Overview of the TREC 2008 Legal Track

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

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

Stephen Tomlinson, stomlins@opentext.com
Open Text Corporation, Ottawa, Ontario, Canada

Jason R. Baron, jason.baron@nara.gov
National Archives and Records Administration
Office of the General Counsel, Suite 3110, College Park, MD 20740, USA

Abstract

TREC 2008 was the third year of the Legal Track, which focuses on evaluation of search technology for discovery of electronically stored information in litigation and regulatory settings. The track included three tasks: Ad Hoc (i.e., single-pass automatic search), Relevance Feedback (two-pass search in a controlled setting with some relevant and nonrelevant documents manually marked after the first pass) and Interactive (in which real users could iteratively refine their queries and/or engage in multi-pass relevance feedback). This paper describes the design of the three tasks and presents the official results.

1 Introduction

The use of information retrieval techniques in law has traditionally focused on providing access to legislation, regulations, and judicial decisions. Searching business records for information pertinent to a case (or "discovery") has also been important, but searching records in electronic form was until recently the exception rather than the norm. The goal of the Legal Track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records, an issue of increasing importance given the vast amount of information stored in electronic form to which access is increasingly desired in the context of current litigation. Ideally, the results of a study of how well comparative search methodologies perform when tasked to execute types of queries that arise in real litigation will serve to better educate the legal community on the feasibility of automated retrieval as well as its limitations. The TREC Legal Track was held for the first time in 2006, when 6 research teams participated in an Ad Hoc retrieval task. In 2007, 13 research teams participated in at least one of the track's three tasks (Ad Hoc, Interactive, and Relevance Feedback). This year, there were a total of 15 participating research teams.

The key goal of the TREC Legal Track is to develop and apply objective criteria for comparing methods for searching large heterogeneous collections using topics that approximate how real lawyers would go about propounding discovery in civil litigation, and to create a large, representative (unstructured and heterogeneous) test collection. Important aspects of this task include a focus on returning sets of documents for subsequent human review (rather than ranked lists), the need to accommodate topics that return relatively

large result sets (which necessitates sampling for assessment), the importance of recall in those result sets (since in real settings many requests for production state that "all" such evidence is to be produced), and the importance of precision in those result sets (to reduce unnecessary review costs).

The 2008 Legal Track includes the same three tasks as in 2007, but with some changes to each. For the Ad Hoc task, the most significant changes were the introduction of a new "highly relevant" category, the use of the balanced F measure as a way of simultaneously reflecting the importance of recall (for exhaustiveness) and precision (for timeliness and affordability), and a new requirement that participating teams estimate the optimal rank threshold for their system (in 2007, all systems had been compared at the number of documents returned by a "reference Boolean run"). The same changes were also made for the Relevance Feedback task. The Interactive task was completely redesigned to more closely model actual practice in e-discovery settings.

The increased visibility of the TREC Legal Track and its importance to the greater legal community were in evidence in 2008. The introduction of a completely redesigned Interactive task this year was accompanied by a signed open letter to the legal profession from Ellen Voorhees and the leadership of The Sedona Conference, urging participation this year by legal service providers and other interested parties [11]. Also, for the first time, in May 2008 the TREC Legal Track was expressly discussed in a U.S. federal court opinion involving the failure of a party to use an adequate search protocol in connection with filtering out potentially privileged documents in litigation. Judge Grimm, writing in Victor Stanley v. Creative Pipe [13], went on to make this rather extraordinary set of observations about discovery of Electronically Stored Information (ESI):

"[T]here is room for optimism that as search and information retrieval methodologies are studied and tested, this will result in identifying those that are most effective and least expensive to employ for a variety of ESI discovery tasks. Such a study has been underway since 2006, when the National Institute of Standards and Technology (NIST), an agency within the U.S. Department of Commerce, embarked on a cooperative endeavor ... to evaluate the effectiveness of a variety of search methodologies. This project, known as the Text Retrieval Conference (TREC) ... Legal Track, [is] a research effort aimed at studying the e-discovery review process to evaluate the effectiveness of a wide array of search methodologies. This evaluative process is open to participation by academics, law firms, corporate counsel and companies providing ESI discovery services. ... The goal of the project is to create industry best practices for use in electronic discovery. This project can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes."

Whether or not the results of the TREC Legal Track to date can be said to meet the judiciary's expectations, it is nevertheless the case that the opinion in Victor Stanley is only one published decision in a growing body of court precedent acknowledging greater sophistication in information retrieval techniques, and calling for parties to collaborate over appropriate search protocols. (We also note that, in another recent opinion, Judge Scheindlin cited data from the 2006 Legal Track in support of her decision on one question at dispute in the lawsuit [10]). These cases have arisen in the aftermath of the changes to the Federal Rules of Civil Procedure involving "electronically stored information" that went into effect in December 2006.

The remainder of this paper is organized as follows. Section 2 describes the Ad Hoc task, Section 3 describes the Relevance Feedback task, Section 4 describes the Interactive task, and Section 5 concludes the paper.

2 Ad Hoc Task

In the Ad Hoc task, the participants were given requests to produce documents, herein called "topics," and a set of documents to search. The following sections provide more details, but an overview of the differences from the previous year is as follows:

- The main evaluation measure this year was F_1@K, where K was specified by the participating system for each topic. This new requirement gave systems the opportunity to show that they could produce a closer set to the optimal set of R relevant documents than the reference Boolean run (for which

K=B, where B is the number of documents matched by the final negotiated Boolean query). It also modeled a real operational requirement of e-discovery systems to return a set of documents, not just an unbounded ranked list.

- Participating teams were allowed to submit up to 100,000 documents for each topic (up from 25,000 in 2007). The maximum B value (the number of documents matched by the final negotiated Boolean query) was likewise increased to 100,000 this year.
- The concept of "highly relevant" documents as a third category for purposes of assessment was introduced (in addition to last year's "relevant" and "not relevant"). This was an experiment for investigating the problem of isolating a set of "hot" or "material" documents for use in later phases of discovery (e.g., depositions) and at trial from a large set of potentially merely tangentially relevant documents, which remains a key concern for the legal profession. Participating systems could specify a different K_h value than K value for each topic for targeting a set of just highly relevant documents.
- Some topics had longer negotiation histories (i.e., a greater number of Boolean queries) than last year.

2.1 Document Collection

The 2008 Legal Track used the same collection as the 2006 and 2007 Legal Tracks, the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0 (referred to here as "IIT CDIP 1.0") which is based on documents released under the tobacco "Master Settlement Agreement" (MSA). The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents [9]. The IIT CDIP 1.0 collection is based on a snapshot, generated between November 2005 and January 2006, of the MSA subcollection of the LTDL. The IIT CDIP 1.0 collection consists of 6,910,192 document records in the form of XML elements. See the 2006 TREC Legal Track overview paper for additional details about the IIT CDIP 1.0 collection [5].

2.2 Topics

Topic development in 2008 continued to be modeled on U.S. civil discovery practice. In the litigation context, a "complaint" is filed in court, outlining the theory of the case, including factual assertions and causes of action representing the legal theories of the case. In a regulatory context, often formal letters of inquiry serve a similar purpose by outlining the scope of the proposed investigation. In both situations, soon thereafter one or more parties create and transmit formal "requests for the production of documents" to adversary parties, based on the issues raised in the complaint or letter of inquiry. See the TREC 2006 Legal Track overview for additional background [5].

For the TREC 2008 Legal Track, three new hypothetical complaints were created by members of the Sedona Conference® Working Group on Electronic Document Retention and Production, a nonprofit group of lawyers who play a leading role in the development of professional practices for e-discovery. These complaints described: (1) a combined wrongful death, negligence, and medical malpractice action against a corporate owner of a factory making fire-resistant products, and the hospital at which the fictional worker died; (2) a fictional U.S. regulatory agency's investigation into a variety of incidents in the Asian trade market involving violations of binding trade agreements between the United States and its Asian trading partners, and (3) a shareholder class action suit alleging securities fraud advertising in connection with a fictional tobacco company's "We're Smokin"' campaign. As in the past two years of the Legal Track, in using fictional names and jurisdictions, the track coordinators attempted to ensure that no third party would mistake the academic nature of the TREC Legal Track for an actual lawsuit involving real-world companies or individuals, and any would-be link or association with either past or present real litigation was entirely unintentional.

For each complaint, a set of topics (formally, "requests to produce") were initially created by the creator of the complaint, and revised by the track coordinators. "Boolean negotiations" to arrive at a consensus

search string for purposes of a baseline Boolean search were thereafter conducted among selected Sedona Conference members. The final topic set contained 45 topics, numbered 102 to 151, of which topics 102–104 were used in the Interactive task. Those three topics were run by Ad Hoc task participants, but were not sampled or scored as part of the Ad Hoc task. An XML formatted version of the topics (fullL08.xml) was created for (potentially automated) use by the participants.

2.3 Participation

Participating teams were allowed to submit up to 8 runs; additional runs could be scored locally. A total of 10 research teams submitted 64 runs for this year's Ad Hoc task. The teams experimented with a wide variety of techniques including the following:

- Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA): Terrier (TERabyte RetrIEveR) Information Retrieval platform, DFRee model, Bo1 (Bose-Einstein statistics) term weighting models, Boolean re-rank, query lexicon, query performance prediction, Z-Score.
- Open Text Corporation: negotiated Boolean queries, defendant Boolean, rank-based merging of vector results with the reference Boolean results, blind feedback, fusion.
- RMIT University: OCR error minimization, noise term removal, text de-hyphenation, ispell dictionary, Zettair search engine, Dirichlet-smoothed language model.
- Sabir Research, Inc: SMART ltu_Lnu vector run, basic blind feedback.
- University of Amsterdam (Kamps): Thresholding a Ranked List, score-distributional threshold optimization (s-d), Probability Thresholds, Bayes' rule, Truncated Normal-Exponential Model, Theoretical Truncation, Technical Truncation, Expectation Maximization, Apache's Lucene.
- University of Iowa (ICTS): custom query analyzer handling complex proximity expressions, pseudo-relevance feedback, authors, mentions, recipients and prodbox.
- University of Iowa (Srinivasan): Lucene StandardAnalyzer, Okapi-BM25, query expansion, wildcard expansions, pseudo-relevance feedback, WordNet, weighted CombSum method, Reference Run boost.
- University of Maryland, College Park: metadata-based query enrichment, author and recipient fields, social network, blind relevance feedback, iterative improvement from the reference Boolean run, hill climbing process, Indri retrieval model.
- University of Waterloo (UWIR): fusion IR methods, stepwise logistic regression, Wumpus search engine, cover density ranking, Okapi BM25, character 4-grams, MultiText, CombMNZ combination method, linear and logarithmic transfer functions.
- Ursinus College: Latent Semantic Indexing (LSI), Essential Dimensions of Latent Semantic Indexing (EDLSI), Distributed EDLSI, BM25 weighting, power normalization technique, singular value decomposition (SVD), log-entropy weighting, OCR error detection, automatic query expansion.

2.4 Reference runs

The track coordinators also created 5 reference runs. 4 of these were the list of documents matching the following Boolean queries for each topic:

- xrefL08D: defendant Boolean (from the ProposalByDefendant field of the .xml topic file)
- xrefL08P: plaintiff Boolean (from the RejoinderByPlaintiff field of the .xml topic file)
- xrefL08C: original consensus Boolean (from the Consensus1 field of the .xml topic file, or the FinalQuery field if no Consensus1 was listed)

- refL08B: final negotiated Boolean query (from the FinalQuery field of the .xml topic file)

Note that the participants were provided with the refL08B run at the time of topic release, but the other 3 Boolean reference runs (xref runs) were not available in time for participants to use them.

Also note that in some cases, the xrefL08P and xrefL08C runs included more than 100,000 documents for a topic (which was not allowed for participant runs). The most was 1,194,522 matches for the plaintiff query of topic 133. Also, in some cases, the xrefL08D run matched 0 documents, in which case the first document (aaa00a00) was submitted as a placeholder (it was always judged non-relevant).

The 5th reference run, called "randomL08", consisted of, for each topic, 100,000 randomly chosen documents from the set of documents not submitted nor in another reference run for the topic. (Last year's random reference run only included 100 documents per topic.)

2.5 Evaluation

Affordable evaluation of large result sets requires sampling and estimation. The deep sampling method we used last year (the "L07 Method" [15]) turned out to be very similar to the "statAP" method evaluated by Northeastern University in the TREC 2007 Million Query Track [4]. (The common ancestor was the (original) "infAP" method [16], which also came from Northeastern.) Both methods associate a probability with each document judgment. Our approach used deeper sampling and more judgments per topic, but used many fewer topics than the Million Query Track. The methods also assigned sampling probabilities differently and targeted different measures.

Like last year, we chose the sampling probabilities to support evaluation of both early precision and deep recall measures, as described below.

2.5.1 Pooling

Like last year, we formed a pool of documents for each topic consisting of all of the documents submitted by any run. Each of the 64 runs submitted for the Ad Hoc task run included as many as 100,000 documents, sorted in a putative best-first order, for each of the topics. The 4 Boolean reference runs were also fully included in the pool, even the plaintiff Boolean run that sometimes matched more than 1 million documents for a topic. The random reference run was created after the other runs were pooled, then itself added to the pool (an additional 100,000 documents). The final pool sizes, before sampling, ranged from 618,756 (for topic 119) to 1,634,012 (for topic 141).

Note that, like traditional TREC pooling, our deep sampling method still implicitly assumes that documents not included in the pool are not relevant for purposes of the recall calculation. (The random reference run allows us to separately analyze the accuracy of this assumption.)

2.5.2 Sampling

Our deep sampling method had a minor adjustment this year in how the sampling probabilities were chosen. A topic's B value was not a factor in the formula this year in order to provide better coverage across the range of K values that a participant might choose (as this year participants could choose a K value that differed from B). The floor on the probability was 5/100000 this year (instead of 5/25000) to compensate for the deeper submission limit this year.

The probability of judging each document d in the pool for a topic was:

```
If (hiRank(d) <= 5) { p(d) = 1.0; }
Else { p(d) = min(1.0, ((5/100000)+(C/hiRank(d)))); }
```

where hiRank(d) is the highest (i.e., best) rank at which any included run retrieved document d, and C is chosen so that the sum of all p(d) (for all submitted documents d) was the number of documents that could be judged (typically 500).

Note: for the 4 Boolean reference runs, which were unranked, the applicable rank was set to the number of documents retrieved (e.g., if 75,000 documents were retrieved for a topic, then all documents for that

topic were considered to be of rank 75,000 for that run; of course, if some other participant run retrieved one of the documents at a higher rank (e.g., 15) the hiRank would be 15 instead of 75,000 for that document). The random reference run was treated as an ordinary ranked run.

The above formula caused the first judging bin of 500 documents to contain the top-5 documents from each run, and it caused measures at depth 100,000 to have the accuracy of approximately 5+C simple random sample points. Measures at depth K have the accuracy of approximately (at least) (5K/100000)+C simple random sample points. The C values this year ranged from 1.70 for topic 113 to 4.41 for topic 105. (The final C values for each topic are listed in the Appendix of these proceedings.) These C values are fairly low, indicating that substantial estimation errors are possible on individual topics. Mean scores (over 24 or 26 topics) should be somewhat more reliable than the estimates for individual topics.

2.5.3 Binning

Like last year, to allow for the possibility that some assessors could judge more than 500 documents, the sampling process was enhanced to have a first bin of approximately 500 documents and 5 additional bins of approximately 100 documents each, using the following approach. The C values were set so that the p(d) values would sum to 1,000, and an initial draw of approximately 1000 documents was done. Then the C values were set so that the p(d) values would sum to 900, and approximately 900 documents were drawn from the initial draw of 1000 (using the ratio of the probabilities); the approximately 100 documents that were not drawn became "bin 6". This process was repeated to create "bin 5", "bin 4", "bin 3" and "bin 2". The approximately 500 documents drawn in the last step became "bin 1".

When the judgments were received from the assessors (as described in the next section), the final p(d) values were based on how many bins the assessor had completed (e.g., if 3 bins had been completed, then the p(d) values from choosing C so that the p(d) sum to 700 were used). If there had been partial judging of deeper bins, the judged documents from these bins were also kept, but with their p(d) reset to 1.0. Note that if the 1st bin was not completed, the topic had to be discarded. For each completed topic, the final number of assessed documents and corresponding C values are listed in the Appendix of these proceedings.

2.6 Relevance Judgments

As in 2007, we primarily sought out second-year and third-year law students who would be willing to volunteer as assessors in order to fulfill a law school requirement or expectation to perform some form of pro bono service to the larger community. A total of 34 Ad Hoc task topics were assigned to assessors, but judgments for 7 of those topics were not available in time for use in the evaluation, so the number of assessed topics was 27 (and one of these could not be used for evaluation because no relevant documents were found for it). Most of the assessors were law students from at least 17 returning and new institutions to the TREC Legal track, with the largest contingent, for the second year running, representing Loyola-LA law school. In addition, participants included several recent graduates of law schools, as well as experienced paralegals and litigation specialists.[1]

As in 2007, the assessors used a Web-based platform developed by NIST that was developed by Ian Soboroff and hosted at the University of Maryland to view scanned documents and to record their relevance judgments. Each assessor was given a set of approximately 500 documents to assess, which was labeled "Bin 1." Additional bins 2 through 6, each consisting of 100 documents, were available for optional additional assessment, depending on willingness and time. (It turned out that 5 Ad Hoc task assessors completed at least one of the optional bins, with one completing all five optional bins.) In total, 14,771 judgments were produced for the 27 topics.

As in 2007, we provided the assessors with a"How To Guide" that explained that the project was modeled on the ways in which lawyers make and respond to real requests for documents, including in electronic form. Assessors were told to assume that they had been requested by a senior partner, or hired by a law firm or

[1]Completed topics were received from individuals representing the following law schools and law firms: Boston U., Cleveland-Marshall, Florida Coastal, Golden Gate, Indiana U-Indianapolis, U. of Alabama, U. of Baltimore, U.C. Hastings, U. of Dayton, U. of Maine, Williamette, Baudino Law Group, Bullivant Houser Bailey, and McCarter & English.

another company, to review a set of documents for "relevance." No special, comprehensive knowledge of the matters discussed in each complaint was expected (e.g., no need to be an expert in federal election law, product liability, etc.). The heart of the exercise was to look for relevant and nonrelevant documents within a topic. Relevance was to be defined broadly. Special rules were to be applied for any document of over 300 pages. The same process was used for assessment for the interactive and relevance feedback tasks (which had different topics, as described below). See the TREC 2006 Legal Track overview for additional background (including measurement of inter-assessor agreement for that year's topics) [5].

This year, for the first time, we asked assessors to identify some documents as "highly relevant." Each reviewed document was judged highly relevant, judged relevant, judged non-relevant, or left as "gray." (Our "gray" category includes all documents that were presented to the assessor, but for which a judgment could not be determined. Among the most common reasons for this were documents that were too long to review (more than 300 pages, according to our "How To Guide") or for which there was a technical problem with displaying the scanned document image.)

The survey returns reveal a considerable variance reported among assessors in their ability to distinguish between "relevant" and "highly relevant" documents: many reported "no" difficulty in so distinguishing; one said "the highly relevant documents came few and far between ... as such, they jumped off the page when I saw them"; others reported comments such as: "it was difficult to decipher the scientific language in order to determine how relevant the information was to my topic"; "there were one or two documents where I wasn't entirely sure ... [so] I erred on the side of inclusiveness (highly relevant)"; "this was the most challenging aspect of the project"; "It takes time to gather a feel for the documents ... If I were working with a team of attorneys, I would create document samples ... so that 'key' or 'highly relevant' documents could more easily [be] identified".

