e-Discovery Team at TREC 2016 Total Recall Track

Ralph C. Losey*
National e-Discovery Counsel
Jackson Lewis P.C.
e-DiscoveryTeam.com
Ralph.Losey@gmail.com

Jim Sullivan, Tony Reichenberger, Levi Kuehn, Jani Grant
Sr. Discovery Services Consultants,
Kroll Ontrack, Inc.
eDiscovery.com
JSullivan@kollontrack.com
TReichenberger@krollontrack.com
LKuehn@krollontrack.com
Jani.Grantz@krollontrack.com

ABSTRACT

The *e-Discovery Team* participated in the 2016 TREC Total Recall Track, *Athome* division, where thirty-four prejudged topics were considered using 290,099 emails of former Florida Governor Jeb Bush. The *Team* participated in TREC 2016 primarily to test the effectiveness of the standard search methodology it uses commercially to search for relevant evidence in legal proceedings: ***Predictive Coding 4.0 Hybrid Multimodal IST***. The *Team's* method uses a *hybrid* approach to continuous active learning with both manual searches and active machine learning based document ranking searches. This is a systematic process involving implementation of a variety of search functions by skilled searchers. The *Team* calls this type of search *multimodal* because all types of search methods are used. A single expert reviewer was used in each topic along with Kroll Ontrack's search and review software, *eDiscovery.com Review* (*EDR*). The *Team* classified 9,863,366 documents as either relevant or irrelevant in all 34 review projects. A total of 34,723 documents were correctly classified as Relevant, as per the *Team's* judgment and corrected standard. The 34,723 relevant documents were found by manual review of 6,957 documents, taking a total of 234.25 man-hours. This represent an average project time of 6.89 hours per topic. The *Team* thus reviewed and classified documents at an average speed of 42,106 files per hour. The *Team's* attained an **average 88% Recall** score across all 34 topics using the corrected standard. The *Team* also attained **F1 scores** of greater than 90% in twelve topics, including two perfect scores of 100% F1.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Search process, relevance feedback, supervised learning, best practices, legal search.
Keywords: Hybrid Multimodal; AI-enhanced review; predictive coding; predictive coding 4.0; electronic discovery; e-discovery; active machine learning; continuous active learning; Intelligent Spaced Training; IST; Computer-assisted review; CAR; Technology-assisted review; TAR; relevant irrelevant training ratios; keyword search.

* The views expressed here by the author of this report, Ralph Losey, are solely his own and should not be attributed to his firm or its clients.

1.0 INTRODUCTION.

The Total Recall track offered multiple pre-judged topics for search in two different divisions, *Athome* and *Sandbox.* The *Sandbox* participants were only permitted to use fully automated systems and the data remained on TREC administrator computers. They searched the same Jeb Bush dataset as *Athome*, plus another dataset not included in the *Athome* division due to confidentiality restrictions. The *Sandbox* participants were prohibited from any manual review of documents or *ad hoc* search adjustments.[1] Even after the submissions ended, the *Sandbox* participants *never look at any documents*, even the unrestricted *Athome* Jeb Bush data.

In the *Athome* experiments the data was loaded onto the participants' own computers and there were no restrictions on the types of searches that could be performed. In the *Sandbox* division data could not be loaded onto the participants' own computers and only fully automated searches were permitted.

The *Team* only participated in the *Athome* experiment, which had thirty-four prejudged topics. This was the only division where the *e- Discovery Team* could use its standard *Predictive Coding 4.0 Hybrid Multimodal IST* method*,* which employs both manual review and machine learning.

The *At Home* and *Sandbox* participants both used a computer "jig" (TREC's quaint term) set up by TREC whereby instant feedback was provided to a participant as whether each document submitted as relevant was in fact previously judged to have been relevant by TREC assessors. When a participant determined that a reasonable effort had been made to find all relevant documents required, which is important in legal search and represents a stopping point for further machine training and document review, they would notify TREC of this supposition and "Call Reasonable." Continued submissions were made after that point so that all documents were classified as either relevant or irrelevant. The goal was to submit as many relevant documents as possible before the Reasonable call, and thereafter to have all false negatives appear in submissions as soon after the Reasonable Call as possible.

The *Athome* group searched the dataset of 290,099 emails of former Florida Governor Jeb Bush. In the version of the Jeb Bush emails used by TREC almost all metadata of these emails has been removed. Moreover, the associated attachments and images were not present. Other collections of the Jeb Bush email exist from PST files that include more information, but the *Team* did not utilize this information and limited its efforts and attention to the official TREC collection. The *Team* normally searches datasets with full metadata included, and all attachments and images. Their searches normally include metadata fields and family associations (relationships between emails and attachments). These omissions in the Jeb Bush dataset increased the difficulty of the *Team's* search, which normally includes a mixture of metadata specific searches.

A significant percentage of the Bush emails were *form type* lobbying emails from constituents, which repeated the same language with little of no variance. The unusually high prevalence of near-duplicate emails made search of many of the Bush topics easier than is typical in legal search.

This same Jeb Bush email collection was used by the Total Recall Track in 2015 for ten topics in which the *Team* also participated. In 2015 Losey searched all ten of these ten topics. None of these search topics was repeated in 2016. For this and other reasons, namely that Losey is a

life-long resident of Florida, very familiar with Jeb Bush and his governance of the state, he was very familiar with this dataset in 2016 and with most of the topics presented.

1.1 Summary of *Team's* Efforts.

The *e-Discovery Team's* 2016 Total Recall Track *Athome* project started June 3, 2016, and concluded on August 31, 2016. Using a single expert reviewer in each topic the *Team* classified 9,863,366 documents in thirty-four review projects.

The topics searched in 2016 and their issue names are shown in the chart below. Also included are the first names of the *e-Discovery Team* member who did the review for that topic, the total time spent by that reviewer and the number of documents manually reviewed to find all of the relevant documents in that topic. The total time of all reviewers on all projects was 234.25 hours. All relevant documents, totaling 34,723 by *Team* count, were found by manual review of 6,957 documents. The thirteen topics in red were considered mandatory by TREC and the remaining twenty-one were optional. The *e-Discovery Team* did all topics.

Topic	Name	Reviewer	Hours Spent	Documents Reviewed
401	Summer Olympics	Ralph	8	363
402	Space	Tony	11	401
403	Bottled Water	Ralph	7	200
404	Eminent Domain	Tony	12	326
405	Newt Gingrich	Ralph	4	67
406	Felon Disenfranchisement	Ralph	7	359
407	Faith Based Initiatives	Ralph	15	479
408	Invasive Species	Tony	8	145
409	Climate Change	Levi	6	87
410	Condominiums	Tony	7	13
411	Stand Your Ground	Ralph	5	274
412	2000 Recount	Tony	10.5	34
413	James V. Crosby	Jim	3	194
414	Medicaid Reform	Tony	11	26
415	George W. Bush	Jim	3.5	156
416	Marketing	Jim	7	72
417	Movie Gallery	Ralph	4	66
418	War Preparations	Tony	8.25	150
419	Lost Foster Child Rilya Wilson	Levi	5	75
420	Billboards	Jim	4	309
421	Traffic Cameras	Jim	2	70
422	Non Resident Aliens	Tony	6	61
423	National Rifle Association	Tony	9	305
424	Gulf Drilling	Levi	6	0
425	Civil Rights Act of 2003	Ralph	8	384
426	Jeffrey Goldhagen	Ralph	5	159
427	Slot Machines	Jim	4.25	235
428	New Stadiums and Arenas	Levi	5	74
429	Elian Gonzalez	Jim	6.25	385
430	Restraints and Helmets	Jani	12	1,033
431	Agency Credit Ratings	Tony	6	82
432	Gay Adoption	Jani	10	766
433	Abstinence	Jim	3.5	44
434	Bacardi Trademark	Ralph	5	83

They were all one-person, solo efforts, although there was coordination and communications between *Team* members on the Subject Matter Expert (SME) type issues encountered. This pertained to questions of true relevance and errors found in the gold standard for many of these topics. A detailed description of the search for each topic is contained in the Appendix.

In each topic the assigned *Team* attorney personally read and evaluated for true relevance every email that TREC returned as a relevant document, and every email that TREC unexpectedly returned as Irrelevant. Some of these were read and studied multiple times before we made our final calls on true relevance, determinations that took into consideration and gave some deference to the TREC assessor adjudications, but were not bound by them. Many other emails that the *Team* members considered irrelevant, and TREC agreed, were also personally reviewed as part of their search efforts. As mentioned, there was sometimes consultations and discussion between *Team* members as to the unexpected TREC opinions on relevance.

This contrasts sharply with participants in the *Sandbox* division. They never make any effort to determine where their software made errors in predicting relevance, or for any other reasons. They accept as a matter of faith the correctness of all TREC's prior assessment of relevance. To these participants, who were all academic institutions, the ground truth itself as to relevance or not, was of no relevance. Apparently, that did not matter to their research.

All thirty-four topics presented search challenges to the *Team* that were easier, some far easier, than the *Team* typically face as attorneys leading legal document review projects. (If the Bush email had not been altered by omission of metadata, the searches would have been even easier.) The details of the searches performed in each of the thirty-four topics are included in the Appendix. The search challenges presented by these topics were roughly equivalent to the most simplistic challenges that the *e-Discovery Team* might face in projects involving relatively simple legal disputes. A few of the search topics in 2016 included quasi legal issues, more than were found in the 2015 Total Recall Track. This is a revision that the *Team* requested and appreciated because it allowed some, albeit very limited testing of legal judgment and analysis in determination of true relevance in these topics. In legal search relevancy, legal analysis skills are obviously very important. In most of the 2016 Total Recall topics, however, no special legal training or analysis was required for a determination of true relevance.

At Home participants were asked to track and report their manual efforts. The *e-Discovery Team* did this by recording the number of documents that were *human reviewed* and classified prior to submission. More were reviewed after submission as part of the Team's TREC relevance checking. Virtually all documents human reviewed were also classified, although all documents classified were not used for active training of the software classifier. The *Team* also tracked effort by number of attorney hours worked as is traditional in legal services. Although the amount of time varied somewhat by topic, the average time spent per topic was only **6.89 hours**. The average review and classification speed for each project was **42,106 files per hour** (9,863,366/234.25).

1.2 *e-Discovery Team* Members.

The *Team* is composed of five legal search experts Ralph Losey, Jim Sullivan, Tony Reichenberger, Levi Kuehn, Jani Grantz -- and one "robot," *Mr. EDR* (the software they used). The *Team* members are not scientists or in academia. Most are lawyers who spend their

working hours looking for evidence in large, chaotic datasets, such as email. They typically assist other attorneys in lawsuits and legal investigations. Their work includes the identification, review, analysis, classification, production, and admission of Electronically Stored Information (ESI) as evidence in courts in the United States and elsewhere.

The *Team* leader and report author is Ralph C. Losey, J.D., a full-time practicing attorney, principal and *National e-Discovery Counsel* of Jackson Lewis P.C., a U.S. law firm with over 800 attorneys and fifty-five offices. He has over 37 years of experience doing legal document reviews. Losey is also a blogger at *e-DiscoveryTeam.com* where he has written over two million words on e-discovery, including six books and over sixty articles on document review.[2] The past six years Losey has participated in multiple public and private experiments, some competitive, to test and prove various predictive coding methods.

Jim Sullivan, J.D., Tony Reichenberger, J.D.,and Jani Grantz J.D., are attorney search and review specialists who work for Kroll Ontrack, Inc. (KO). Levi Kuehn is a non-attorney search and review specialists who works for KO. Kroll Ontrack is the primary e-discovery vendor used by Losey and his law firm. It is a global e-Discovery software, processing and project management company (eDiscovery.com). The *Team* robot, *Mr. EDR*, is the *Team's* personalization of KO's software, *eDiscovery.com Review* (EDR). Losey, Sullivan and Reichenberger participated in the 2015 TREC Total Recall Track. So too did a prior version of Mr. EDR, which is in a process of constant enhancement.

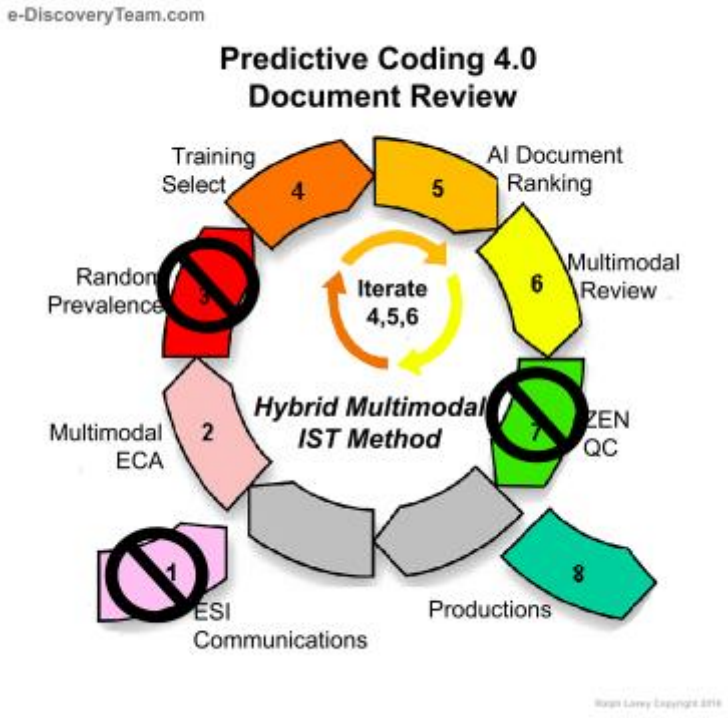
2.*0 E-DISCOVERY TEAM'S* SEARCH METHOD.

The *e-Discovery Team* uses what they call a *Predictive Coding 4.0 Hybrid Multimodal IST* method for search and review of large document collections.[3] This method is a type of continuous active learning text retrieval system that employs supervised machine learning and a variety of manual search methods.[4] The various types of searches included in the *Team's* multimodal approach are shown in the search pyramid, below.

Linear review refers to an SME's examination of all documents by certain key witnesses in a lawsuit during certain time frames critical to the disputed facts in a lawsuit. Keyword search in our methodology refers to the use of terms originating from legal and document analysis, and

from witness interviews. Judgmental sampling and verification by SMEs are also used to test the terms before they are used throughout a document collection. Our keyword search also includes a variety of Boolean functions and parametric targeting, wherein searches are limited to certain metadata fields of an electronic document. Similarity and concept searches refer to a variety of *passive* machine learning analytic search techniques. The AI search at the top of the pyramid refers to the use of active machine learning. The EDR KO software uses a proprietary type of logistic regression algorithm.

The standard eight-step workflow normally used by the *Team* in legal search projects is shown in the diagram below. To meet the *Team's* self-imposed time requirements of completing every review project with minimal time efforts, the standard steps Three and Seven were omitted as will be further explained. Further, due to the set-up of the TREC experiments, the first step of our workflow, ESI Communications, was severely constrained to the point of being practically meaningless, as will also be further explained. The *Team's* standard workflow was thus reduced from eight to five steps as shown below. Also, the amount of time the *Team* normally spends on each step was also limited.



In the first step of ESI Communications *Team* members on a legal review project typically spend hours in discussion and analysis of scope of relevance and the target documents. The communications often include hundreds of written exchanges, both informal, such as emails and chats, and formal, such as (1) detailed requests for information contained in court documents such a *subpoenas* or *Request For Production*; (2) input from a qualified SME, who is typically a legal expert with deep knowledge of the factual issues in the case, and thus deep knowledge of what the presiding judge in the legal proceeding will hold to be relevant and discoverable; and, (3) dialogues with the party requesting the production of documents to clarify the search target, and other parties. The ESI communications may lead to formal motions

with the governing court, legal memorandums, hearings before the presiding judge and opinions rendered by one or more judges on the scope of relevance.

The only ESI communications in the TREC experimental set-up was a very short, one sentence description of relevance for each topic. Two topics had a two-sentence description (410-Condominiums and 423-National Rifle Association). The only other type of ESI communications in this TREC Track were the automated, instant returns of all documents submitted as to whether TREC considered them to be relevant or not. There were no appeals or other procedures set-up for *Athome* division participants who actually examined the documents for true relevance to challenge obvious errors in judgment. The *Sandbox* division participants who search the same topics and dataset never actually look at any documents or make any relevance decisions; it is a fully automated process for them. They only train based on the automatic feedback from TREC's assessor judgments.

3.0 RELATED WORK

It is generally accepted in the legal search community that the use of *predictive coding* type search algorithms can improve the search and review of documents in legal proceedings.[5] The use of predictive coding has also been approved, and even encouraged by various courts around the world, including numerous courts in the U.S.[6]

Although there is agreement on use of predictive coding, there is controversy and disagreement as to the most effective ***methods*** of use.[7] There are proponents for a variety of different methods to find training documents for predictive coding. Some advocate for the use of chance selection alone, others for the use of top ranked documents alone, others for a combination of top ranked and mid-level ranked documents where classification is unsure.[8] The *e-Discovery Team* uses a method that includes a combination of all three of these selection processes and more.

Some attorneys and predictive coding software vendors advocate for the use of predictive coding search methods alone, and forego other search methods when they do so, such as keyword search, concept searches, similarity searches and linear review. The *e-Discovery Team* members reject that approach and instead advocate for a *hybrid multimodal* approach they call *Predictive Coding 4.0*.[9] This method uses an approach to active machine learning that the *Team* calls ***IST***, standing for "***Intelligently Spaced Training***." Under IST the attorney in charge decides exactly when to train. This is different from other systems where the machine retrains after each document is coded, or certain predetermined number, and the human trainer has no discretion as to timing.[10]

The *e-Discovery Team* approach includes all types of search methods (thus the term *multimodal*) to find relevant documents, with primary reliance placed on predictive coding. The *Team* also uses a variety of methods to find suitable training documents for predictive coding, including high ranking documents, and all other search methods. This is a fundamental difference with other methods that rely entirely on predictive coding to find relevant documents, and rely entirely upon high-ranking documents for training. Grossman and Cormack have scientifically tested these high-ranking training methods, and measured their effectiveness, but this does not mean that they endorse them as an exclusive tool, nor claim this to be their own preferred method.[11]

4.0 *E-Discovery Team's* Four Research Questions and Short Answers.

4.1 Primary Question (repeat from 2015).

What Recall, Precision and Effort levels will the *e-Discovery Team* attain in TREC test conditions over all thirty-four topics using the *Team's* ***Predictive Coding 4.0 Hybrid Multimodal IST*** search methods and Kroll Ontrack's software, *eDiscovery.com Review* (EDR).

Short Answer: Again, as in the 2015 Total Recall Track, the *Team* attained very good results with high levels of Recall and Precision in almost all topics, including perfect or near perfect results in several topics using the *corrected* gold standard, and very little human effort.

4.2 Second Question.

What is the impact of incorrect Subject Matter Expert ("SME") judgments by the TREC assessors on Recall and Precision. (Unplanned question that unfortunately arose out of the circumstances encountered.)

Short Answer: This had a substantial impact on many topics where there were many errors in the standard, and only minor impact on topics where the disagreements were small.

4.3 Third Question.

What is the most effective search method from the *Team's* multimodal tool-set for retrieval of relevant documents in the relatively simplistic search challenges presented by most, but not all, of the thirty-four topics. (Unplanned question that arose out of the circumstances encountered.)

Short Answer: For the easy topics what the *Team* calls "tested, parametric, Boolean keyword search" was the most effective search method to find relevant documents.

4.4 Fourth Question.

What is the role of active machine learning in retrieval of relevant documents in the simplistic search challenges presented by many of the thirty-four topics.

Short Answer: The *Team* found that for the easiest topics in the 2016 *Total Recall Track* the primary role of active machine learning was reduced to a quality assurance function.

5.0 EXPERIMENTS AND DISCUSSIONS

The *e-Discovery Team* sought to answer the four previously listed Research Questions in its experiments at the 2016 TREC Total Recall Track.

5.1 First and Primary Research Question.

What Recall, Precision and Effort levels will the *e-Discovery Team* attain in TREC test conditions over all thirty-four topics using the *Team's Predictive Coding 4.0* hybrid multimodal search methods and Kroll Ontrack's software, *eDiscovery.com Review* (EDR).

Again, as in the 2015 Total Recall Track, the *Team* attained very good results with high levels of Recall and Precision in all topics, including perfect or near perfect results in several topics using the *corrected* gold standard. The *Team* did so even though it only used five of the eight steps in its usual methodology, *intentionally severely constrained the amount of human effort expended on each topic* and worked on a dataset stripped of metadata. The *Team's* enthusiasm for the record setting results, which were significantly better than its 2015 effort, is tempered by the fact that the search challenges presented in most of the topics in 2016 were not difficult and the TREC relevance judgments had to be corrected in most topics.

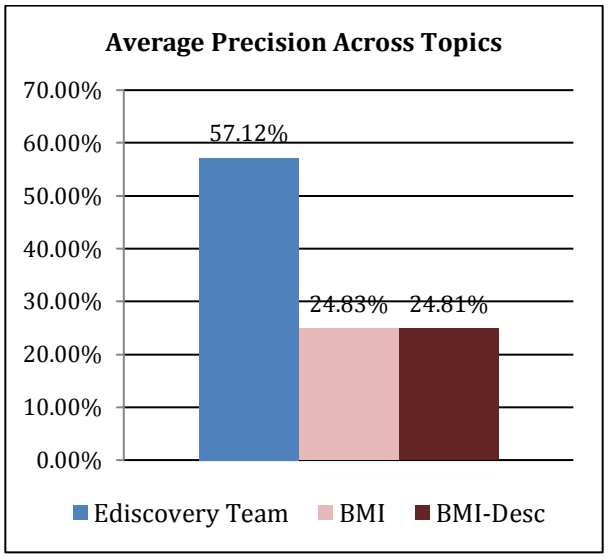
Even using the given uncorrected TREC standard for scoring, and even though in most topics we did not train on the TREC returned-relevant documents that the *Team* considered

irrelevant, the *Team* overall still attained excellent results. Under the corrected standard, the results were much better. The following chart compares the *Team's* Recall, Precision and F-Measure for each *Athome* topic with the results obtained by TREC's BMI and BMI-Desc runs. These comparative statistics show the scores at the time of reasonable call. This first chart uses the uncorrected defective standard and is thus of limited value in the topics that had many mistakes.

