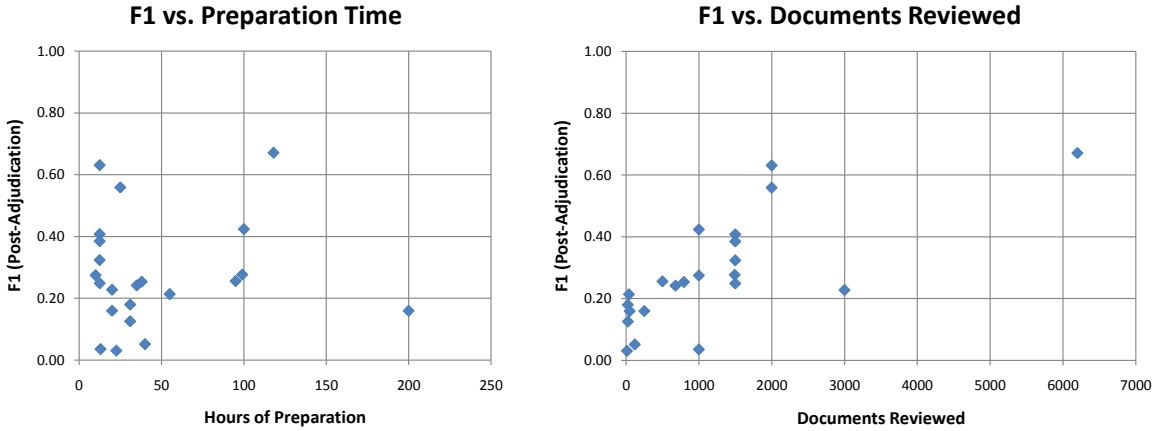

Figure 12: Interactive runs — F_1 vs. Preparation Time and F_1 vs. Documents Reviewed.

Simple visual inspection of the two charts finds that, if any of these measures of resource utilization is to correlate well with retrieval effectiveness, that measure is likely to be documents reviewed: the data points on the documents-reviewed chart pattern more tightly in a linear trajectory than do the data points on the preparation-time chart. The hypothesis prompted by visual inspection is borne out by the calculation of estimates and confidence intervals for the Pearson product-moment correlation coefficient. For the correlation between preparation time and F_1, we estimate the coefficient to be 0.079, with a 95% confidence interval of (-0.354, 0.485): the data are not evidence of a significant correlation between preparation time and effectiveness (as measured by F_1). For the correlation between documents reviewed and F_1, on the other hand, we estimate the coefficient to be 0.721, with a 95% confidence interval of (0.430, 0.876): the data are evidence of a significant correlation between documents reviewed and effectiveness (as measured by F_1).

In this section, together with the previous, we have looked at three variables that might be expected to correlate, positively, with the effectiveness of a submission: time spent interacting with the Topic Authority, overall time spent in preparing a submission, and the number of documents manually reviewed in the course of preparing a submission. Of these three, only one, documents reviewed, has in fact been found to have a significant correlation with the F_1 realized by s submission.

This is not to say that the other two factors (interaction with the Topic Authority, preparation time) have no impact on the effectiveness of a retrieval effort; it is rather to say that, if these factors do make a contribution to effectiveness, we have to capture the nature of that contribution in something other than minutes spent with the Topic Authority or hours spent in preparation. We may have to look at the quality, rather than the quantity, of time spent.

5.4.3 Adjudication

The results of the adjudication process always merit further analysis, as such analysis can provide insights into how to make the process more efficient and effective. A full analysis of the results, however, is beyond the scope of this overview; for purposes of this report, we briefly touch on two aspects of the 2010 adjudication results.

First, with regard to the results for the appealed messages that were included in the adjudication sets, we noted above (Section 5.3.4) that a fair number of these saw a change in assessment as a result of the adjudication process: on average, the assessments on about 38% of the messages included in the adjudication sets via appeal were overturned. We also noted above, that, for the 2010 exercise, we did not provide the Topic Authorities with participant-prepared documentation of the grounds for their appeals. Now, in previous years, we did provide the Topic Authorities with such documentation, and, in previous years, we also observed higher rates of overturn on appealed documents, with overturn rates regularly exceeding 70% (see the 2009 Track Overview [6]). A question for further study, therefore, is whether the absence of appeals documentation, in the 2010 exercise, resulted, on occasion, in a Topic Authority's missing a salient feature of a document to which such documentation could have directed his or her attention.