Another difference from 2007 is that the posted Word files with the background complaints and requests for the assessors did not include a copy of the Boolean negotiations this year, to reduce the chance that knowledge of the Boolean strings might somehow influence the assessing. This change was suggested by [6].

A qrelsL08.normal file was created in the common trec_eval qrels format. Its 4th column was a 2 (judged highly relevant), 1 (judged relevant), 0 (judged non-relevant), -1 (gray) or -2 (gray). (In the assessor system, -1 was "unsure" (the default setting for all documents) and -2 was "unjudged" (the intended label for gray documents).) A qrelsL08.probs file was also created, which was the same as qrelsL08.normal except that there was a 5th column which listed the p(d) for the document (i.e., the probability of that document being selected for assessment from the pool of all submitted documents). qrelsL08.probs can be used with the l07_eval utility to estimate a run's scores (such as F_1, precision and recall) from the judged samples.

We asked assessors to record how much time they spent on their task. Past review rates averaged to 25 documents per hour in 2006 and 20 documents per hour in 2007. Based on partial survey results for 2008, assessors reported that it took a collective 631.15 hours to review 13,543 documents, or a rate of 21.5 documents per hour. As this result is in line with past years, it appears that the new highly relevant category did not substantially affect the assessment rate.

Overall, survey returns contained uniformly positive reviews for the experience of being a volunteer assessor in 2008, including such statements as "the project was a lot of fun"; "interesting and educational"; "enjoyed participating"; "[the coordinators were] very helpful, courteous, and gracious at all times, which made this sometimes tedious project seem much more engaging, exciting, and purposeful"; "a great way to get [pro bono] hours for evening students with limited availability"; "Once in a lifetime opportunity to help with a fascinating e-Discovery project."

2.7 Computing Evaluation Measures

The formulas for estimating the number of relevant, non-relevant and gray documents in the pool for each topic, and also for estimating precision and recall, were the same as last year [15].

The new F_1 measure this year was estimated as follows:

Define $estF_1@k$ to be the estimated F_1 of S at depth k:

$$estF_1@k = \frac{2 * estPrec@k * estRecall@k}{estPrec@k + estRecall@k} \quad (1)$$

Note: we define $estF_1@k$ as 0 if both $estPrec@k$ and $estRecall@k$ are 0.

The K and B values are integers and hence can be substituted for k in the above formulas. R, however, can be fractional, hence we provide the following additional definition:

Define $F_1@R = F_1@R_{ceil}$

where R_{ceil} is the ceiling of R (i.e., the smallest integer greater than or equal to R).

For runs that did not contribute to the pools, the same estimation process can be used, albeit with the same limitations as in traditional TREC pooling (in particular, the assumption that unpooled documents are not relevant), and possibly larger sampling errors if the run would have influenced the hiRank() settings that were used to set the sampling probabilities of the documents. Note that the Interactive task, described below, did more detailed sampling of the entire collection for 3 test topics, providing another option for evaluating a novel technique.

2.8 Results

For the Ad Hoc task, 27 topics were assessed. However, one topic (#130) had no relevant judgments, leaving 26 useful topics. Furthermore, 2 topics had no "highly relevant" judgments (topics 136 and 142) leaving 24 useful topics for measures just counting "highly relevant" documents.

2.8.1 Number of Relevant Documents

Applying the cited formulae, the estimated number of relevant documents in the pool, on average per topic, was 82,403, almost 5x more than last year (16,904). The number varied considerably by topic, from 110 (for topic 137) to 658,399 (for topic 131). Unfortunately, six of the topics had more than 100,000 relevant documents, i.e., more relevant documents than the participant runs were allowed to retrieve for a topic. (Last year, the most relevant documents for a topic was 77,467 (topic 71).) Perhaps one should remove these six topics for future training (though further study is needed to determine how best to deal with this issue). For this paper we have used all of the available topics in the scoring.

An explanation offered for the increase in the number of relevant documents is that the topic formulators had been instructed in previous years to try to keep the requests narrow because of concerns about the shallow pooling traditionally used at TREC. With the deeper sampling approach now in use, this concern went away, resulting in more broadly worded topics. However, if this is what happened, it was not a planned change, and the participants were not advised that this year's topics might tend to be broader (though the higher B values of the reference Boolean run this year (see next section) may have been a tip-off).

Over the 24 topics with highly relevant judgments, the estimated number of highly relevant documents in the pool, on average per topic, was 11,542. This number ranged from 1 (for topic 109) to 51,313 (for topic 145). Hence 100% recall of highly relevant documents was possible within the constraint of retrieving at most 100,000 documents for each topic. A concern though is the small numbers of highly relevant documents for some topics (e.g., if a topic has just 1 highly relevant document, then recall for that topic can only be 0% or 100%). Perhaps one should remove topics of small numbers of highly relevant documents for future training, but for this paper we have used all of the available topics in the scoring.

2.8.2 Boolean Negotiation Results

The average B value was 40,402 for the 26 evaluation topics. For the subset of 24 topics that have highly relevant judged documents, the average B value was almost the same (39,930). These values are about eight times larger than last year's average B value of 5,004 (i.e., the final negotiated Boolean query matched about eight times as many documents this year, on average).

This year, the average recall of the final negotiated Boolean query was 24%, close to last year's average recall of 22%. Once again, the recall varied widely by topic, ranging from 0.1% (for topic 150) to 88% (for

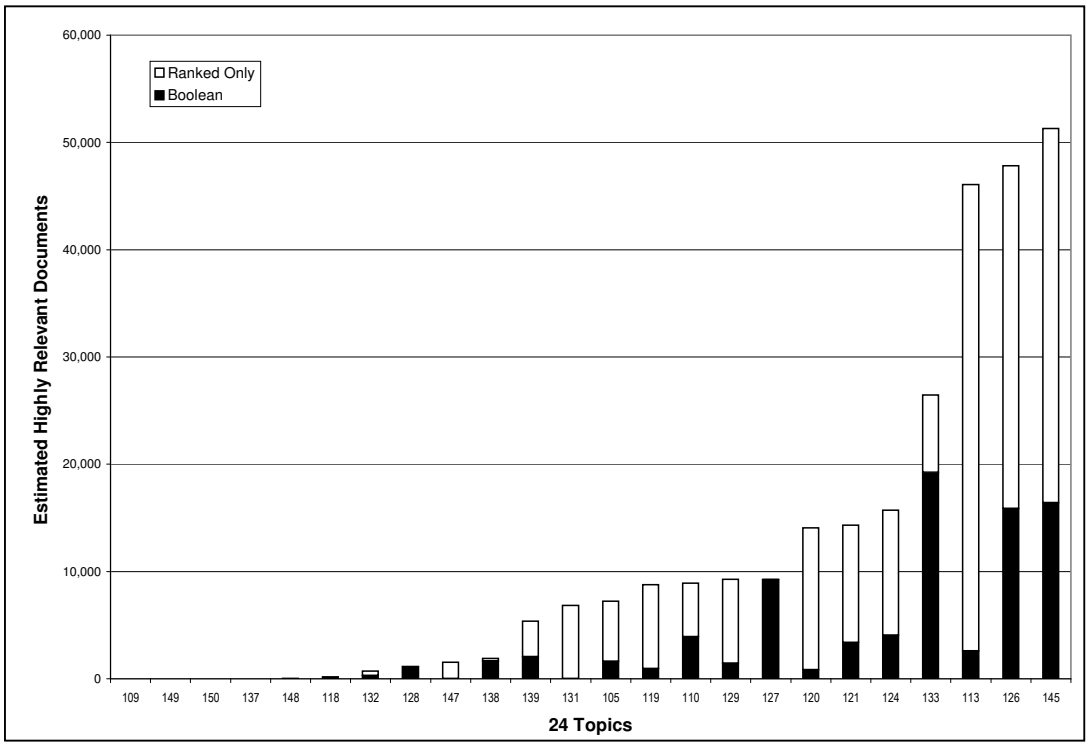

Figure 1: Highly relevant documents not found by the consensus Boolean run.

topic 148). For highly relevant documents, the average recall of the final negotiated Boolean query was 33%. It ranged from 0% (for topics 109, 147 and 150) to 100% (for topics 128 and 137).

The average F_1 of the final negotiated Boolean query was 16%, close to last year's average F_1 of 14%. The F_1 varied by topic, ranging from 0.2% (for topic 150) to 38% (for topic 128). For highly relevant documents, the average F_1 of the final negotiated Boolean query was 9%. It ranged from 0% (for topics 109, 147 and 150) to 38% (for topic 145).

Table 1 lists the mean scores of the negotiated Boolean queries[2]. It shows that the initial proposal by the defendant produced a relatively small set with relatively high precision, but had relatively low recall. In contrast, the rejoinder by the plaintiff produced a relatively large set with relatively high recall (57% of highly relevant documents on average), but more modest precision. The original consensus of the negotiators decreased the recall but did not gain much in precision on average. Final adjustments needed for some queries

[2]While F_1 must always be between P and R for individual topics (since it is a harmonic mean, and all means lie between the extreme values), Table 1 shows that the *arithmetic mean* of F_1 is not the same as the F_1 of the arithmetic mean of P and the arithmetic mean of R. This is because the computation of F_1 is nonlinear. Consider the following two-topic example: Topic 1 (P=0.01, R=0.99, F_1=0.02), Topic 2 (P=0.99, R=0.01, F_1=0.02). F_1 is always close to the smaller value, and in this example the mean F_1 is 0.02, but, mean P is 0.5, as is mean R. So we have the mean F_1 on the per-topic values outside the bounds of the mean P and the mean R.

All Relevant (26 topics)	Retrieved	Precision	Recall	F_1
Defendant	3,180	0.41	0.04	0.05
Plaintiff	219,606	0.23	0.43	0.19
Consensus1	93,190	0.24	0.33	0.20
Final	40,402	0.28	0.24	0.16
	Avg. K			
Median (23 request runs)	14,363	0.26	0.12	0.10
Median (41 other runs)	40,402	0.28	0.25	0.16
Highly Relevant only	(24 topics)			
Defendant	3,445	0.14	0.06	0.06
Plaintiff	234,016	0.08	0.57	0.09
Consensus1	97,259	0.07	0.42	0.09
Final	39,930	0.08	0.33	0.09
	Avg. K_h			
Median (23 request runs)	5,838	0.10	0.22	0.05
Median (41 other runs)	19,965	0.09	0.34	0.08

Table 1: Mean scores of the Negotiated Boolean Queries and Median Mean Scores of the Participant Runs.

to put B in the 100 to 100,000 range likewise decreased the recall without much gain in precision. On balance, the F_1 scores were highest with the original consensus of the negotiators, though the plaintiff query produced a similar mean F_1 score. Before trying to draw firm conclusions about the effect of different stages in the Boolean negotiation history, we will want to look at the negotiation results on a topic-by-topic basis.

Figure 1 shows graphically that many of the highly relevant documents that are estimated to exist were not found by the consensus Boolean run (xrefL08C).

2.8.3 Participant Results

Table 2 shows the estimated F_1 at K for each submitted participant run (along with the 5 reference runs). Several participating teams submitted runs that achieved higher mean estimated F_1 values (across all topics) than the reference Boolean run.

Unlike last year, several participant runs also had a higher Recall@B than the reference Boolean run, including some runs that are known to have used the same techniques as last year. It is not yet clear whether this results from topics that are different in important ways (as the larger Boolean result sets might suggest) or from some other factor (e.g., subtle system improvements, or differences in the way the final Boolean query was constructed).

The highest scoring run in mean F_1@K (wat7fuse) just set K=100,000 (the maximum allowed) for all topics, which probably isn't a generally applicable thresholding approach. The number of relevant documents (more than 80,000 on average per topic, including several topics of more than 100,000) was unexpectedly much larger than last year, so the submission cutoff of 100,000 may not have allowed enough flexibility to really test the thresholding ability of the systems.

The 100,000 cutoff issue would not seem to affect evaluation on highly relevant documents as much since the number of highly relevant documents was less than 52,000 for every topic. However, the top-scoring run in F_1@K_h (wat6fuse in Table 3) just set K_h to a constant 12,500 for all topics (close to the average number of highly documents per topic (11,542)). It should be noted though that the participants did not have any training data for the highly relevant category, so this year's results may not represent what could be done with further study.

(A glossary for Tables 2 and 3 appears in Section 2.8.7.)

Run	Fields	Ret.	Avg. K	(P@K, R@K)	F_1@K	F_1@R	S1J, P5	R@B, R@ret
wat7fuse	br	99999	99999	(0.210, **0.555**)	**0.220**	0.243	19/26, 0.754	**0.329**, 0.555
CTFgge10kBr0	bdprBM	100000	100000	(0.218, 0.552)	0.216	0.215	13/26, 0.567	0.292, 0.552
otL08fbe	bmBM	100000	75228	(0.241, 0.409)	0.215	**0.246**	17/26, 0.654	0.272, 0.451
otL08frw	brmBM	100000	64232	(0.239, 0.380)	0.207	0.220	**21/26**, **0.769**	0.269, 0.461
wat8fuse	brv	99999	40402	(0.324, 0.329)	0.201	0.243	19/26, 0.754	**0.329**, 0.555
(xrefL08C)	cmM	93190	93190	(0.244, 0.333)	0.196			0.333
(xrefL08P)	pmM	219606	219606	(0.231, 0.425)	0.191			0.425
otL08fv	bmM	100000	46369	(0.243, 0.345)	0.190	0.186	13/26, 0.485	0.254, 0.447
CTFggeBkBr1	bdprBM	100000	44397	(0.320, 0.292)	0.186	0.220	13/26, 0.567	0.289, 0.552
CTFgge4kBr0	bdprBM	100000	40000	(0.267, 0.336)	0.185	0.215	13/26, 0.567	0.292, 0.552
otL08rv	rmM	100000	52081	(0.242, 0.311)	0.185	0.216	15/26, 0.592	0.268, 0.422
CTFggeBkBr0	bdprBM	100000	44397	(0.309, 0.288)	0.180	0.215	13/26, 0.567	0.292, 0.552
otL08rvl	rmM	100000	81826	(0.190, 0.402)	0.179	0.215	16/26, 0.585	0.278, 0.443
CTFggeRkBr0	bdprBM	100000	40402	(0.308, 0.292)	0.178	0.215	13/26, 0.567	0.292, 0.552
wat6fuse	br	99999	25000	(0.310, 0.282)	0.175	0.243	19/26, 0.754	**0.329**, 0.555
wat2text	r	99999	25000	(0.289, 0.234)	0.167	0.231	17/26, 0.615	0.246, 0.445
IowaSL0805b	bdprB	100000	40402	(0.273, 0.294)	0.164	0.228	18/26, 0.700	0.294, **0.559**
IowaSL0808b	bdporB	100000	40402	(0.268, 0.288)	0.163	0.214	17/26, 0.692	0.288, 0.518
IowaSL0804b	bdprB	100000	40402	(0.284, 0.289)	0.162	0.221	18/26, 0.723	0.289, 0.551
IowaSL0804	bdpr	100000	40402	(0.274, 0.292)	0.162	0.208	15/26, 0.677	0.292, 0.551
UMDCRC40	bdprmB	40442	40402	(0.282, 0.246)	0.162	0.155	12/26, 0.427	0.246, 0.246
UMDCRP3	bdprmB	41009	40402	(0.278, 0.272)	0.162	0.153	12/26, 0.427	0.272, 0.274
IowaSL0808m2	bdprB	100000	40402	(0.281, 0.293)	0.161	0.222	18/26, 0.715	0.293, 0.553
UMDAURCC40	bdprmB	40442	40402	(0.278, 0.241)	0.161	0.146	10/26, 0.339	0.241, 0.241
IowaSL0808m3	bdporB	100000	40402	(0.277, 0.288)	0.161	0.218	18/26, 0.708	0.288, 0.558
refL08B	bvmBM	40402	40402	(0.280, 0.240)	0.161			0.240, 0.240
otL08fb	bvmBM	40402	40402	(0.280, 0.240)	0.161	0.165	17/26, 0.577	0.240, 0.240
IowaSL0805	bdprB	100000	40402	(0.256, 0.294)	0.160	0.219	15/26, 0.631	0.294, **0.559**
UMDAURCP3	bdprmB	41313	40402	(0.274, 0.250)	0.159	0.141	10/26, 0.339	0.250, 0.251
RMITrp2	r	100000	14363	(0.298, 0.185)	0.159	0.217	16/26, 0.581	0.263, 0.447
RMITrp1	r	100000	13876	(0.309, 0.171)	0.158	0.216	16/26, 0.562	0.262, 0.434
wat3nobool	brB	99999	99999	(0.159, 0.352)	0.157	0.174	16/26, 0.523	0.194, 0.352
CTFggeSkBr0	bdprBM	100000	25084	(0.344, 0.211)	0.154	0.215	13/26, 0.567	0.292, 0.552
wat4fuse	br	99999	13842	(0.373, 0.195)	0.154	0.243	19/26, 0.754	**0.329**, 0.555
UMDSTD	rm	97539	40332	(0.240, 0.213)	0.151	0.187	17/26, 0.529	0.213, 0.323
CTFrtSkBr0	rB	100000	25832	(0.294, 0.193)	0.135	0.182	13/26, 0.488	0.264, 0.416
SabL08ab1	bdporm	100000	20000	(0.260, 0.243)	0.131	0.233	12/26, 0.585	0.298, 0.512
wat1fuse	br	99999	7419	(0.416, 0.158)	0.130	0.243	19/26, 0.754	**0.329**, 0.555
CTFrtSk	r	100000	25832	(0.258, 0.178)	0.128	0.174	13/26, 0.385	0.213, 0.360
uva-xconst	r	100000	16904	(0.271, 0.171)	0.126	0.171	16/26, 0.485	0.204, 0.347
otL08rvlq	rmM	100000	50865	(0.186, 0.253)	0.126	0.160	14/26, 0.415	0.217, 0.352
UIowa08Lega	bm	84969	18153	(0.284, 0.215)	0.125	0.173	9/26, 0.431	0.233, 0.388
SabL08arbn	bdporm	100000	20000	(0.243, 0.242)	0.123	0.208	13/26, 0.577	0.271, 0.524
RMITrp3	r	92550	8043	(0.313, 0.109)	0.113	0.178	13/26, 0.385	0.217, 0.350
IowaSL08Ref	r	100000	40402	(0.177, 0.169)	0.112	0.152	9/26, 0.223	0.169, 0.368
uva-xb	r	100000	17301	(0.261, 0.137)	0.103	0.171	16/26, 0.485	0.204, 0.347
SabL08ar2	rm	100000	20000	(0.224, 0.130)	0.097	0.148	9/26, 0.331	0.208, 0.345
UrsinusBM25b	r	100000	10168	(0.462, 0.090)	0.091	0.187	16/26, 0.581	0.180, 0.314
UIowa08LegE0	r	99988	20000	(0.194, 0.118)	0.087	0.157	4/26, 0.254	0.186, 0.350
RMITbp1	b	98873	4767	(0.293, 0.087)	0.070	0.148	8/26, 0.404	0.215, 0.402
uva-xk	r	100000	7447	(0.265, 0.070)	0.069	0.171	16/26, 0.485	0.204, 0.347
RMITbp3	b	100000	5057	(0.298, 0.085)	0.068	0.158	11/26, 0.469	0.219, 0.418
uvabase	r	100000	5852	(0.304, 0.067)	0.065	0.173	13/26, 0.506	0.203, 0.353
RMITbp2	b	100000	4852	(0.284, 0.083)	0.065	0.137	10/26, 0.396	0.207, 0.377
UrsinusPwrB	r	100000	3292	(0.390, 0.048)	0.055	0.182	7/26, 0.308	0.179, 0.372
wat5fuse	br	99999	1004	(0.529, 0.057)	0.053	0.243	19/26, 0.754	**0.329**, 0.555
otL08db	dmM	3180	3180	(0.407, 0.035)	0.050	0.037	16/26, 0.469	0.034, 0.035
xrefL08D	dmM	3180	3180	(0.407, 0.035)	0.050			0.035
UCEDLSIa	r	100000	15225	(0.115, 0.056)	0.046	0.111	4/26, 0.112	0.146, 0.296
UIowa08LegE1	b	8910	6119	(0.180, 0.044)	0.042	0.042	6/26, 0.225	0.066, 0.066
UIowa08LegE2	b	8910	6119	(0.180, 0.044)	0.042	0.042	6/26, 0.225	0.066, 0.066
UrsinusPwrA	r	100000	905	(0.403, 0.021)	0.025	0.168	7/26, 0.310	0.191, 0.375
UrsinusPwrC	r	100000	850	(0.396, 0.020)	0.024	0.167	7/26, 0.310	0.183, 0.374
UrsinusBM25a	r	100000	3157	(**0.565**, 0.025)	0.021	0.153	16/26, 0.500	0.152, 0.345
UCEDLSIb	r	100000	1535	(0.139, 0.013)	0.009	0.094	2/26, 0.065	0.123, 0.242
UIowa08Leg3	bm	65376	13077	(0.044, 0.004)	0.008	0.032	1/26, 0.039	0.020, 0.037
UrsinusVa	r	100000	802	(0.050, 0.007)	0.005	0.069	4/26, 0.067	0.073, 0.202
randomL08		100000	20000	(0.013, 0.001)	0.002	0.010	1/26, 0.023	0.008, 0.010
UIowa08LegE4	bm	7692	1539	(0.000, 0.000)	0.000	0.000	0/26, 0.000	0.000, 0.000

Table 2: Mean scores for submitted Ad Hoc task runs, using All Relevant documents.