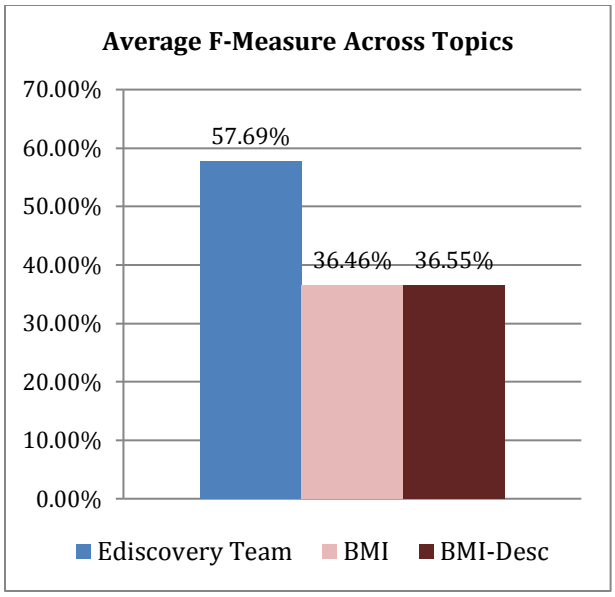
COMPARISONS AT TIME OF REASONABLE CALL USING UNCORRECTED TREC STANDARDS

		Recall			Precision			F-Measure		
		Ediscovery Team	BMI	BMI-Desc	Ediscovery Team	BMI	BMI-Desc	Ediscovery Team	BMI	BMI-Desc
athome401	**Summer Olympics**	41.05%	91.70%	**92.58%**	**73.44%**	15.31%	15.45%	**52.66%**	26.23%	26.48%
athome402	Space	72.57%	**91.07%**	90.28%	22.04%	**30.86%**	30.59%	33.81%	**46.09%**	45.70%
athome403	Bottled Water	7.16%	**97.71%**	**97.71%**	**80.41%**	37.49%	37.49%	13.14%	**54.18%**	**54.18%**
athome404	Eminent Domain	22.94%	91.74%	**91.93%**	**64.43%**	26.55%	26.61%	33.83%	41.19%	**41.27%**
athome405	**Newt Gingrich**	95.08%	**99.18%**	98.36%	**28.09%**	9.82%	9.74%	**43.36%**	17.87%	17.73%
athome406	**Felon Disenfran**	73.23%	**92.91%**	**92.91%**	**66.91%**	9.58%	9.58%	**69.92%**	17.37%	17.37%
athome407	Faith Based Initiatives	31.02%	91.80%	**91.99%**	**68.72%**	41.86%	41.95%	42.75%	57.50%	**57.62%**
athome408	**Invasive Species**	55.17%	**83.62%**	**83.62%**	**64.65%**	7.87%	7.87%	**59.53%**	14.39%	14.39%
athome409	**Climate Change**	84.65%	**95.05%**	94.06%	**40.71%**	13.99%	13.85%	**54.98%**	24.40%	24.14%
athome410	**Condominiums**	95.10%	**99.48%**	99.03%	**46.13%**	42.59%	42.40%	**62.12%**	59.64%	59.38%
athome411	**Stand Your Ground**	66.29%	70.79%	**84.27%**	**67.05%**	5.70%	6.09%	**66.67%**	10.55%	11.36%
athome412	2000 Recount	57.38%	91.35%	**92.48%**	**49.18%**	40.97%	41.48%	52.96%	**56.57%**	57.27%
athome413	**James V. Crosby**	96.34%	99.08%	**99.27%**	**89.00%**	28.73%	28.78%	**92.52%**	44.55%	44.63%
athome414	Medicaid Reform	91.66%	96.90%	**97.26%**	**35.32%**	35.10%	35.23%	51.01%	51.54%	**51.73%**
athome415	**George W. Bush**	**94.08%**	63.39%	67.08%	**91.04%**	61.09%	58.66%	**92.53%**	62.22%	62.59%
athome416	Marketing	60.30%	94.19%	**95.57%**	42.08%	43.32%	**43.96%**	49.57%	59.35%	**60.22%**
athome417	**Movie Gallery**	99.61%	**99.81%**	99.66%	**99.38%**	57.28%	57.19%	**99.49%**	72.79%	72.67%
athome418	**War Preparations**	39.57%	93.05%	**93.58%**	**50.34%**	12.68%	12.76%	**44.31%**	22.32%	22.45%
athome419	Lost Foster Child Rilya Wilson	**98.84%**	93.06%	93.61%	15.04%	48.13%	**48.41%**	26.10%	63.44%	**63.82%**
athome420	**Billboards**	92.54%	**99.46%**	99.32%	**92.16%**	31.65%	31.61%	**92.35%**	48.02%	47.95%
athome421	**Traffic Cameras**	90.48%	**100.00%**	**100.00%**	**12.50%**	1.90%	1.90%	**21.97%**	3.73%	3.73%
athome422	Non Resident Aliens	93.55%	**100.00%**	**100.00%**	0.90%	**2.81%**	**2.81%**	1.79%	5.46%	5.46%
athome423	**National Rifle Association**	51.05%	**99.65%**	**99.65%**	**33.18%**	18.68%	18.68%	**40.22%**	31.46%	31.46%
athome424	Gulf Drilling	99.60%	**100.00%**	**100.00%**	22.76%	**26.39%**	**26.39%**	37.05%	**41.76%**	**41.76%**
athome425	**CivilRights Act 2003**	91.32%	**98.60%**	**98.60%**	**96.59%**	33.70%	33.70%	**93.88%**	50.23%	50.23%
athome426	**Jeffrey Goldhagen**	70.00%	**94.17%**	**94.17%**	**87.50%**	9.17%	9.17%	**77.78%**	16.72%	16.72%
athome427	**Slot Machines**	89.21%	**96.68%**	**96.68%**	**35.77%**	16.98%	16.98%	**51.07%**	28.89%	28.89%
athome428	New Stadiums	93.10%	**98.49%**	**98.49%**	17.81%	**26.95%**	**26.95%**	29.91%	**42.31%**	**42.31%**
athome429	**Elian Gonzalez**	**94.20%**	99.27%	99.27%	**92.41%**	35.45%	35.45%	**93.29%**	52.24%	52.24%
athome430	**Restraints & Helmets**	71.95%	94.25%	**94.65%**	**65.00%**	36.40%	36.55%	**68.30%**	52.52%	52.74%
athome431	**Agency Credit Rate**	75.69%	**99.31%**	**99.31%**	**47.60%**	11.61%	11.61%	**58.45%**	20.78%	20.78%
athome432	**Gay Adoption**	85.00%	**98.57%**	**98.57%**	**86.23%**	11.20%	11.20%	**85.61%**	20.12%	20.12%
athome433	**Abstinence**	99.11%	**100.00%**	**100.00%**	**66.07%**	9.09%	9.09%	**79.29%**	16.67%	16.67%
athome434	**Bacardi Trademark**	86.84%	**100.00%**	**100.00%**	**91.67%**	3.44%	3.44%	**89.19%**	6.65%	6.65%

In the precision category, which in Legal Search is the *money shot* that has the greatest impact on the cost of a document review project, the *e-Discovery Team* dominated, even using the uncorrected TREC standard. It had the highest precision level on 28 of the 34 topics (82%). They are highlighted in blue in the above chart. The e-Discovery *Team's* average precision score was 57.1%. The average precision of both BMI and BMI-Desc was 24.8%. Thus the *Team's* precision score was on average ***more two and a quarter times higher*** than that of the BMI standards.



In the F1-measure, which is the standard value used in legal search to evaluate overall precision and recall of a project, the *e-Discovery Team* again dominated. This is somewhat surprising in view of the fact that these measurements were based on the uncorrected TREC standard. The *Team* had the highest F1 scores on 23 of the 34 topics (68%). They are highlighted in blue in the above chart. The *e-Discovery Team's* average F1 score was 57.69%. The average F1 of BMI and BMI-Desc was 36.5%. Thus the *Team's* F1 score was on average ***more than 58% higher*** *than that of the BMI standards*.



Even using TREC's often erroneous standards, the *Team* still attained higher recall than both the BMI and BMI-Desc standards on two topics: topic 415 *George Bush* with a score of **94.08%**; and, topic 419 *Lost Foster Child Rilya Wilson* with a score of **98.84%**. Moreover, the *Team* attained recall levels in excess of 90% at the time of reasonable call in the following additional topics:

- **95.08%** on topic 406 Felon Disenfranchisement;
- **95.10%** on topic 410 Condominiums;
- **96.34%** on topic 413 James V. Crosby;
- **99.61%** on topic 417 Movie Gallery;
- **92.54%** on topic 420 Billboards;
- **90.48%** on topic 421 Traffic Cameras;
- **93.55%** on topic 422 Non Resident Aliens;
- **99.60%** on topic 424 Gulf Drilling;
- **91.32%** on topic 425 Civil Rights Act of 2003;
- **93.10%** on topic 428 New Stadiums and Arenas;
- **94.20%** on topic 429 Elian Gonzalez;
- **99.11%** on topic 433 Abstinence.

In summary, even with the uncorrected TREC standards, where in most topics the *Team* did not use all documents returned as relevant for all of its training documents, it attained Recall scores greater than 90% in fourteen of the thirty-four topics. The *Team* attained Recall scores of 80% or higher in four additional topics. The average results obtained across all thirty-four topics at the time of reasonable call were as follows:

- 75.46% Recall
- 57.12% Precision
- 57.69% F1
- 121 Docs Reviewed Effort

The *Team,* composed as it is of trained attorneys who engage in relevance analysis on a daily basis in the context of actual lawsuits, believes strongly in the idea of a ground truth of relevance, in other words, True Facts, not Alternate Facts. The *Team's* work depends on an objective, consistent assessment of true relevant documents. The boundaries of true relevance or irrelevance is a judgment call based on somewhat subjective factors, but once the border is established, it must be consistently followed in legal search. For that reason the measurements of the effectiveness of the *Team* performance based on a defective, inconsistent standard, is of little interest to the *Team*. We consider the only significant measurement of our results to arise out of use of the corrected gold standard. These are described next.

This next chart uses the corrected standard. It is the primary reference chart we use to measure our results. Unfortunately, it is not possible to make any comparisons with BMI standards because we do not know the order in which the BMI documents were submitted.

			Revised Gold Standard				
Topic	Name	Reviewer	Total Relevant	Relevant Found	Recall	Precision	F1 score
401	Summer Olympics	Ralph	137	126	91.971%	98.438%	95.094%
402	Space	Tony	679	489	72.018%	38.054%	49.796%
403	Bottled Water	Ralph	123	96	78.049%	98.969%	87.273%
404	Eminent Domain	Tony	519	182	35.067%	93.814%	51.052%
405	Newt Gingrich	Ralph	123	123	100.000%	29.782%	45.896%
406	Felon Disenfranchisement	Ralph	203	197	97.044%	100.000%	98.500%
407	Faith Based Initiatives	Ralph	1,654	1,465	88.573%	62.634%	73.378%
408	Invasive Species	Tony	168	86	51.190%	86.869%	64.419%
409	Climate Change	Levi	224	198	88.393%	47.143%	61.491%
410	Condominiums	Tony	1,317	1,314	99.772%	47.351%	64.223%
411	Stand Your Ground	Ralph	59	59	100.000%	67.045%	80.272%
412	2000 Recount	Tony	850	747	87.882%	45.410%	59.880%
413	James V. Crosby	Jim	600	581	96.833%	98.308%	97.565%
414	Medicaid Reform	Tony	844	783	92.773%	35.917%	51.786%
415	George W. Bush	Jim	12,267	11,554	94.188%	92.358%	93.264%
416	Marketing	Jim	1,485	911	61.347%	43.967%	51.223%
417	Movie Gallery	Ralph	5,945	5,945	100.000%	100.000%	100.000%
418	War Preparations	Tony	141	114	80.851%	77.551%	79.167%
419	Lost Foster Child Rilya Wilson	Levi	1,982	1,964	99.092%	15.022%	26.089%
420	Billboards	Jim	739	707	95.670%	95.541%	95.605%
421	Traffic Cameras	Jim	54	52	96.296%	34.211%	50.485%
422	Non Resident Aliens	Tony	48	48	100.000%	1.493%	2.941%
423	National Rifle Association	Tony	190	147	77.368%	33.409%	46.667%
424	Gulf Drilling	Levi	495	493	99.596%	22.667%	36.929%
425	Civil Rights Act of 2003	Ralph	718	653	90.947%	96.171%	93.486%
426	Jeffrey Goldhagen	Ralph	98	91	92.857%	94.792%	93.814%
427	Slot Machines	Jim	263	249	94.677%	41.431%	57.639%
428	New Stadiums and Arenas	Levi	476	447	93.908%	18.433%	30.817%
429	Elian Gonzalez	Jim	844	819	97.038%	97.153%	97.095%
430	Restraints and Helmets	Jani	1,013	735	72.557%	67.001%	69.668%
431	Agency Credit Ratings	Tony	149	120	80.537%	52.402%	63.492%
432	Gay Adoption	Jani	137	125	91.241%	90.580%	90.909%
433	Abstinence	Jim	141	141	100.000%	83.929%	91.262%
434	Bacardi Trademark	Ralph	38	38	100.000%	100.000%	100.000%
	AVERAGE		1,021	935	88.169%	64.937%	69.152%
	TOTALS		34,723				

The average results obtained across all thirty-four topics at the time of reasonable call using the corrected standard are shown below in bold. The average scores using the uncorrected standard are shown for comparison in parenthesizes.

- **88.17% Recall** (75.46%)
- **64.94% Precision** (57.12%)
- **69.15% F1** (57.69%)
- **124 Docs Reviewed Effort** (124)

At the time of reasonable call the *Team* had **recall scores** greater than 90% in twenty-two of the thirty-four topics and greater than 80% in five more topics. Recall of greater than 95% was attained in fourteen topics. These Recall scores under the corrected standard are shown in the below chart. The results are far better than we anticipated, including six topics with total recall – 100%, and two topics with both total recall and perfect precision, topic 417 Movie Gallery and topic 434 Bacardi Trademark.

Name	Recall
Summer Olympics	91.971%
Newt Gingrich	100.000%
Felon Disenfranchisement	97.044%
Faith Based Initiatives	88.573%
Climate Change	88.393%
Condominiums	99.772%
Stand Your Ground	100.000%
2000 Recount	87.882%
James V. Crosby	96.833%
Medicaid Reform	92.773%
George W. Bush	94.188%
Movie Gallery	100.000%
War Preparations	80.851%
Lost Foster Child Rilya Wilson	99.092%
Billboards	95.670%
Traffic Cameras	96.296%
Non Resident Aliens	100.000%
Gulf Drilling	99.596%
Civil Rights Act of 2003	90.947%
Jeffrey Goldhagen	92.857%
Slot Machines	94.677%
New Stadiums and Arenas	93.908%
Elian Gonzalez	97.038%
Agency Credit Ratings	80.537%
Gay Adoption	91.241%
Abstinence	100.000%
Bacardi Trademark	100.000%

At the time of reasonable call the *Team* had **precision scores** greater than 90% in thirteen of the thirty-four topics and greater than 75% in three more topics. Precision of greater than 95% was attained in nine topics. These Precision scores under the corrected standard are shown in the below chart. Again, the results were, in our experience, incredibly good, including three topics with perfect precision at the time of the reasonable call.

Name	Precision
Summer Olympics	98.438%
Bottled Water	98.969%
Eminent Domain	93.814%
Felon Disenfranchisement	100.000%
Invasive Species	86.869%
James V. Crosby	98.308%
George W. Bush	92.358%
Movie Gallery	100.000%
War Preparations	77.551%
Billboards	95.541%
Civil Rights Act of 2003	96.171%
Jeffrey Goldhagen	94.792%
Elian Gonzalez	97.153%
Gay Adoption	90.580%
Abstinence	83.929%
Bacardi Trademark	100.000%

At the time of reasonable call the *Team* had **F1 scores** greater than 90% in twelve of the thirty-four topics and greater than 75% in two more. F1 of greater than 90% was attained in eight topics. These F1 scores under the corrected standard are shown in the below chart. Note there were two topics with a perfect score, Movie Gallery (100%) and Bacardi Trademark (100%) and three more that were near perfect: Felon Disenfranchisement (98.5%), James V. Crosby (97.57%), and Elian Gonzalez (97.1%).

Name	F1 score
Summer Olympics	95.094%
Felon Disenfranchisement	98.500%
Stand Your Ground	80.272%
James V. Crosby	97.565%
George W. Bush	93.264%
Movie Gallery	100.000%
War Preparations	79.167%
Billboards	95.605%
Civil Rights Act of 2003	93.486%
Jeffrey Goldhagen	93.814%
Elian Gonzalez	97.095%
Gay Adoption	90.909%
Abstinence	91.262%
Bacardi Trademark	100.000%

We were lucky to attain two perfect scores in 2016 (we attained one in 2015), in topic 417 Movie Gallery and topic 434 Bacardi Trademark. The perfect score of 100% F1 was obtained in topic 417 by locating all 5,945 documents relevant under the corrected standard after reviewing only 66 documents. This topic was filled with form letters and was a fairly simple search.

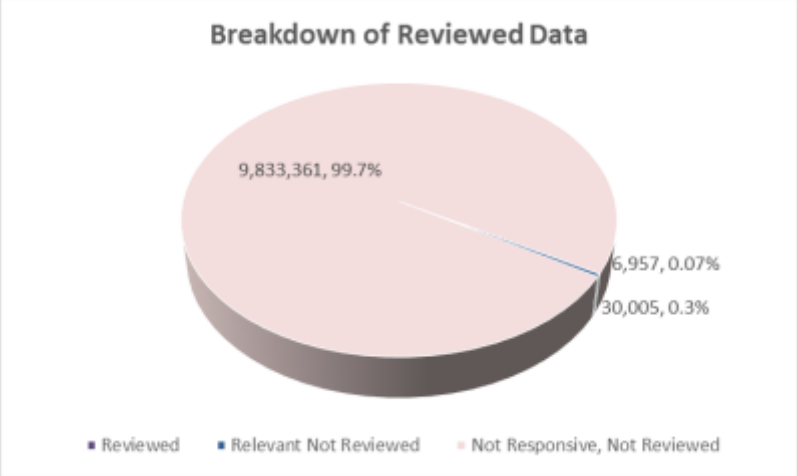
The perfect score of 100% F1 was obtained in topic 434 Bacardi Trademark by locating all 38 documents relevant under the corrected standard after reviewing only 83 documents. This topic had some legal issues involved that required analysis, but the reviewing attorney, Ralph Losey, is an SME in trademark law so this did not pose any problems. The issues were easy and not critical to understand relevance. This was a simple search involving distinct language and players. All but one of the 38 relevant documents were found by tested, refined keyword search. One additional relevant document was found by a similarity search. Predictive coding searches were run after the keywords searches and nothing new was uncovered. Here machine learning merely performed a quality assurance role to verify that all relevant documents had indeed been found.

The *Team* proved once again, as it did in 2015, that perfect recall and perfect precision is possible, albeit rare, using the *Team's* methods and fairly simple search projects.

The *Team's* top ten projects attained remarkably high scores with an average **Recall of 95.66%, average Precision of 97.28% and average F-Measure: 96.42%.** The top ten are shown in the chart below.

		Revised Gold Standard					
Topic	Name	Total Relevant	Docs Submitted	Relevant Found	Recall	Precision	F1 score
401	Summer Olympics	137	128	126	91.971%	98.438%	95.094%
406	Felon Disenfranchisement	203	197	197	97.044%	100.000%	98.500%
413	James V. Crosby	600	591	581	96.833%	98.308%	97.565%
415	George W. Bush	12,267	12,510	11,554	94.188%	92.358%	93.264%
417	Movie Gallery	5,945	5,945	5,945	100.000%	100.000%	100.000%
420	Billboards	739	740	707	95.670%	95.541%	95.605%
425	Civil Rights Act of 2003	718	679	653	90.947%	96.171%	93.486%
426	Jeffrey Goldhagen	98	96	91	92.857%	94.792%	93.814%
429	Elian Gonzalez	844	843	819	97.038%	97.153%	97.095%
434	Bacardi Trademark	38	36	38	100.000%	100.000%	100.000%
	AVERAGE				95.655%	97.276%	96.442%

In addition to Recall, Precision and F1, the *Team* per TREC requirements also measured the ***effort*** involved in each topic search. We measured effort by the number of documents that were actually human-reviewed prior to submission and coded relevant or irrelevant. We also measured effort by the total human time expended for each topic. Overall, the *Team human-reviewed* only 6,957 documents to find *all the 34,723 relevant documents* within the overall corpus of 9,863,366 documents. The total time spent by the *Team* to review the 6,957 documents, and do all the search and analysis and other work using our *Hybrid Multimodal Predictive Coding 4.0* method, was 234.25 hours.