Second, with regard to the results for the non-appealed messages that were included in the adjudication sets, we noted above that the assessments on some of these were also overturned: averaging across the four topics, the assessments on about 23.5% of non-appealed messages included in the adjudication sets were changed as a result of the adjudication process. We now look at how the overturn rates vary by the specific source from which the non-appealed messages were drawn.

Recall that the non-appealed messages included in the adjudication sets were drawn from five sources: (i) dual-assessment conflicts; (ii) dual-assessment agreements; (iii) single-assessment R's from the All-N stratum; (iv) single-assessment N's from the All-N stratum; and (v) single-assessment R's and N's from the R strata. Table 21 breaks down the adjudication results by source. The table shows the overturn rate observed on messages from each source, as well as, in parentheses, the number of overturns over the number of messages adjudicated. Results are shown for each topic and for the aggregate results of all four topics.

As can be seen from the table, of the non-appealed messages included in the adjudication sets, those with the highest overturn rate are the dual-assessment conflicts: in aggregate, over 50% of the messages included from this source saw a change in assessment as a result of adjudication. This is not surprising, given that a conflicting assessment had already been rendered on each of these. The messages with the second highest

Source	Overturn Rate (Overturned/Adjudicated)				
	Topic 301	**Topic 302**	**Topic 303**	**Topic 304**	**Aggregate**
Dual – Conflict	0.683 (28/41)	0.444 (4/9)	0.415 (17/41)	0.487 (37/76)	0.515 (86/167)
Dual – Agree	0.150 (3/20)	0.033 (2/60)	0.100 (2/20)	0.100 (2/20)	0.075 (9/120)
Single – All-N R	0.200 (4/20)	0.000 (0/0)	0.400 (2/5)	0.000 (0/5)	0.200 (6/30)
Single – All-N N	0.100 (2/20)	0.000 (0/30)	0.000 (0/20)	0.050 (2/40)	0.036 (4/110)
Single – R Strata	0.253 (25/99)	0.162 (16/99)	0.208 (21/101)	0.297 (47/158)	0.239 (109/457)
Total	0.310 (62/200)	0.111 (22/198)	0.225 (42/187)	0.294 (88/299)	0.242 (214/884)

Table 21: Non-Appealed Adjudications — Results by Source.

overturn rate are the single-assessed messages from the R strata: in aggregate, the assessments on about 24% of the messages drawn from this source were changed as a result of adjudication. This too is perhaps not surprising, given that, in most cases, a message in an R stratum will have been found Relevant by at least one participant and Not Relevant by at least one other participant. Indeed, of the 214 non-appealed messages that saw a change in assessment as a result of adjudication, 195 came from one of the two sources just noted. Fewer messages were drawn from the other sources, but they also had lower rates of overturn, with those rates, on the aggregated results, ranging from 20% (for the All-N R's) down to less than 4% (for the All-N N's; note also that, for two of the four topics, the overturn rate on messages from this source was 0%). Understanding the implications of these results will require further study.

We look forward to continuing the analysis of the adjudication data, and of the 2010 Interactive task more generally, in other papers and venues.

6 Conclusion

This has been the fifth year of the TREC Legal Track, and our second year of building test collections based on Enron email [3, 6, 7, 8]. Relevance judgments are now available for 11 topical production requests and now also for privilege. The Legal Track will continue in 2011 with an expanded Learning task.

Acknowledgments

The TREC Legal Track could not exist without the countless hours of pro bono effort afforded by the Topic Authorities, and professional and law student reviewers. Special thanks to John M. Horan, who acted as a surrogate Topic Authority for all eight topics of the Learning task, and who spent countless hours supervising some fifty students in the completion of 135 review batches. Thanks are due to the students of Loyola Law School, who took on the majority of those assignments. The coordinators recognize Denise Tirrell of Loyola Law School as the reviewer with the highest accuracy on any batch of documents in the Learning task. In a batch of 496 documents, Denise correctly identified 194 of 198 relevant documents, and 297 of 298 non-relevant documents—an overall accuracy of 99%. For their contributions to the interactive task, we owe special thanks to the professional review firms that provided the pro bono review resources needed to assess the evaluation samples (BIA, Huron, Daegis, and Aphelion Legal Solutions) and are especially grateful to our dedicated Topic Authorities, who gave generously of their time both in providing guidance to participants and in adjudicating assessments (Mira Edelman, John F. Curran, Robert E. Singleton, and Michael Roman Geske).

References

[1] TREC-2010 Legal Track – Complaint K, 2010.
Available at http://trec-legal.umiacs.umd.edu/LT10_Complaint_K_final-corrected.pdf.