Run	Fields	Ret.	Avg. K_h	(P@K_h, R@K_h)	F_1@K_h	F_1@R_h	S1J, P5	R@B, R@ret
wat6fuse	br	99999	12500	(0.128, 0.400)	**0.106**	0.160	9/24, **0.342**	0.473, 0.663
wat4fuse	br	99999	7123	(0.161, 0.298)	0.106	0.160	9/24, **0.342**	0.473, 0.663
wat8fuse	brv	99999	19965	(0.117, 0.324)	0.105	0.160	9/24, **0.342**	0.473, 0.663
wat7fuse	br	99999	50000	(0.071, 0.572)	0.100	0.160	9/24, **0.342**	0.473, 0.663
wat2text	r	99999	12500	(0.121, 0.319)	0.098	0.132	7/24, 0.275	0.429, 0.619
(xrefL08C)	cmM	97259	97259	(0.074, 0.418)	0.095			0.418
otL08frw	brmBM	100000	21799	(0.100, 0.316)	0.095	0.164	8/24, 0.333	0.385, 0.643
CTFggeRkBr0	bdprBM	100000	39930	(0.080, 0.438)	0.095	0.101	5/24, 0.227	0.438, 0.655
wat1fuse	br	99999	3820	(0.183, 0.223)	0.093	0.160	9/24, **0.342**	0.473, 0.663
UMDAURCC40	bdprmB	39970	39930	(0.079, 0.358)	0.092	0.129	3/24, 0.125	0.358, 0.358
UMDCRP3	bdprmB	40530	39930	(0.078, 0.436)	0.092	0.162	2/24, 0.117	0.436, 0.437
UMDCRC40	bdprmB	39970	39930	(0.078, 0.433)	0.092	0.162	2/24, 0.117	0.433, 0.433
otL08fb	bvmBM	39930	39930	(0.078, 0.335)	0.091	0.104	3/24, 0.192	0.335, 0.335
refL08B	bvmBM	39930	39930	(0.078, 0.335)	0.091			0.335, 0.335
UMDAURCP3	bdprmB	40805	39930	(0.078, 0.351)	0.091	0.119	3/24, 0.125	0.351, 0.353
(xrefL08P)	pmM	234016	234016	(0.078, 0.568)	0.091			0.568
UMDSTD	rm	97537	39855	(0.066, 0.423)	0.088	0.117	6/24, 0.200	0.423, 0.552
IowaSL0804b	bdprB	100000	19965	(0.100, 0.352)	0.088	0.148	5/24, 0.250	0.407, 0.668
otL08fbe	bmBM	100000	14124	(0.105, 0.267)	0.086	0.117	4/24, 0.267	0.349, 0.531
IowaSL0808b	bdporB	100000	19965	(0.103, 0.348)	0.085	0.141	3/24, 0.233	0.423, 0.641
SabL08ab1	bdporm	100000	10000	(0.090, 0.360)	0.084	0.085	1/24, 0.125	0.435, 0.640
CTFggeBkBr1	bdprBM	100000	44745	(0.087, 0.460)	0.082	0.102	5/24, 0.227	0.433, 0.655
CTFgge4kBr0	bdprBM	100000	40000	(0.066, 0.476)	0.082	0.101	5/24, 0.227	0.438, 0.655
IowaSL0808m3	bdporB	100000	19965	(0.097, 0.345)	0.081	0.141	5/24, 0.242	0.402, 0.681
CTFggeSkBr0	bdprBM	100000	25275	(0.095, 0.387)	0.081	0.101	5/24, 0.227	0.438, 0.655
CTFggeBkBr0	bdprBM	100000	44745	(0.085, 0.446)	0.080	0.101	5/24, 0.227	0.438, 0.655
CTFgge10kBr0	bdprBM	100000	100000	(0.050, **0.655**)	0.079	0.101	5/24, 0.227	0.438, 0.655
otL08rv	rmM	100000	30039	(0.100, 0.372)	0.079	0.165	**10/24**, 0.300	**0.500**, **0.696**
IowaSL0808m2	bdprB	100000	19965	(0.094, 0.344)	0.079	0.148	6/24, 0.250	0.409, 0.678
IowaSL0805b	bdprB	100000	19965	(0.093, 0.345)	0.079	0.146	5/24, 0.250	0.411, 0.679
wat3nobool	brB	99999	50000	(0.056, 0.309)	0.077	0.084	2/24, 0.183	0.218, 0.340
IowaSL0805	bdprB	100000	19965	(0.082, 0.332)	0.075	0.111	4/24, 0.192	0.415, 0.679
uva-xconst	r	100000	8452	(0.097, 0.293)	0.073	0.147	8/24, 0.250	0.458, 0.567
CTFrtSkBr0	rB	100000	26440	(0.082, 0.370)	0.072	0.111	5/24, 0.160	0.373, 0.553
IowaSL0804	bdpr	100000	19965	(0.078, 0.347)	0.071	0.117	4/24, 0.200	0.409, 0.668
UIowa08Lega	bm	88741	9841	(0.094, 0.262)	0.069	0.065	2/24, 0.142	0.358, 0.486
uva-xb	r	100000	8549	(0.094, 0.230)	0.066	0.147	8/24, 0.250	0.458, 0.567
SabL08arbn	bdporm	100000	10000	(0.081, 0.340)	0.066	0.072	1/24, 0.125	0.452, 0.680
otL08rvl	rmM	100000	43531	(0.059, 0.441)	0.065	0.150	7/24, 0.325	0.464, 0.645
otL08fv	bmM	100000	15389	(0.076, 0.316)	0.064	0.086	3/24, 0.108	0.354, 0.519
RMITrp1	r	100000	604	(0.298, 0.063)	0.064	**0.177**	7/24, 0.242	0.443, 0.644
RMITrp2	r	100000	647	(**0.300**, 0.063)	0.063	0.164	7/24, 0.271	0.441, 0.616
otL08db	dmM	3445	3445	(0.143, 0.062)	0.063	0.057	7/24, 0.192	0.062, 0.062
xrefL08D	dmM	3445	3445	(0.143, 0.062)	0.063			0.062
uva-xk	r	100000	5838	(0.117, 0.246)	0.060	0.147	8/24, 0.250	0.458, 0.567
UrsinusBM25b	r	100000	5668	(0.190, 0.160)	0.060	0.136	8/24, 0.267	0.407, 0.604
uvabase	r	100000	4542	(0.101, 0.239)	0.052	0.130	8/24, 0.263	0.454, 0.552
UrsinusPwrB	r	100000	678	(0.151, 0.085)	0.049	0.115	3/24, 0.142	0.271, 0.509
CTFrtSk	r	100000	26440	(0.043, 0.348)	0.048	0.084	4/24, 0.100	0.362, 0.531
SabL08ar2	rm	100000	10000	(0.054, 0.224)	0.048	0.076	2/24, 0.175	0.304, 0.463
UrsinusBM25a	r	100000	1571	(0.220, 0.073)	0.045	0.110	8/24, 0.192	0.345, 0.525
IowaSL08Ref	r	100000	19965	(0.049, 0.259)	0.041	0.093	5/24, 0.083	0.309, 0.504
UIowa08LegE0	r	99987	10000	(0.053, 0.237)	0.041	0.087	1/24, 0.100	0.314, 0.463
UCEDLSIa	r	100000	7113	(0.057, 0.087)	0.034	0.042	0/24, 0.000	0.202, 0.380
wat5fuse	br	99999	521	(0.225, 0.132)	0.034	0.160	9/24, **0.342**	0.473, 0.663
otL08rvlq	rmM	100000	8891	(0.063, 0.242)	0.033	0.108	6/24, 0.133	0.299, 0.541
RMITrp3	r	92819	243	(0.235, 0.055)	0.028	0.119	4/24, 0.217	0.351, 0.553
UrsinusPwrA	r	100000	231	(0.125, 0.067)	0.026	0.106	3/24, 0.148	0.274, 0.518
UrsinusPwrC	r	100000	175	(0.124, 0.067)	0.025	0.105	3/24, 0.148	0.273, 0.518
RMITbp2	b	100000	212	(0.118, 0.089)	0.022	0.059	3/24, 0.075	0.256, 0.532
UIowa08LegE2	b	9568	4219	(0.036, 0.039)	0.020	0.022	1/24, 0.058	0.088, 0.088
UIowa08LegE1	b	9568	4219	(0.036, 0.039)	0.020	0.022	1/24, 0.058	0.088, 0.088
RMITbp1	b	98779	212	(0.107, 0.088)	0.018	0.060	3/24, 0.075	0.263, 0.536
RMITbp3	b	100000	143	(0.102, 0.085)	0.012	0.064	5/24, 0.125	0.324, 0.612
UrsinusVa	r	100000	491	(0.019, 0.001)	0.001	0.030	0/24, 0.000	0.090, 0.270
UCEDLSIb	r	100000	293	(0.011, 0.000)	0.000	0.043	0/24, 0.000	0.163, 0.340
UIowa08Leg3	bm	70824	7084	(0.001, 0.000)	0.000	0.001	0/24, 0.008	0.001, 0.008
UIowa08LegE4	bm	8333	834	(0.000, 0.000)	0.000	0.000	0/24, 0.000	0.000, 0.000
randomL08		100000	10000	(0.000, 0.000)	0.000	0.007	0/24, 0.000	0.008, 0.008

Table 3: Mean scores for submitted Ad Hoc task runs, using only Highly Relevant documents.

2.8.4 Estimated Gray Percentages

As previously mentioned, "gray" documents are those documents that were presented to the assessor but could not be judged non-relevant nor relevant (including highly relevant). The assessors were not expected to read through documents longer than 300 pages (though they were still asked to search such documents for relevant passages, in which case the document could be judged relevant). Sometimes there was a technical problem displaying the document in the assessor system.

At depth B, the reference Boolean run had the highest percentage of gray documents (averaged over the 26 topics) at 2.5%. At depth K, only the RMITrp3 run had a higher gray percentage than the reference Boolean run (and it still rounded to 2.5%). At depth 5, runs UrsinusVa and UCEDLSIa had the highest gray percentages at 5.4%. At depth K_h (averaged over 24 topics), the reference Boolean run had the highest gray percentage (2.7%). These gray percentages seem low enough to not adversely affect the comparability of mean precision, recall and F_1 estimates in this study.

2.8.5 Random Run Results

Like last year, we created a "random run" (this year named randomL08) in hopes of estimating the number of relevant documents that may have been outside of the pooled results of participating systems. By design, the randomL08 run had no overlap with any of the participant runs; for each topic, it was a random sample of just the unsubmitted documents.

The number of assessed random run documents varied by topic depending on the size of the pool, the number of bins judged by the assessor, and the randomness inherent in the sampling process. On average, 35 random run documents were assessed per topic, ranging from as low as 20 (for topics 113 and 118) to as high as 55 (for topic 124). More than three times as many random run documents were judged per topic this year as in 2007 (when 10 random run documents were assessed on average per topic).

One topic had several more random run documents judged relevant than the others. Topic 131 had 11 relevant judgments for its 33 random run documents. No other topic had more than 2 relevant judgments for the random run documents. We consider the judgments of this outlier topic to be suspect (in particular, none of its relevant documents that we (some of the track coordinators) have reviewed have looked relevant to us). For the rest of this subsection, we exclude topic 131.

Over the remaining 25 topics, there were 889 random run documents assessed. Of these, 8 were judged relevant (including 1 judged highly relevant), 864 were judged non-relevant, and 17 were left as gray. So, like last year, approximately 1% of the unsubmitted documents were judged relevant. For the new "highly relevant" category, only 0.1% of the unsubmitted documents were judged highly relevant.

Last year, when we reviewed the 3 random run documents that were judged relevant, they did not appear to be relevant to us, suggesting that the 1% number may actually be an assessor false positive rate rather than the percentage of unsubmitted documents that are relevant. This year, we have just reviewed the one random run document that was judged highly relevant (document nhx30c00 of topic 126). Again, it did not appear to be relevant to us (let alone highly relevant). Last year, an analysis of "assessor blunders" by a participant concluded that "the number of assessing disagreements due to blunders is still much less than the number of assessing disagreements due to scope of relevance" [6]. The Interactive task (described below) looked at the impact of adjudicating judgments this year.

2.8.6 Marginal Precision Rates to Depth 100,000

Table 4 shows how precision falls with retrieval depth for the Ad Hoc task runs. The table includes the median and highest estimated marginal precision rates of the Ad Hoc task runs for depths 1-25,000, 25,001-50,000, 50,001-75,000 and 75,001-100,000. The median run was still maintaining more than 10% precision at the deepest stratum (depths 75,001-100,000), indicating that depth 100,000 was likely not deep enough to cover all of the relevant documents that a run could potentially find. This result seems consistent with the earlier finding that 6 topics had more than 100,000 estimated relevant documents. Perhaps it would be better for reusability to discard these 6 topics, but additional analysis will be needed before we can draw firm conclusions.

All Relevant (26 topics)	Depths 1-25000	Depths 25001-50000	Depths 50001-75000	Depths 75001-100000
Median (23 request runs)	0.232	0.166	0.122	0.115
Median (41 other runs)	0.274	0.187	0.172	0.127
Highest (all 64 runs)	0.325	0.223	0.214	0.242
Highly Relevant only (24 topics)				
Median (23 request runs)	0.063	0.033	0.021	0.015
Median (41 other runs)	0.073	0.038	0.040	0.017
Highest (all 64 runs)	0.112	0.057	0.085	0.106
Ratio of Highly Relevant to Relevant				
of Medians (23 request runs)	27%	20%	17%	13%
of Medians (41 other runs)	27%	20%	23%	13%

Table 4: Median and Highest Estimated Marginal Precision Rates

Table 4 also shows that the percentage of relevant documents judged highly relevant also tends to fall with retrieval depth. For the median runs, approximately 27% of the relevant documents were judged highly relevant in the highest stratum (depths 1-25,000) while just 13% of the relevant documents were judged highly relevant in the deepest stratum (depths 75,001-100,000). This is another indicator that the collection may have better coverage of highly relevant documents than relevant documents.

2.8.7 Table Glossary

The following glossary explains the codes used in Tables 2 and 3.

"Fields": The topic fields used by the run: 'b' Boolean query (final negotiated), 'C' complaint, 'd' defendant Boolean (initial proposal), 'i' instructions and definitions, 'p' plaintiff Boolean (rejoinder query), 'o' other negotiation history (Defendant2, Plaintiff2, etc.), 'c' original consensus Boolean (or final Boolean if the Consensus1 field was not used), 'r' request text, 'v' B value, 'm' metadata fields were indexed, 'B' reference Boolean run was used, 'M' manual processing was involved, 'F' feedback run (old relevance assessments were used, applicable to RF task only).

"Ret.": The Average Number of Documents Retrieved per Topic.

"Avg. K": The Average K value.

"P@K" and "R@K": Estimated Precision and Recall at Depth K.

"F_1@K": Estimated F_1 at Depth K.

"F_1@R": Estimated F_1 at Depth R (where R is the estimated number of relevant documents).

"S1J": Success of the First Judged Document.

"P5": Estimated Precision at Depth 5.

"R@B": Estimated Recall at Depth B.

"R@ret": Estimated Recall of the full retrieval set.

"K_h": K value when just counting Highly relevant documents as relevant.

"R_h": Estimated number of Highly relevant documents.

"Ret_r": The Average Number of "Residual" Documents Retrieved per Topic (RF task only).

"K_r": The Average "Residual" K value (RF Task only).

Table 2 counts all relevant documents as relevant (averaged over 26 topics). Table 3 shows the mean scores when just counting highly relevant documents as relevant (averaged over 24 topics).

Parentheses are used for the 2 reference runs (xrefL08C and xrefL08P) which sometimes retrieved more than 100,000 documents for a topic (which was not allowed for participant runs).

For the 4 reference Boolean runs, only measures at the retrieval depth are shown since a specific ordering of Boolean results is not defined.

The Appendix of these proceedings lists more detailed information for each topic, including median and high F_1 scores for each topic.

3 Relevance Feedback Task

The objective in the Relevance Feedback task was to automatically discover previously unknown relevant documents by augmenting the evidence available from the topic description with evidence available from a limited number of existing relevance assessments. This task provides a simple and well controlled model for assessing the utility of a two-pass search process. The 2007 Relevance Feedback task relied on 2006 relevance assessment pools that had (generally) been drawn from near the top of submitted ranked retrieval runs. For the 2008 Relevance Feedback task, relevance assessments sampled from throughout the runs submitted in 2007 were available. We therefore selected some Ad Hoc topics from each year for use in the 2008 Relevance Feedback task. Teams could use positive and/or negative judgments in conjunction with the metadata for and/or full text from the judged documents to refine their models.

The same document collection was used in the Ad Hoc and Relevance Feedback tasks, so participation in both tasks did not require indexing a second collection.

3.1 Topic Selection

40 topics were selected from among those used in 2006 and 2007. These topics were chosen by the track coordinators based on a variety of factors, as follows:

Topics were rejected if any of the following applied: the residual B_r value was less than 100 (where B_r is the number of documents matching the final negotiated Boolean query after documents judged in previous years were omitted); the residual B_r value was greater than 100,000; or the topic had been used in last year's Relevance Feedback task.