It is typical in legal search to try to measure the efficiency of a document review by the number of documents classified by an attorney in an hour. For instance, a typical contract review attorney can read and classify an average of 50 documents per hour. The *Team* classified 9,863,366 documents by review of 6,957 documents taking a total time of 234.25 hours. The

Team's overall review rate for the entire corpus was thus 42,106 files per hour (9,863,366/234.25).

In legal search it is also typical, indeed mandatory, to measure the costs of review and bill clients accordingly. If we here assume a high attorney hourly rate of $500 per hour, then the total cost of the review of all 34 Topics would be $117,125. That is a cost of just over $0.01 per document. In a traditional legal review, where a lawyer reviews one document at a time, the cost would be far higher. Even if you assume a low attorney rate of $50 per hour, and review speed of 50 files per hour, the total cost to review every document for every issue would be $9,863,366. That is a cost of $1.00 per document, which is actually low by legal search standards.[13]

Analysis of project duration is also very important in legal search. Instead of the 234.25 hours expended by our *Team* using *Predictive Coding 4.0*, traditional linear review would have taken 197,267 hours (9,863,366/50). In other words, the review of thirty-four projects, which we did in our part-time after work in one Summer, would have taken a team of two lawyers using traditional methods, 8 hours a day, every day, over 33 years! These kinds of comparisons are common in Legal Search.

Detailed descriptions of the searches run in all thirty-four topics are included in the Appendix.

5.2 Research Question No. 2.

What is the impact of multiple errors in SME judgments by the TREC assessors on Recall and Precision.

The impact of assessor errors on Recall and Precision was significant, depending in part upon the number of errors made by TREC assessors in a particular topic. The importance of the computer maxim, "Garbage In, Garbage Out – *GIGO,"* was shown to have direct application to machine learning and text retrieval. The impact seen here is, however, exaggerated by the presence of numerous near duplicate form emails in the Bush collection. More research on this question is needed to try to quantify the impact of SME errors using *Predictive Coding 4.0 Hybrid Multimodal IST* methods.

After the *Team* encountered numerous errors on the first topics undertaken, we were forced to create our own gold standard of true relevant documents for each topic. The *Team's* new gold standard corrected for the obvious errors seen in TREC's assessments of relevance. In all close questions on relevance the judgment of TREC's assessors was accepted as accurate.

The errors and inconsistencies seen by the *Team's* close study of the documents were not accepted. In most, but not all topics, the *Team* did not use the documents with obvious errors for its machine training. In all topics the *Team* created its own standard and made comparative recall, precision and F1 calculations based thereon. The observation and correction of TREC errors in gold standard became a collaborative effort among the *Team* to peer review and verify our corrected standard. Most of these efforts, many of which occurred after the conclusion of the Track in August, were not included in the time reports of efforts expended by attorneys in the search.

The *Team* was very reluctant to take this step. It meant a lot more work and make everything much more complicated. We would certainly have let pass a few errors or mere differences of opinion. We recognize that no standard is ever perfect. As lawyers the *Team* understands all too well that some, perhaps many judgments on relevance are subjective.

Again, in all close questions on relevance the judgments of TREC's assessors were accepted, even though we personally disagreed.

The *Team* means no disrespect by the creation of an alternate gold standard. We appreciate and respect the efforts made by the TREC assessors and organizers. Still, the volume of obvious errors encountered forced us to take this action. The integrity of our primary research question to test the effectiveness of our hands-on type multimodal hybrid methods demanded that we do so. We understand that the impact on other *Total Recall* Participants, ones that never actually examine documents, may be far less, perhaps even negligible. Still, there could be an impact, even for them, in some topics

where more than an insignificant number of the same or similar documents were inconsistently judged.

The decision to not accept the errors seen, and to instead create our own gold standard, resulted in substantial additional work for the *Team*. In some topics, described in the Appendix, we even took the step of making two "reasonable calls." One was for TREC, and the second call, which always took place on the next submission, was for our own internal tracking. In the second call we would include emails that we knew from prior submissions of the same or similar document would again be incorrectly considered irrelevant by TREC. We knew they were true relevant and so waited until after our public *reasonable call* to TREC to submit them and then we make our own internal *reasonable call*. We were attempting to, in effect, *play two games at once*, and maximize our score in each game. Keeping track of two standards added an unexpected layer of difficulty to our work and we did not bother to do so in most topics.

In some topics the difference between the two standards was substantial. In a few topics only minor differences were seen. Disagreements on relevance are not unexpected in any standard involving at least somewhat subjective mass relevance adjudications. We do not intend to engage in a criticism of the specific gold standard creation methods used in 2016 *Total Recall Track*, except to note that the appeals procedure included in the 2008 and 2009 TREC Legal Tracks could have improved the accuracy of the results for the *Total Recall Track Athome* participants.[12] Further, the *Team* understands that the TREC assessors work was much more time constrained than was the work of the *Team*. Moreover, unlike the *Team*, the TREC assessors did not have the benefit of SME input from a native Floridian lawyer (Losey) who was familiar with Florida politics and Governor Bush and, since 2015, had put substantial time reviewing this email collection.

The following chart contains a detailed comparison of recall, precision and F1 the *Team* attained based under both the TREC and *Team* assessments. Again, the Appendix search descriptions include a few examples of the kind of classification errors encountered. Again, the *Team* recognizes that no gold standard is ever perfect, including its own revised standards. The *Team* invites input from other participants and organizers of the Total Recall Track concerning relevance of any document. Upon request and agreement we will provide any participant or organizer with a confidential spreadsheet listing the *Team's* gold standard for each topic by identification of TREC ID Document Numbers. We invite any challenges and questions concerning relevance. The *Team* continues to believe in meaningfulness of relevance, true facts and the importance of a correct gold standard to any text retrieval experiment.

				TREC Standard				Revised Gold Standard				
Topic	Name	Reviewer	Total Relevant	Relevant Found	Recall	Precision	F1 Score	Total Relevant	Relevant Found	Recall	Precision	F1 score
401	Summer Olympics	Ralph	229	94	41.048%	73.438%	52.661%	137	126	91.971%	98.438%	95.094%
402	Space	Tony	638	463	72.571%	22.037%	33.808%	679	489	72.018%	38.054%	49.796%
403	Bottled Water	Ralph	1,090	78	7.156%	80.412%	13.142%	123	96	78.049%	98.969%	87.273%
404	Eminent Domain	Tony	545	125	22.936%	64.433%	33.829%	519	182	35.067%	93.814%	51.052%
405	Newt Gingrich	Ralph	122	116	95.082%	28.087%	43.364%	123	123	100.000%	29.782%	45.896%
406	Felon Disenfranchisement	Ralph	127	93	73.228%	66.906%	69.925%	203	197	97.044%	100.000%	98.500%
407	Faith Based Initiatives	Ralph	1,586	492	31.021%	68.715%	42.745%	1,654	1,465	88.573%	62.634%	73.378%
408	Invasive Species	Tony	116	64	55.172%	64.646%	59.535%	168	86	51.190%	86.869%	64.419%
409	Climate Change	Levi	202	171	84.653%	40.714%	54.984%	224	198	88.393%	47.143%	61.491%
410	Condominiums	Tony	1,346	1,280	95.097%	46.126%	62.121%	1,317	1,314	99.772%	47.351%	64.223%
411	Stand Your Ground	Ralph	88	59	66.292%	67.045%	66.667%	59	59	67.045%	67.045%	67.045%
412	2000 Recount	Tony	1,410	809	57.376%	49.179%	52.962%	850	747	87.882%	45.410%	59.880%
413	James V. Crosby	Jim	546	526	96.337%	89.002%	92.524%	600	581	96.833%	98.308%	97.565%
414	Medicaid Reform	Tony	839	770	91.657%	35.321%	50.992%	844	783	92.773%	35.917%	51.786%
415	George W. Bush	Jim	12,106	11,389	94.077%	91.039%	92.533%	12,267	11,554	94.188%	92.358%	93.264%
416	Marketing	Jim	1,446	872	60.304%	42.085%	49.574%	1,485	911	61.347%	43.967%	51.223%
417	Movie Gallery	Ralph	5,931	5,908	99.612%	99.378%	99.495%	5,945	5,945	100.000%	100.000%	100.000%
418	War Preparations	Tony	187	74	39.572%	50.340%	44.311%	141	114	80.851%	77.551%	79.167%
419	Lost Foster Child Rilya Wilson	Levi	1,989	1,966	98.844%	15.037%	26.104%	1,982	1,964	99.092%	15.022%	26.089%
420	Billboards	Jim	737	682	92.537%	92.162%	92.349%	739	707	95.670%	95.541%	95.605%
421	Traffic Cameras	Jim	21	19	90.476%	12.500%	21.965%	54	52	96.296%	34.211%	50.485%
422	Non Resident Aliens	Tony	31	29	93.548%	0.902%	1.786%	48	48	100.000%	1.493%	2.941%
423	National Rifle Association	Tony	286	146	51.049%	33.182%	40.220%	190	147	77.368%	33.409%	46.667%
424	Gulf Drilling	Levi	497	495	99.598%	22.759%	37.051%	495	493	99.596%	22.667%	36.929%
425	Civil Rights Act of 2003	Ralph	714	652	91.317%	96.593%	93.880%	718	653	94.364%	96.171%	95.259%
426	Jeffrey Goldhagen	Ralph	120	84	70.000%	87.500%	77.778%	98	91	92.857%	94.792%	93.814%
427	Slot Machines	Jim	241	215	89.212%	35.774%	51.069%	263	249	94.677%	41.431%	57.639%
428	New Stadiums and Arenas	Levi	464	432	93.103%	17.814%	29.907%	476	447	93.908%	18.433%	30.817%
429	Elian Gonzalez	Jim	827	779	94.196%	92.408%	93.293%	844	819	97.038%	97.153%	97.095%
430	Restraints and Helmets	Jani	991	713	71.948%	64.995%	68.295%	1,013	735	72.557%	67.001%	69.668%
431	Agency Credit Ratings	Tony	144	109	75.694%	47.598%	58.445%	149	120	80.537%	52.402%	63.492%
432	Gay Adoption	Jani	140	119	85.000%	86.232%	85.612%	137	125	91.241%	90.580%	90.909%
433	Abstinence	Jim	112	111	99.107%	66.071%	79.286%	141	141	100.000%	83.929%	91.262%
434	Bacardi Trademark	Ralph	38	33	86.842%	91.667%	89.189%	38	38	100.000%	100.000%	100.000%
	AVERAGE		1,056	881	75.461%	57.121%	65.022%	1,021	935	87.300%	64.937%	68.815%
	TOTALS		36,962					34,723				

The topics we found that had the largest assessor errors, and thus the largest changes in **Recall** measure at the time of reasonable call, are:

- Topic 401 Summer Olympics: 41.05% to 91.97%.
- Topic 403 Bottled Water: 7.16% to 78.05%.
- Topic 404 Eminent Domain: 22.94% to 35.07%
- Topic 406 Felon Disenfranchisement: 73.23% to 97.04%.
- Topic 407 Faith Based Initiatives: 31.02% to 88.57%.
- Topic 412 2000 Recount: 57.37% to 87.88%.
- Topic 418: War Preparations: 39.57% to 80.85%.
- Topic 421 Traffic Cameras: 90.48% to 96.30%.
- Topic 422 Non Resident Aliens: 94.55% to 100%.
- Topic 423 National Rifle Association: 51.05% to 77.37%.
- Topic 426 Jeffery Goldhagen: 70.00% to 92.86%.
- Topic 432 Gay Adoption: 85.00% to 91.24%.

- Topic 434 Bacardi Trademark: 86.84% to 100%.

The standards with the highest changes in recall measure are shown below with the percent of recall change for each and the percent of error in recall measurement. The large error rate seen in Topic 403 is an anomaly explained by the presence of one contested form email (*Protect Florida's Springs*) that had 913 near duplicates.[14] The error rates in other topics were also magnified to varying degrees for the same reason, the high prevalence of forms emails in the Jeb Bush collection.

- Topic 403 Bottled Water: 7.16% to 78.05%.
    - Change of 70.89%.
    - Error of 990%.
- Topic 407 Faith Based Initiatives: 31.02% to 88.57%.
    - Change of 57.55%.
    - Error of 186%.
- Topic 401 Summer Olympics: 41.05% to 91.97%.
    - Change of 50.92%.
    - Error of 124%.
- Topic 418 War Preparations: 39.57% to 80.85%.
    - Change of 41.28%.
    - Error of 104%.
- Topic 412 2000 Recount: 57.37% to 87.88%.
    - Change of 30.51%.
    - Error of 53%.
- Topic 423 National Rifle Association: 51.05% to 77.37%.
    - Change of 26.32%.
    - Error of 52%.
- Topic 426 Jeffery Goldhagen: 70.00% to 92.86%.
    - Change of 22.86%.
    - Error of 33%.

This data shows the importance of correctly judged gold standards and the impact of erroneous, inconsistent SME judgments upon the effectiveness of any search. The impact of the SME type errors seen here is exaggerated by the fact that the Bush collection contains an unusually high number of form emails. Further work on this research question is needed.

5.3 Research Question No. 3.

What is the most effective search method from the Team's multimodal tool-set for retrieval of relevant documents in the relatively simplistic search challenges presented by most, but not all, of the thirty-four topics.

For most of the topics in 2016 the *Team's* use of what it calls "tested, parametric, Boolean keyword search" was the most effective search method to find relevant documents.[15] The *Team* was surprised by how well a sophisticated use of keywords could locate nearly all the target relevant documents in many of the topics. This shows the continued importance of a multimodal approach to legal search, including especially keyword search, when done properly,[16] especially in simple lawsuits involving relatively easy search issues.

In *post hoc* research the *Team* ran keyword only searches across all topics. We did so to calculate the scores that the *Team* **would have accrued** in each topic, **if the *Team* had only run keyword searches**, and had not supplemented these searches with other types of search, including similarity, concept and predictive coding based searches. Below is a chart showing a comparison of the BMI (pure machine learning) results to the Keyword-only results. The uncorrected standard is here used because comparisons are not possible under the corrected standard. Comparisons under the corrected standard are not possible because no information has been provided by TREC as to the order of BMI document submissions. Without that information the BMI results under the corrected standard cannot be calculated. Since uncorrected data is used for the standard, the specific measurements here are not perfect, although we think these comparisons still provide useful information.

	BMI Results			Search Results		
	Recall	Precision	F1	Recall	Precision	F1
Summer Olympics	91.70%	15.31%	26.23%	79.91%	33.15%	46.86%
Space	91.07%	30.86%	46.09%	71.16%	17.61%	28.23%
Bottled Water	97.71%	37.49%	54.18%	93.76%	62.70%	75.15%
Eminent Domain	91.74%	26.55%	41.19%	51.93%	15.57%	23.95%
Newt Gingrich	99.18%	9.82%	17.87%	92.62%	63.48%	75.33%
Felon Disenfranchisement	92.91%	9.58%	17.37%	76.38%	7.39%	13.47%
Faith Based Initiatives	91.80%	41.86%	57.50%	87.70%	84.71%	86.18%
Invasive Species	83.62%	7.87%	14.39%	56.90%	32.35%	41.25%
Climate Change	95.05%	13.99%	24.40%	41.58%	15.03%	22.08%
Condominiums	99.48%	42.59%	59.64%	86.18%	35.80%	50.59%
Stand Your Ground	70.79%	5.70%	10.55%	51.69%	12.64%	20.31%
2000 Recount	91.35%	40.97%	56.57%	28.16%	13.49%	18.24%
James V. Crosby	99.08%	28.73%	44.55%	98.17%	69.34%	81.27%
Medicaid Reform	96.90%	35.10%	51.54%	63.53%	18.87%	29.10%
George W. Bush	63.39%	61.09%	62.22%	85.73%	86.87%	86.30%
Marketing	94.19%	43.32%	59.35%	42.81%	5.68%	10.04%
Movie Gallery	99.81%	57.28%	72.79%	99.51%	99.49%	99.50%
War Preparations	93.05%	12.68%	22.32%	43.85%	1.64%	3.17%
Lost Foster Child Rilya Wilson	93.06%	48.13%	63.44%	33.48%	34.33%	33.90%
Billboards	99.46%	31.65%	48.02%	84.26%	67.43%	74.91%
Traffic Cameras	100.00%	1.90%	3.73%	61.90%	13.40%	22.03%
Non Resident Aliens	100.00%	2.81%	5.46%	54.84%	25.00%	34.34%
National Rifle Association	99.65%	18.68%	31.46%	36.01%	45.58%	40.23%
Gulf Drilling	100.00%	26.39%	41.76%	67.00%	50.53%	57.61%
Civil Rights Act of 2003	98.60%	33.70%	50.23%	75.91%	87.42%	81.26%
Jeffrey Goldhagen	94.17%	9.17%	16.72%	65.00%	81.25%	72.22%
Slot Machines	96.68%	16.98%	28.89%	82.16%	25.65%	39.09%
New Stadiums and Arenas	98.49%	26.95%	42.31%	65.95%	32.24%	43.31%
Elian Gonzalez	99.27%	35.45%	52.24%	87.91%	66.45%	75.69%
Restraints and Helmets	94.25%	36.40%	52.52%	66.09%	30.66%	41.89%
Agency Credit Ratings	99.31%	11.61%	20.78%	65.97%	14.48%	23.75%
Gay Adoption	98.57%	11.20%	20.12%	77.14%	53.20%	62.97%
Abstinence	100.00%	9.09%	16.67%	99.11%	73.51%	84.41%
Bacardi Trademark	100.00%	3.44%	6.65%	81.58%	13.84%	23.66%
AVERAGES	**94.54%**	**24.83%**	**36.46%**	**69.29%**	**40.91%**	**47.72%**

As shown in the above chart, machine learning provided a substantially better recall almost across the board in comparison to keyword alone (it had a smaller recall on only one of the thirty-four topics). However, machine learning alone improved on precision on only ten of the topics versus Keyword, and improved on F-measure on only 11. This would be indicative of a typically broad classifier, in need of narrowing its scope. It suggests that keywords can play a beneficial role in the initial searches (Step Two in the *Team's* eight-step process, *Multimodal ECA*).

Keyword search is shown to have its own drawbacks. They were often far too narrow and could be adversely impacted by context of the terms. To that end, machine learning exceeds and excels at expanding the scope of documents to consider and returning only those sets that are pertain to the issue at hand.

Going beyond the *post hoc* experiment results, and based on our general experience, we see a contrast between a pure machine learning approach, and a hybrid multi-modal approach, that is described by *Team* member Tony Reichenberger as follows:

> A machine learning process takes the whole document set and seeks to narrow it down to find documents of relevance. A hybrid multi-modal approach starts by narrowly focusing on relevant documents to fuel machine learning, and then expands the set of documents to consider for relevance based on machine feedback.

5.4 Research Question No. 4.

The *Team* found that for the seven easiest topics in the 2016 *Total Recall Track* the primary role of active machine learning was reduced to a quality assurance function:

Topic 422 Non-Resident Aliens
Topic 413 James V. Crosby
Topic 417 Movie Gallery
Topic 434 Bacardi Trademark
Topic 426 Jeffrey Goldhagen
Topic 405 Newt Gingrich
Topic 411 Stand Your Ground

Predictive coding based searches of high ranking documents would in some of these topics uncover a few relevant documents not already located by keyword search, or concept and similarity search, and thus improve recall somewhat. In some active machine learning searches we did not find *any* new relevant documents. Instead the predictive coding searches only confirmed that all relevant documents had already been found by the other methods. Again, the description of those searches in the Appendix provides further details.

6. CONCLUSIONS

The *Team* has shown that it's standard method of document review, *Predictive Coding 4.0 Hybrid Multimodal* using continuous *Intelligently Spaced Training*, is extremely effective by all objective measures, including Recall, Precision, F1, project speed and effort. The *Team* method of finding relevant emails took an average of only 6.89 hours per project by review of an average of 124 documents reviewed per topic.

The *Team* classified 9,863,366 documents as either relevant or irrelevant in thirty-four review projects. A total of 34,723 were correctly classified as Relevant, as per the *Team's* judgment and corrected standard. The 34,723 relevant documents were found by manual

review of 6,957 documents, taking a total of 234.25 man-hours. The *Team* thus reviewed and classified documents at an average speed of 42,106 files per hour.

Even at these speeds and reviewer time limitations, and even with the handicap of having to omit three of the *Team* standard eight-step protocol (1-ESI Communications, 3- Random Prevalence, 7-ZEN QC), the *Team*'s average score across all thirty-four topics was: 88.17% Recall, 64.94% Precision and 69.15% F1. The *Team's* top ten projects attained remarkably high scores with an average of 95.66% Recall, 97.28% Precision and 96.42% F1. The *Team* attained an **average 88% Recall** score across all 34 topics using the corrected standard. The *Team* also attained **F1 scores** of greater than 90% in twelve topics, including two perfect scores of 100% F1. The *Team* cautions that these high scores in a short amount of time and other handicaps were only possible because of the ease of the searches and simplicity of the Bush email.

The *Team* found that the proper use of multimodal search, including especially keyword search, can, in the right case, with the right data, easy targets, and a skilled searcher and SME, be very effective, even *without* the use of active machine learning. For easy search challenges, such as those presented in the 2016 *Total Recall Track* topics, the primary role of active machine learning is reduced to a quality assurance function. Predictive coding can be used to verify that the other multimodal search methods have already found all relevant documents.

The success of the other methods alone, without predictive coding, was not expected. The *Team* knew from its experience in Legal Search that keyword search alone, even when done properly and even when supplemented by various passive analytic based searches, does not usually work well to attain high recall in search projects with complex relevance issues or with complex "dirty" data. These are the kind of searches that the *Team* typically works with every day in Legal Search. For complex projects active machine learning is required. In the more complex and difficult projects, using keyword search alone would be a significant danger. It can be very imprecise and can easily miss unexpected word usage and misspellings. That is one reason the *e-Discovery Team* always supplements keyword search with a variety of other search methods, including predictive coding. Still, our research in 2016 TREC has shown that tested, parametric Boolean keyword search alone can attain good recall and precision when there is simple data, clear targets and a skilled reviewer.