[2] Jason R. Baron, Bruce Hedin, Douglas W. Oard, and Stephen Tomlinson. Interactive Task Guidelines – TREC-2008 Legal Track, 2008. Available at http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf.

[3] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 Legal Track Overview. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, pages 79–98, 2007.

[4] Viv Bewick, Lic Cheek, and Jonathan Ball. Statistics review 13: Receiver operating characteristics curves. *Critical Care*, 8(6):508–512, 2004.

[5] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Interactive Task Guidelines – TREC-2010 Legal Track, 2010. Available at http://trec-legal.umiacs.umd.edu/itg10_final.pdf.

[6] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 Legal Track. In *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2010.

[7] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 Legal Track. In *The Seventeenth Text REtrieval Conference (TREC 2008)*, 2009.

[8] Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, and Paul Thompson. Overview of the TREC 2007 Legal Track. In *The Sixteenth Text Retrieval Conference (TREC 2007)*, 2008.

[9] William Webber, Douglas W. Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, 2010.

A Sampling & Assessment Tables—Interactive Task

In this appendix, we present tables that summarize, for each of the four Interactive topics, the results of the sampling and assessment process followed in the 2010 exercise. Each table shows: (i) total messages in each stratum (in the full collection); (ii) total messages sampled from each stratum; (iii) total sampled messages observed to be assessable; (iv) total sampled messages observed to be assessable and relevant (pre-adjudication); and (v) total sampled messages observed to be assessable and relevant (post-adjudication). It is on the basis of the data contained in these tables that we arrived at the estimates of the message-level recall, precision, and F_1 attained in each run.

Each table is structured as follows. The leftmost columns represent the relevance values (R = Relevant; N = Not Relevant) from the participant submissions that define each stratum. The right-hand columns show the counts of messages in each stratum; more specifically, the columns show the following data:

N = total messages in the stratum;

n = total messages sampled from the stratum;

a = total sampled messages observed to be assessable;

r_1 = total sampled messages observed to be assessable and relevant (pre-adjudication);

r_2 = total sampled messages observed to be assessable and relevant (post-adjudication).

The tables for the four Interactive topics follow.

Stratum					Counts (Messages)				
CS	**IS**	**IT**	**SF**	**UW**	N	n	a	r_1	r_2
R	R	R	R	R	18	2	2	2	2
R	R	R	R	N	50	5	5	5	5
R	R	R	N	R	6	2	2	1	1
R	R	R	N	N	8	2	2	2	2
R	R	N	R	R	5	2	2	2	2
R	R	N	R	N	102	9	9	7	7
R	R	N	N	R	17	2	2	1	1
R	R	N	N	N	42	4	4	2	4
R	N	R	R	R	60	5	5	3	5
R	N	R	R	N	131	12	12	9	9
R	N	R	N	R	71	6	6	3	4
R	N	R	N	N	301	27	27	23	23
R	N	N	R	R	68	6	6	2	3
R	N	N	R	N	2,145	231	228	132	142
R	N	N	N	R	38	3	3	1	1
R	N	N	N	N	2,366	251	251	93	115
N	R	R	R	R	2	2	1	1	0
N	R	R	R	N	2	2	2	0	2
N	R	R	N	R	0	0	0	0	0
N	R	R	N	N	0	0	0	0	0
N	R	N	R	R	1	1	1	0	1
N	R	N	R	N	142	13	13	9	12
N	R	N	N	R	1	1	1	1	1
N	R	N	N	N	197	18	18	8	15
N	N	R	R	R	35	3	3	3	3
N	N	R	R	N	831	91	91	38	33
N	N	R	N	R	97	9	9	8	7
N	N	R	N	N	11,558	780	780	215	198
N	N	N	R	R	59	5	5	2	2
N	N	N	R	N	19,863	1,260	1,259	169	151
N	N	N	N	R	141	13	13	4	3
N	N	N	N	N	417,092	3,075	3,075	97	72
TOTAL					**455,449**	**5,842**	**5,837**	**843**	**826**

Table 22: Sampling & assessment — Topic 301.