The above criteria left 64 topics to choose from. Grouped by complaint, remaining were 4 topics from 2006-A, 7 topics from 2006-B, 5 topics from 2006-C, 4 topics from 2006-D, 4 topics from 2006-E, 13 topics from 2007-A, 9 topics from 2007-B, 8 topics from 2007-C, and 10 topics from 2007-D.

To get to 40, we chose to balance the number from each complaint of a given year, which led to choosing 3 topics from each complaint of 2006 (a total of 15 from 2006) and 6 or 7 topics from each complaint of 2007 (7 from A, 6 from the others, to make a total of 25 from 2007). From each complaint, the topics were chosen randomly.

At least the first bin for 12 of those topics was assessed by volunteers. (The assessed topics are listed in the Appendix of these proceedings.)

3.2 Participation

Participating teams were allowed to submit up to 8 runs; additional runs could be scored locally. A total of 5 research teams submitted 29 runs for this year's Relevance Feedback task. The teams experimented with a variety of techniques including the following:

- Open Text Corporation: baseline runs, relevance feedback, pure feedback run, ranked-based fusion, sampling-based thresholds.
- Sabir Research, Inc: basic Rocchio feedback, all judged docs, add 40 terms, SMART ltu_Lnu vector run, all Boolean query negotiation terms.
- University of Iowa (Srinivasan): relstrings, WEKA API, WEKA SMO, Platt's sequential minimal optimizing algorithm, support vector machine, polynomial kernel, logistic regression, relevance probability estimates, classifier models.

- University of Missouri-Kansas City: VSM, BM25, LM, Expand15, CombMNZ.
- Ursinus College: BM25 baseline, Power Norm baseline, 5 terms over weight 5 added, 10 terms over weight 5 added.

3.3 Evaluation

29 Relevance Feedback runs were submitted by 5 research teams. Participating teams were allowed to submit up to 101,000 documents per topic. "Residual evaluation" was used for the Relevance Feedback task. Hence, before pooling, any documents that were already judged (of which there were at most 1000 per topic) were removed from the Relevance Feedback runs. Also, any documents past 100,000 residual documents retrieved were discarded before pooling.

The pools were then enriched before judgment with four additional runs:

- refRF08B (the final negotiated Boolean query results)
- randomRF08 (100,000 randomly selected residual documents from the unpooled documents for each topic)
- oldrel08 (10 (or as many as available) randomly chosen relevant documents from past judging of the topic)
- oldnon08 (10 (or as many as available) randomly chosen non-relevant documents from past judging of the topic).

The p(d) formula for the Relevance Feedback task was the same as for the Ad Hoc task except that the p(d) was set to 1.0 for all of the documents in oldrel08 and oldnon08 (small assessor-consistency study). Also, the first bin to judge was typically just 400 documents instead of 500 (because fewer documents needed to be judged to maintain the same accuracy (C value) as in the Ad Hoc task).

3.4 Relevance Assessment

Relevance assessments for the Relevance Feedback task were performed using exactly the same process as for the Ad Hoc task. A total of 12 topics were completed (3 of those 12 assessors completed at least one additional bin, 2 of those assessors completed all 5 additional bins).

3.5 Results

Of the 12 topics for which assessments are available, all 12 had some judgments of "relevant," but 3 topics had no "highly relevant" judgments (topics 36, 47 and 83). Thus there are 9 useful topics for measures that focus on "highly relevant" documents.

3.5.1 Number of Relevant Documents

The estimated number of (residual) relevant documents in the pool, on average per topic, was 23,536. The number varied considerably by topic, from 107 (for topic 14) to 101,197 (for topic 73).

Table 5 compares the estimated number of relevant documents for the 7 topics assessed in both the 2007 Ad Hoc task and this year's 2008 Relevance Feedback task. The estimates vary considerably; it's not immediately clear how much of the differences are from assessor inconsistency, or sampling error, or differences in the participating runs, or differences in the pooling depth (this year's runs were pooled 4x deeper, 100,000 vs. 25,000).

Over the 9 topics with highly relevant judgments, the estimated number of highly relevant documents in the pool, on average per topic, was 2,640. This number ranged from 22 (for topic 85) to 12,246 (for topic 73).

Topic	Judged Rel. in 2007	Est. Rel. in 2007	Est. Resid. Rel. in 2008
60 (2007-A-9)	10	83.2	36,821.0
73 (2007-B-5)	72	31,894.5	101,196.9
79 (2007-C-1)	35	1,486.6	56,162.3
80 (2007-C-2)	391	38,649.9	46,094.8
83 (2007-C-5)	44	13,987.5	830.6
85 (2007-C-7)	96	3,890.7	746.7
89 (2007-D-1)	78	6,083.6	11,660.8
Avg.	104	13725.1	36216.2

Table 5: Comparison of Estimated Numbers of Relevant Documents (2007 vs. 2008).

All Relevant (12 topics)	Retrieved	Precision	Recall	F_1
Reference Boolean	3,488	0.37	0.23	0.14
	Avg. K_r			
Median (10 baseline runs)	2,965	0.22	0.18	0.09
Median (19 feedback runs)	3,519	0.23	0.12	0.06
Highly Relevant only	(9 topics)			
Reference Boolean	3,870	0.11	0.34	0.12
	Avg. K_{hr}			
Median (10 baseline runs)	1,714	0.08	0.23	0.05
Median (19 feedback runs)	3,894	0.04	0.24	0.04

Table 6: Mean scores (Boolean and Participant Medians) for the Relevance Feedback task.

3.5.2 Baseline vs. Feedback Results

Table 6 compares the scores of the reference Boolean run, the median of 10 participant baseline runs, and the median of 19 participant feedback runs. Whether counting all relevant or just highly relevant documents, the mean F_1 score was higher for the reference Boolean run than for either the median baseline or feedback run. Furthermore, the median feedback run actually scored lower in mean F_1 than the median baseline run. However, the number of topics is small.

Table 8 shows the results for the 29 Relevance Feedback runs (and 2 reference runs). The highest mean F_1@K_r score came from the reference Boolean run. Several runs had a higher mean F_1@R_r than the Boolean run's mean F_1, suggesting that thresholding the retrieval set remains a challenge. In last year's R@B_r measure, few runs scored a higher mean R@B_r than the Boolean run.

Table 9 shows the results just counting highly relevant documents. A few runs did have a higher mean F_1@K_{hr} than the reference Boolean run, but (as per the medians) the majority did not. We should note that the groups did not have any training data for the highly relevant category this year.

3.5.3 Assessor Consistency Results

Assessor agreement on the (up to) 10 documents judged relevant and 10 documents judgment non-relevant can shed some light on the cause of the larged observed differences in our estimates. Table 7 shows these results. The "Previously Judged Relevant" column shows how this year's assessor judged the (up to) 10 documents that were judged relevant when the topic was used in the 2006 or 2007 Ad Hoc task (as per

Topic	Previously Judged Relevant (oldrel08)	Prev. Judged Non-relevant (oldnon08)
14 (2006-A-9)	tot=10, hrel=4, orel=1, non=3, gr=2	tot=10, hrel=0, orel=0, non=10, gr=0
28 (2006-C-4)	tot=10, hrel=9, orel=0, non=1, gr=0	tot=10, hrel=3, orel=1, non=6, gr=0
31 (2006-C-7)	tot=10, hrel=6, orel=3, non=1, gr=0	tot=10, hrel=1, orel=1, non=8, gr=0
36 (2006-D-3)	tot=10, hrel=0, orel=7, non=3, gr=0	tot=10, hrel=0, orel=0, non=10, gr=0
47 (2006-E-6)	tot=6, hrel=0, orel=2, non=4, gr=0	tot=10, hrel=0, orel=2, non=8, gr=0
60 (2007-A-9)	tot=10, hrel=2, orel=3, non=1, gr=4	tot=10, hrel=1, orel=2, non=7, gr=0
73 (2007-B-5)	tot=10, hrel=1, orel=0, non=9, gr=0	tot=10, hrel=1, orel=3, non=6, gr=0
79 (2007-C-1)	tot=10, hrel=4, orel=3, non=3, gr=0	tot=10, hrel=1, orel=2, non=7, gr=0
80 (2007-C-2)	tot=10, hrel=0, orel=8, non=2, gr=0	tot=10, hrel=0, orel=2, non=8, gr=0
83 (2007-C-5)	tot=10, hrel=0, orel=4, non=6, gr=0	tot=10, hrel=0, orel=0, non=10, gr=0
85 (2007-C-7)	tot=10, hrel=0, orel=0, non=10, gr=0	tot=10, hrel=0, orel=1, non=9, gr=0
89 (2007-D-1)	tot=10, hrel=2, orel=7, non=1, gr=0	tot=10, hrel=0, orel=1, non=9, gr=0
Totals	tot=116, hrel=28, orel=38, non=44, gr=6	tot=120, hrel=7, orel=15, non=98, gr=0

Table 7: Consistency of Previous and New Judgments for the 12 RF Topics (tot=total, hrel=highly relevant, orel=other relevant, non=non-relevant, gr=gray).

the oldrel08 run). The "Prev. Judged Non-relevant" column shows the same information for 10 documents previously judged non-relevant (as per the oldnon08 run). The labels are "tot" for total judged (which was 10 except when less than 10 documents were judged relevant previously), "hrel" for highly relevant, "orel" for other relevant, "non" for non-relevant, and "gr" for gray. We see that just 58% of previously judged relevant documents were judged relevant again this year; almost half of these were judged highly relevant (note that in previous years, the "highly relevant" category was not available). 18% of previously judged non-relevant documents were judged relevant this year; note that these non-relevant documents may have been rated highly by past search engines (which boosted their chance of being in the previous judging pool in the first place). For the previously judged relevant documents, we do not see perfect agreement for any topic. For the previously judged non-relevant documents, there are 3 topics for which both assessors agreed that all 10 documents were non-relevant. There were 2 topics (73 and 85) for which this year's assessor found that more of the previously judged non-relevant documents were relevant than of the previously judged relevant documents.

Past assessor agreement studies typically have found a lot of assessor disagreements, but generally retrieval systems are rated similarly regardless of which assessor's judgments are used [8]. We have not to date attempted to quantify whether our levels of disagreement are more or less than the norm. Note that none of these double-assessed documents were used in this year's evaluation (as residual evaluation excludes previously judged documents).

4 Interactive Task

In 2008, the Legal Track introduced a completely redesigned Interactive Task. The purpose of the redesign was to arrive at a task that modeled more completely and accurately the objectives and conditions of e-discovery in the real world. It was hoped that featuring such a task would further advance the Legal Track toward its goal of fostering greater communication and collaboration among the legal, scientific, and e-discovery communities. In the following, we (1) review key features of the design of the task, (2) provide a description of the procedures whereby task submissions were evaluated, (3) review specific parameters that defined this year's exercise, (4) summarize the results obtained, and (5) provide some further analysis of the results of the task.

Run	Fields	Ret_r	K_r	(P@K_r, R@K_r)	$F_1\text{@K}_r$	$F_1\text{@R}_r$	S1J, P5	R@B_r, R@ret_r
refRF08B	bvmBM	3488	3488	(0.367, 0.228)	**0.142**			0.228, 0.228
otRF08fb	bvmBM	3488	3488	(0.367, 0.228)	**0.142**	0.151	9/12, 0.656	0.228, 0.228
otRF08rvl	rmM	100000	87738	(0.126, **0.607**)	0.134	0.185	6/12, 0.450	0.096, 0.640
UMKCTL08RF6	F-bBM	100000	3476	(0.367, 0.212)	0.131	**0.309**	8/12, 0.667	0.212, 0.690
otRF08fv	bmM	100000	70255	(0.099, 0.495)	0.119	0.202	8/12, 0.533	0.103, 0.596
otRF08frw	brmBM	100000	39726	(0.140, 0.549)	0.116	0.307	**10/12**, **0.800**	0.236, 0.659
UMKCTL08RF3	F-bM	100000	859	(**0.377**, 0.142)	0.109	0.239	7/12, 0.583	0.154, 0.662
IowaSL08RF3B		50663	2442	(0.249, 0.207)	0.092	0.209	8/12, 0.650	0.199, 0.545
UMKCTL08RF2	F-bM	100000	880	(0.371, 0.108)	0.091	0.206	8/12, 0.667	0.120, 0.626
UMKCTL08RF1	F-bM	100000	875	(0.327, 0.124)	0.088	0.225	8/12, 0.629	0.148, 0.669
otRF08fbF	F-bmBM	100000	36620	(0.118, 0.392)	0.084	0.261	8/12, 0.633	**0.261**, 0.472
UMKCTL08RF5	F-bBM	100000	3519	(0.348, 0.149)	0.083	0.237	8/12, 0.667	0.144, 0.670
otRF08fbFR	F-bmBM	100000	8776	(0.266, 0.261)	0.082	0.261	8/12, 0.633	**0.261**, 0.472
IowaSL08RF1B		75148	2442	(0.287, 0.150)	0.082	0.204	8/12, 0.589	0.135, **0.694**
otRF08F	F-mM	100000	36716	(0.109, 0.327)	0.075	0.156	7/12, 0.600	0.118, 0.446
SabL08rf1	F-bdporm	99609	3547	(0.318, 0.145)	0.075	0.248	9/12, 0.717	0.145, 0.606
IowaSL08RF3A	F	50663	2442	(0.227, 0.148)	0.062	0.150	3/12, 0.383	0.163, 0.545
SabL08rfbase	bdporm	100000	3517	(0.313, 0.111)	0.060	0.167	5/12, 0.422	0.110, 0.579
IowaSL08RFTr	F-bdpr	100000	2302	(0.214, 0.116)	0.058	0.160	6/12, 0.650	0.118, 0.617
IowaSL08RF1A	F	75148	2354	(0.150, 0.121)	0.058	0.197	5/12, 0.367	0.132, 0.689
IowaSL08RF2B		15944	2288	(0.231, 0.104)	0.053	0.099	6/12, 0.517	0.085, 0.300
IowaSL08RF2A	F	15944	2344	(0.209, 0.098)	0.051	0.084	4/12, 0.283	0.080, 0.300
UCBM25T10Th5	F-r	100000	21406	(0.257, 0.086)	0.048	0.126	3/12, 0.392	0.069, 0.395
otRF08FR	F-mM	100000	8826	(0.203, 0.129)	0.047	0.156	7/12, 0.600	0.118, 0.446
UCPwrT10Th5	F-r	100000	13233	(0.172, 0.113)	0.036	0.092	3/12, 0.267	0.066, 0.445
IowaSL08RF2C	F	15944	2351	(0.176, 0.086)	0.034	0.091	4/12, 0.333	0.083, 0.300
UCBM25T5Th5	F-r	100000	6517	(0.240, 0.037)	0.021	0.121	6/12, 0.433	0.070, 0.377
UCPwrT5Th5	F-r	100000	11880	(0.110, 0.101)	0.020	0.079	2/12, 0.267	0.062, 0.442
UCRFPwrBL	r	100000	1547	(0.113, 0.034)	0.017	0.089	1/12, 0.117	0.054, 0.420
UCRFBM25BL	r	100000	564	(0.200, 0.014)	0.013	0.128	3/12, 0.283	0.074, 0.432
randomRF08		100000	20000	(0.005, 0.002)	0.003	0.002	0/12, 0.000	0.002, 0.018

Table 8: Mean scores for submitted Relevance Feedback task runs, using All Relevant documents.

Run	Fields	Ret_r	K_{hr}	(P@K_{hr}, R@K_{hr})	$F_1\text{@K}_{hr}$	$F_1\text{@R}_{hr}$	S1J, P5	R@B_r, R@ret_r
UMKCTL08RF2	F-bM	100000	879	(0.109, 0.259)	**0.136**	0.126	3/9, 0.222	0.181, 0.607
otRF08frw	brmBM	100000	2392	(0.124, 0.336)	0.129	0.145	2/9, 0.289	0.347, 0.647
UMKCTL08RF3	F-bM	100000	857	(0.121, 0.269)	0.128	0.158	2/9, 0.222	0.237, 0.699
UMKCTL08RF6	F-bBM	100000	3857	(0.117, 0.358)	0.124	**0.204**	4/9, 0.311	0.359, 0.694
UMKCTL08RF1	F-bM	100000	872	(0.098, 0.286)	0.121	0.157	1/9, 0.222	0.250, 0.693
refRF08B	bvmBM	3870	3870	(0.112, 0.336)	0.118			0.336, 0.336
otRF08fb	bvmBM	3870	3870	(0.112, 0.336)	0.118	0.148	2/9, 0.259	0.336, 0.336
UMKCTL08RF5	F-bBM	100000	3894	(0.083, 0.239)	0.092	0.146	3/9, 0.222	0.239, 0.694
IowaSL08RF3B		55146	1035	(0.105, 0.284)	0.084	0.123	**6/9**, 0.378	0.346, 0.540
IowaSL08RFTr	F-bdpr	100000	923	(**0.146**, 0.157)	0.076	0.080	4/9, **0.422**	0.159, 0.583
IowaSL08RF1A	F	79742	946	(0.057, 0.187)	0.066	0.093	1/9, 0.111	0.243, 0.794
IowaSL08RF1B		79742	1035	(0.094, 0.169)	0.065	0.104	**6/9**, 0.393	0.223, **0.795**
UCBM25T10Th5	F-r	100000	18905	(0.052, 0.257)	0.054	0.096	0/9, 0.111	0.117, 0.607
otRF08fbFR	F-bmBM	100000	10150	(0.096, 0.400)	0.050	0.202	3/9, 0.267	**0.367**, 0.661
SabL08rfbase	bdporm	100000	3896	(0.064, 0.146)	0.049	0.053	1/9, 0.148	0.144, 0.686
otRF08fbF	F-bmBM	100000	15305	(0.032, **0.516**)	0.044	0.202	3/9, 0.267	**0.367**, 0.661
IowaSL08RF2B		16229	912	(0.096, 0.120)	0.044	0.093	4/9, 0.222	0.147, 0.459
SabL08rf1	F-bdporm	99628	3926	(0.034, 0.176)	0.042	0.103	1/9, 0.200	0.175, 0.703
IowaSL08RF3A	F	55146	1035	(0.040, 0.204)	0.037	0.095	2/9, 0.244	0.287, 0.540
UCPwrT10Th5	F-r	100000	8934	(0.032, 0.174)	0.032	0.041	0/9, 0.000	0.186, 0.728
otRF08rvl	rmM	100000	39342	(0.018, 0.418)	0.029	0.036	3/9, 0.200	0.133, 0.718
otRF08FR	F-mM	100000	10213	(0.016, 0.311)	0.028	0.089	2/9, 0.244	0.237, 0.658
IowaSL08RF2A	F	16229	956	(0.029, 0.110)	0.027	0.046	1/9, 0.089	0.126, 0.459
IowaSL08RF2C	F	16229	963	(0.020, 0.115)	0.025	0.078	1/9, 0.133	0.131, 0.459
otRF08fv	bmM	100000	22248	(0.018, 0.371)	0.025	0.085	1/9, 0.156	0.185, 0.686
UCPwrT5Th5	F-r	100000	4557	(0.029, 0.044)	0.020	0.031	0/9, 0.067	0.095, 0.737
otRF08F	F-mM	100000	15390	(0.009, 0.444)	0.015	0.089	2/9, 0.244	0.237, 0.658
UCBM25T5Th5	F-r	100000	3980	(0.022, 0.025)	0.015	0.079	0/9, 0.133	0.109, 0.609
UCRFPwrBL	r	100000	265	(0.021, 0.008)	0.007	0.019	0/9, 0.022	0.069, 0.505
UCRFBM25BL	r	100000	323	(0.051, 0.013)	0.006	0.026	1/9, 0.100	0.106, 0.568
randomRF08		100000	10000	(0.000, 0.000)	0.000	0.000	0/9, 0.000	0.000, 0.000

Table 9: Mean scores for submitted Relevance Feedback task runs, using only Highly Relevant documents.