Finally, we found that a high number of errors made in relevance judgments by reviewers and SMEs, regardless of whether due to human carelessness or lack of expertise, can have a significant impact on the metrics evaluating the efficiency and effectiveness of a project. We do not have enough information yet to quantify this impact. Still, the data at hand confirms the commonsense *GIGO* notion that the impact of training errors can be significant and that the degree of impact varies according to the type and number of assessor errors. Much more research is needed in this area.

The assessor errors may have little or no impact on the metrics of the automatic *Sandbox* division participants in the Recall Track, where they anyway never look at documents, and are not concerned with *true relevance*, just with *matching* the TREC standard. Still, errors in TREC gold standard may also impact participants in the *Sandbox* division in some topics. Without a reliable standard, one that mirrors true relevance, and is so certified by diligent skilled humans, the auto-search exercises appear to be equivalent to *a snake eating its own tail*, an ***Ouroboros***.[17] Without a proper gold standard, the auto runs in the impacted topics may only measure the ability of one software program to follow and match another. It is like a deluded,

self-serving snake eating its own tail. This is a kind of *blind leading the blind* negative feedback loop. It does not measure the ability of the software to attain true recall of the target documents. It just measures the ability of one program to follow another.

7. ACKNOWLEDGMENTS

The *e-Discovery Team* would like to thank Kroll Ontrack, Inc. and Jackson Lewis P.C. for their generous support of this project. We would also like to thank the employees at Kroll Ontrack who pitched in behind the scenes and on weekends to help make this happen. Losey also thanks his wife, Molly, for once again sacrificing a summer vacation so he would have time to participate in this project.

8. REFERENCES (Endnotes)

[1] The Total Recall Track fully automated method uses a type of *monomodal* search method where only certain defined high-ranking documents are used for training. This method is more fully described in paper by the Total Recall Track Administrators, Grossman & Cormack, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*, CoRR abs/1504.06868 (2015). They call the method "*Autonomous TAR*." *Id.* at pg. 6.

[2] *E-Discovery For Everyone*, Ralph Losey; Foreword Judge Paul Grimm (ABA 2016*); Perspectives On Predictive Coding And Other Advanced Search Methods for the Legal Practitioner*; Editors: Jason R. Baron, Ralph C. Losey, Michael Berman; Foreword by Judge Andrew Peck (ABA 2016); *Adventures in Electronic Discovery* (West Thomson Reuters, 2011); *Electronic Discovery: New Ideas, Trends, Case Law, and Practices* (West Thomson Reuters, 2010); *Introduction to E-Discovery: New Cases, Ideas, and Techniques* (ABA 2009); *e-Discovery: Current Trends and Cases* (ABA 2008). *Also see Predictive Coding Articles by Ralph Losey,* (collection of over 60 articles by Ralph Losey further describing the hybrid multimodal approach) found at https://e-discovery*team*.com/doc-review/.

[3] Losey, R., Predictive Coding 4.0 (e-Discovery Team, 2016) found at https://e-discoveryteam.com/doc-review/predictive-coding-4-0/.

[4] The *e-Discovery Team's* hybrid multimodal approach relies upon and encourages participation of skilled reviewers in the search process, the *hybrid* approach. Our aim is *augmentation* of skilled attorneys to perform legal search, not *automation*, not replacement. In these respects the *e-Discovery Team* follows the teachings of Gary Marchionini, Dean of the School of Information and Library Sciences of U.N.C. at Chapel Hill, who explained in *Information Seeking in Electronic Environments* (Cambridge 1995) that information seeking expertise is a critical skill for successful search. Professor Marchionini argues, and we agree, that: "*One goal of human-computer interaction research is to apply computing power to amplify and augment these human abilities*." We also follow the teachings of UCLA Professor Marcia J. Bates who has advocated for a multimodal approach to search since 1989. Bates, Marcia J., *The Design of Browsing and Berrypicking Techniques for the Online Search Interface*, Online Review 13 (October 1989): 407-424. As Professor Bates explained in 2011 in *Quora*:

> *"An important thing we learned early on is that successful searching requires what I called "berrypicking." … Berrypicking involves 1) searching many different places/sources, 2) using different search techniques in different places, and 3) changing your search goal as you go along and learn things along the way. This may seem fairly obvious when stated this way, but, in fact, many searchers erroneously think they will find everything they want in just one place, and second, many information systems have been designed to permit only one kind of searching, and inhibit the searcher from using the more effective berrypicking technique."*

Also see: White & Roth, *Exploratory Search: Beyond the Query-Response Paradigm* (Morgan & Claypool, 2009).

[5] *Predictive Coding* is defined by The Grossman-Cormack Glossary of Technology-Assisted Review, 2013 Fed. Cts. L. Rev. 7 (January 2013) (*Grossman-Cormack Glossary*) as: "An industry-specific term generally used to describe a Technology Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant Documents, based on Subject Matter Expert(s) Coding of a Training Set of Documents. " A Technology Assisted Review process is defined as: "A process for Prioritizing or Coding a Collection of electronic Documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of Documents and then extrapolates those judgments to the remaining Document Collection. … TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness." *Also see*: Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, Richmond Journal of Law and Technology, Vol. XVII, Issue 3, Article 11 (2011).

[6] Da Silva Moore v. Publicis Groupe 868 F. Supp. 2d 137 (SDNY 2012) and numerous cases later citing to and following this landmark decision by Judge Andrew Peck, including another more recent opinion by Judge Peck, *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125 (S.D.N.Y. 2015). Losey was defense counsel in charge of the predictive coding review in *Da Silva.*

[7] Grossman & Cormack, Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, SIGIR'14, July 6–11, 2014; Grossman & Cormack, Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review", 7 Federal Courts Law Review 286 (2014); Herbert Roitblat, series of five OrcaTec blog posts (1, 2, 3, 4, 5), May-August 2014; Herbert Roitblat, Daubert, Rule 26(g) and the eDiscovery Turkey OrcaTec blog, August 11th, 2014; Hickman & Schieneman, The Implications of Rule 26(g) on the Use of Technology-Assisted Review, 7 FED. CTS. L. REV. 239 (2013); Losey, R. Predictive Coding 3.0, part one (e-Discovery Team 10/11/15).

[8] *Id*.; Webber, Random vs active selection of training examples in e-discovery (Evaluating e-Discovery blog, 7/14/14).

[9] Losey, R., Predictive Coding 4.0 – Nine Key Points of Legal Document Review and an Updated Statement of Our Workflow (e-Discovery Team, 9/12/16) (Part One of an Eight Part Series explaining the recent advancements from our Predictive Coding method from version 3.0 to version 4.0).

[10] The merits of the *Team's* approach to the timing of machine learning are detailed in *Predictive Coding 4.0* Part Two.

[11] Grossman & Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery,* SIGIR'14, July 6–11, 2014.

[12] Participant appeal rights could have mitigated the errors seen in 2016, but this can be burdensome and, as seen in those Tracks in 2008 and 2009, can create their own issues. *See*: Oard, Hedlin, Tomlinson, Baron, *Overview of the TREC 2008 Legal Track,* found at http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf; and Oard, Hedlin, Tomlinson, Baron, Oard, *Overview of the TREC 2009 Legal Track* found at: http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf.

[13] *See In re Fannie Mae Sec. Litig*., 552 F.3d 814 (D.C. Cir. 2009) ($9.09 per file cost for a privilege review, using contract lawyers and linear method. Total cost of $6,000.000 to review 660,000); Losey, *E-Discovery For Everyone* (ABA 2016), Chapter Three *Perspective on Legal Search and Document Review*.

[14] The full description of relevance for 403 is: "**403-Bottled Water** - All documents concerning the extraction of water in Florida for bottling by commercial enterprises." We disagreed with 1,038 TREC relevance classifications on this topic. We found that 1,001 documents they coded as relevant, were actually irrelevant under that definition, and 37 documents they coded as irrelevant, were actually relevant. The total count of relevant documents according to TREC was 1,089. In fact the *Team* found only 125 relevant documents. We found 121 of those relevant documents before reasonable was called. Four more documents were found after the call. The TREC SME assessors made only a few errors, but the errors were magnified because they were in near duplicate form emails. The primary error seen pertained to omission of the following relevance restriction statement on some, but not all, documents: "*for bottling by commercial enterprises*." TREC correctly judged some emails that *concerned the extraction of water in Florida*, but did not pertain to bottling, to be irrelevant. But TREC also TREC incorrectly judged some emails that *concerned the extraction of water in Florida*, but did not pertain to bottling, to be relevant. One such was the mentioned form email (*Protect Florida's Springs*), with over 913 near duplicates. This form only pertained to the use of water for commercial development, Florida springs and protection of Manatees (a perennial Florida favorite). The form email was unrelated to commercial bottling. There are only a couple of commercial bottlers in Florida and it is easy to identify them, if you know this. The TREC assessor sometimes ignored the commercial bottling qualifier, and sometimes did not. It was not a relevance decision. The single error on the *Protect Florida's Springs* form emails was magnified because of the number of copies (913) of this form. That explains the high error rate anomaly seen in topic 403, which was otherwise a very low prevalence topic with only 125 relevant documents. Without this one error the judging on the topic would not have been that bad. Since most datasets do not have so many form emails in them, this kind of multiplying error would not usually happen.

[15] For a detailed description see the section, *Tested, Parametric Boolean Keyword Search* in Predictive Coding 4.0 (e-Discovery Team, 2016).

[16] In the legal profession keyword searches are often performed by unskilled attorneys in a very unsophisticated "improper" manner. They frequently simply guess as to what words are important and do not first test the words nor study the dataset. Also, they rarely used Boolean logic, nor limit the searches to specific document parameters. *Child's Game of "Go Fish" is a Poor Model for e-Discovery Search*, Losey, R., *Adventures in Electronic Discovery, 209-211* (West 2011) at pgs. 204-210. *Also See: William A. Gross Constr. Assocs., Inc. v. Am. Mutual Mfrs. Ins. Co.*, 256 F.R.D. 134, 134 (S.D.N.Y. 2009).

[17] The *Ouroboros* is an ancient symbol of continuing self-reference and recursivity, and thus an apt symbol, although not necessarily positive, for the iterative cycles in active machine learning. This symbol has been used before in machine learning. *See eg.:* Knud Thomsen, *The Ouroboros Model in the light of venerable criteria*, Journal Neurocomputing archive, Vol. 74 Issue 1-3, December, 2010, pgs. 121-128; Thomsen, *Flow of Activity in the Ouroboros Model*, arXiv:0903.5054 [cs.AI] (2009) found at https://arxiv.org/pdf/0903.5054v1.pdf. *Also see*: *Wikipedia*, *Self-reference* found at: https://en.wikipedia.org/wiki/Self-reference, and *Recursivity* found at https://en.wikipedia.org/wiki/Recursion. The first description in the West of the *ouroboros* can be found in Plato in the Dialogue *Timaeus.* The *ouroboros* is described as the first living thing created in the universe which:

> had no need of eyes because there was nothing outside of him to be seen; nor of ears because there was nothing to be heard; ... his own waste providing his own food, and all that he did or suffered taking place in and by himself. For the Creator conceived that a being which was self-sufficient would be far more excellent than one which lacked anything.

Plato, *Timaeus*, found at http://classics.mit.edu/Plato/timaeus.html. This is a danger inherent in any fully automated document ranking system. Losey, *Why the 'Google Car' Has No Place in Legal Search* (e-Discovery Team, February 2016) (caution against over reliance on fully automated methods of active machine learning) found at: https://e-discoveryteam.com/2016/02/24/why-the-google-car-has-no-place-in-legal-search/.

APPENDIX

TREC Total Recall Track 2016
e-Discovery Team
Ralph C. Losey

E-Discovery Team *Narrative Report*
of All Thirty-Four Topic Searches

This Appendix Narrative Report describes the search of all thirty-four Total Recall topics in TREC 2016 using the *e-Discovery Team's* Hybrid Multimodal method. The searches are reported here numerically by Topic number, except for topic 434 Bacardi Trademark. We did not review the topics in numerical order. The first project was started on June 7, 2016 by Losey. It was *topic 434 Bacardi Trademark*. The last *Topic 415 George W Bush,* concluded on August 30, 2016 by Sullivan. We report on the first topic we reviewed first to provide a background and further information as to why we went to the drastic step of correcting the standard.

The summaries were prepared by the attorney who ran that topic.

At the beginning of each Topic the results are reported for that Topic. Each has the same form and discloses metrics at the times when: (1) the Reasonable call was made; and, (2) the point where 97.5% Recall was attained. They are summarized along with a variation of a standard *Confusion Matrix*, a/k/a *Contingency Table*. The Confusion Matrix itself is highlighted in blue. It is followed by a list of the key the values attained: **Recall, Precision, F1 Measure, Accuracy, Error, Elusion and Fallout**.

Due to the poor judging by TREC Assessors as to relevant documents in some topics, we were forced to try to note the documents incorrectly judged in all topics. We provide a very short discussion of the some of the errors. We also provide corrected statistics of these topics to show how our *Team* did when a correct standard was used. The true, corrected measures were dramatically different in some topics.

The actual review counts shown in these counts do not include documents reviewed after submission. Each document returned by TREC with an unexpected coding was examined to try to guess the scope of relevance used in a topic, or determine if the adjudication was in error, the later being an all too frequent experience for *Team* members.

Topic 434 - Bacardi Trademark
Total Documents: 290,099
Total Relevant: 38
Total Prevalence: 0.01%

Confusion Matrix - Bacardi Trademark

	@Reasonable	@90% Recall	@95% Recall
True Positives	*38*	*35*	*37*
True Negatives	*290,061*	*290,061*	*290,061*
False Positives	*0*	*0*	*0*
False Negatives	*0*	*3*	*1*
Recall	100.00%	92.11%	97.37%
Precision	100.00%	100.00%	100.00%
F1 Measure	100.00%	95.89%	98.67%
Accuracy	100.00%	99.9990%	99.9997%
Error	0.0%	0.0010%	0.0003%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	0.00%

Topic 434 - Bacardi Trademark - UNCORRECTED
Total Documents: 290,099
Total Relevant: 38
Total Prevalence: 0.01%

Confusion Matrix - Bacardi Trademark

	@Reasonable	@90% Recall	@95% Recall
True Positives	*33*	*35*	*37*
True Negatives	*290,058*	*290,058*	*289,659*
False Positives	*3*	*3*	*402*
False Negatives	*5*	*3*	*1*
Recall	86.84%	92.11%	97.37%
Precision	91.67%	92.11%	8.43%
F1 Measure	89.19%	92.11%	15.51%
Accuracy	100.00%	100.00%	99.86%
Error	0.00%	0.00%	0.14%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	0.14%

Summary

The TREC Total Recall project commenced on June 7, 2016 with work on Topic 434 Bacardi Trademark. This topic was run by Losey. He completed work on June 8, 2016 after spending a total of four hours on the project. In the course of the project he reviewed a total of 107 documents.

Errors in Gold Standard

Unfortunately, multiple obvious errors in TREC's judging of relevant documents were immediately encountered. Although there only 38 relevant documents found, a quick review of the 38 documents TREC called relevant shows that three are not relevant. They have nothing whatsoever to do with this topic. These three (two including a duplicate) obviously irrelevant documents have TREC ID number: 119771, 005283 (duplicate of 119771), 147890. Three more documents (two including a partial duplicate chain email) are relevant to this topic, but were called irrelevant by TREC. Their TREC ID numbers are: 110559, 110507 (same chain as 110559), 126174.

The error in calling two documents relevant, that are obviously irrelevant, suggests a failure of quality control and over-reliance on software. Since there were so few relevant documents – 38 - it would only have taken a few minutes to review them all. Anyone would quickly see that three (two plus a duplicate) of the documents were erroneously identified by the software to be relevant. We understand the assessors used Sofia-ml software to find the relevant documents, or software close thereto, just like most of the auto-run participants. The TREC assessors also supposedly verified the software's predictions with quality control efforts. We assume this meant a human actually looking at the documents. Obviously this human review control check did not happen here for some reason or they would have seen that 119771, 005283 (duplicate of 119771), 147890 were not relevant.

The failure of the assessors and *Sofia-ml software* (this software was used in 2015 and we assume was used again in 2016) to find the three relevant documents missed (actually only two, plus a chain) is easier to understand. That is simply a failure of the search software and the human search expert, the TREC assessors, who directed the search (assuming that there was in fact human assessor involvement, and TREC did not simply rely on automated procedures). An error in finding relevant documents is a result of skill and software deficiencies, not carelessness. Still, the net result in a low prevalence project like this of six errors is very significant – 16% (6/38).

It is important to note that these errors are not merely disagreements as to relevance. In other topics we did encounter close calls that we disagreed with, but we could see had a rational basis. They were not obvious mistakes. We did not adjust the standards for such opinion divergences. In other topics we encountered many documents where duplicates or near duplicates of the same document were coded inconsistently. There is no question that some of them were coded incorrectly.

The differences in judgment reported here are all obvious errors or errors of consistency. All close calls were granted to TREC, as is appropriate, but these obvious bloopers should not stand. The *e-Discovery Team* protested the many obvious errors it saw in the 2015 Total Recall Track, and made some public comments thereon in its reports. We participated again in 2016 based on assurances that the quality control and judgments would be improved. We are unhappy to report that although there has been some improvement, it appears to be very spotty. Errors in gold-standard judgments were again made in 2016 that have consequences on metrics, especially in the low prevalence topics that are common in the Total Recall Track.

These errors have little or no impact on the metrics of the automatic group participants, where they anyway never look at documents, and are not concerned with true relevance, just with matching the TREC standard. Still, a flawed gold standard does impact the validity of comparisons between ad hoc participants, such as our *Team*, where human searchers actually look at and evaluate the relevance of documents, and the auto run participant results. Moreover, without a valid objective standard, one that corrects for computer errors, the auto-search exercise would just be like a dog chasing its own tail. All it measures is the ability of one software program to follow and match another. It does not measure the ability of the software to attain true recall of the target documents.

In Losey's view the Bacardi Trademark issue was a relatively simple search, as explained further below. After correcting for the six obvious errors described above, Losey actually scored a perfect run on this issue with 100% Recall and 100% Precision as shown below.

Topic 434 - Bacardi Trademark

Total Documents: 290,099
Total Relevant: 38
Total Prevalence: 0.01%

Confusion Matrix - Bacardi Trademark

	@Reasonable	**@90% Recall**	**@95% Recall**
True Positives	*38*	*35*	*37*
True Negatives	*290,061*	*290,061*	*290,061*
False Positives	*0*	*0*	*0*
False Negatives	*0*	*3*	*1*
Recall	100.00%	92.11%	97.37%
Precision	100.00%	100.00%	100.00%
F1 Measure	100.00%	95.89%	98.67%
Accuracy	100.00%	99.9990%	99.9997%
Error	0.0%	0.0010%	0.0003%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	0.00%

Description of Search Process

Although it may seem fast to some readers to see a review of 290,099 documents completed by one attorney in only four hours, please note that this time did not include time spent prior to the search and outside of this topic. This includes time on such things as general set-up, procedures, project orientation, and communication protocols. The time reported also does not include the time note taking and report creations.

Aside from encountering several obvious errors in judging this topic, this was an interesting search project. The only information provided by TREC of Topic 434 was as follows:

> **Bacardi Trademark Lobbying** - Documents related to the Jeb Bush administration's involvement in a trademark dispute between Bacardi and the U.S. Patent and Trademark Office.

Losey chose this topic as he assumed it would be an easy topic for him to start with. Losey is an attorney in Florida with 36 years of legal experience, including a background in trademark law and analysis. Also, he is a native and sixty-five year resident of Florida who remembers well the Jeb Bush years and is familiar with many of the characters and issues mentioned in the Jeb Bush email.

Based on the description of this issue Losey hoped that the search would require some legal analysis and background. As it turned out, only a limited amount of such legal analysis and knowledge of trademark law and procedures was required, but it did help, especially in his full understanding of the relevant documents. From his perspective, this was a relatively easy search, even without legal or local knowledge. He found it comparable to legal search project in a simple, one issue lawsuit that had an easily defined target.

Losey began the project with a 30 minute Google search. Actually, the search itself took 3 minutes. The remaining 27 minutes were spent studying a political newspaper article that Losey knew from experience would likely be authoritative and complete. This provided important background information and was the equivalent to the Step One in the *Team's* standard Hybrid Multimodal workflow.