Stratum						Counts (Messages)				
IN	**IS**	**LA**	**MM**	**UM**	**UW**	N	n	a	r_1	r_2
R	R	R	R	R	R	2	2	2	1	2
R	R	R	R	R	N	5	5	5	0	5
R	R	R	R	N	R	2	2	2	2	2
R	R	R	R	N	N	4	4	4	1	4
R	R	R	N	R	R	1	1	1	1	1
R	R	R	N	R	N	1	1	1	0	0
R	R	R	N	N	R	0	0	0	0	0
R	R	R	N	N	N	1	1	0	0	0
R	R	N	R	R	R	2	2	2	2	2
R	R	N	R	R	N	1	1	1	1	1
R	R	N	R	N	R	0	0	0	0	0
R	R	N	R	N	N	2	2	2	2	2
R	R	N	N	R	R	0	0	0	0	0
R	R	N	N	R	N	2	2	1	0	0
R	R	N	N	N	R	0	0	0	0	0
R	R	N	N	N	N	4	4	4	1	0
R	N	R	R	R	R	1	1	1	1	1
R	N	R	R	R	N	0	0	0	0	0
R	N	R	R	N	R	0	0	0	0	0
R	N	R	R	N	N	3	3	3	0	0
R	N	R	N	R	R	3	3	3	2	3
R	N	R	N	R	N	1	1	1	1	0
R	N	R	N	N	R	0	0	0	0	0
R	N	R	N	N	N	12	12	12	3	4
R	N	N	R	R	R	2	2	2	1	1
R	N	N	R	R	N	14	14	14	5	4
R	N	N	R	N	R	1	1	1	1	1
R	N	N	R	N	N	24	24	23	3	2
R	N	N	N	R	R	3	3	3	3	3
R	N	N	N	R	N	21	21	21	9	11
R	N	N	N	N	R	17	17	17	12	12
R	N	N	N	N	N	4,354	1,550	1,544	28	6
N	R	R	R	R	R	10	10	9	6	9
N	R	R	R	R	N	4	4	4	1	3
N	R	R	R	N	R	1	1	1	1	1
N	R	R	R	N	N	5	5	5	4	4
N	R	R	N	R	R	3	3	3	3	3
N	R	R	N	R	N	1	1	1	0	0
N	R	R	N	N	R	0	0	0	0	0
N	R	R	N	N	N	3	3	2	1	0
N	R	N	R	R	R	3	3	3	1	3
N	R	N	R	R	N	1	1	1	0	0
N	R	N	R	N	R	0	0	0	0	0
N	R	N	R	N	N	4	4	4	2	4
N	R	N	N	R	R	1	1	1	0	1
N	R	N	N	R	N	5	5	3	2	2
N	R	N	N	N	R	0	0	0	0	0
N	R	N	N	N	N	20	20	13	8	3
N	N	R	R	R	R	1	1	1	0	1
N	N	R	R	R	N	1	1	0	0	0
N	N	R	R	N	R	0	0	0	0	0
N	N	R	R	N	N	1	1	1	0	1
N	N	R	N	R	R	1	1	1	1	1
N	N	R	N	R	N	4	4	3	3	1
N	N	R	N	N	R	2	2	2	2	1
N	N	R	N	N	N	48	30	29	2	5
N	N	N	R	R	R	3	3	3	1	1
N	N	N	R	R	N	15	15	15	6	7
N	N	N	R	N	R	6	6	6	3	2
N	N	N	R	N	N	46	30	30	7	2
N	N	N	N	R	R	20	20	19	9	13
N	N	N	N	R	N	132	95	94	30	26
N	N	N	N	N	R	50	30	30	15	20
N	N	N	N	N	N	450,575	3,800	3,800	4	3
TOTAL						**455,449**	**5,779**	**5,754**	**192**	**184**

Table 23: Sampling & assessment — Topic 302.