4.1 Task Design

The goal of the Interactive Task is to model as accurately as possible the real-world conditions in which companies and law firms, and the e-discovery firms they engage, must meet their document-retrieval objectives and obligations. Pursuant to that goal, four key features were introduced into the 2008 design of the task: (1) the designation of a single individual (an attorney) to act as the authority for defining the intent and scope of a topic; (2) a provision that allowed participants to engage with that authority for purposes of clarifying relevance to a topic; (3) the specification of the task objective to be, for each topic, a binary assessment (relevant, not relevant) of all documents in the target collection; and (4) the provision for an appeal and adjudication process as a corrective on possible errors in sample assessments. In the following, we elaborate on each of these features.[3]

Topic Authority. When the lead attorney on a matter oversees a document production, he or she will have formed, or be in the process of forming, an opinion as to what is responsive to the requests for production and what is not. In forming that conception of responsiveness, the attorney will take into account both considerations of substantive relevance and considerations related to case strategy (e.g., whether to take a broad view of responsiveness, thereby minimizing the risk of being challenged for underproduction or to take a narrow view, thereby restricting what is produced to just that which has to be produced). When that attorney employs the products or services of an e-discovery firm, or, for that matter, the services of a traditional manual-review team, he or she does so with the goal of efficiently applying that conception of responsiveness across the full document population implicated by the matter. The review team is not asked to consider, weigh, and resolve differences between all possible conceptions of relevance; the review team is asked to replicate, across the document population, one conception of responsiveness, that of the senior attorney who has hired the firm and who bears ultimate responsibility for the validity of the production. The goal, therefore, of the review team or e-discovery firm engaged to assist in a document production effort is to replicate the responsiveness assessments the senior litigator in the matter would make, if he or she had the time and leisure needed to review for responsiveness every document in the population.

In order to model this aspect of real-world e-discovery, we introduced a new role into the Interactive task, that of the "Topic Authority." The role that the Topic Authority plays in the Interactive task is modeled on that played by the lead attorney in a lawsuit, the attorney who must form a conception of what is and is not responsive to a request for production and who must then communicate that conception to the team or vendor who will be asked to replicate that conception of responsiveness across the target document population. In the Interactive task, it is the role of the Topic Authority to define what is and is not relevant to a topic and it is the objective of the teams participating in the task to retrieve documents that match the Topic Authority's definition of relevance.

In keeping with this role, the Topic Authority is essential to the execution of three key elements of the task. The first is topic clarification. While teams are going about their efforts to retrieve documents relevant to a topic, it is the role of the Topic Authority to give the teams guidance when they seek clarification as to the intent and scope of the topic. The second is review oversight. In order to be able to obtain valid measures of effectiveness, it is essential that the samples reviewed for purposes of evaluation be assessed in accordance with the Topic Authority's conception of relevance; it is the role of the Topic Authority to provide assessors with guidance as to what is and is not relevant. The third is final adjudication. As an additional measure to ensure the quality of the assessments in the evaluation samples, we provide teams with the opportunity to appeal any sample assessments they believe were made in error; it is the role of the Topic Authority to render final judgment on all appealed assessments.

Topic clarification. 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 better understanding of the Topic Authority's conception of relevance. In the 2008 Interactive task, this provision took the following form. Each team could ask for up to 10 hours of a Topic Authority's time for purposes of clarifying a topic. A team could call upon a Topic Authority at any point in the exercise, from the kickoff of the task to the deadline for the

[3]For a full description of the task protocol, see the task guidelines posted on the Legal Track website [3].

submission of results. How a team made use of the Topic Authority's time was largely unrestricted: a team could ask the Topic Authority to pass judgment on example documents; a team could submit questions to the Topic Authority by email; a team could arrange for conference calls to discuss aspects of the topic. The one constraint (apart from the 10-hour maximum) we did place on communication between the teams and their designated Topic Authorities was introduced in order to minimize the sharing of information developed by one team with another; while we instructed the Topic Authorities to be free in sharing the information they had about their topics, we also asked that they avoid volunteering to one team specific information that was developed only in the course of interaction with another team.

Submission of results. When an attorney vouches for the validity of a document production, he or she is vouching for the accuracy of a binary classification of the document population implicated by the litigation, a classification into the subset of the population that is responsive to the requests for production and the subset that is not. When an e-discovery firm supports an attorney in this effort, it must make a similar relevance determination. The 2008 Interactive task, modeling this requirement, specified that each team's final deliverable be a binary classification of the full population for relevance to each target topic. Teams were of course free to use relevance ranking as a means to arrive at their result sets, but the final deliverable was a single binary classification (relevant/not relevant) of the full population of documents.

Appeal and adjudication. Assessors can make errors; as an additional quality-control check on the sample assessments, we introduced an appeal and adjudication phase to the task. Once sample review was complete, participating teams were given access to sample results, allowing them to review any mismatches between their assessments and those of the assessors; teams were not, at this stage, given access to the results submitted by any other team. Teams were permitted to appeal any assessments they believed were directly and specifically contradicted by information given them by the Topic Authority in the course of their communications regarding the topic. Teams were not permitted to appeal assessments that represented differences in interpretation. The "court of appeal" and the final arbiter was the Topic Authority.

4.2 Evaluation

The metrics used to gauge the effectiveness of each team's efforts were recall, precision, and F_1. In this section, we briefly describe the sampling, assessment, and measurement procedures whereby estimates of the target metrics were obtained.

4.2.1 Sampling

A separate evaluation sample was drawn for each topic targeted in the Interactive Task. The sampling design was fairly straightforward, its salient features being results-based stratification, fairly large sample sizes, and moderately disproportionate representation of strata. Specifics of the sampling design are as follows.

Stratification. For purposes of drawing each sample, the document population was partitioned into strata, with strata being defined by the cross-classification of results submitted by each of the teams whose performance was to be measured via the sample. In the case of a topic for which three teams submitted results for evaluation, for example, the collection would be partitioned into eight strata, one for each of the possible combinations of binary assessments (R/R/R, R/R/N, R/N/R, R/N/N, and so on). The full evaluation sample was created by drawing samples of documents from each of the resulting strata.

Sample Size. The samples drawn for each topic were fairly large (ranging from 2,500 documents to 6,500 documents). The provision for larger sample sizes made it possible to draw a sufficient number of documents from each possible result combination (stratum) and to obtain fairly precise estimates of the target metrics (even in the case of low-yielding topics). The drawing of larger samples was in part enabled by the fact that the Interactive Task, being in other regards a fairly time-intensive exercise, targets a relatively small number of topics, thereby allowing a greater amount of assessor resources to be concentrated on each topic.

Allocation. In constructing the sample, strata were largely represented in proportion to their full-population proportions. The exception to the rule is that very large strata (such as the "All-N" stratum, the stratum containing documents no team assessed as relevant), though represented in larger numbers than the smaller strata, were not represented in the numbers strict proportionality would have dictated. This

departure from strict proportionality enabled the inclusion in the sample of a greater number of each team's positive assessments, and, in particular, cases in which one team's assessments were at variance with those of all other teams.

The actual results of constructing evaluation samples in accordance with this sampling design are detailed below (section 4.4.3).

4.2.2 Assessment

Once the samples were drawn, the documents they contained were reviewed for relevance to the target topics. Document assessment followed a two-step procedure, whereby volunteer assessors made a first-pass review of the sample and then the participating teams, after reviewing the results of the assessors' first pass, had the opportunity to appeal initial assessments to the Topic Authority for final adjudication.

First-Pass Assessment. The task of the volunteer assessors was twofold. In the first instance, each assessor had to make a threshold decision as to the assessability of each document that had been assigned to him or her.[4] Primary reasons for a document's having been deemed not assessable were (1) length (in excess of 300 pages), (2) substantial non-English content, and (3) failure to load properly into the review platform; the vast majority (98.8%) of documents in the samples met the threshold criterion of assessability.

Documents that met the assessability threshold were then reviewed for relevance to the target topic. For purposes of making relevance assessments, assessors were provided with topic-specific guidelines that documented the criteria the Topic Authority wanted to be applied in assessing the relevance of documents to the target topic; these guidelines were essentially compilations of all the guidance that the Topic Authority had given teams in the course of the exercise to that point. When assessors encountered a document the status of which was insufficiently determined by these topic-specific guidelines, they had the opportunity to seek further guidance (via email) from the appropriate Topic Authority. The final topic-specific guidelines (including additional guidance given by the Topic Authorities to the assessors) can be found on the Legal Track Home Page [3].

Appeal & Adjudication. Recognizing that, although the topic-specific guidelines and the opportunity for further clarification could be expected to go some way down the path to ensuring that sample assessments were aligned with the Topic Authority's conception of relevance, those provisions in themselves could not be expected to eliminate all scope for error, the coordinators included, as a further corrective on sample assessments, a provision for appeal and adjudication of the first-pass assessments.

The mechanism for submitting an appeal of a first-pass assessment was fairly straightforward. Once the first-pass review of the samples had been completed, teams were provided both with their initial (pre-adjudication) recall, precision, and F_1 scores and with lists recording all assessable documents in the sample, the assessor's assessment of each of those documents, and their own submitted assessment of the same documents. Teams did not have access, at this stage, to the assessments submitted by any other team. Teams were also provided with each document's probability of selection into the sample (information that a team could use to prioritize its appeal efforts). As a further aid to their review of the assessments, teams were given the topic-specific guidance that had been provided to the assessors (supplemented with the additional guidance the Topic Authority had provided the assessors in the course of the sample review).

Upon comparing the first-pass assessments with their own submitted assessments, a team could decide to appeal initial assessments it believed the manual assessor had made in error. The circumstances in which a team could lodge an appeal were not unconstrained; a team could lodge an appeal only in one of the following circumstances.

- An appeal could be made in cases in which a team believed that a sample assessment was directly contradicted by specific guidance already provided by the Topic Authority.

[4]Completed review batches were received from individuals representing the following law schools and non-academic institutions: U. of Baltimore, Georgetown, Loyola Law School Los Angeles, U. of Maine, Rutgers School of Law—Camden, Texas Wesleyan, Anchors Smith Grimsley, Parker, Bunt & Ainsworth, Redgrave Daley Ragan & Wagner, Stafford Frey Cooper, Chevron Corporation, and the National Archives and Records Administration.

- An appeal could be made in cases in which a team believed that it was immediately apparent that one sample assessment was inconsistent with another (e.g., a set of duplicate documents that had been inconsistently assessed).

A team could not appeal a case in which previously obtained guidance was insufficiently specific to decide between competing assessments (the task made it the responsibility of the team to obtain, through interaction with the Topic Authority, guidance of the specificity required to decide such cases).

The number of documents a team could appeal was unrestricted (although teams were encouraged, in the interest of efficiency, to be judicious in selecting the documents they wished to appeal).

Teams were asked to consolidate all their appeals into a single document, including the following information (where appropriate) for each document the assessment of which was being appealed:

- document ID;
- current assessment (that of the first-pass assessor);
- proposed revised assessment;
- specific reason for the revision;
- excerpt(s) from the document supporting the case for revision; and
- additional notes or comments.

The appeals documents were not anonymized before being submitted to the Topic Authority for final adjudication (meaning the Topic Authority did know which appeals had been made by which teams). Teams were encouraged to make their appeals documentation as complete as possible, but were permitted, if the Topic Authority agreed, to arrange time to discuss their appeals by telephone, if they believed that that would make the process more efficient.

The final decision on all appeals rested with the Topic Authority. There was no second round of appeals.

4.2.3 Metrics

Once we have final assessments for all documents (that is, once we have completed a first-pass review of all sampled documents, allowed teams to appeal any first-pass assessments they wish to dispute, and obtained the Topic Authority's final judgment on all appealed documents), we are in a position to obtain estimates of the metrics by which we will gauge each team's effectiveness in performing the task (recall, precision, and F_1). The procedures for obtaining those estimates are largely a matter of (i) obtaining stratified estimates of the inputs to the target metrics, then (ii) combining those inputs in the appropriate manner to obtain estimates of the metrics themselves. The specifics of these estimation procedures are reviewed in Appendix A to this document.

4.3 Task Specifics

With the posting of the final guidelines for the 2008 Interactive Task (on June 22, 2008), teams were able to begin their work. In this section, we describe some of the specific elements that defined this year's running of the task.

4.3.1 Document Collection

The document collection used for the Interactive task was be the same as that used for the Ad Hoc and Relevance Feedback tasks, the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0. For more on the document collection, see section 2.1 above.

4.3.2 Topics

Three topics were selected as the retrieval targets for the Interactive task (the resource-intensive nature of the task, both for teams and for Topic Authorities, constrained the number of topics we could accommodate). A participating team was free to take on one, two, or all three topics, as it chose.

All three topics were associated with the same mock complaint (a modified version of a complaint used in the 2006 Legal Track). Two of the topics (102, 103) were entirely new for 2008; one (104) was used in a prior years (in the 2006 Ad Hoc task and in the 2007 Relevance Feedback task). Even the previously-used topic, however, was essentially "new," due to the fact that modifications to the complaint and the addition of the Topic Authority's guidance effectively reoriented the topic.

The specific topics are as follows.

- **Topic 102.** Documents referring to marketing or advertising restrictions proposed for inclusion in, or actually included in, the Master Settlement Agreement ("MSA"), including, but not limited to, restrictions on advertising on billboards, stadiums, arenas, shopping malls, buses, taxis, or any other outdoor advertising.
- **Topic 103.** All documents which describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes.
- **Topic 104.** All documents discussing or referencing payments to foreign government officials, including but not limited to expressly mentioning "bribery" and/or "payoffs."

Topics were selected with an eye to representing the sorts of challenges typically encountered in real-world document discovery and were, as is typical of real-world document requests, underspecified as to scope and intent. Among the questions left open by the statement of Topic 103, for example, include what set of marketing practices constitute a "campaign" and where to draw the boundaries around specifically "retail" marketing campaigns. Topic 104 raises the question of whether just illegitimate payments are in-scope or legitimate payments are to be considered within the scope of the request as well. Each of the topic statements, upon further inspection, will be found to raise a number of such questions, questions the answers to which will depend on the outlook the producing party (represented, in our task, by the Topic Authority) has on the issues being litigated, the specific request for production, and his or her discovery obligations.

4.3.3 Topic Authorities

To guide the teams in addressing these sorts of questions, a single Topic Authority was assigned to each of the topics. The Topic Authorities for the 2008 Interactive Task were as follows.

- **Topic 102.** Joe Looby (of FTI Consulting).
- **Topic 103.** Maura Grossman (of Wachtell, Lipton, Rosen & Katz).
- **Topic 104.** Conor Crowley (of Daley Crowley LLP).

Archives of the topic-clarification guidance the Topic Authorities provided both to the teams and to the relevance assessors over the course of the exercise have been maintained and will be made available by NIST for future use by researchers.

4.3.4 Participating Teams

Four teams submitted results for the Interactive task. The teams and the topics for which they submitted results are as follows.

- **University at Buffalo** ("UB"). Submitted results for Topic 103.
- **Clearwell Systems** ("CS"). Submitted results for Topics 102, 103, and 104.

- **H5** ("H5"). Submitted results for Topic 103.
- **University of Pittsburgh** ("UP"). Submitted results for Topics 102 and 103.

As can be seen from the list, participants in the task included both teams from academic institutions and teams from industry. While the task has been designed to be fair and accessible to all participants, it is also fair to recognize, in light of the fact that there is a mix of academic and industry participants (a mix that is very welcome), that different teams will bring different resources to the task.

In addition to the results submitted by these teams, who participated in all phases of the task, results were also submitted, for all three topics, by participants in the Ad Hoc task, who, however, did not interact with the Topic Authorities in preparing their submissions. For evaluation purposes, we created an additional benchmark result set, the Ad Hoc Pool ("AH"), formed by pooling each of the 64 Ad Hoc submissions, to a maximum depth of 100,000, along with the results of each of the 4 associated Boolean queries (for more on the Boolean reference runs, see Section 2.4 above). The random run was not used.

4.4 Results

Task guidelines and topics were released on June 22, 2008. Teams submitted their results on or before September 12, 2008. The evaluation protocol outlined above was carried out in the weeks following. In this section, we review the results of the Interactive Task.

4.4.1 Team-TA Interaction

As noted above, teams were permitted to call on up to 10 hours of a Topic Authority's time for purposes of clarifying the scope and intent of a topic. Figure 2 summarizes 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 for each team for each topic); the grey 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 to the left of each bar). The contributors to the Ad Hoc Pool did not participate in the topic-clarification phase of the task, and so recorded zero minutes of Topic Authority time.

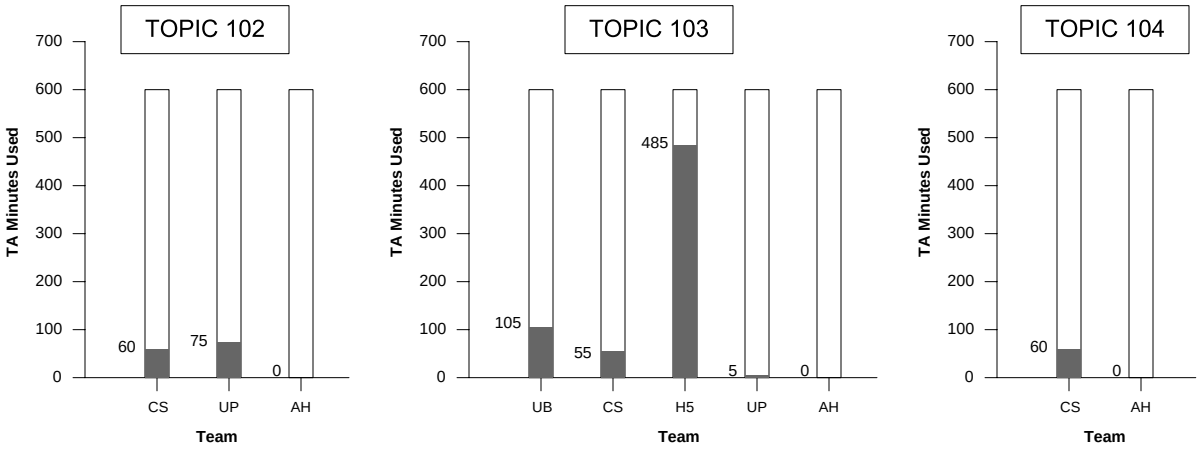

Figure 2: Team-TA Interaction Time.

A few initial observations can be made regarding these data. First, there is considerable range in the amount of time teams spent engaging with their Topic Authorities. For Topic 103, for example, one team (Pittsburgh) used just 5 of their permitted 600 Topic Authority minutes, while another (H5) used 485

minutes. Second, apart from the H5 team, participants generally used only a small portion of their permitted Topic Authority time; on average, teams (apart from H5) used about 60 minutes of a Topic Authority's time for purposes of clarifying the definition of a target topic (about 10% of the time allowed for this purpose). In analyzing the results, we are therefore interested in seeing whether there is a correlation between up-front time spent with the Topic Authority and retrieval effectiveness (as measured by recall, precision, and F_1).

4.4.2 Submissions

Submissions for each topic consisted of the submissions of teams fully participating in the task as well as the pooled submissions of the Ad Hoc participants. Table 10 summarizes the total number of documents submitted by each team for each topic.