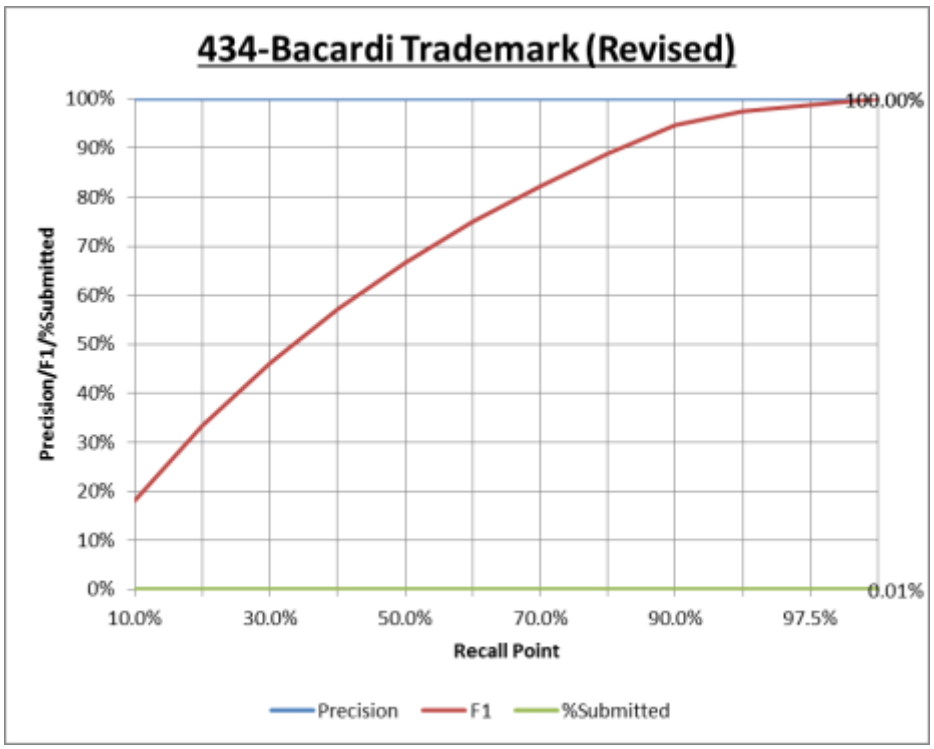
Based on this one newspaper article Losey identified the key persons involved, the time-line, and the key words likely to appear in any relevant documents, Based on that he formulated multiple keyword searches. The next day, June 8, 2016, he began Step Two, *Multimodal Search Reviews*. Losey spent two hours using parametric Boolean keyword searches. The searches were refined and new terms added based upon the documents seen. In this step 2 multimodal search review Losey found 37 of the 38 relevant documents found. A similarity search found one additional document. A concept search led to nothing new.

To summarize, the initial keyword and similarity searches conducted in step 2 found all 38 of the relevant documents in this collection. Losey spent another 1.5 hours in the submission process running multiple active machine learning training sessions, which is steps 4, 5 and 6 in our standard workflow. These did not lead to the discovery of any new

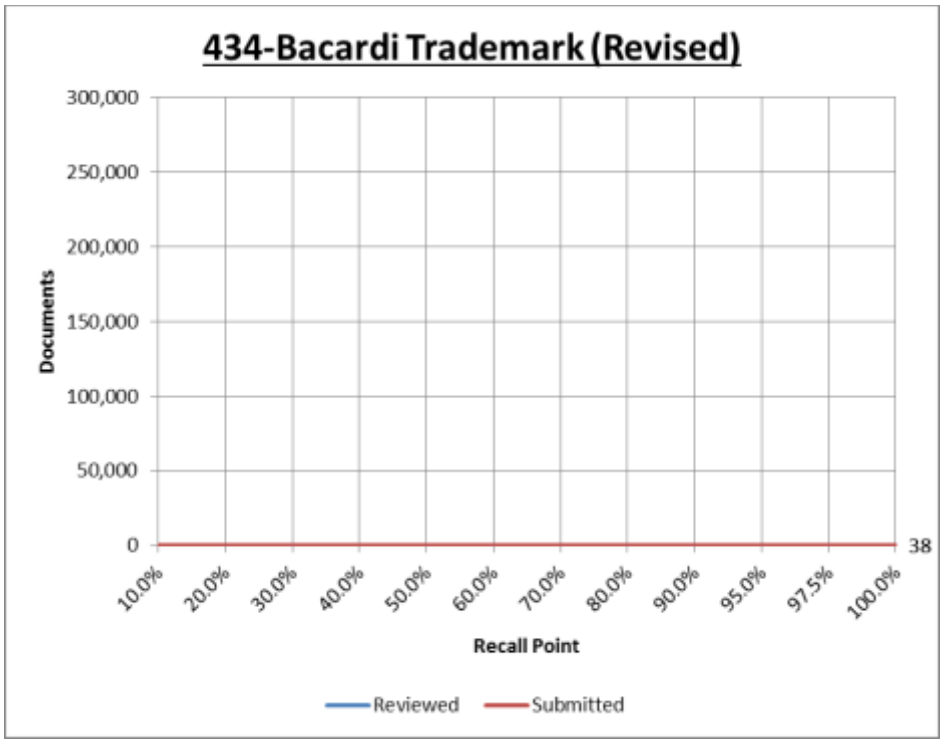
documents, but did serve as an expedited quality control measure to verify that the keyword and similarity searches had in fact uncovered all relevant documents. Steps 3 and 7 were skipped for three reasons: (1) to save time; (2) because Losey did not consider these additional quality control-assurance steps to be necessary in this simple project; and, (3) the predictive coding document-ranking work, where high-ranking documents were reviewed by Losey and coded as irrelevant, served as an effective quality assurance measure.

Graphs

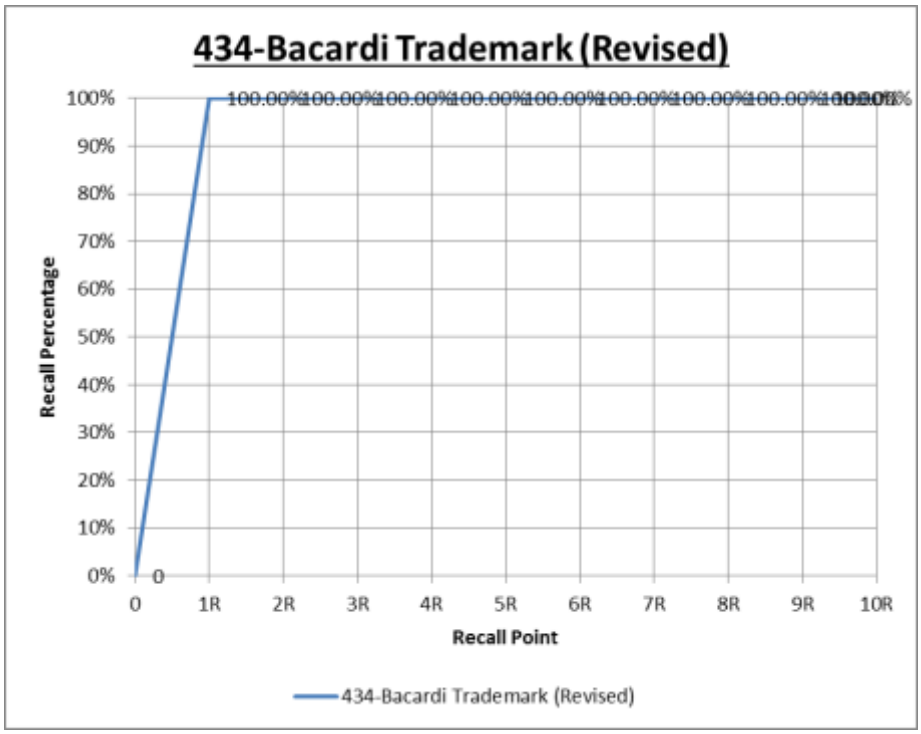
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Bacardi Trademark topic, the 90% recall threshold had been attained by submitting only 0.01%% of the corpus, 35 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 401 - Summer Olympics
Total Documents: 290,099
Total Relevant: 137
Total Prevalence: 0.05%

Confusion Matrix - Summer Olympics

	@Reasonable	@90% Recall	@95% Recall
True Positives	*126*	*124*	*131*
True Negatives	*289,960*	*289,960*	*289,950*
False Positives	*2*	*2*	*12*
False Negatives	*11*	*13*	*6*
Recall	91.97%	90.51%	95.62%
Precision	98.44%	98.41%	91.61%
F1 Measure	95.09%	94.30%	93.57%
Accuracy	99.9955%	99.9948%	99.9938%
Error	0.0045%	0.0052%	0.0062%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	0.00%

Topic 401 - Summer Olympics - UNCORRECTED
Total Documents: 290,099
Total Relevant: 229
Total Prevalence: 0.08%

Confusion Matrix - Summer Olympics

	@Reasonable	@90% Recall	@95% Recall
True Positives	*94*	*207*	*218*
True Negatives	*289,836*	*272,397*	*173,073*
False Positives	*34*	*17,473*	*116,797*
False Negatives	*135*	*22*	*11*
Recall	41.05%	90.39%	95.20%
Precision	73.44%	1.17%	0.19%
F1 Measure	52.66%	2.31%	0.37%
Accuracy	99.94%	93.97%	59.74%
Error	0.06%	6.03%	40.26%
Elusion	0.05%	0.01%	0.01%
Fallout	0.01%	6.03%	40.29%

Summary

Topic 401 was run by Losey, who started on July 15th, 2016 and ended on August 5th, 2016. He manually categorized 319 documents and studied 261 documents during the course of the 8 hours he spent on this project. The review was very much an *on and off again* type of project extending over three weeks. This is a poor way to do document review, necessitated by time demands at work, and probably did impact the results.

The full description provided as a relevance guide for this topic is: **Summer Olympics - All documents concerning a bid to host the Summer Olympic Games in Florida.**

Losey found this topic very interesting. The 2016 Olympics were on television at the same time. And he was fascinated that Florida had even made the attempt of Florida to bid on the 2012 Olympics back in 2001 because he had never heard of that. This was an effort by Tampa that received very poor press and only lukewarm political support by Central Florida, where Losey lives. It was interesting to learn from the Bush emails that the main reason Tampa lost the bid, and was disqualified early on, was the threat of Hurricanes. This is turn was triggered by the fact that Hurricane Cassandra threatened when the site committee was visiting. The two finalists were San Francisco and NYC, and NYC was selected as the bid City for the US. Of course, it did not get the 2012 Summer Olympics either. London did.

Multimodal review was done as usual, primarily by keywords (i.e, - "Olympi*"), similarity and predictive coding. The keyword searches were very effective in this topic in part because the main organizer of the Olympic bid was a man named *Turanchik*, which is novel name in Florida. Also, many of the emails with the word Olympic were relevant, but far from all. Losey would usually focus on ranking searches seen in the keyword folders.

There were several twists and turns that make the relevance hunt somewhat challenging (not totally simplistic, like many of the other topics). Mr. EDR has a role to play here, although I think most of what Losey found could have been found via keyword, and the rest by brute force by well-trained reviewers. Still, the AI made it much more efficient and is served as a good QC pushing up the scores attained here.

By these method Losey found a total of 127 documents at the time of reasonable call. Losey had submitted 129 documents as probable relevant at that point. Two of these submissions were later seen to be irrelevant and thus mistakes on Losey's part. The reasonable call was made after the eighth submission. The reporting for some reason is in error on this topic as it only shows 126 relevant found by that time, not 127. There were nine more submissions were made after the reasonable call. In these post call submissions 10 documents were returned by TREC as relevant that were relevant, or at least arguably so, and were not previously found by my search. The record incorrectly says 11 were found post call. The actual recall here was 92.7%, not the 91.97 shown above, but this error was found too late to correct and is anyway very minor.

For an example of two documents that Losey first considered them to be irrelevant, but later changed his mind, consider the emails bearing our Control # 3006405 and 3006419.

Based upon TREC's classification of these documents as relevant, we determined that Losey had made a mistake to classify them a relevant. The emails do not mention the Olympics, but do mention the Florida organizer, Turanchik. Upon closer study it is apparent that the emails did pertain to the Summer Olympics site committee, and so these two emails should be relevant. TREC got those emails right, but the errors usually went the other way.

TREC made many errors on this topic. As an example, many emails directly relevant to the Florida Olympics bid had to do with building certain trains and roads. The construction was needed for Olympic hosting infrastructure. TREC would often classify as relevant other emails concerning road and train construction, even though they had nothing to do with the Olympics. A human would have understood the difference, but these emails were obviously never read by a human assessor, just predicted by TREC's AI. We would run into errors like this all of the time in some topics like this, such that we began to play a game to hold our interest to try to figure out why the TREC AI made classification mistakes. It is sort of like reverse engineering from the often errors seen. We found that many of the obvious bloopers TREC made concerned relevant information *not present at the beginning* of an email. Instead, the relevant sections were found in the middle or end of a document. TREC's classifier algorithm seems to be front-ended, plus we suspect the human quality control did not look past the first couple of sentences either.

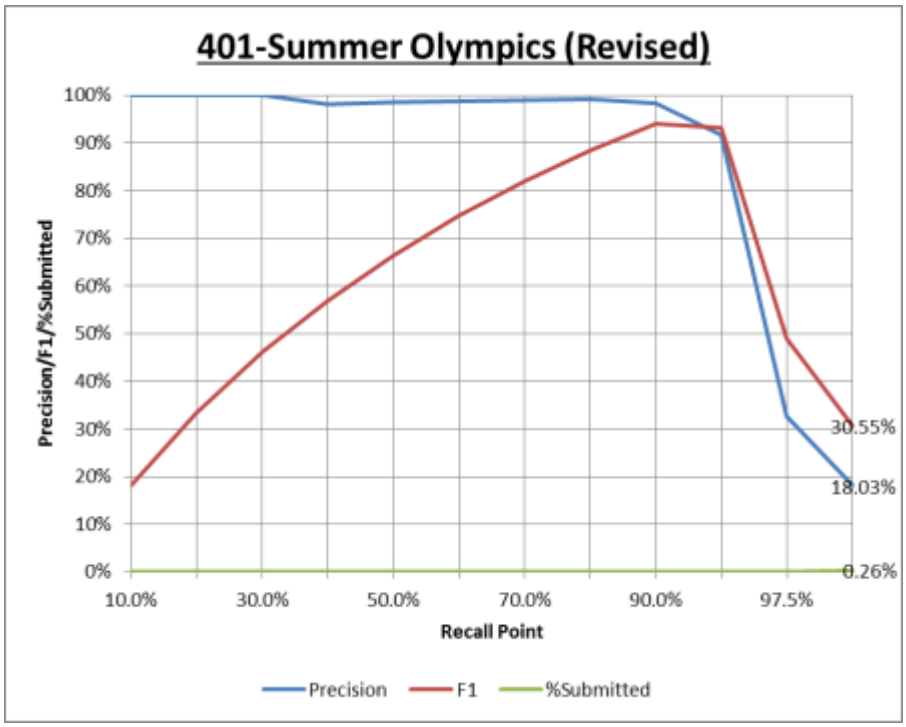
Another TREC error seen many times is the classification of an email as relevant, just because it had the word Olympics (especially near the front of an email), even though the word did not refer to the topic of Summer Olympics as required. An example is a reference seen many times to the Special Olympics, an event that did take place in Florida, but at a different time and place.

As an example of inconsistent coding by TREC, consider Control # 4600522 and Control # 4600409. The first is an email report on a Senate Bill - SB 1806 - that pertains to an aspect of funding related to the Olympic Committee. TREC correctly called the email relevant, which was a good catch. But then TREC incorrectly classified as irrelevant Jeb's email reply to the report, which simply said "thanks Pam" but otherwise included the original email from Pam giving the legislative report. We frequently ran into things like that.

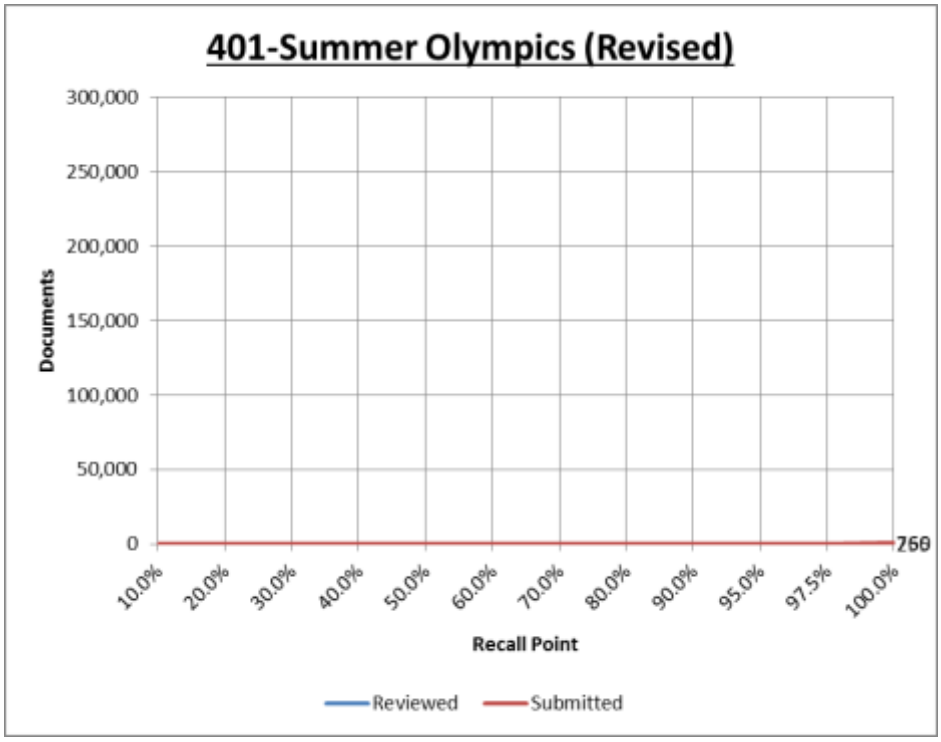
One error made by TREC assessors on the gold standard here was somewhat funny. It is an email on Project **Olympus**, dated in 2003. This is long after Florida gave up on the Summer Olympics (2001), and of course, its Olympus, not Olympics. Turns out it pertains to a Boeing airplane assembly plant they were trying to get in Jacksonville. Lots of similar language as in getting the Summer Olympics venue, but this had to do with getting Boeing to build a plant. Any human who actually read the email would see the error right away, but this was beyond the grasp of the machine learning TREC employed here.

Graphs

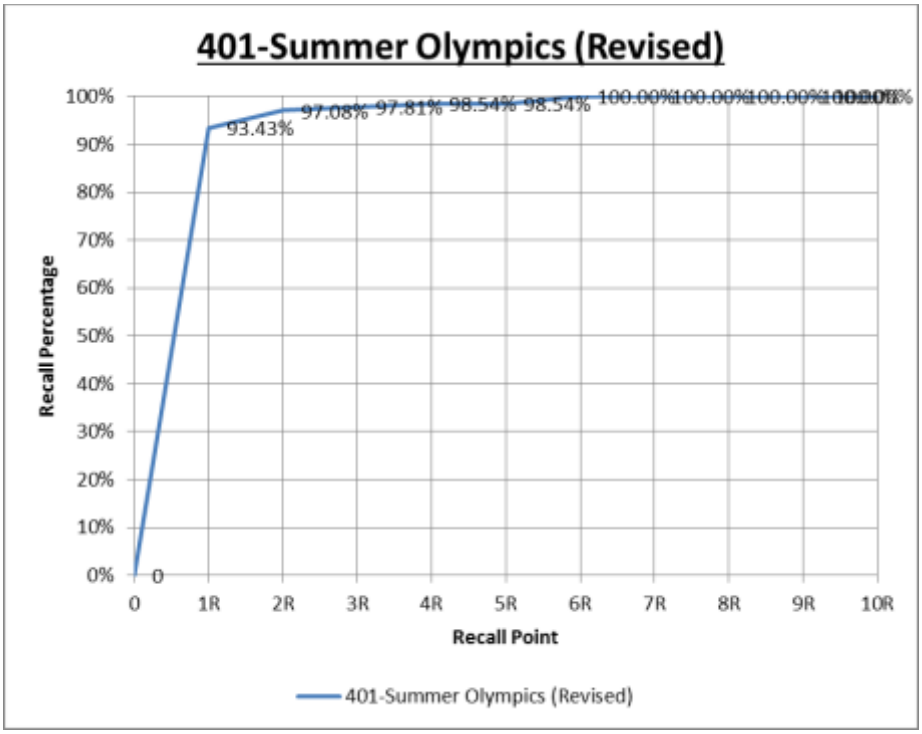
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Summer Olympics topic, the 90% recall threshold had been attained by submitting only 0.04%% of the corpus, 126 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc.).



