# TREC Legal Track – Learning Task Final Guidelines *(revision 0.0)*

*Gordon V. Cormack*
*Maura R. Grossman*

**Abstract**

In the learning task, participants are given a "seed set" of documents from a larger collection that have previously been assessed by TREC as responsive or non-responsive to a legal discovery request. Using this information, participants must (a) rank the documents in the larger collection from most likely to least likely to be responsive; and (b) for each document, estimate the likelihood of responsiveness as a probability. The ranking will be evaluated by how well it places responsive documents before non-responsive ones, as assessed by TREC. The likelihood estimate will be evaluated by how well it estimates recall throughout the ranked list.

## 1 e-Discovery Context

The learning task models the use of automatic or semi-automatic methods to guide review strategy for the first or later passes of a multi-stage document review effort, organized as follows:

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

2. **Learning by example.** A learning method is used to rank the documents in the collection from most to least likely to be responsive to the production request, and to estimate this likelihood for

each document. The input to the learning method consists of the seed set, the assessments for the seed set, and the unranked collection; the output is a ranked list consisting of the document identifer and a probability of responsiveness for each document in the collection.

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

3. **Review process.** A review process may be conducted, with strategy guided by the ranked list. One possible strategy is to review documents in order, thus discovering as many responsive documents as possible for a given amount of effort. Another possible strategy is triage: to review only mid-ranked documents, deeming, without further review, the top-ranked ones to be responsive, and the bottom-ranked ones to be non-responsive.

   Review strategy may be guided not only by the order of the ranked list, as outlined above, but also by the estimated effectiveness of various alternatives. Consider the strategy of reviewing the top-ranked documents. Where should a *cut* be made so that documents above the cut are reviewed and documents below are not? For triage, where should the two necessary cuts be made?

Practically every review strategy decision boils down to the question,

> *Of this particular set of documents, how many are responsive and how many are not?*

An informed choice of where to make the cut demands that we know how many documents both above and below the cut are responsive. There is no single correct answer – the best choice depends on the relative risks, costs and benefits of missing responsive documents or of reviewing or producing non-responsive documents. In discovery efforts,

the cost of failing to produce responsive documents is typically much higher than the cost of producing non-responsive ones (unless they are privileged!). The specific tradeoff will be different in almost every matter.

For this reason, we require not that the learning method make the cut, but that it provide the necessary information to make the best cut for the particular situation. To this end, we require an estimate of the probability that any particular document is responsive. If the probability estimates are accurate, it follows that the number of responsive documents in *any* set will be the sum of the probabilities associated with the documents in the set.

Suppose the cost of missing a responsive document is estimated to be $95 and the cost of reviewing a non-responsive one is $5. It therefore follows that the best strategy is to review whose documents whose probability of relevance is above $5\%$, and none below. One simply has to sum the estimates to find the cut that makes this tradeoff.

When opposing counsel demands to know how many documents were missed, or why they weren't identified, the estimate yields a defensible answer.

## 2 TREC Context

This task is part of the TREC 2010 Legal Track. TREC – The Text Retrieval Conference – has the following goals:

- to encourage research in information retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

The Legal Track, in its 5th year, pursues these goals within the context of electronic discovery. The Learning task is a successor to the batch task of TREC 2009. The principal differences are in focus, the requirement to rank every document in the corpus, and the requirement to provide a likelihood estimate for every document in the corpus. In addition, the task uses the same collection and some of the same production requests as its sister task, the TREC 2010 Legal Interactive Task, to facilitate comparison.

## 3 Task Details

There will be eight production requests, each using the Enron document collection. For each request, the participating team will be given the text of the request, guidelines for its interpretation, a seed set of documents, and assessments for the seed documents.

The results for all eight requests must be encoded in a text file according to the standard TREC format, where each line contains:

- *requestid* Q0 *docid rank estP runid*

*requestid* is a number assigned by TREC identifying the production request. Q0 is a historical artifact of the TREC format. *docid* is a TREC-assigned document identifier. *rank* is the ranking of the document by estP, where 1 is the most likely relevant document for the request. estP is a probability estimate between 0.0 and 1.0. *runid* is a unique identifier for the submission, formed by joining

- a sequence of 3 or 4 characters identifying the team (composed by participant)
- a sequence of 3 or 4 charecters identifying the method (composed by participant)
- Capital “A” if the method is fully automated; capital “M” if the method is fully manual, meaning the seed documents are not used as input to any program; capital “T” if both automated and manual methods are used.