Team	**Topic 102**	**Topic 103**	**Topic 104**
ublegal08	n/a	67,334	n/a
Clearwell08	13,695	175,455	549
H52008	n/a	608,807	n/a
PittSIST1	4,505	25,816	n/a
adhocpool08	546,126	837,889	689,548

Table 10: Submitted Results.

As can be seen from the table, there was, even within the same topic, considerable variation in the numbers of documents participants identified as relevant. For Topic 103, for example, setting aside the set of pooled Ad Hoc results, submissions ranged from a low of 25,816 documents (Pittsburgh) to a high of 608,807 documents (H5). The goal of evaluation would be to see how these submissions lined up with the Topic Authorities' conceptions of relevance

4.4.3 Sampling & Assessment

The evaluation protocol described above (Section 4.2) was followed, with samples being drawn, assessed, and adjudicated. The results of the sample assessment process are summarized in Tables 11 – 13; in the tables, the column labels are defined as follows:

N = total documents in the stratum;

n = total documents sampled from the stratum;

a = total sampled documents observed to be assessable;

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

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

A few further notes on the sampling and assessment data follow.

Sample sizes. As can be seen from the tables, larger samples were drawn for some topics than were drawn for others. Generally speaking, the greater the number of participants who submitted results for a topic, the larger the sample that was drawn; this was done to ensure that, even for a topic with several participants, we would be able to sample a meaningful number of documents from each stratum. Topic 103, for example (see Table 12), saw participation by five teams (including the Ad Hoc Pool), making for 32 possible strata (as defined by cross-classifying team submissions), and so called for a larger sample (6,500 documents); Topic 104, on the other hand (see Table 13), saw participation by two teams (again including the Ad Hoc Pool), making for four possible strata, and so was covered with a smaller sample (2,500 documents).

CS	UP	AH	N	n	a	r_1	r_2
R	R	R	2,015	215	214	203	203
R	R	N	1	1	1	1	1
R	N	R	10,608	1,041	1,038	665	666
R	N	N	1,071	108	102	15	17
N	R	R	2,435	249	246	199	199
N	R	N	54	11	11	4	4
N	N	R	531,068	1,125	1,110	354	352
N	N	N	6,362,940	1,750	1,713	106	106
TOTAL			**6,910,192**	**4,500**	**4,435**	**1,547**	**1,548**

Table 11: Sampling & Assessment – Topic 102.

Allocation among strata. As noted above (Section 4.2.1), samples were composed by sampling from each stratum, with stratum-specific sample sizes being largely proportionate to the stratum's size in the full collection; an exception was made in the case of very large strata, from which fewer documents were drawn than strict proportionality would dictate, so as to ensure that even small strata would have some representation. For Topic 103, for example, the "All-N" stratum (the stratum containing documents no team considered relevant) contained, in the full collection, 5,708,286 documents, or 82.6% of the collection; from this stratum, we sampled a large number of documents (1,625) but a number smaller than the full-collection proportion would dictate (the 1,625 represented 25.0% of the 6,500-document sample). This under-representation of the "all-N" stratum enabled us to bring more positively assessed documents into the sample and to obtain a clearer view of where teams differed from each other, while still obtaining a good measure of the rate at which relevant documents were missed by all participants collectively.

First-Pass Assessments. A total of 22 volunteer assessors participated in the first-pass review of the evaluation samples. Most of the assessors were students at law schools; others were practicing attorneys or, in some cases, paralegals. Documents were reviewed in 500-document batches; most assessors completed a single batch, although some took on additional batches after completing their first. In carrying out their task, all assessors were, as described above (Section 4.2.2), supported by guidance provided by the Topic Authority. In the tables, the column labeled a reports, for each stratum, the number of documents the reviewers found to be assessable (as noted above (Section 4.2.2), nearly 99% of sampled documents were found to be assessable); the column labeled r_1 reports the number of documents the reviewers found to be both assessable and relevant.

Appeal & Adjudication. The column labeled r_2 reports the post-adjudication counts of relevant documents, that is, for each stratum, the number of documents found to be both assessable and relevant after teams had had the opportunity to appeal any first-pass assessments they believed were inconsistent with the Topic Authority's guidance and after the Topic Authority had rendered a final assessment on those appealed documents (see Section 4.2.2). In all, 966 assessments were appealed (aggregating across all three topics). Of these, the Topic Authority agreed with the appealing team (or teams; there were some cases of overlapping appeals) on 762 (78.9%); the Topic Authority denied the appeal (maintained the first-pass assessment) on 204 (21.1%). The topic that saw the most appeals was Topic 103 (950 appealed assessments, compared to 10 for Topic 102 and 6 for Topic 104), and this is the topic for which the appeal and adjudication mechanism had the greatest impact on results. We provide some further analysis of the appeal and adjudication process below (Section 4.5.3).

4.4.4 Metrics

Once the sampling/assessment/adjudication process had been completed, it was possible to obtain, via the estimation procedures described in Appendix A to this document, estimates both of the yield of (actually) relevant documents for each topic and of the effectiveness of each of the participating teams in retrieving

UB	CS	H5	UP	AH	N	n	a	r_1	r_2
R	R	R	R	R	5,727	46	46	38	43
R	R	R	R	N	24	5	5	4	5
R	R	R	N	R	11,965	98	98	78	94
R	R	R	N	N	995	9	9	9	9
R	R	N	R	R	131	5	5	3	2
R	R	N	R	N	0	0	0	0	0
R	R	N	N	R	1,547	13	13	2	3
R	R	N	N	N	220	5	5	2	2
R	N	R	R	R	1,901	15	15	11	13
R	N	R	R	N	46	5	5	2	4
R	N	R	N	R	17,082	145	145	111	130
R	N	R	N	N	10,291	84	84	61	73
R	N	N	R	R	176	5	5	1	0
R	N	N	R	N	19	5	5	2	1
R	N	N	N	R	7,679	62	61	23	17
R	N	N	N	N	9,531	77	77	17	14
N	R	R	R	R	8,068	65	65	49	59
N	R	R	R	N	101	5	5	2	3
N	R	R	N	R	73,280	541	540	393	481
N	R	R	N	N	28,409	235	235	146	186
N	R	N	R	R	1,185	10	10	4	1
N	R	N	R	N	37	5	4	3	3
N	R	N	N	R	23,688	193	193	84	44
N	R	N	N	N	20,078	171	164	57	44
N	N	R	R	R	5,321	43	43	33	41
N	N	R	R	N	371	5	5	2	3
N	N	R	N	R	151,787	800	795	552	672
N	N	R	N	N	293,439	1,100	1,095	621	822
N	N	N	R	R	2,253	18	18	6	5
N	N	N	R	N	456	5	5	2	2
N	N	N	N	R	526,099	1,100	1,087	234	145
N	N	N	N	N	5,708,286	1,625	1,579	111	60
TOTAL					**6,910,192**	**6,500**	**6,421**	**2,663**	**2,981**

Table 12: Sampling & Assessment – Topic 103.

those documents.

Yields. Estimates (post-adjudication) of the full-collection yields of relevant documents for each topic are summarized in Table 14. Reported in the table are, for both count and percentage, the point estimate of the yield and the 95% confidence interval associated with the estimate.

Metrics. Estimates (post-adjudication) of the participants' effectiveness (as measured by recall, precision, and F_1), in retrieving those relevant documents are summarized in Table 15. Reported in the table are, for each metric, the point estimate of the metric and the 95% confidence interval associated with the estimate.

With regard to the results reported in the table, a few initial observations are possible.

We see, first, that the set formed by aggregating the 64 Ad Hoc submissions (to a depth of 100,000) along with the results of the four Boolean reference queries (the Ad Hoc Pool), generally was able to achieve recall and precision scores in the 0.30 - 0.40 (30% - 40%) range. The one exception to the rule is the low precision achieved on Topic 104; this exception, however, is likely an effect of the manner in which the set was constructed (going to a depth of 100,000, where possible, for each run) combined with the fact that Topic 104 was very low-yielding. These recall and precision numbers were sufficient to allow the Ad Hoc Pool, on two topics (102 and 104), to rank highest in overall effectiveness (as measured by F_1).

We see, second, that the one team that fully availed itself of the opportunity to engage with the Topic

CS	AH	N	n	a	r_1	r_2
R	R	527	265	263	62	64
R	N	22	15	15	0	0
N	R	689,021	970	963	24	22
N	N	6,220,622	1,250	1,242	7	6
TOTAL		**6,910,192**	**2,500**	**2,483**	**93**	**92**

Table 13: Sampling & Assessment – Topic 104.

	Document Count		% of Collection	
Topic	**Est.**	**95% C.I.**	**Est.**	**95% C.I.**
Topic 102	562,402	(489,837, 634,967)	8.1%	(7.1%, 9.2%)
Topic 103	786,862	(732,679, 841,045)	11.4%	(10.6%, 12.2%)
Topic 104	45,614	(20,913, 70,314)	0.7%	(0.3%, 1.0%)

Table 14: Estimated Yields.

Authority, the team from H5 (which submitted results for Topic 103 only), was able to achieve substantially higher recall than the other entrants (including the Ad Hoc Pool) while at the same time achieving high precision (in a statistical tie with the Pittsburgh team for highest precision). This result enabled the H5 team, on the topic for which it submitted results, to rank highest in overall effectiveness (as measured by F_1), and that by a considerable margin (over 0.30 (30 percentage points) higher than the next-highest entry, that of the Ad Hoc Pool).

We see, third, that the other participants, who took less advantage of the opportunity to engage with the Topic Authority, generally submitted results that scored high on precision but low on recall. Whether this result is to be attributed (a) to incomplete topic clarification, (b) to a drawback of the retrieval methods applied, or (c) to a combination of both, is a question for further analysis; the participants' own papers on this year's task will undoubtedly add further information on this question.

In the next section, we provide some further analysis of these results. Before turning to that analysis, however, we first provide a visual summary of the immediate post-adjudication results; Figure 3 plots each team's post-adjudication scores on precision-recall diagrams.

4.5 Further Analysis

We have seen the immediate post-adjudication metrics. There remain, however, a number of important questions that merit further study if we are to gain a proper understanding of the significance of the task results. In this section, we provide some further analysis (i) of the performance of individual Ad Hoc runs; (ii) of the correlation between Team-TA interaction and effectiveness; (iii) of the impact of the adjudication process on assessments; and (iv) of the effect of the state of the OCR in the test collection.

4.5.1 Individual Ad Hoc Runs

To this point in our discussion, we have included reference to the Ad Hoc submissions, but only as an aggregated reference set. It is also possible to assess the effectiveness of individual Ad Hoc runs (as well as the Boolean reference runs); in this section, we take a look at some of the more interesting of these results.

Table 16 summarizes the effectiveness (as measured by recall, precision, and F_1) attained by select Ad Hoc and Boolean runs. The table does not provide results for all 64 ad hoc runs; these can be found in the Appendix to the Proceedings. Instead, the table focuses on the following runs:

Topic	Team	Recall		Precision		F_1	
		Est.	**95% C.I.**	**Est.**	**95% C.I.**	**Est.**	**95% C.I.**
Topic 102	Ad Hoc Pool	0.314	(0.266, 0.362)	0.328	(0.301, 0.355)	0.321	(0.293, 0.349)
	Clearwell	0.016	(0.014, 0.018)	0.652	(0.629, 0.674)	0.031	(0.027, 0.035)
	Pittsburgh	0.007	(0.006, 0.008)	0.866	(0.836, 0.896)	0.014	(0.012, 0.015)
Topic 103	H5	0.624	(0.579, 0.668)	0.810	(0.795, 0.824)	0.705	(0.676, 0.734)
	Ad Hoc Pool	0.403	(0.371, 0.434)	0.382	(0.368, 0.396)	0.392	(0.375, 0.408)
	Clearwell	0.158	(0.146, 0.169)	0.711	(0.692, 0.730)	0.258	(0.243, 0.274)
	Buffalo	0.061	(0.056, 0.066)	0.716	(0.689, 0.743)	0.113	(0.105, 0.121)
	Pittsburgh	0.026	(0.024, 0.029)	0.804	(0.763, 0.844)	0.051	(0.047, 0.055)
Topic 104	Ad Hoc Pool	0.345	(0.111, 0.580)	0.023	(0.014, 0.032)	0.043	(0.026, 0.060)
	Clearwell	0.003	(0.001, 0.004)	0.234	(0.198, 0.269)	0.006	(0.002, 0.009)

Table 15: Post-Adjudication Metrics.

- the highest scoring (as measured by F_1) of the individual Ad Hoc runs ("High Individual Ad Hoc");
- Defendant's proposed Boolean query ("Boolean—Defendant");
- Plaintiff's proposed Boolean query ("Boolean—Plaintiff");
- the final negotiated Boolean query ("Boolean—Final"); and
- the full test collection ("Full Collection").

For each run, the table presents the total number of documents retrieved by the run and the estimated recall, precision, and F_1 attained by that set of retrieved documents. Scores are based on post-adjudication assessments.

The data in the table occasion a few observations. With regard to Topic 102, recall that, for this topic, the set that, in our earlier analysis, proved most effective was the Ad Hoc Pool (attaining an F_1 score of 0.321), with the two active participants in the topic scoring somewhat lower. We see from Table 16 that a few of the individual runs now under consideration (High Individual Ad Hoc (in this case, Open Text's *otL08frw*), Boolean—Plaintiff, and Boolean—Final) retrieved sets that attained F_1 scores higher than those attained by the active participants' submissions (but still lower than the score attained by the Ad Hoc Pool); and the High Individual Ad Hoc could well have achieved a higher score, had it not been subject to the 100,000-document submission constraint. In this instance, it appears that the active participants were not able to take full advantage of the opportunity to engage with the Topic Authority as a means of improving their effectiveness (and it should be noted that they utilized only a small portion of the time available to them for this purpose).

With regard to Topic 103, recall that the set that, in our earlier analysis, proved most effective was that turned in by one of the active participants (H5), a set that attained an F_1 score of 0.705. None of the individual runs currently under consideration were able to approach that level of effectiveness, the closest being Boolean—Plaintiff which attained an F_1 score of 0.326. As can also be seen from the table, for Topic 103, as for Topic 102, a rather sizeable F_1 score can be attained simply by submitting the entire collection (0.209 for Topic 103; 0.153 for Topic 102); the high yields of these topics make for a high "floor" for precision (and thereby for F_1) for a full-collection submission.

With regard to Topic 104, recall that, in our earlier analysis, it was the Ad Hoc Pool that scored highest on the F_1 metric (0.043), but that neither set that was evaluated achieved particularly high scores. Of the runs reviewed in Table 16, one, the High Individual Ad Hoc (in this case, Waterloo's *wat7fuse*), was able to achieve a higher F_1 (0.061) than that attained by the Ad Hoc Pool. For this topic, however, all runs,

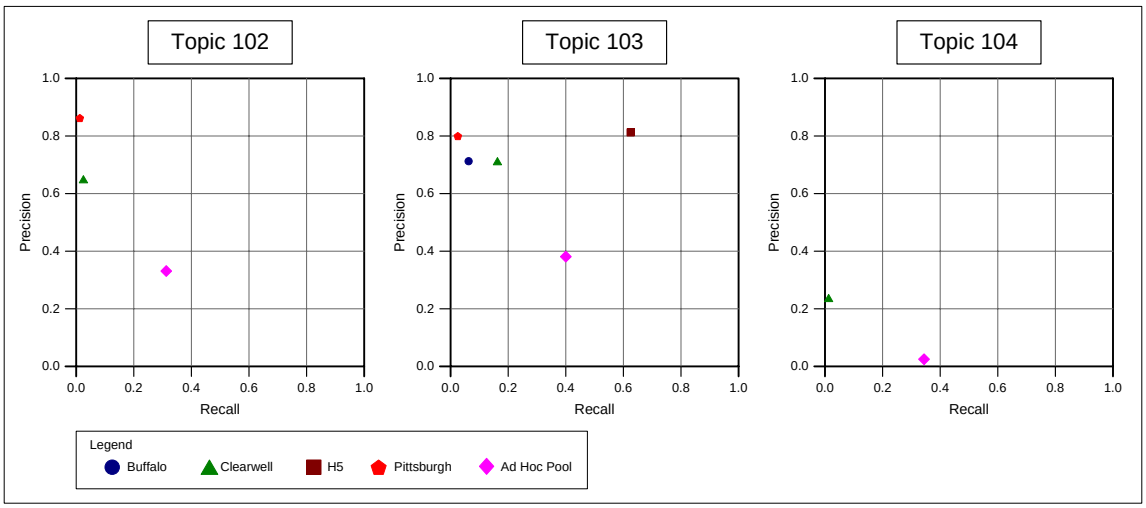

Figure 3: Recall & Precision — Post-Adjudication.

including the High Individual Ad Hoc just mentioned, were found to have attained very low scores, indicating that the participants' conceptions of relevance were not well aligned with the Topic Authority's conception of what should be retrieved for the topic.

4.5.2 Team-TA Interaction

Earlier (Section 4.4.1), 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 five minutes in one instance to 485 minutes in another. Such variation naturally prompts the question of whether there is a correlation between the amount of time spent with a Topic Authority and retrieval effectiveness.

Figure 4 plots retrieval effectiveness (as measured by post-adjudication F_1 scores) against time spent with the Topic Authority on the topic-clarification portion of the task. Results for all participating teams and topics are represented on the diagram; we also include the results we get for the Ad Hoc Pool (which, of course, represents the aggregated results of 68 runs, none of which made use of Topic Authority time).

From the diagram, we see that, while there are not a large number of data points (10, if we include the results for the Ad Hoc Pool), and while there are no data points that correspond to the "middle" segment of the time range, there does appear to be a correlation between effectiveness and time spent with the Topic Authority. Submissions that resulted in low F_1 scores tend to have come from approaches that made little use of the Topic Authority's time; the team that made the most use of the Topic Authority's time achieved a very high F_1 score.

The impression is borne out by correlation measures. Looking just at the results turned in by the active participants in the task (i.e., setting aside the results of the Ad Hoc Pool), we obtain a Pearson product-moment correlation coefficient of 0.927 with a 95% confidence interval of (0.577, 0.989). Even including the results of the Ad Hoc Pool, we find evidence of a positive correlation: $r = 0.699$ (0.124, 0.923).

If there is a correlation between retrieval effectiveness and time spent interacting with the Topic Authority, the next question is whether there are some modes of interaction that are more effective than others. After all, it would be surprising if effectiveness were simply a function of time spent with the Topic Authority, regardless of how that time was used; we would expect that there are some approaches to gathering the required information that work better than others. This is a question that merits further study. The participants' papers on their approaches to this year's task may shed some light on the question; the question

Topic	Run	Retrieved	R	P	F_1
Topic 102	High Individual Ad Hoc	100,000	0.121	0.710	0.207
	Boolean—Defendant	980	0.002	0.930	0.003
	Boolean—Plaintiff	113,796	0.113	0.640	0.192
	Boolean—Final	86,742	0.099	0.693	0.173
	Full Collection	6,910,192	1.000	0.083	0.153
Topic 103	High Individual Ad Hoc	100,000	0.103	0.762	0.181
	Boolean—Defendant	35,290	0.030	0.634	0.058
	Boolean—Plaintiff	280,383	0.218	0.649	0.326
	Boolean—Final	80,225	0.074	0.731	0.135
	Full Collection	6,910,192	1.000	0.117	0.209
Topic 104	High Individual Ad Hoc	100,000	0.096	0.045	0.061
	Boolean—Defendant	16	0.000	0.571	0.000
	Boolean—Plaintiff	2,682	0.002	0.085	0.004
	Boolean—Final	2,680	0.002	0.085	0.004
	Full Collection	6,910,192	1.000	0.007	0.013

Table 16: Individual Ad Hoc and Boolean Runs.

should also be borne in mind as we look ahead to the 2009 running of the task.