Topic 402 – Space
Total Documents: 290,099
Total Relevant: 679
Total Prevalence: 0.23%

Confusion Matrix - Space

	@Reasonable	@90% Recall	@95% Recall
True Positives	*489*	*612*	*646*
True Negatives	*288,624*	*286,907*	*285,667*
False Positives	*796*	*2,513*	*3,753*
False Negatives	*190*	*67*	*33*
Recall	72.02%	90.13%	95.14%
Precision	38.05%	19.58%	14.69%
F1 Measure	49.80%	32.18%	25.44%
Accuracy	99.6601%	99.1106%	98.6949%
Error	0.3399%	0.8894%	1.3051%
Elusion	0.07%	0.02%	0.01%
Fallout	0.28%	0.87%	1.30%

Topic 402 – Space - UNCORRECTED
Total Documents: 290,099
Total Relevant: 638
Total Prevalence: 0.22%

Confusion Matrix – Space

	@Reasonable	@90% Recall	@95% Recall
True Positives	*463*	*575*	*607*
True Negatives	*287,823*	*285,622*	*277,407*
False Positives	*1,638*	*3,839*	*12,054*
False Negatives	*175*	*63*	*31*
Recall	72.57%	90.13%	95.14%
Precision	22.04%	13.03%	4.79%
F1 Measure	33.81%	22.76%	9.13%
Accuracy	99.38%	98.65%	95.83%
Error	0.62%	1.35%	4.17%
Elusion	0.06%	0.02%	0.01%
Fallout	0.57%	1.33%	4.16%

Summary

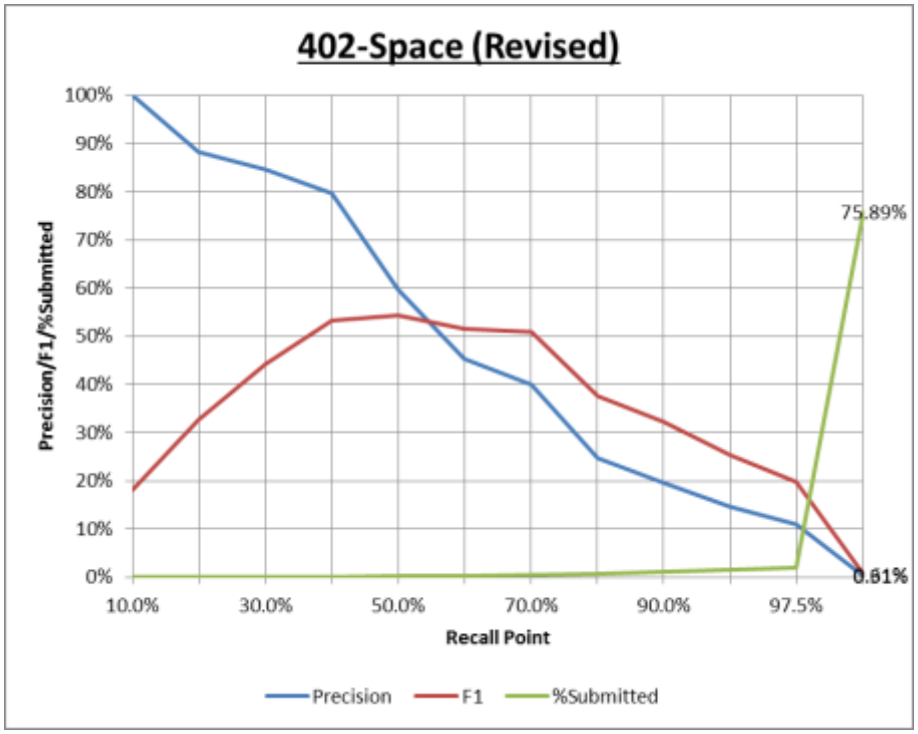
This project was conducted by Tony Reichenberger. The full description of the topic is: **Space-All documents concerning the space industry, the space program, space travel (whether manned or unmanned, public or private), and the study or exploration of space in Florida.**

The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. When scores dropped to 5%, a final search for "space" was submitted another learning session run, and documents were submitted in probability order.

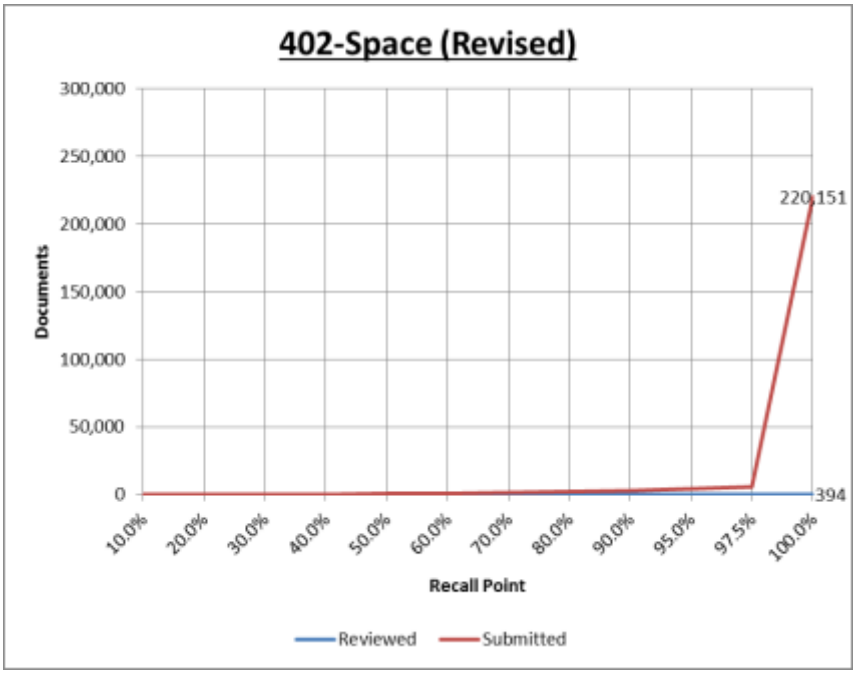
The reasonable call was made when following a learning session all remaining documents had scores less than 12.5%.

Graphs

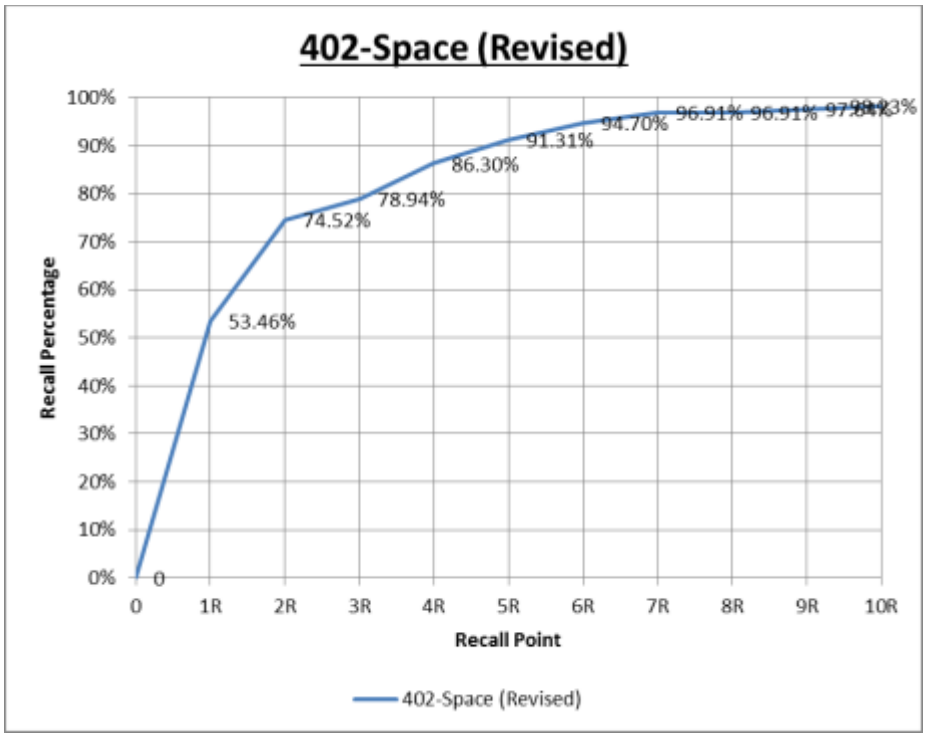
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Space topic, the 90% recall threshold had been attained by submitting only 1.08%% of the corpus, 3,125 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc.).



Topic 403 - Bottled Water
Total Documents: 290,099
Total Relevant: 123
Total Prevalence: 0.04%

Confusion Matrix - Bottled Water

	@Reasonable	@90% Recall	@95% Recall
True Positives	*96*	*111*	*117*
True Negatives	*289,975*	*289,975*	*289,975*
False Positives	*1*	*1*	*1*
False Negatives	*27*	*12*	*6*
Recall	78.05%	90.24%	95.12%
Precision	98.97%	99.11%	99.15%
F1 Measure	87.27%	94.47%	97.10%
Accuracy	99.9903%	99.9955%	99.9976%
Error	0.0097%	0.0045%	0.0024%
Elusion	0.01%	0.00%	0.00%
Fallout	0.00%	0.00%	0.00%

Topic 403 - Bottled Water - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,090
Total Prevalence: 0.38%

Confusion Matrix - Bottled Water

	@Reasonable	@90% Recall	@95% Recall
True Positives	*78*	*981*	*1,036*
True Negatives	*288,990*	*288,870*	*288,866*
False Positives	*19*	*139*	*143*
False Negatives	*1,012*	*109*	*54*
Recall	7.16%	90.00%	95.05%
Precision	80.41%	87.59%	87.87%
F1 Measure	13.14%	88.78%	91.32%
Accuracy	99.64%	99.91%	99.93%
Error	0.36%	0.09%	0.07%
Elusion	0.35%	0.04%	0.02%
Fallout	0.01%	0.05%	0.05%

Summary

This project was run by Losey from June 11th to June 15th 2016. He spent six hours on the project, personally reviewed 218 documents and manually categorized 1,126. He called reasonable after nine submissions and made a total of nineteen submissions.

The full description for the topic is: **Bottled Water - All documents concerning the extraction of water in Florida for bottling by commercial enterprises.** Again this topic was interesting to Losey because the extraction of Florida's precious water aquifer from spring water, for the purpose of sales of bottled water around the world, takes place near where he lives in Florida. He is also politically opposed to this since Nestle does so without payment for the water, just because they own land near a spring, and he contends it should be preserved for Floridians, or at the very least, Nestle should be charge full value for the state's critical resource. In spite of general familiarity with the situation, Losey began his work by Google searches to find out the names and other details of this controversial topic.

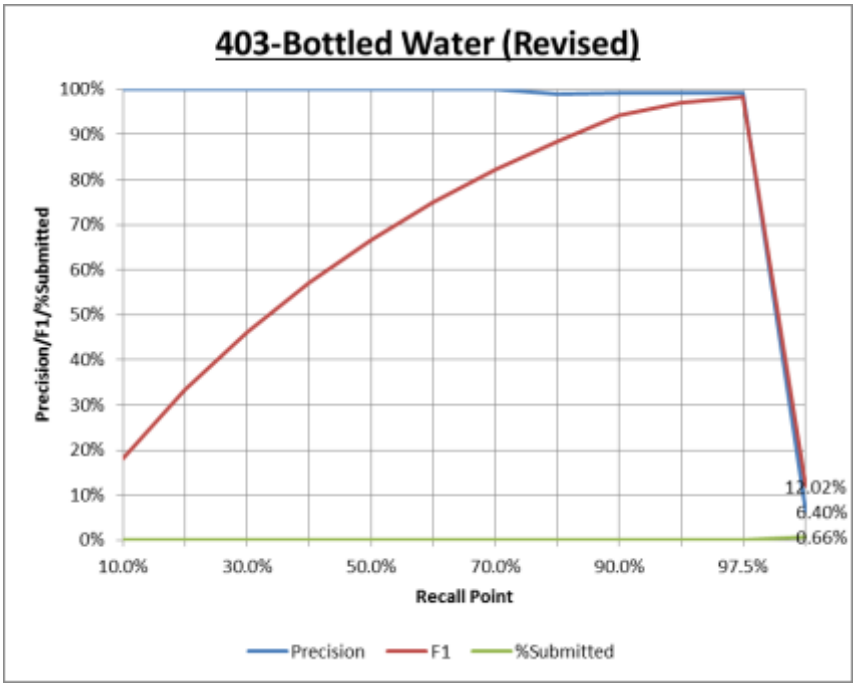
Usual Multimodal approach was used in what proved to be a simple keyword search type project. The people involved in this issue were well defined and distinct. No AI was used except for quality assurance purposes.

As described in the *Team's* Final Report (fn 14) the large error rate seen in Topic 403 is an anomaly explained by the wrong call of one contested form email (*Protect Florida's Springs*) that had 913 near duplicates. Losey knew this form email had that many copies and so submitted a test submission before submitting the rest. He submitted a test expecting it to come back irrelevant because the email did not pertain to bottling. In the test the form came back as irrelevant, as it should have. But, as it turned out, that test was deceiving, because on most copies of this form TREC incorrectly classified it a relevant.

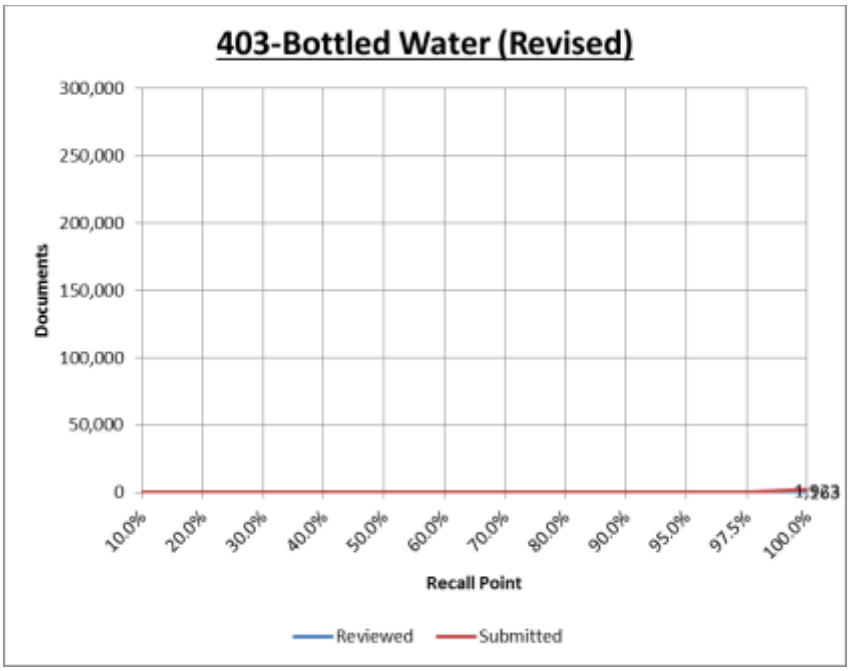
TREC's many other errors in judging this project appeared to be either completely off, just random error, or based upon calling a document relevant just because it mentioned extraction of water from Florida, even though the extraction was not for purposes of *bottling by commercial enterprises.*

Graphs

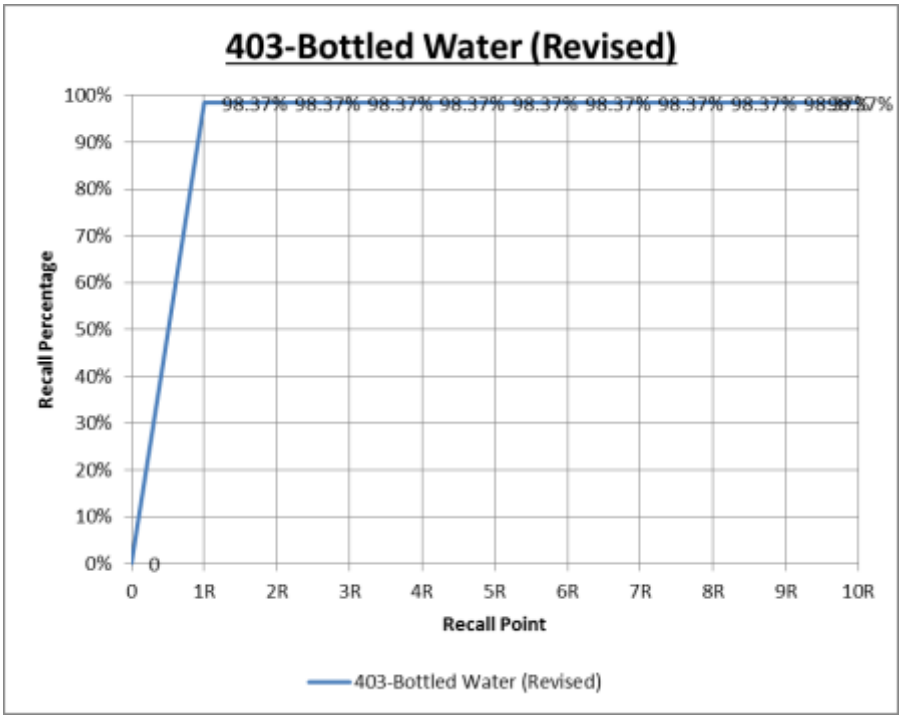
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Bottled Water topic, the 90% recall threshold had been attained by submitting only 0.04%% of the corpus, 112 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


403-Bottled Water (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
98.37% 98.37% 98.37% 98.37% 98.37% 98.37% 98.37% 98.37%
0
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
403-Bottled Water (Revised)

Topic 404 - Eminent Domain
Total Documents: 290,099
Total Relevant: 519
Total Prevalence: 0.18%

Confusion Matrix - Eminent Domain

	@Reasonable	@90% Recall	@95% Recall
True Positives	*182*	*468*	*494*
True Negatives	*289,568*	*287,446*	*285,864*
False Positives	*12*	*2,134*	*3,716*
False Negatives	*337*	*51*	*25*
Recall	35.07%	90.17%	95.18%
Precision	93.81%	17.99%	11.73%
F1 Measure	51.05%	29.99%	20.89%
Accuracy	99.8797%	99.2468%	98.7104%
Error	0.1203%	0.7532%	1.2896%
Elusion	0.12%	0.02%	0.01%
Fallout	0.00%	0.74%	1.28%

Topic 404 - Eminent Domain - UNCORRECTED
Total Documents: 290,099
Total Relevant: 545
Total Prevalence: 0.19%

Confusion Matrix - Eminent Domain

	@Reasonable	@90% Recall	@95% Recall
True Positives	*125*	*491*	*518*
True Negatives	*289,485*	*283,179*	*249,999*
False Positives	*69*	*6,375*	*39,555*
False Negatives	*420*	*54*	*27*
Recall	22.94%	90.09%	95.05%
Precision	64.43%	7.15%	1.29%
F1 Measure	33.83%	13.25%	2.55%
Accuracy	99.83%	97.78%	86.36%
Error	0.17%	2.22%	13.64%
Elusion	0.14%	0.02%	0.01%
Fallout	0.02%	2.20%	13.66%

Summary

The project was run by Tony Reichenberger. The full description of this topic is: **Eminent Domain-All documents concerning the legality or morality of expropriating land in Florida for commercial development.**

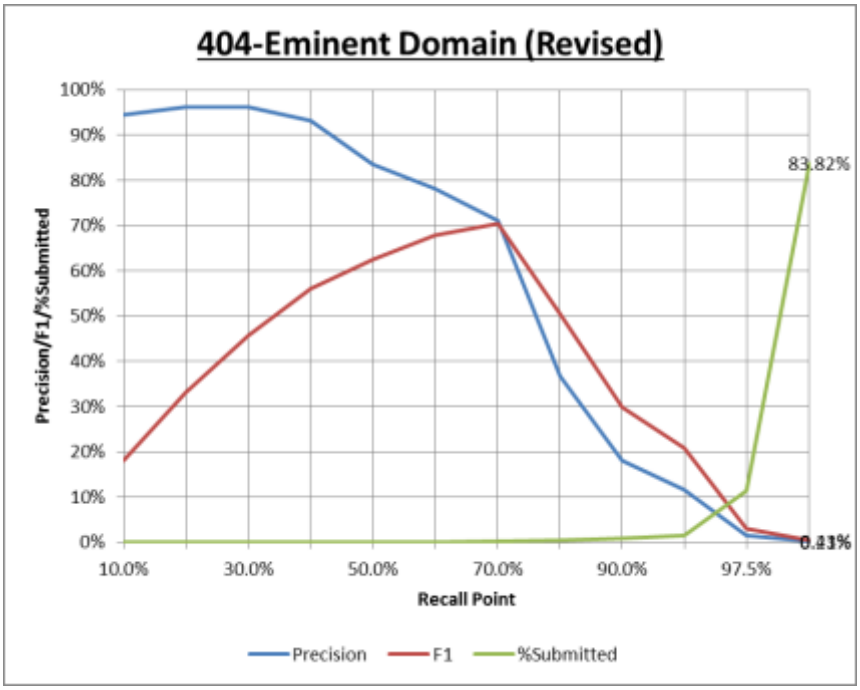
The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken.

The reasonable call was made when following a learning session after all keyword hits had been exhausted.

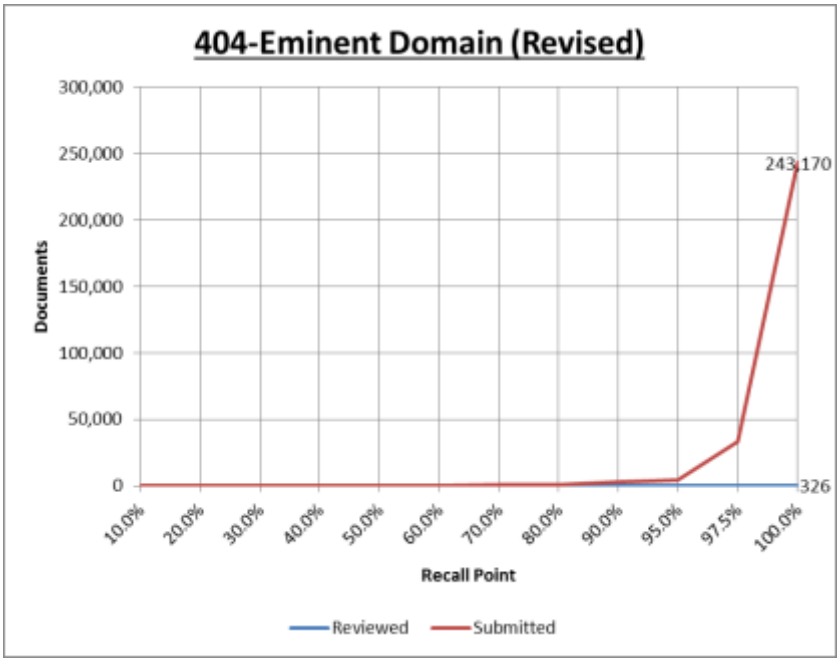
With this topic, the assessors seemed to treat any land acquisition (or even suggestion of it) by the state as "eminent domain," even if it did not apply. For instance, a situation where the state actively sought a private purchaser of an amusement park (Cypress Gardens) was found to be relevant even though this is not eminent domain. Likewise, a situation where people protested the state turning an airstrip in the Everglades previously belonging to Homestead Air Force Base into a commercial airport is not eminent domain related. As such, this was an issue that the standard (particularly for lawyers who know the issue) was inherently flawed, and therefore was not really representative of comparisons between human-only or hybrid reviewers and machine learning auto-runs.