The task will use the EDRM Enron v2 collection, which may be downloaded without restriction from the Web. [1] in either EDRM XML or PST formats. The document ids will be those from the EDRM XML format. Mapping files will be provided to convert information derivable from the PST-format documents to document ids.

For the learning task, a document is considered to be either an email message as a whole, or a particular attachment to an email message. The EDRM XML version contains both native and text versions of each document, each with a unique identifier. The PST version contains the complete email messages; participants will need to extract the attachments from these messages and treat them as separate documents.

The Enron corpus contains many duplicate documents. For efficiency, submissions should refer to only the canonical document from each set of duplicates, as defined by TREC.[2] A mapping is provided from all document identifiers to canonical documents, so participants may eliminate duplicates prior to processing, or after the fact.

The EDRM download is approximately 100GB in size. As a convenience, TREC is providing to participants a download for the text version of all canonical documents as a separate download (609 MB). Furthermore, TREC is providing the native version of all canonical attachments (7 GB).

## 4 Manner of participation

We hope to attract participants that use various combinations of manual effort and technology. It is possible and desirable that a participating team employ two or more of the strategies, provided that the efforts are organized to conform to the information flow constraints. That is, a group might complete a fully automated review, and then undertake a technology-assisted review; or, a group might complete a manual review, and then undertake a technology-assisted review. Or a group might undertake manual and automated reviews concurrently, provided no input to the automated system, including tuning, was done by any individual having seen the production request or associated materials.

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

[2] http://durum0.uwaterloo.ca/trec/legal10/

Teams are free – and encouraged – to participate also in the TREC 2010 Legal Interactive Task, which will have three production requests and a privilege review.

## 5 Evaluation Details

The effectiveness of ranking will be evaluated separately from the effectiveness of estimation. Several measures will be computed for each. The principal measures being considered to evaluate ranking are:

- Receiver operating characteristic (ROC) curves to display the trade-off between missed responsive documents and produced non-responsive documents (false negatives and false positives) for all posible cuts.
- Area under the ROC curve (AUC) to serve as a summary measure of ranking effectiveness. AUC can be interpreted in terms of a very simple game (see below).
- Precision and recall at specific cuts, such as 10, 1000, 10000 documents.
- R-Precision.
- Average Precision.

For estimation the principal measures being considered are:

- Root-mean-squared Average Recall Error (RMSRE). The RMS difference between estimated recall and actual recall, at all recall levels.
- Information Gain (IG). An information-theoretic measure that captures both the accuracy of the estimates and the effectiveness of the ranking. IG can be interpreted in terms of a very simple game (see below).
- F1. The actual F1@K measure (used in other legal tasks) achieved when K is the cut value that optimizes apparent F1@K.
- Apparent F1. The value of F1@K, if the estimates were accurate.

# 6 The Information Gain Game

We illustrate the notion of information gain with a game. It is not necessary to understand the theory behind it, as the rules and strategy are simple.

- You are given a sequence of documents to review, in some arbitrary order;
- For each document, you must guess a number $estP$ between 0 and 1, which is your estimate of the likelihood that the document is responsive. (It would be inadvisable to guess 0 or 1 exactly, as the downside of an incorrect guess would be infinite.)
- If the document turns out to be responsive, your score is

$$IG = 1 + \log_2 estP$$

- If the document turns out to be non-responsive your score is

$$IG = 1 + \log_2(1 - estP)$$

A strategy to play well has the following elements:

- If you think the document is responsive, choose $estP > \frac{1}{2}$
- If you are pretty confident the document is responsive, choose $estP \gg \frac{1}{2}$
- Don't be overconfident.
- Never choose $estP = 1$.
- If you think the document is non-responsive, choose $estP < \frac{1}{2}$
- If you are pretty confident the document is non-responsive, choose $estP \ll \frac{1}{2}$
- Don't be overconfident.
- Never choose $estP = 0$.

The best possible score, on average, is achieved if $estP$ is in fact an accurate estimate of the probability. If you think the document is responsive and you think there's a $65\%$ chance you're right, guess $estP = 0.65$.

If your guesses are sensiblee, you'll get a positive average $IG$ score. If you get a negative $IG$ score, you would have been better off flipping a coin!

## 7 The AUC Game

The AUC game works like this:

- You are given a stack of documents, that you arrange with the most likely to be responsive on top, and so on down to the least likely on the bottom.
- For every responsive document and for every non-responsive document in the stack, if the responsive document is above the non-responsive one, you score a point; otherwise you score 0.
- Your AUC score is
$$AUC = \frac{points}{R \times N}$$
were $points$ is the total number of points you scored, $R$ is the number of responsive documents, and $N$ is the number of non-responsive documents.