Stratum						Counts (Messages)				
CB1	**CB2**	**EQ**	**IT**	**UB**	**UW**	N	n	a	r_1	r_2
R	R	R	R	R	R	321	27	27	24	25
R	R	R	R	R	N	1,128	113	113	96	104
R	R	R	R	N	R	27	2	2	2	2
R	R	R	R	N	N	114	9	9	9	9
R	R	R	N	R	R	328	27	27	26	25
R	R	R	N	R	N	4,121	421	414	315	310
R	R	R	N	N	R	24	2	2	2	2
R	R	R	N	N	N	452	37	37	28	30
R	R	N	R	R	R	14	2	2	2	1
R	R	N	R	R	N	63	5	5	0	0
R	R	N	R	N	R	0	0	0	0	0
R	R	N	R	N	N	6	2	2	1	1
R	R	N	N	R	R	10	2	2	1	1
R	R	N	N	R	N	657	69	68	39	17
R	R	N	N	N	R	8	2	2	2	2
R	R	N	N	N	N	127	11	11	9	4
R	N	R	R	R	R	0	0	0	0	0
R	N	R	R	R	N	7	2	2	2	2
R	N	R	R	N	R	0	0	0	0	0
R	N	R	R	N	N	1	1	1	1	1
R	N	R	N	R	R	3	2	2	0	1
R	N	R	N	R	N	80	7	7	4	2
R	N	R	N	N	R	0	0	0	0	0
R	N	R	N	N	N	7	2	2	2	2
R	N	N	R	R	R	0	0	0	0	0
R	N	N	R	R	N	2	2	2	0	0
R	N	N	R	N	R	0	0	0	0	0
R	N	N	R	N	N	0	0	0	0	0
R	N	N	N	R	R	0	0	0	0	0
R	N	N	N	R	N	32	3	3	0	0
R	N	N	N	N	R	0	0	0	0	0
R	N	N	N	N	N	4	2	2	1	0
N	R	R	R	R	R	94	8	8	8	8
N	R	R	R	R	N	220	18	18	14	14
N	R	R	R	N	R	7	2	2	2	2
N	R	R	R	N	N	50	4	4	3	4
N	R	R	N	R	R	79	7	7	5	6
N	R	R	N	R	N	1,093	110	103	69	70
N	R	R	N	N	R	31	3	3	2	2
N	R	R	N	N	N	357	30	29	25	23
N	R	N	R	R	R	5	2	2	0	0
N	R	N	R	R	N	38	3	3	3	3
N	R	N	R	N	R	0	0	0	0	0
N	R	N	R	N	N	12	2	2	2	2
N	R	N	N	R	R	13	2	2	2	0
N	R	N	N	R	N	447	37	34	9	3
N	R	N	N	N	R	4	2	2	2	1
N	R	N	N	N	N	175	14	14	8	2
N	N	R	R	R	R	224	19	19	16	17
N	N	R	R	R	N	589	59	58	33	34
N	N	R	R	N	R	62	5	5	5	5
N	N	R	R	N	N	267	22	22	9	10
N	N	R	N	R	R	249	21	21	15	17
N	N	R	N	R	N	3,369	344	326	143	143
N	N	R	N	N	R	113	9	9	8	5
N	N	R	N	N	N	3,828	392	384	77	72
N	N	N	R	R	R	48	4	4	3	3
N	N	N	R	R	N	990	102	102	16	11
N	N	N	R	N	R	64	5	5	5	5
N	N	N	R	N	N	7,311	745	743	28	22
N	N	N	N	R	R	121	10	10	4	3
N	N	N	N	R	N	15,840	1,616	1,553	135	47
N	N	N	N	N	R	252	21	21	6	8
N	N	N	N	N	N	411,961	2,750	2,750	40	7
TOTAL						**455,449**	**7,120**	**7,009**	**1,263**	**1,090**

Table 24: Sampling & assessment — Topic 303.

Stratum					Counts (Messages)				
CB1	**CB2**	**CB3**	**CB4**	**IN**	N	n	a	r_1	r_2
R	R	R	R	R	568	27	27	12	15
R	R	R	R	N	8,464	519	510	175	184
R	R	R	N	R	0	0	0	0	0
R	R	R	N	N	2	2	2	1	1
R	R	N	R	R	5	2	2	0	0
R	R	N	R	N	5	2	2	0	0
R	R	N	N	R	0	0	0	0	0
R	R	N	N	N	3	2	2	1	1
R	N	R	R	R	205	10	10	1	5
R	N	R	R	N	3,236	198	198	41	39
R	N	R	N	R	3	2	2	2	2
R	N	R	N	N	54	3	3	0	0
R	N	N	R	R	0	0	0	0	0
R	N	N	R	N	1	1	1	0	0
R	N	N	N	R	0	0	0	0	0
R	N	N	N	N	8	2	2	0	0
N	R	R	R	R	266	13	13	6	8
N	R	R	R	N	4,090	253	252	117	118
N	R	R	N	R	0	0	0	0	0
N	R	R	N	N	1	1	1	0	0
N	R	N	R	R	10	2	2	1	2
N	R	N	R	N	312	15	15	2	3
N	R	N	N	R	0	0	0	0	0
N	R	N	N	N	11	2	2	0	0
N	N	R	R	R	837	39	39	20	21
N	N	R	R	N	23,341	1,441	1,439	372	370
N	N	R	N	R	42	2	2	1	0
N	N	R	N	N	1,451	88	88	12	12
N	N	N	R	R	311	15	15	7	11
N	N	N	R	N	13,276	816	815	98	96
N	N	N	N	R	714	34	34	5	9
N	N	N	N	N	398,233	3,275	3,275	59	44
TOTAL					**455,449**	**6,766**	**6,753**	**933**	**941**

Table 25: Sampling & assessment — Topic 304.