Graphs

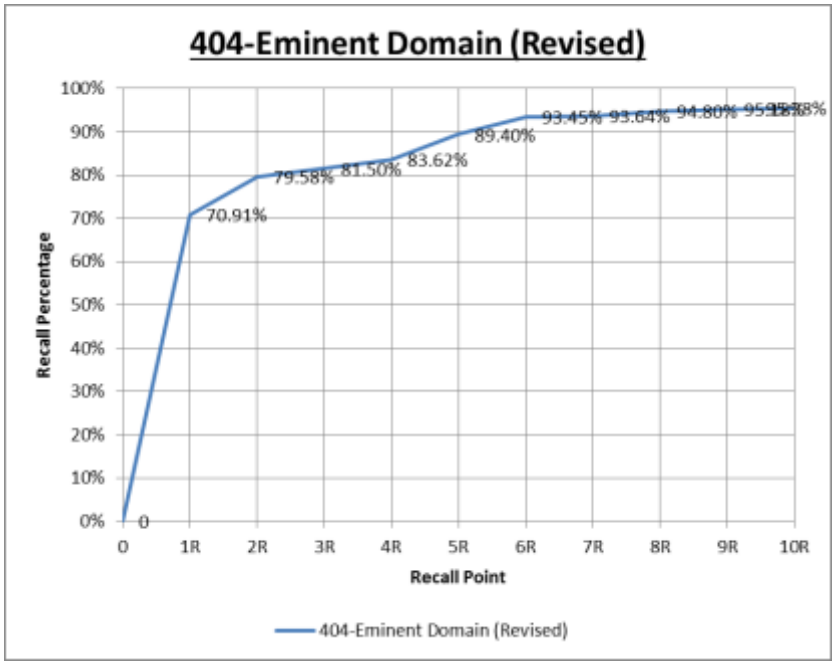
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Eminent Domain topic, the 90% recall threshold had been attained by submitting only 0.90%% of the corpus, 2,602 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


404-Eminent Domain (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
70.91%
79.58%
81.50%
83.62%
89.40%
93.45%
93.64%
94.80%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
404-Eminent Domain (Revised)

Topic 405 - Newt Gingrich
Total Documents: 290,099
Total Relevant: 123
Total Prevalence: 0.04%

Confusion Matrix - Newt Gingrich

	@Reasonable	@90% Recall	@95% Recall
True Positives	*123*	*111*	*117*
True Negatives	*289,686*	*289,924*	*289,922*
False Positives	*290*	*52*	*54*
False Negatives	*0*	*12*	*6*
Recall	100.00%	90.24%	95.12%
Precision	29.78%	68.10%	68.42%
F1 Measure	45.90%	77.62%	79.59%
Accuracy	99.9000%	99.9779%	99.9793%
Error	0.1000%	0.0221%	0.0207%
Elusion	0.00%	0.00%	0.00%
Fallout	0.10%	0.02%	0.02%

Topic 405 - Newt Gingrich - UNCORRECTED
Total Documents: 290,099
Total Relevant: 122
Total Prevalence: 0.04%

Confusion Matrix - Newt Gingrich

	@Reasonable	@90% Recall	@95% Recall
True Positives	*116*	*110*	*116*
True Negatives	*289,680*	*289,920*	*289,910*
False Positives	*297*	*57*	*67*
False Negatives	*6*	*12*	*6*
Recall	95.08%	90.16%	95.08%
Precision	28.09%	65.87%	63.39%
F1 Measure	43.36%	76.12%	76.07%
Accuracy	99.90%	99.98%	99.97%
Error	0.10%	0.02%	0.03%
Elusion	0.00%	0.00%	0.00%
Fallout	0.10%	0.02%	0.02%

Summary

The project was run by Losey from July 8th to July 15th 2016. He spent four hours, reviewed 66 documents and manually classified 432. He called Reasonable after 11 submissions and then did just one more submission (12 total).

The full description of this topic was: **All documents concerning House Speaker Newt Gingrich or any entities or personnel associated with Newt Gingrich.**

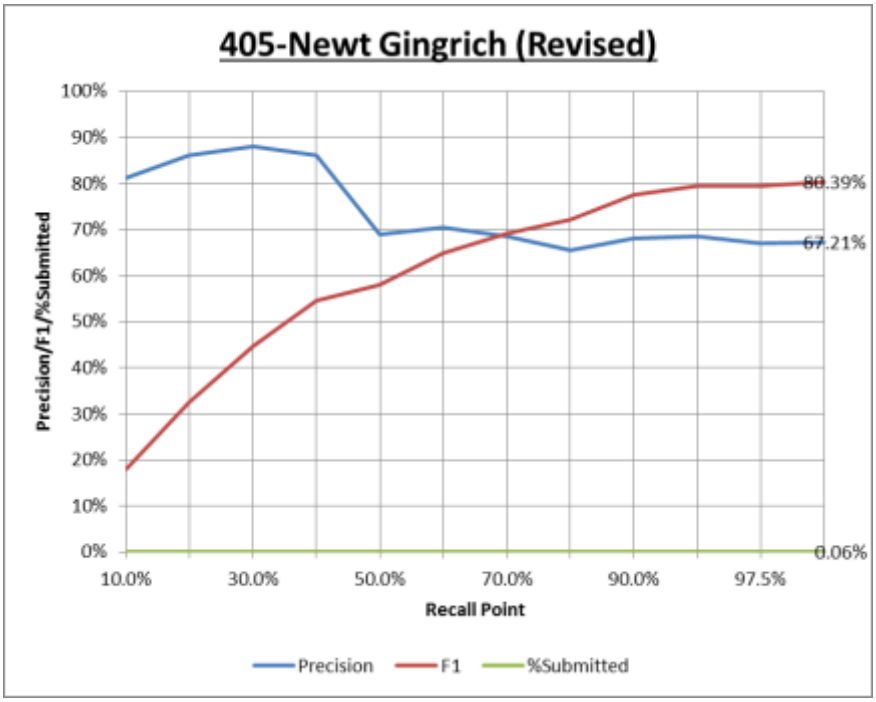
This was a fairly simple search because, fortunately for Florida, Newt Gingrich and his company had only limited impact on Florida and Governor Bush. Seventeen keyword search folders were created at the beginning of the project and tested. That took most of the time here. The work went easier than most topics because there were very few TREC errors seen.

The very first submission of documents to TREC located all but five of the relevant documents. They were all found by keyword search *Newt OR Gingrich**. I only looked at two documents in that search folder and saw they were obviously relevant. So I assumed all of the others with hits were relevant too, since this is such an unusual name, and did not bother to review them before classifying them. In "real life" we would spend more time verifying, of course. We would look at all 183 docs, as this is a small number. But part of our experiment here was to see how little effort we could put into these searches and still do reasonably well. AI ranking based searches were used after the first searches and first submission to find the rest. Again, this was an experiment to see how well we could do in an easy project like this with minimal human efforts after an initial discovery of the easy to find documents by keywords.

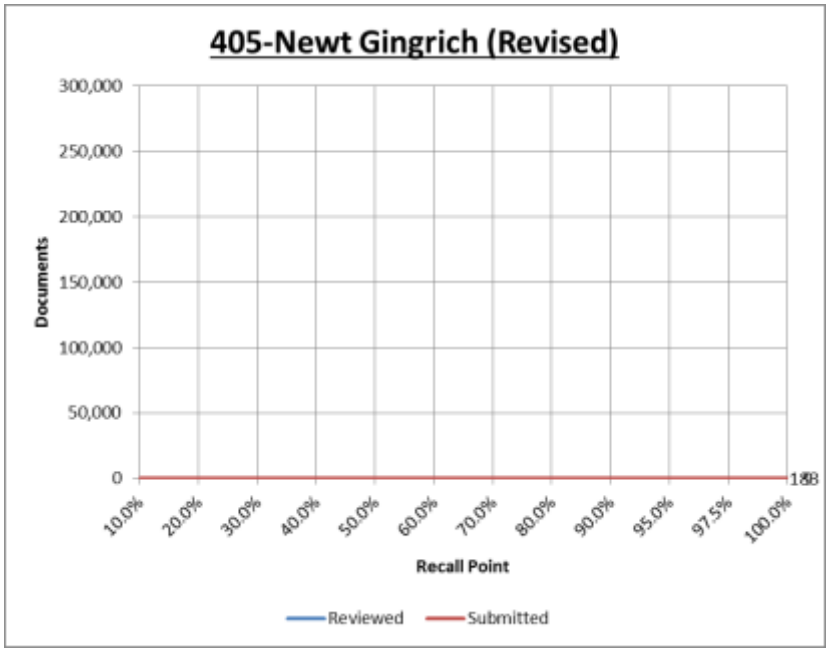
After that first submission Losey decided not to look at any documents in this topic or manually search. Instead he relied on just AI ranking and simply trained high ranking documents. He just assumed the predicted coding was right and used all of Mr. EDR's top ranked documents without inspection. He did so with many small submissions of the *unique*, most highly ranked documents. This was done to allow training to continue to improve. The only slight effort here was to differentiate *unique* docs, and only submit the top 25 *unique* ones. If an Email had the same subject line, it was presumed "NON-Unique" and Losey would skip down to the next ranked document that did not have the exact same subject line. He continued this pattern until all documents with a 50% or higher probable relevance had been submitted and then called reasonable.

Graphs

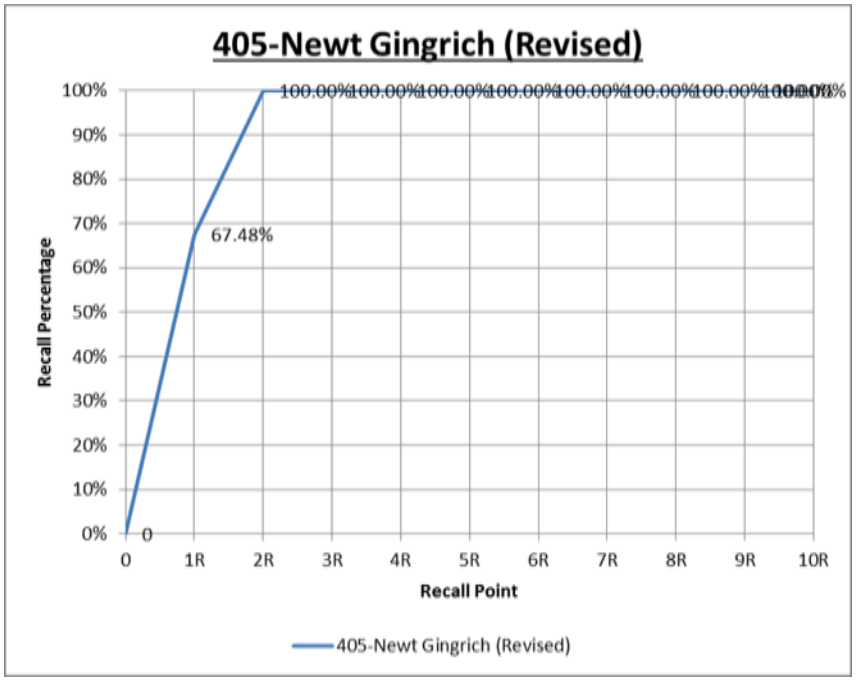
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Newt Gingrich topic, the 90% recall threshold had been attained by submitting only 0.06%% of the corpus, 163 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


405-Newt Gingrich (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
67.48%
0
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
405-Newt Gingrich (Revised)

Topic 406 - Felon Disenfranchisement
Total Documents: 290,099
Total Relevant: 203
Total Prevalence: 0.07%

Confusion Matrix - Felon Disenfranchisement

	@Reasonable	@90% Recall	@95% Recall
True Positives	*197*	*183*	*193*
True Negatives	*289,896*	*289,896*	*289,896*
False Positives	*0*	*0*	*0*
False Negatives	*6*	*20*	*10*
Recall	97.04%	90.15%	95.07%
Precision	100.00%	100.00%	100.00%
F1 Measure	98.50%	94.82%	97.47%
Accuracy	99.9979%	99.9931%	99.9966%
Error	0.0021%	0.0069%	0.0034%
Elusion	0.00%	0.01%	0.00%
Fallout	0.00%	0.00%	0.00%

Topic 406 - Felon Disenfranchisement - UNCORRECTED
Total Documents: 290,099
Total Relevant: 127
Total Prevalence: 0.04%

Confusion Matrix - Felon Disenfranchisement

	@Reasonable	@90% Recall	@95% Recall
True Positives	*93*	*115*	*121*
True Negatives	*289,926*	*260,205*	*196,906*
False Positives	*46*	*29,767*	*93,066*
False Negatives	*34*	*12*	*6*
Recall	73.23%	90.55%	95.28%
Precision	66.91%	0.38%	0.13%
F1 Measure	69.92%	0.77%	0.26%
Accuracy	99.97%	89.73%	67.92%
Error	0.03%	10.27%	32.08%
Elusion	0.01%	0.00%	0.00%
Fallout	0.02%	10.27%	32.09%

Summary

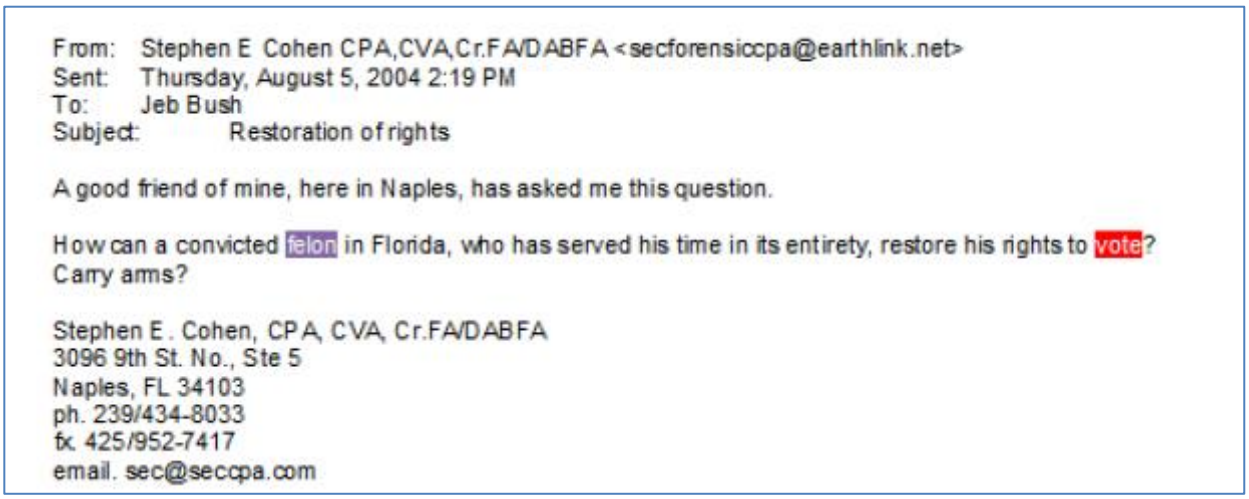
This project was run by Losey from August 20th to 23rd 2016. He expended at least seven hours on the project, probably longer (his record on his time here is uncertain). He reviewed 209 documents and categorized 232. He made a total of 17 submissions and called reasonable after the 9th submission.

The full description of the topic is: **Felon Disenfranchisement-All documents concerning the right of felons to vote in Florida, including but not limited to voter purges and reinstatement of voter rights. Individual clemency cases in Florida are not relevant.**

The rules in play here on relevance were hard to follow, including the clemency exclusion. That, and the presence of many borderline, ambiguous documents, made this a relatively difficult search. Several hours of unreported time, in addition to the seven recorded, were expended in post submission analysis of TREC's return documents.

Multimodal was used, with some keyword search up front, but there was special emphasis placed in this topic on the use of AI features and document ranking searches. This was done intentionally as an experiment and to make the review easier in this relatively difficult topic. Review of the top ranked documents was the primary search used. The AI ranked document review was improved by going lower on the keyword hit folders, where *hidden gems* of relevance were found low at lower than expected ranks. AI ranking searches were not only used as QC of other searches, but also to speed up the review and make it more efficient. The next-doc search and keyword list functions were also used this topic to maximize efficiency.

The usual high number of TREC errors were seen on this topic, including many obvious mistakes, and inconsistencies. Below is an example, just to give an idea on the inconsistent coding. The first inquiry email was called **irrelevant** by TREC and the second reply email by Bush was called **relevant**.

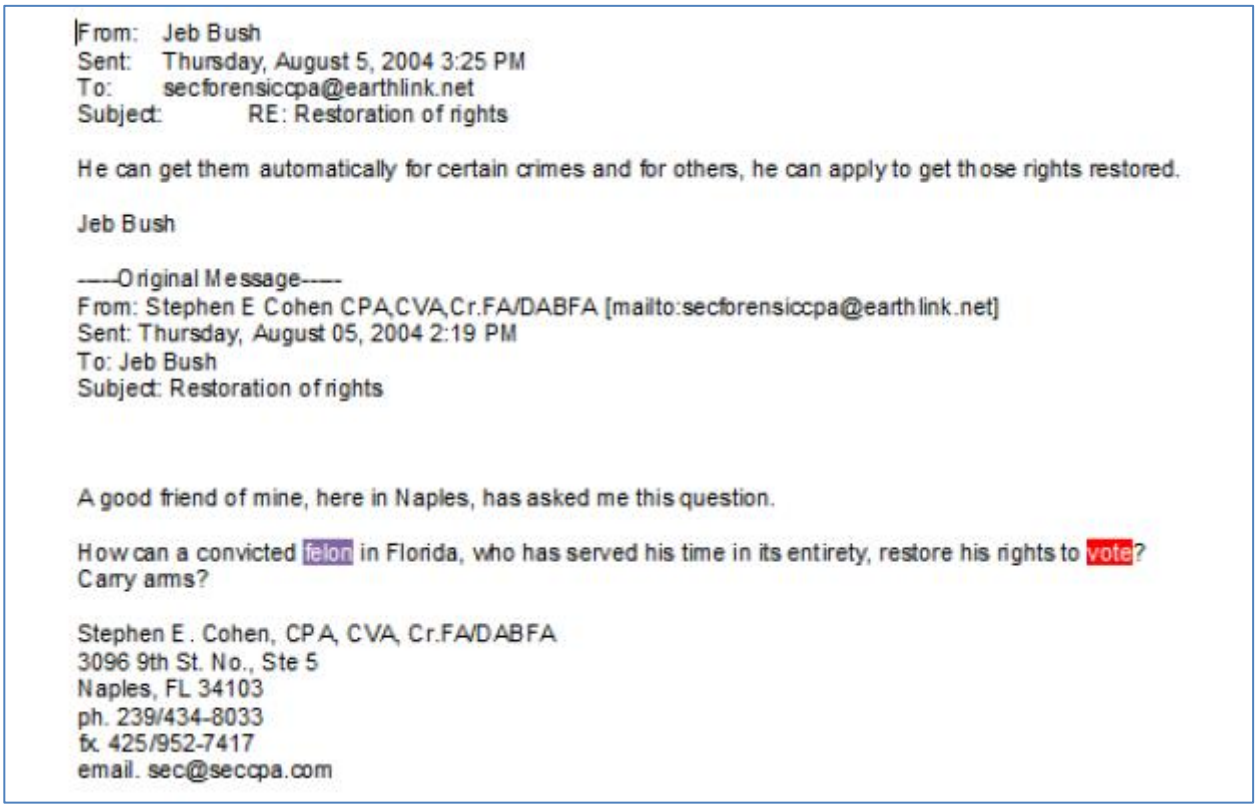


From: Stephen E Cohen CPA,CVA,Cr.FA/DABFA <secforensiccpa@earthlink.net>
Sent: Thursday, August 5, 2004 2:19 PM
To: Jeb Bush
Subject: Restoration of rights

A good friend of mine, here in Naples, has asked me this question.

How can a convicted felon in Florida, who has served his time in its entirety, restore his rights to vote?
Carry arms?

Stephen E. Cohen, CPA, CVA, Cr.FA/DABFA
3096 9th St. No., Ste 5
Naples, FL 34103
ph. 239/434-8033
fx. 425/952-7417
email. sec@seccpa.com



From: Jeb Bush
Sent: Thursday, August 5, 2004 3:25 PM
To: secforensiccpa@earthlink.net
Subject: RE: Restoration of rights

He can get them automatically for certain crimes and for others, he can apply to get those rights restored.

Jeb Bush

-----Original Message-----
From: Stephen E Cohen CPA,CVA,Cr.FA/DABFA [mailto:secforensiccpa@earthlink.net]
Sent: Thursday, August 05, 2004 2:19 PM
To: Jeb Bush
Subject: Restoration of rights

A good friend of mine, here in Naples, has asked me this question.