The data we have been considering suggests that making effective use of the Topic Authority's time (or, translating to the real-world task being modeled, the time of the senior litigator responsible for a document production) is useful for retrieving the set of documents the Topic Authority wants retrieved. The importance of this interaction is also underlined by the 2008 Topic Authorities themselves, who, in their reflections on the 2008 Interactive Task, note the following [7].

> The successful outcome of an information retrieval task is highly dependent on the amount of time — and the quality use of the time — spent with the person or persons tasked with the ultimate responsibility for defining relevance. It is not possible to replicate subjective judgment calls without spending time with the subjects who are ultimately responsible for making those determinations.

A good lesson for teams looking ahead to 2009.

4.5.3 Assessment & Adjudication

As noted above (Section 4.4.3), of the 13,339 sampled documents (aggregating across all three topics) that were found to be assessable, and so received a first-pass Relevant or Non-Relevant assessment, 966 were appealed; of the 966 first-pass assessments that were appealed, nearly 80% were in fact overturned by the Topic Authority. Changes in assessments on that order obviously can have a substantial impact on results. In this section, we take a closer look at some of the effects of the appeal and adjudication process.

Distribution across topics. We note, first, that the appeals were not evenly distributed across the three topics. There were many more appeals of Topic 103 assessments than there were of either Topic 102 or Topic 104 assessments. Now, Topic 103 had the greatest number of participants and had the largest evaluation sample; even allowing for such considerations, however, it remains true that the appeal/adjudication mechanism was utilized more heavily for Topic 103 than it was for the other two topics, and it is on the results for Topic 103 that we would expect the mechanism to have the strongest effect.

Impact on F_1 scores. As expected, the appeal and adjudication mechanism had little effect on the F_1 scores of participants in Topics 102 or 104. For these topics, the greatest Pre- to Post-Adjudication

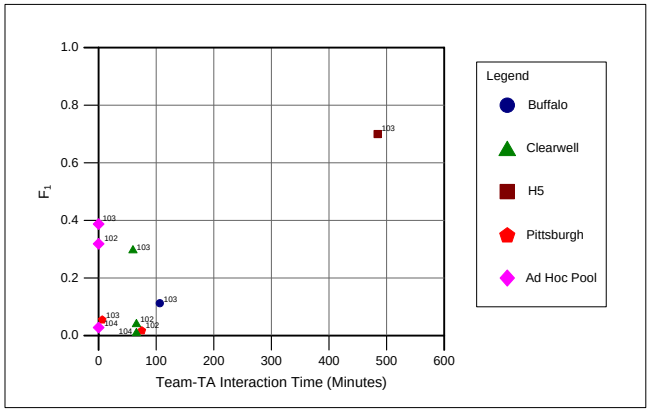

Figure 4: F_1 *vs.* Team-TA Interaction Time.

difference was a 0.4% drop in the F_1 score reported for the Ad Hoc Pool; the appeals for these topics were simply too few to have a large impact. The appeal and adjudication mechanism did have an effect on the F_1 scores of participants in Topic 103. All active participants (and the Ad Hoc Pool) saw an improvement in their F_1 scores as a result of the appeals process. For the active participants, the Pre- to Post-Adjudication differences ranged, in absolute terms, from 0.013 (Pittsburgh) to 0.277 (H5). Expressed as a proportion of their Pre-Adjudication scores, however, the differences were more equal; all active participants saw their F_1 scores improve by 28% to 38% relative to their Pre-Adjudication values. The Ad Hoc Pool, on the other hand, realized a smaller gain, a 6% gain on its Pre-Adjudication value.

Impact on ranking of results. For no topic did the results of the appeal and adjudication process affect the ranking of submissions on F_1. Even for Topic 103, where, as we have just seen, the appeal and adjudication process had an impact on absolute F_1 values, the Pre- to Post-Adjudication changes in scores did not affect the relative ordering of participants on the metric.

4.5.4 Generalizability of Results — the Impact of OCR Quality

The Legal Track's Interactive Task, like all TREC tasks, is intended to serve a number of purposes. It is a research exercise, intended to advance our understanding both of the increasingly challenging text retrieval needs of the legal community and of the retrieval methods and capabilities best suited to meet those needs. It is a forum for the development of task and evaluation designs that will answer the questions the research, legal, and vendor communities have about the application of search technologies to e-discovery problems. It is a resource to which the legal community can turn to gain a better understanding of the sorts of approaches that are most likely to prove effective at performing the e-discovery tasks they most urgently need assistance with.

With an eye to the latter purpose (serving as a resource for attorneys looking to evaluate approaches), we have tried to make the Interactive Task as realistic as possible. While some compromise on realism is always going to be required if a task is to enable valid approach-to-approach comparisons, we have endeavored, in this year's Interactive Task, to model as closely as possible the conditions and objectives of real-world responsive review. There is, however, one component of the task that is decidedly anomalous to what one could expect in a typical responsive review of 2009: the document collection, or, more specifically, the quality of the OCRing of the document collection.

The limited accuracy of the OCR text in the IIT CDIP test collection has been a subject of conversation since we started using the collection in the first year of the Legal Track. The document text that, together with the document metadata, teams rely on in conducting their retrieval efforts, is the result of an OCR process applied to images of the original documents; the output of that process is often faulty (i.e., a

garbled rendering of the content of the original document), either because the source image is not susceptible to accurate OCRing (e.g., is characterized by poor resolution, small type, or predominantly handwritten content) or because the OCR process that was used was less accurate than might now be achievable. This aspect of the collection is atypical of what would be expected in most current collections that are likely to be the focus of litigation, in that the latter would be sourced almost entirely from born-digital files and so would have minimal OCR issues.

In order to enable the results of the task to be more generalizable to conditions typically encountered in the real world, we would like to be able to take into account the possible impact of the state of the OCR in the test collection. While it would be beyond the scope of this overview to address all the questions that would have to be addressed if we were to attempt to arrive at a precise quantification of the OCR effect, we provide the following analysis as a simple but informative gauge of the possible impact on performance of the state of the OCR.

OCR Score. To begin with, we need some method of objectively quantifying the quality of the OCR. While a number of measures are discussed in the literature, the non-stopword accuracy measure advocated by researchers at the Information Science Research Institute (ISRI) at University of Nevada, Las Vegas [12] is a reasonably straightforward measure that will suit our purposes. Under this approach, a document's OCR score is, generally stated, the proportion of correctly rendered lexical words out of the sum of correctly and incorrectly rendered lexical words. The higher the OCR score (the closer to 1.0) the better the quality of the OCRing; the lower the OCR score (the closer to 0.0) the worse the quality of the OCRing.

More specifically, for purposes of the calculation, a correctly rendered word is a token that is found in a reference dictionary; an incorrectly-rendered word is a token that is not found in the dictionary. Certain types of tokens are ignored for the purposes of the calculation (e.g., stopwords, proper names), making the score a function primarily of lexical words likely to be found in a dictionary.

The specific token types we exclude from the calculation are the following:

- any token likely to be a proper noun or acronym;
- any token likely to be a stopword;
- any token of 3 characters or less; and
- any token that contains numerics;

These exclusions are the same as those excluded by the ISRI team. The one addition we made was the exclusion of numeric strings; the accuracy of the rendering of these (which occur with great frequency in the test collection) cannot be tested by reference to a dictionary, so they were excluded from the calculation. The reference dictionary we used was the single-word component of the Moby Words wordlist, a list of 354,984 English single words that includes archaic words and variant spellings [2]. Counted as a token likely to be a proper noun or acronym was any string beginning with an upper case alphabetic. Counted as a stopword was any word on the list of 319 stopwords made available by the Information Retrieval Group at the University of Glasgow [1].

Retrieved vs. Not-Retrieved Documents. The intuition that prompts a closer look at the effect of OCR is that teams will find it more difficult to retrieve documents with poorly rendered text than they will documents with accurately rendered text; to be sure, in some cases of poor quality OCR, metadata values can be drawn on, but, given the subtlety and complexity of the target topics, metadata alone will, in the majority of cases, not be able to compensate for the inaccurate rendering of the document text. Because the manual assessors who review the evaluation samples base their assessments on the source images of the documents, rather than on the OCR text, it would not be surprising if there were some number of documents that the assessors found relevant but, due to the poor state of the OCR, the teams were, collectively, unable to retrieve.

With our evaluation samples and our rough-and-ready OCR score at hand, it is fairly easy to test whether the intuition is borne out by the data. Using the evaluation samples, we compare, for each topic, the mean OCR score of relevant documents retrieved by at least one team to the mean OCR score of relevant documents

not retrieved by any team. If the motivating intuition is correct, we should find that the latter score is lower than the former. Table 17 presents the results.

Collection Subset	**Topic 102**	**Topic 103**	**Topic 104**
Strata other than All-N Stratum	0.868	0.730	0.857
All-N Stratum	0.617	0.462	0.533

Table 17: Mean OCR Scores — Relevant Documents.

As can be seen from the table, the data corroborate our intuition. The mean OCR scores are evidence that, for all three topics, the OCR quality of relevant documents that have been found by at least one team tends to be better than the OCR quality of relevant documents that have been missed by all teams.

OCR-Adjusted Scores. If, on the one hand, the OCR effect is real (i.e., affecting the likelihood with which a document will be successfully retrieved), and if, on the other hand, the collections of documents that typically figure in current litigation are, unlike the test collection, largely free from OCR issues, we would like to have some way to control for the OCR effect, so that those seeking to generalize from task conditions and results to real-world conditions and results would be better equipped to do so.

The approach we take is reasonably straightforward: we confine our view to those parts of the collection characterized by more accurate OCRing (as indicated by OCR scores) and see how participants performed on just those parts of the collection. More specifically, what we do is select certain threshold values for the OCR-accuracy score, then, relying on our already-adjudicated samples, obtain estimates of the recall, precision, and F_1 achieved by each participant on documents at or above those threshold values.

In selecting thresholds at which to obtain these adjusted metrics, we begin with an OCR-accuracy threshold of ≥ 0.95 (the minimum level of non-stopword accuracy deemed acceptable for the IR application that is the focus of the ISRI study noted above), then drop down to three additional lower-accuracy thresholds: ≥ 0.85, ≥ 0.75, and ≥ 0.50. The estimated proportions of the full collection (or manually-assessable part of it; see Section 4.2.2 above) and proportions of actually relevant documents included at each threshold are summarized in Table 18.

		OCR Threshold				
Topic	**Attribute**	$\geq$ **0.95**	$\geq$ **0.85**	$\geq$ **0.75**	$\geq$ **0.50**	$\geq$ **0.00**
Topic 102	Assessable Documents	1,236,267	2,913,871	3,661,652	4,518,658	6,768,452
	% of Total Assessable	18.3%	43.1%	54.1%	66.8%	100.0%
	Relevant Documents	157,004	309,492	375,099	423,127	562,402
	% of Total Relevant	27.9%	55.0%	66.7%	75.2%	100.0%
Topic 103	Assessable Documents	1,293,287	3,045,992	3,796,267	4,598,831	6,739,015
	% of Total Assessable	19.2%	45.2%	56.3%	68.2%	100.0%
	Relevant Documents	200,330	387,039	451,655	562,720	786,862
	% of Total Relevant	25.5%	49.2%	57.4%	71.5%	100.0%
Topic 104	Assessable Documents	1,318,793	2,934,061	3,677,679	4,608,247	6,865,404
	% of Total Assessable	19.2%	42.7%	53.6%	67.1%	100.0%
	Relevant Documents	12,846	21,426	22,144	34,234	45,614
	% of Total Relevant	28.2%	47.0%	48.5%	75.1%	100.0%

Table 18: Thresholding the Collection Based on OCR Scores.

From Table 18, we see that, although the collection, as a whole, is characterized by poor quality OCR,

there is a sizeable subset that is characterized by more accurate rendering of the source text: nearly 20% of the collection is estimated to have an OCR-accuracy score of 0.95 or better and nearly 45% (nearly 3 million documents) a score of 0.85 or better; there is some topic-to-topic variation in the proportions due to that fact that these are sample-based estimates of total assessable documents and the fact that different manual assessors may have had slightly different interpretations of the assessability criteria. We also see that, as with assessable documents in general, so with relevant documents in particular, a substantial proportion are included in the subsets characterized by high-quality OCR: generalizing across topics, about 25% of relevant documents have an OCR score of 0.95 or better and 50% a score of 0.85 or better.

Table 19 summarizes, for each topic, the estimates of recall, precision, and F_1 we obtain for each participant at each OCR-accuracy threshold; of course, if we dropped the threshold all the way to 0.00, we would obtain the unadjusted post-adjudication metrics already reported (Table 15).

		Threshold = 0.95			**Threshold = 0.85**			**Threshold = 0.75**			**Threshold = 0.50**		
Topic	**Team**	**R**	**P**	F_1	**R**	**P**	F_1	**R**	**P**	F_1	**R**	**P**	F_1
102	AH	**0.513**	0.411	0.457	**0.447**	0.388	0.415	**0.408**	0.373	0.390	**0.390**	0.355	0.372
	CS	**0.036**	0.686	0.069	**0.026**	0.667	0.050	**0.022**	0.661	0.043	**0.021**	0.657	0.040
	UP	**0.017**	0.885	0.033	**0.011**	0.878	0.022	**0.010**	0.870	0.019	**0.009**	0.869	0.018
103	H5	**0.772**	0.826	0.798	**0.786**	0.830	0.808	**0.776**	0.833	0.803	**0.699**	0.829	0.759
	AH	**0.517**	0.398	0.450	**0.528**	0.422	0.469	**0.520**	0.416	0.463	**0.470**	0.396	0.430
	CS	**0.240**	0.692	0.356	**0.236**	0.733	0.358	**0.222**	0.724	0.339	**0.192**	0.724	0.304
	UB	**0.097**	0.696	0.170	**0.085**	0.697	0.152	**0.081**	0.697	0.145	**0.073**	0.694	0.132
	UP	**0.042**	0.779	0.080	**0.042**	0.823	0.080	**0.039**	0.819	0.075	**0.033**	0.810	0.063
104	AH	**0.613**	0.032	0.061	**0.535**	0.027	0.051	**0.551**	0.024	0.047	**0.419**	0.025	0.047
	CS	**0.004**	0.246	0.009	**0.005**	0.254	0.010	**0.005**	0.248	0.010	**0.004**	0.245	0.007

Table 19: OCR-Adjusted Metrics.

From the table, we see that precision is for the most part not affected by the quality of the OCR; increasing the number of poorly-OCRed documents in the test set does not appear to result in a higher proportion of false positives. Recall, on the other hand, is affected: as you include more poorly OCRed documents (and those of increasingly poor quality) into the set on which you measure performance, you find that recall tends to decrease, for all participants and all topics (and continues to drop, down to the unadjusted post-adjudication numbers). Looked at another way, if you confine your attention to the part of the collection characterized by higher-quality OCRing (that part most like a collection likely to be targeted in contemporary litigation), you find that all participants are estimated to have achieved higher levels of recall than they are estimated to have achieved when the poorly OCRed documents are included in the test set; indeed, for the team that had the highest unadjusted post-adjudication recall (H5), the OCR-based adjustments point to the achievement of recall in the neighborhood of 80%. We see, finally, that the ordering of results, in terms of F_1, is unaffected by the thresholding; the changes to collection characteristics brought about by thresholding on OCR scores result in across-the-board upward adjustments to recall, so the relative rankings based on F_1 do not change.

In this section, we have taken a brief look at the possible effect of the state of the OCR in the test collection on participants' results. Our analysis has been a deliberately rough-and-ready one, and it could certainly be followed up with additional research and analysis. We could, for example, look more narrowly at specific OCR intervals and see whether there is a particular point at which the state of the OCR begins to have a strong negative effect on recall (and whether that point differs for different approaches). For the purpose of this section, however, which is to provide additional perspective for those seeking to generalize from task conditions and results to real-world conditions and results, the analysis suffices to provide evidence (a) that the state of the OCR in the test collection has a real (depressing) effect on recall and (b) that, if

you adjust for that effect, you will see increases, sometimes substantial, in the levels of recall reported for all participants. We conclude the section with a visual; Figure 5 plots each team's adjusted scores (assuming an OCR-accuracy threshold of 0.85) on precision-recall diagrams.

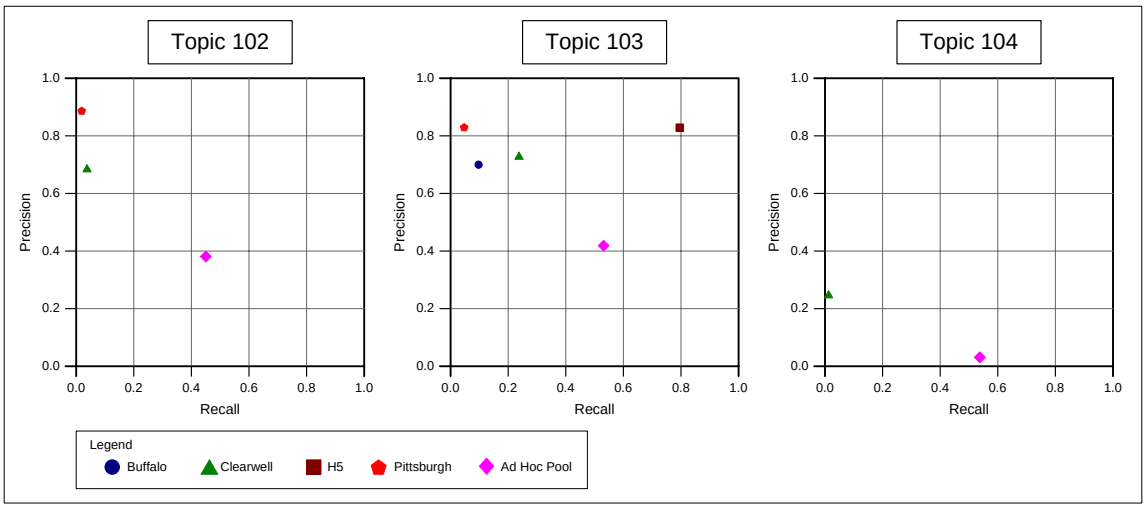

Figure 5: Recall & Precision — OCR-Adjusted (OCR-Accuracy Threshold = 0.85).

4.6 Conclusions from the 2008 Interactive Task

We wrap up the paper with some general observations on the 2008 Legal Track (Section 5). Before turning to that conclusion, however, we summarize a few of the key lessons learned from this year's Interactive Task.

- The Interactive Task is intended to model the conditions and objectives of real-world responsive review as accurately as possible. To that end, we introduced a number of new elements to the design of the task, the most important of these being the role of the Topic Authority. While there are always challenges to implementing a new design, and this task was no exception, the design fostered needed communication between the attorneys and researchers who participated in the task and yielded interesting and meaningful results.
- If one engages with that authority effectively, one can bring about the conditions for effective document retrieval; the one participant in this year's exercise who took full advantage of the opportunity to interact with their designated Topic Authority was able to achieve, simultaneously, high recall and high precision (in the neighborhood of 80% on both metrics, if you adjust for the state of the OCR in the test collection).
- This year's exercise taught us (track coordinators, topic authorities, participating teams, and others who contributed their support) a great deal, lessons that we will build on in what we expect to be a still more productive running of the task in 2009.

5 Conclusion

In the three years of the TREC Legal Track we have constructed a unique test collection that we expect to have enduring value. The collection contains OCR text and metadata for nearly 7 million documents (with the corresponding document images also available) and relevance assessments for a total of 109 richly

structured topics. Moreover, we have iterated to what we believe to be relatively stable evaluation designs for the Ad Hoc, Relevance Feedback, and Interactive tasks. A community of 22 research teams from 21 institutions in seven countries that did not exist three years ago now has experience with one or more of those tasks, and a broad range of results have been reported.

This year's track yielded several important results. Shifting the principal evaluation measure in the Ad Hoc task from recall to F_1 and shifting the reference point from one specified by us (B) to one specified by each participating team (K) resulted in greater task fidelity, modeling the real operational requirement of e-discovery systems to return a set of documents, not just an unbounded ranked list. Five of the 10 participating research teams submitted a run of higher mean $F_1@K$ than the reference Boolean run, an encouraging result given that, in the previous year (2007), no team had outscored the reference Boolean run in that year's main measure (mean Recall@B).

However, we cannot say that the additional flexibility of setting K was the reason for the automated approaches outperforming the negotiated Boolean queries in 2008 because 7 of the 10 participating research teams actually submitted a run of higher mean Recall@B than the reference Boolean run in 2008. Furthermore, the highest mean Recall@B was from a run that used the same automated approach as a run from 2007 which scored lower in Recall@B than the reference Boolean run of 2007. More analysis is needed to understand this result. A possible contributing factor is that, compared to the previous year, this year's topics were found to average five times as many relevant documents (i.e., they were "broader"). The result sets for the final Boolean queries averaged eight times larger than last year, which is also consistent with broader topics. Another possibility is simply natural variation in human performance; the Boolean negotiations that we use as a reference are themselves a human activity that naturally can vary somewhat from session to session.

A second major change this year was the introduction of a "highly relevant" category for relevance assessment, modeling the legal concept of evidence being "material" rather than merely relevant. It is not yet clear how useful this new category will be, in part because we have yet to study inter-annotator agreement for this assessment task. Nonetheless, we found that the consensus Boolean query found 42% of the highly relevant documents, on average per topic, which is better coverage than its recall of all relevant documents (33% this year). However, this result still implies that, on average per topic, 58% of the "highly relevant" documents were not found by the consensus Boolean query, indicating that it is not just tangentially relevant documents that are being missed by the negotiated Boolean approach.

The Relevance Feedback task attracted fewer participants than we had hoped this year. The participants' feedback runs did not generally outperform the baseline runs, which may reflect the difficulty of automatically making use of feedback in the presence of OCR errors. An intriguing observation from this year's task, which re-used topics from previous years, is that the estimated numbers of relevant documents for a topic often differed dramatically depending on which year's assessments were used. This suggests that we should devote some attention to characterizing the extent to which differing in assessor opinion may be a confounding factor when comparing results obtained through sampling and estimation.

The completely redesigned Interactive task attracted research teams from two companies and two universities, a number sufficient to wring out the new evaluation design. No participant undertook a completely manual review (although that would be permissible). The approaches used in this task varied widely, as did the time and resources invested, and the number of topics completed (which varied between 1 and 3). For the one topic completed by all participants, the submitted result sets all had similar precision (71%-81%), but recall varied substantially (from less than 3% to more than 60%). Two of the four submissions substantially outperformed the consensus Boolean query on this topic (which had a precision of 76% and recall of 13%). Relatively rich sampling was done for the topics of this task (up to 6,500 assessments for one topic). We expect that the far richer sampling for these topics will serve as a useful reference point when designing cost-effective sampling strategies in the future. The participants also had the opportunity to appeal the original assessments in this task, further increasing the quality of this resource. As we approach the limit of what relevance assessment by volunteers can support, cost-effective sampling is a matter of increasing urgency.

Although there is surely more to be learned by continuing to work with the IIT CDIP v1.0 test collection,

we are nearing the point of diminishing returns beyond which further investment in this one collection may no longer yield new insights with importance that is in line with the costs to participants and assessors. That is not to say that further work with this test collection would not be justified. Quite to the contrary, a lot of work remains to be done. But the test collection needed to support that work is now for the most part in place, and for that reason we are now considering turning our attention to other collections with new characteristics that would help to extend our investigation of this important problem space.

Acknowledgments

This track would not have been possible without the efforts of a great many people. Our heartfelt thanks go to Ian Soboroff for creating the relevance assessment system; to the dedicated group of pro bono relevance assessors and the pro bono coordinators at the participating law schools; once again to Conor Crowley (Daley Crowley LLP), Joe Looby and Ryan Bilbrey (FTI Consulting), and the team from H5 (Todd Elmer, Jim Donahue, Misti Gerber, and others) for their invaluable assistance with the Ad Hoc task (complaint drafting, topic formulation, and participating in Boolean negotiations); also to Conor Crowley, Maura Grossman, and Joe Looby for their service in the role as Topic Authorities; to Julie Hoff and Dana Novak (Redgrave Daley Ragan & Wagner LLP) for valuable support in the sample assessment phase of the project; and finally, to Richard Braman, Executive Director of The Sedona Conference®, for his continued support of the TREC Legal Track.

A Estimation of Metrics — Interactive Task

As described more fully above (Section 4.2), evaluation in Interactive Task follows a four-step protocol, whereby (i) stratified samples are drawn for each of the target topics, (ii) a first-pass manual review is conducted of each of the samples, (iii) those first-pass assessments are, if a participating team chooses, appealed and adjudicated, and (iv) estimates of each team's performance, as measured by recall, precision, and F_1, are obtained. In this appendix, we review the specifics of the last step, the procedures whereby estimates of our target metrics are obtained.

Arriving at estimates of the target metrics is itself a two-step process. The first step is to obtain estimates of three input components: (i) total documents both assessable and relevant in the full collection; (ii) total documents assessable in the set submitted by the team whose run is being evaluated; and (iii) total documents both assessable and relevant in the set submitted by the team whose run is being evaluated. The second step is to combine the elements in the appropriate way to obtain estimates of Recall, Precision, and F_1. The specifics are as follows (most of the formulae cited below will be found in any standard discussion of stratified sampling, such as that in Thompson (2002) [14]).

Notation. We begin with some notation.

N = total number of documents in the full collection.

τ_{a} = total number of assessable documents in the full collection.

τ_{r} = total number of assessable and relevant documents in the full collection.

L = total number of strata into which the full collection has been partitioned.

N_h = total number of documents in stratum h ($h = 1, 2, \ldots, L$).

$\tau_{\mathrm{a}(h)}$ = total number of assessable documents in stratum h.

$\tau_{\mathrm{r}(h)}$ = total number of assessable and relevant documents in stratum h.

$p_{\mathrm{a}(h)}$ = proportion of documents in stratum h that are assessable.

$p_{\mathrm{r}(h)}$ = proportion of documents in stratum h that are assessable and relevant.

n_h = total number of documents sampled from stratum h.

a_h = of documents sampled from stratum h, total number assessable.

r_h = of documents sampled from stratum h, total number assessable and relevant.

$L_{(\mathrm{A})}$ = total number of strata into which the documents Team A submitted as relevant were partitioned.

$\tau_{\mathrm{a(A)}}$ = of documents Team A submitted as relevant, total number assessable.

$\tau_{\mathrm{r(A)}}$ = of documents Team A submitted as relevant, total number assessable and relevant.

Estimation of inputs. Now, we review the procedures used to obtain estimates of the inputs to the metrics. As noted above, there are three inputs we require: (i) full-collection estimate of total assessable and relevant; (ii) estimate of total assessable in a team's submission; and (iii) estimate of total assessable and relevant in a team's submission.

Component 1: Full-Collection Estimate of Total Assessable and Relevant. We obtain the first component by finding the stratified estimate of the total documents both assessable and relevant in the full population. To obtain this estimate, we first find within-stratum estimates then find full-collection estimates.

Within-stratum estimates are obtained as follows.

1. Obtain estimate of within-stratum proportion of documents both assessable and relevant.

$$\hat{p}_{\mathrm{r}(h)} = \frac{r_h}{n_h} \tag{2}$$

2. Obtain estimate of within-stratum total number of documents both assessable and relevant.

$$\hat{\tau}_{\mathrm{r}(h)} = N_h \hat{p}_{\mathrm{r}(h)} \tag{3}$$

3. Obtain estimate of within-stratum sample variance.

$$s_h^2 = \left(\frac{n_h}{n_h - 1} \right) \hat{p}_{\mathrm{r}(h)} \left(1 - \hat{p}_{\mathrm{r}(h)} \right) \tag{4}$$

4. Obtain estimate of variance of within-stratum total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{r}(h)}) = N_h \left(N_h - n_h \right) \frac{s_h^2}{n_h} \tag{5}$$

Full-collection estimates are obtained as follows.

1. Obtain estimate of full-collection total number of documents both assessable and relevant.

$$\hat{\tau}_{\mathrm{r}} = \sum_{h=1}^{L} \hat{\tau}_{\mathrm{r}(h)} \tag{6}$$

2. Obtain estimate of variance of full-collection total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{r}}) = \sum_{h=1}^{L} \widehat{var}(\hat{\tau}_{\mathrm{r}(h)}) \tag{7}$$

Component 2: Estimate of Total Assessable in a Team's Submission. We obtain the second component by finding the stratified estimate of the total documents assessable in the part of the population that a team identified as relevant. Obtaining this estimate is again a matter of finding the value of a stratified estimator; in this case, however, the strata that figure in the estimate are just those that contain the team's positive assessments. More specifically, the steps (for a team we'll call Team A) are the following.

Within-stratum estimates are obtained as follows.

1. Obtain estimate of within-stratum proportion of documents that are assessable.

$$\hat{p}_{\mathrm{a}(h)} = \frac{a_h}{n_h} \tag{8}$$

2. Obtain estimate of within-stratum total number of documents that are assessable.

$$\hat{\tau}_{\mathrm{a}(h)} = N_h \hat{p}_{\mathrm{a}(h)} \tag{9}$$

3. Obtain estimate of within-stratum sample variance.

$$s_h^2 = \left(\frac{n_h}{n_h - 1} \right) \hat{p}_{\mathrm{a}(h)} \left(1 - \hat{p}_{\mathrm{a}(h)} \right) \tag{10}$$

4. Obtain estimate of variance of within-stratum total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{a}(h)}) = N_h \left(N_h - n_h \right) \frac{s_h^2}{n_h} \tag{11}$$

Full-submission estimates are obtained as follows.

1. Obtain estimate of full-submission total number of documents that are assessable.

$$\hat{\tau}_{\mathrm{a(A)}} = \sum_{h=1}^{L_{(\mathrm{A})}} \hat{\tau}_{\mathrm{a}(h)} \tag{12}$$

2. Obtain estimate of variance of full-submission total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{a(A)}}) = \sum_{h=1}^{L_{(\mathrm{A})}} \widehat{var}(\hat{\tau}_{\mathrm{a}(h)}) \tag{13}$$

Component 3: Estimate of Total Assessable and Relevant in a Team's Submission. We obtain the third component by finding the stratified estimate of the total documents assessable and relevant in the part of the population that a team identified as relevant. Obtaining this estimate is again a matter of finding the value of a stratified estimator; as with the second component, the strata that figure in the estimate are just those that contain the team's positive assessments. More specifically, the steps (for "Team A") are the following.

Within-stratum estimates are obtained as follows.

1. Obtain estimate of within-stratum proportion of documents that are both assessable and relevant.

$$\hat{p}_{\mathrm{r}(h)} = \frac{r_h}{n_h} \tag{14}$$

2. Obtain estimate of within-stratum total number of documents that are both assessable and relevant.

$$\hat{\tau}_{\mathrm{r}(h)} = N_h \hat{p}_{\mathrm{r}(h)} \tag{15}$$

3. Obtain estimate of within-stratum sample variance.

$$s_h^2 = \left(\frac{n_h}{n_h - 1}\right) \hat{p}_{\mathrm{r}(h)} \left(1 - \hat{p}_{\mathrm{r}(h)}\right) \tag{16}$$

4. Obtain estimate of variance of within-stratum total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{r}(h)}) = N_h \left(N_h - n_h\right) \frac{s_h^2}{n_h} \tag{17}$$

Full-submission estimates are obtained as follows.

1. Obtain estimate of full-submission total number of documents that are both assessable and relevant.

$$\hat{\tau}_{\mathrm{r(A)}} = \sum_{h=1}^{L_{(\mathrm{A})}} \hat{\tau}_{\mathrm{r}(h)} \tag{18}$$

2. Obtain estimate of variance of full-submission total estimator.

$$\widehat{var}(\hat{\tau}_{\mathrm{r(A)}}) = \sum_{h=1}^{L_{(\mathrm{A})}} \widehat{var}(\hat{\tau}_{\mathrm{r}(h)}) \tag{19}$$

Estimation of metrics. Finally, we review the procedures used to obtain estimates of the metrics themselves. With estimates of the three inputs in hand, we combine the elements to obtain estimates of Recall, Precision, and F_1. Variances for the estimates of the metrics are obtained by propagating the variances of their component elements (in accordance with the principles of Gaussian Error Propagation).

Recall. We obtain estimates and 95% confidence intervals for recall as follows.

1. Obtain estimate of recall (for "Team A").

$$\hat{R}_{(\mathrm{A})} = \frac{\hat{\tau}_{\mathrm{r(A)}}}{\hat{\tau}_{\mathrm{r}}} \tag{20}$$

2. Obtain estimate of variance for recall.

$$\widehat{var}(\hat{R}_{(\mathrm{A})}) = \hat{R}^2_{(\mathrm{A})} \left(\frac{\widehat{var}(\hat{\tau}_{\mathrm{r(A)}})}{\hat{\tau}^2_{\mathrm{r(A)}}} + \frac{\widehat{var}(\hat{\tau}_{\mathrm{r}})}{\hat{\tau}^2_{\mathrm{r}}} \right) \tag{21}$$

3. Obtain 95% confidence interval for recall.

$$\hat{R}_{(\mathrm{A})} \pm 1.96 \sqrt{\widehat{var}(\hat{R}_{(\mathrm{A})})} \tag{22}$$

Precision. We obtain estimates and 95% confidence intervals for precision as follows.

1. Obtain estimate of precision (for "Team A").

$$\hat{P}_{(\mathrm{A})} = \frac{\hat{\tau}_{\mathrm{r(A)}}}{\hat{\tau}_{\mathrm{a(A)}}} \tag{23}$$

2. Obtain estimate of variance for precision.

$$\widehat{var}(\hat{P}_{(\mathrm{A})}) = \hat{P}^2_{(\mathrm{A})} \left(\frac{\widehat{var}(\hat{\tau}_{\mathrm{r(A)}})}{\hat{\tau}^2_{\mathrm{r(A)}}} + \frac{\widehat{var}(\hat{\tau}_{\mathrm{a(A)}})}{\hat{\tau}^2_{\mathrm{a(A)}}} \right) \tag{24}$$

3. Obtain 95% confidence interval for precision.

$$\hat{P}_{(\mathrm{A})} \pm 1.96 \sqrt{\widehat{var}(\hat{P}_{(\mathrm{A})})} \tag{25}$$

F_1. We obtain estimates and 95% confidence intervals for F_1 as follows.

1. Obtain estimate of F_1 (for "Team A").

$$\hat{F}_{1(\mathrm{A})} = \frac{2}{\frac{1}{\hat{R}_{(\mathrm{A})}} + \frac{1}{\hat{P}_{(\mathrm{A})}}} = \frac{2\hat{R}_{(\mathrm{A})}\hat{P}_{(\mathrm{A})}}{\hat{R}_{(\mathrm{A})} + \hat{P}_{(\mathrm{A})}} \tag{26}$$

2. Obtain estimate of variance for $\left(\frac{1}{\hat{R}_{(\mathrm{A})}} + \frac{1}{\hat{P}_{(\mathrm{A})}}\right)$.

$$\begin{aligned} \widehat{var}\left(\frac{1}{\hat{R}_{(\mathrm{A})}} + \frac{1}{\hat{P}_{(\mathrm{A})}}\right) \quad = \quad & \left(\frac{1}{\hat{R}_{(\mathrm{A})}}\right)^2 \left(\frac{\widehat{var}(\hat{\tau}_{\mathrm{r}})}{\hat{\tau}^2_{\mathrm{r}}} + \frac{\widehat{var}(\hat{\tau}_{\mathrm{r(A)}})}{\hat{\tau}^2_{\mathrm{r(A)}}} \right) + \\ & \left(\frac{1}{\hat{P}_{(\mathrm{A})}}\right)^2 \left(\frac{\widehat{var}(\hat{\tau}_{\mathrm{a(A)}})}{\hat{\tau}^2_{\mathrm{a(A)}}} + \frac{\widehat{var}(\hat{\tau}_{\mathrm{r(A)}})}{\hat{\tau}^2_{\mathrm{r(A)}}} \right) \end{aligned} \tag{27}$$

3. Obtain estimate of variance for F_1.

$$\widehat{var}\hat{F}_{1(\mathrm{A})} = \hat{F}^2_{1(\mathrm{A})} \left(\frac{\widehat{var}\left(\frac{1}{\hat{R}_{(\mathrm{A})}} + \frac{1}{\hat{P}_{(\mathrm{A})}}\right)}{\left(\frac{1}{\hat{R}_{(\mathrm{A})}} + \frac{1}{\hat{P}_{(\mathrm{A})}}\right)^2} \right) \tag{28}$$

4. Obtain 95% confidence interval for F_1.

$$\hat{F}_{1(\mathrm{A})} \pm 1.96\sqrt{\widehat{var}(\hat{F}_{1(\mathrm{A})})} \tag{29}$$

References

[1] Information Retrieval Group, University of Glasgow. http://ir.dcs.gla.ac.uk/resources.html.

[2] Moby Project Home Page. http://icon.shef.ac.uk/Moby/mwords.html.

[3] TREC Legal Track Home Page. http://trec-legal.umiacs.umd.edu/.

[4] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Million Query Track 2007 Overview. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, November 2007. http://trec.nist.gov.

[5] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 Legal Track Overview. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[6] Chris Buckley. Examining Overfitting in Relevance Feedback: Sabir Research at TREC 2007. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, November 2007. http://trec.nist.gov.

[7] Maura R. Grossman, Conor R. Crowley, and Joe Looby. Reflections of the Topic Authorities. Available at http://trec-legal.umiacs.umd.edu/.

[8] Donna K. Harman. The TREC Test Collections. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 21–52, 2005.

[9] H. Schmidt, K. Butter, and C. Rider. Building digital tobacco document libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D-Lib Magazine*, 8(2), 2002.

[10] SEC v. Collins & Aikman Corp., et al. 2009 WL 94311 (S.D.N.Y.).

[11] Sedona Conference Open Letter, dated May 22, 2008. Available at http://trec-legal.umiacs.umd.edu/.

[12] ISRI Staff. Measuring and delivering 95% non-stopword document accuracy. Technical Report 2003-04, Information Science Research Institute, University of Nevada, Las Vegas, September 2003.

[13] Victor Stanley v. Creative Pipe. 250 F.R.D. 251 (D. Md. 2008).

[14] Steven K. Thompson. Sampling, Second Edition, 2002.

[15] 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) Proceedings*, November 2007. http://trec.nist.gov.

[16] Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th International Conference on Information and Knowledge Management (CIKM)*, pages 102–111, 2006.