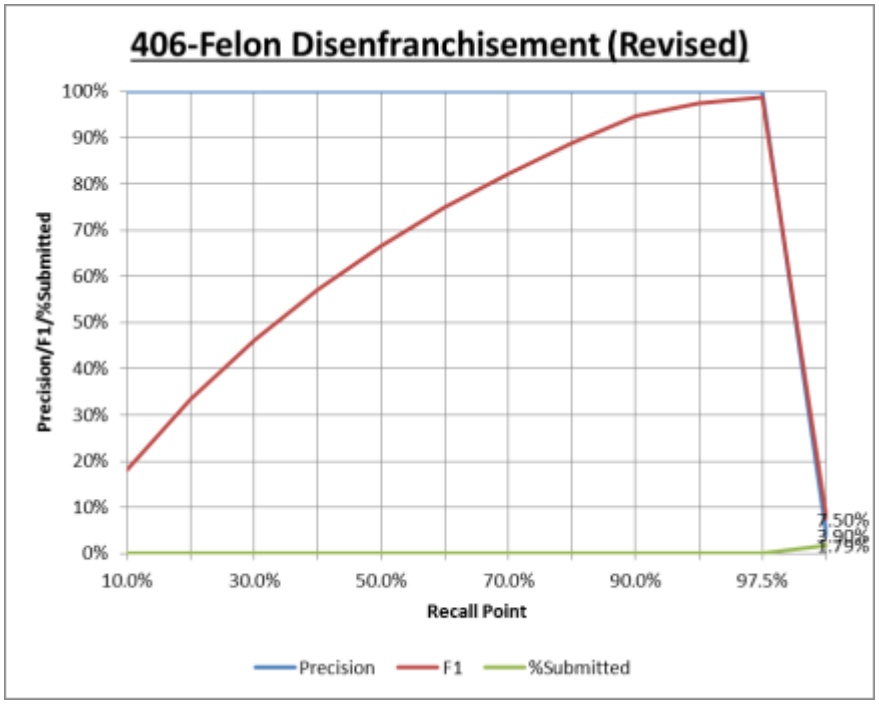
How can a convicted felon in Florida, who has served his time in its entirety, restore his rights to vote?
Carry arms?

Stephen E. Cohen, CPA, CVA, Cr.FA/DABFA
3096 9th St. No., Ste 5
Naples, FL 34103
ph. 239/434-8033
fx. 425/952-7417
email. sec@seccpa.com

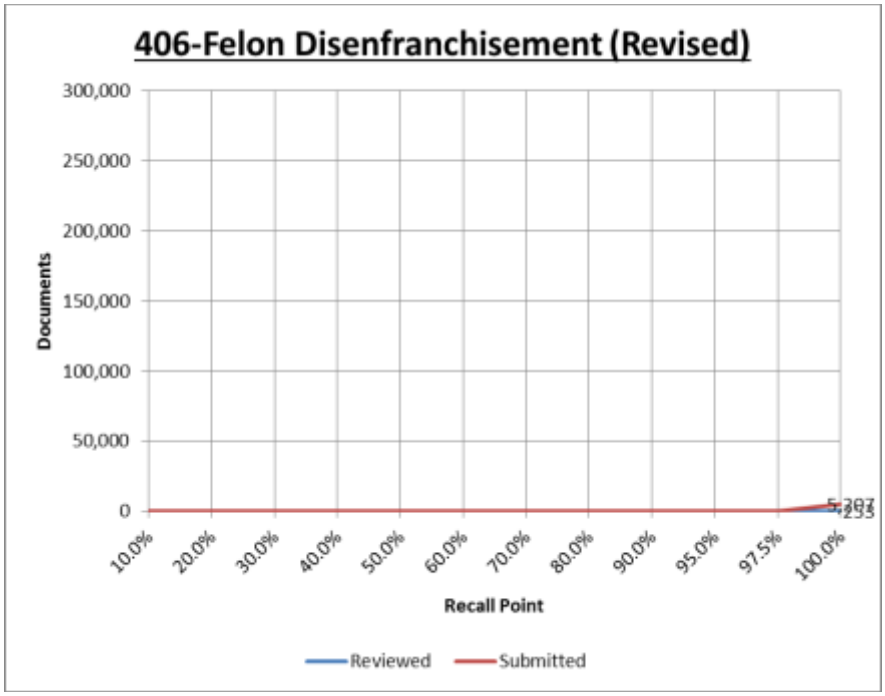
In fact, they were both relevant. We would typically, but not always, include both documents into training and ignore TREC errors. The color you see added in the above emails is not in the originals. It is added by the software per user direction to assist in the quick human review of a document. Typically keywords the user selects are colored. This feature is a terrific time saver and was heavily utilized by all reviewers in all topics.

Graphs

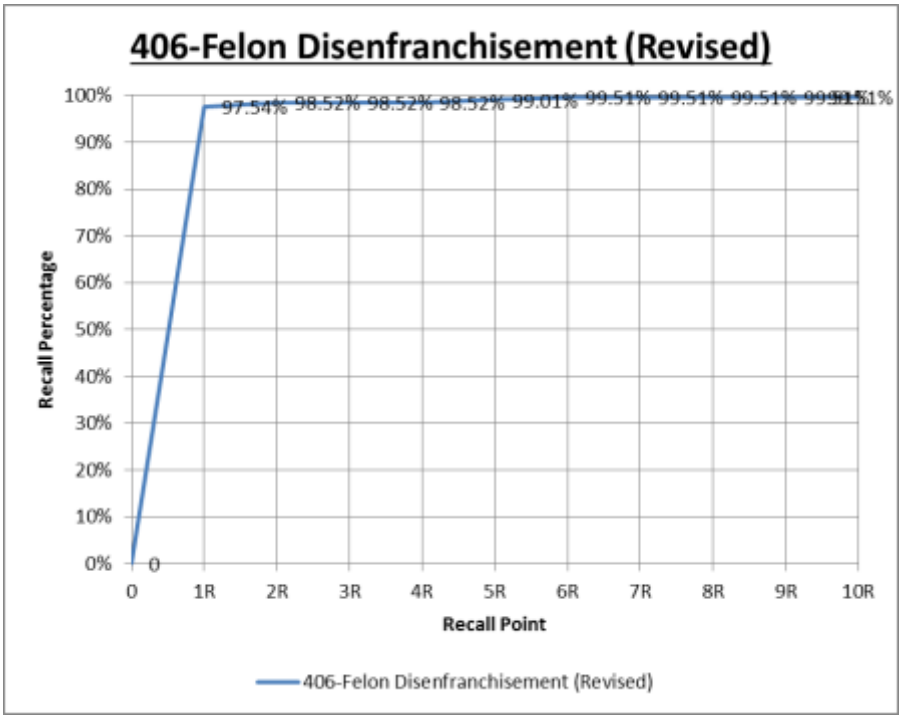
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Felon Disenfranchisement topic, the 90% recall threshold had been attained by submitting only 0.06%% of the corpus, 183 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


406-Felon Disenfranchisement (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
97.54%
0
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
406-Felon Disenfranchisement (Revised)

Topic 407 - Faith Based Initiatives
Total Documents: 290,099
Total Relevant: 1,654
Total Prevalence: 0.57%

Confusion Matrix - Faith Based Initiatives

	@Reasonable	@90% Recall	@95% Recall
True Positives	*1,465*	*1,489*	*1,572*
True Negatives	*287,571*	*287,331*	*281,747*
False Positives	*874*	*1,114*	*6,698*
False Negatives	*189*	*165*	*82*
Recall	88.57%	90.02%	95.04%
Precision	62.63%	57.20%	19.01%
F1 Measure	73.38%	69.96%	31.68%
Accuracy	99.6336%	99.5591%	97.6629%
Error	0.3664%	0.4409%	2.3371%
Elusion	0.07%	0.06%	0.03%
Fallout	0.30%	0.39%	2.32%

Topic 407 - Faith Based Initiatives - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,586
Total Prevalence: 0.55%

Confusion Matrix - Faith Based Initiatives

	@Reasonable	@90% Recall	@95% Recall
True Positives	*492*	*1,428*	*1,507*
True Negatives	*288,289*	*285,930*	*281,388*
False Positives	*224*	*2,583*	*7,125*
False Negatives	*1,094*	*158*	*79*
Recall	31.02%	90.04%	95.02%
Precision	68.72%	35.60%	17.46%
F1 Measure	42.75%	51.03%	29.50%
Accuracy	99.55%	99.06%	97.52%
Error	0.45%	0.94%	2.48%
Elusion	0.38%	0.06%	0.03%
Fallout	0.08%	0.90%	2.47%

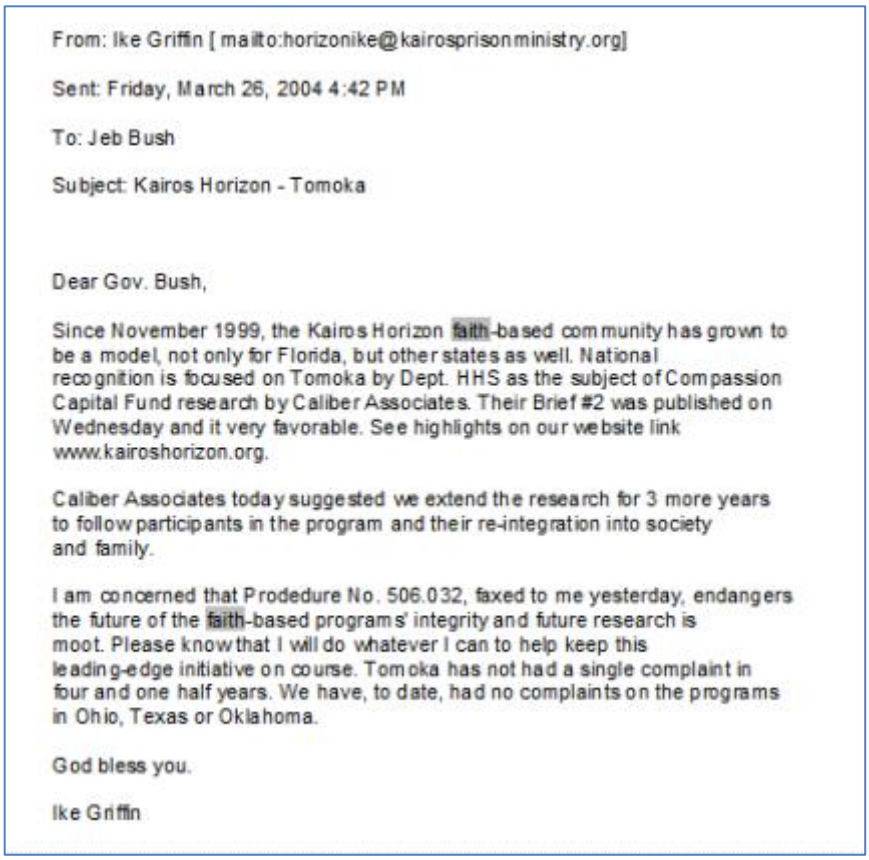
Summary

This topic was run by Losey and was the last topic reviewed from August 28th to 31st 2016. He made 21 submissions and called reasonable after the 14th. He reviewed 400 documents and categorized 1,791. Losey spent far more time on this topic than any of the others, 15 hours.

The full description of this topic is: **Faith-Based Initiatives - All documents concerning grants or other initiatives in Florida to offload social services to so-called faith-based agencies. Services include but are not limited to education, prisons, and emergency relief.**

Losey created 46 different searches and search folders, also a high-volume record that helps explain the 15 hours this topic took to complete. A full multimodal approach was used, not just keywords, as this was a relatively difficult topic. Bush had many emails concerning this topic as this was one of his pet projects as governor. In addition to the many keyword searches, similarity and near duplication searches were use in any correct, TREC verified relevant document. There was also heavy reliance placed on AI ranking searches as the project matured. As an experiment in this topic the relevant documents that were incorrectly labeled as irrelevant by TREC were excluded from training. The result of this alternate strategy was not clear. Of course, no documents incorrectly labeled as relevant by TREC were used in training. We wanted to avoid the avoid the phenomena we had observed many times by this point, and which the *Team* had started calling the *Ouroboros* effect. This is the negative feedback loop where one automated classifier blindly follows another with no regard to ground truth. We saw that as akin to a snake eating its own tail, the *Ouroboros,* that is discussed in the Conclusion to the *Team's* Final Report and Footnote 17.

This topic had many errors by TREC. Some were borderline, so, as we always did, we accepted them as correct, even though they were against our view of relevance. Only the clearly wrong were corrected. Here is an example. The data contained seven copies of the same email, or nearly the same. The emails were all ironically written by a person who lives just a few blocks from his home. Below is one copy.



From: Ike Griffin [mailto:horizonike@kairosprisonministry.org]

Sent: Friday, March 26, 2004 4:42 PM

To: Jeb Bush

Subject: Kairos Horizon - Tomoka

Dear Gov. Bush,

Since November 1999, the Kairos Horizon faith-based community has grown to be a model, not only for Florida, but other states as well. National recognition is focused on Tomoka by Dept. HHS as the subject of Compassion Capital Fund research by Caliber Associates. Their Brief #2 was published on Wednesday and it very favorable. See highlights on our website link www.kairoshorizon.org.

Caliber Associates today suggested we extend the research for 3 more years to follow participants in the program and their re-integration into society and family.

I am concerned that Prodedure No. 506.032, faxed to me yesterday, endangers the future of the faith-based programs' integrity and future research is moot. Please know that I will do whatever I can to help keep this leading-edge initiative on course. Tomoka has not had a single complaint in four and one half years. We have, to date, had no complaints on the programs in Ohio, Texas or Oklahoma.

God bless you.

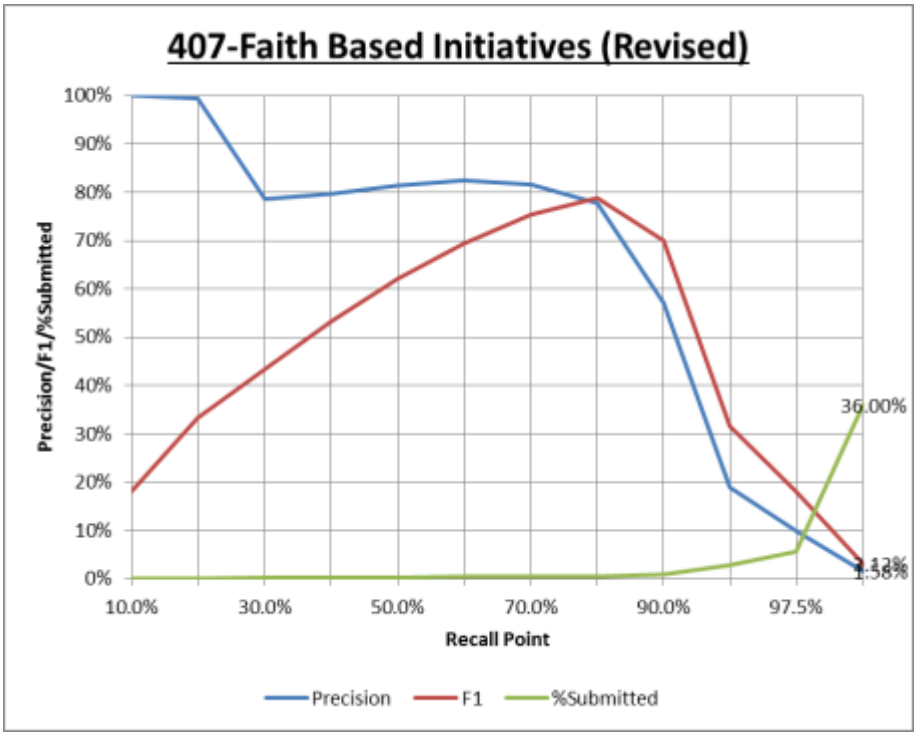
Ike Griffin

Three copies of the emails were classified as relevant by TREC and four were classified by TREC as irrelevant. It is hard to understand how this could happen, but we saw it all the time.

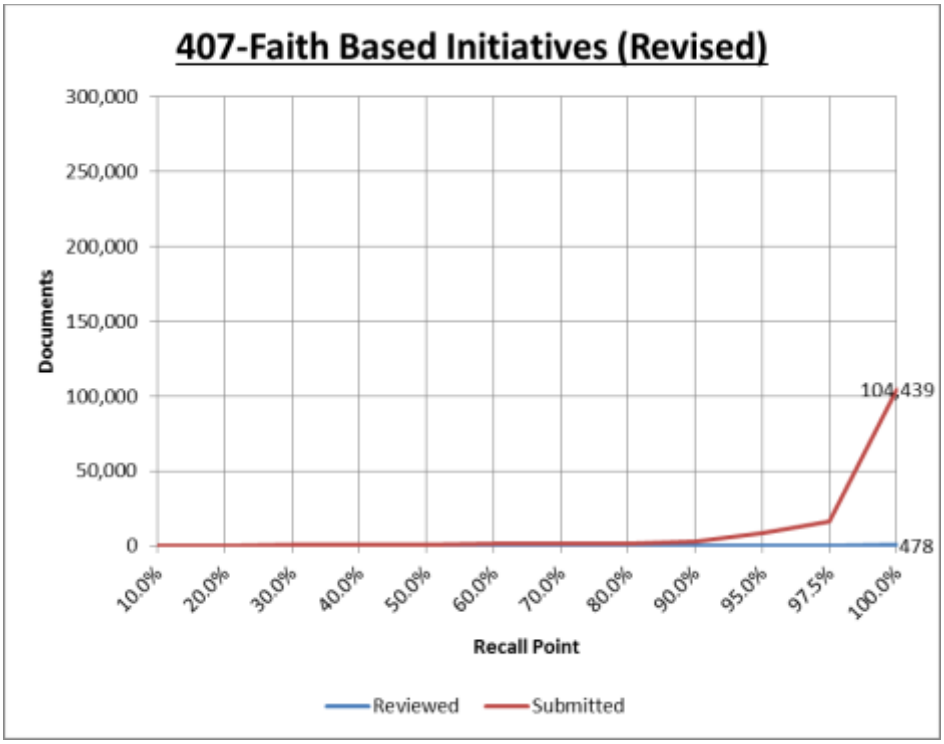
Just before making his personal reasonable call after the 14th submission, Losey submitted the highest ranked documents down to 50%, and select keyword folders documents regardless of rank. He did so with little or no review in the last several submissions, relying on AI ranking alone informed by keyword search folders. Losey noted that he was sure he could find more relevant at that point if he kept reviewing more documents, but, after expending almost 15 hours on this topic already, it would not be a reasonable effort to do so. It would be excessive for all but the largest cases under Rule 26(b)(1) *Federal Rules of Civil Procedure*.

Graphs

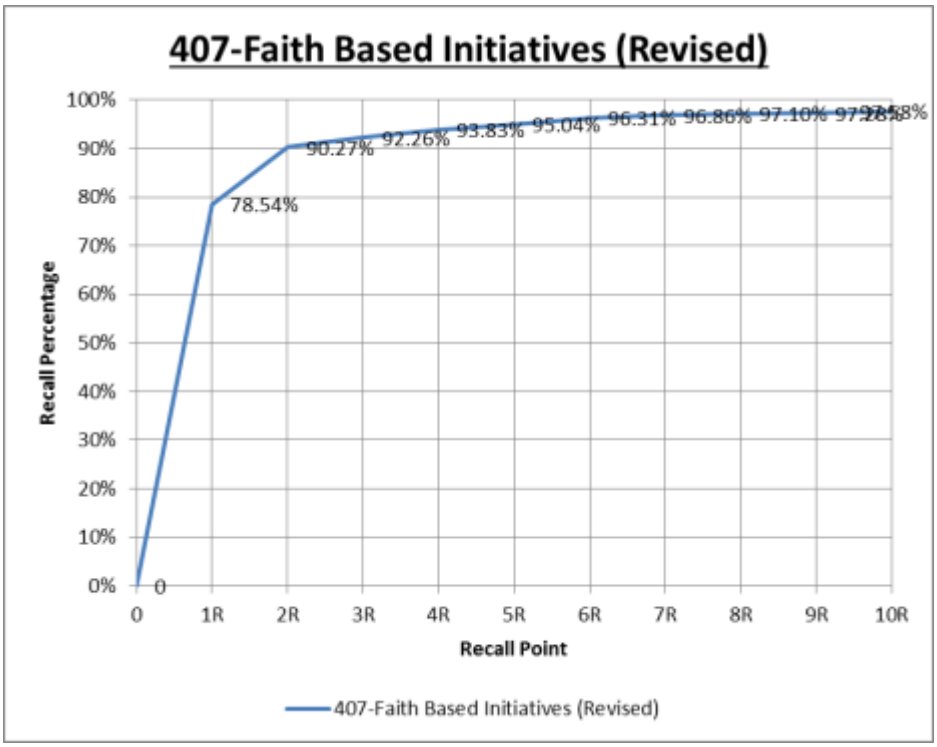
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Faith Based Initiatives topic, the 90% recall threshold had been attained by submitting only 0.90%% of the corpus, 2,603 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 408 - Invasive Species
Total Documents: 290,099
Total Relevant: 168
Total Prevalence: 0.06%

Confusion Matrix - Invasive Species

	@Reasonable	@90% Recall	@95% Recall
True Positives	*86*	*152*	*160*
True Negatives	*289,918*	*289,530*	*263,137*
False Positives	*13*	*401*	*26,794*
False Negatives	*82*	*16*	*8*
Recall	51.19%	90.48%	95.24%
Precision	86.87%	27.49%	0.59%
F1 Measure	64.42%	42.16%	1.18%
Accuracy	99.9673%	99.8563%	90.7611%
Error	0.0327%	0.1437%	9.2389%
Elusion	0.03%	0.01%	0.00%
Fallout	0.00%	0.14%	9.24%

Topic 408 - Invasive Species - UNCORRECTED
Total Documents: 290,099
Total Relevant: 116
Total Prevalence: 0.04%

Confusion Matrix - Invasive Species

	@Reasonable	@90% Recall	@95% Recall
True Positives	*64*	*105*	*111*
True Negatives	*289,948*	*67,601*	*18,751*
False Positives	*35*	*222,382*	*271,232*
False Negatives	*52*	*11*	*5*
Recall	55.17%	90.52%	95.69%
Precision	64.65%	0.05%	0.04%
F1 Measure	59.53%	0.09%	0.08%
Accuracy	99.97%	23.34%	6.50%
Error	0.03%	76.66%	93.50%
Elusion	0.02%	0.02%	0.03%
Fallout	0.01%	76.69%	93.53%

Summary

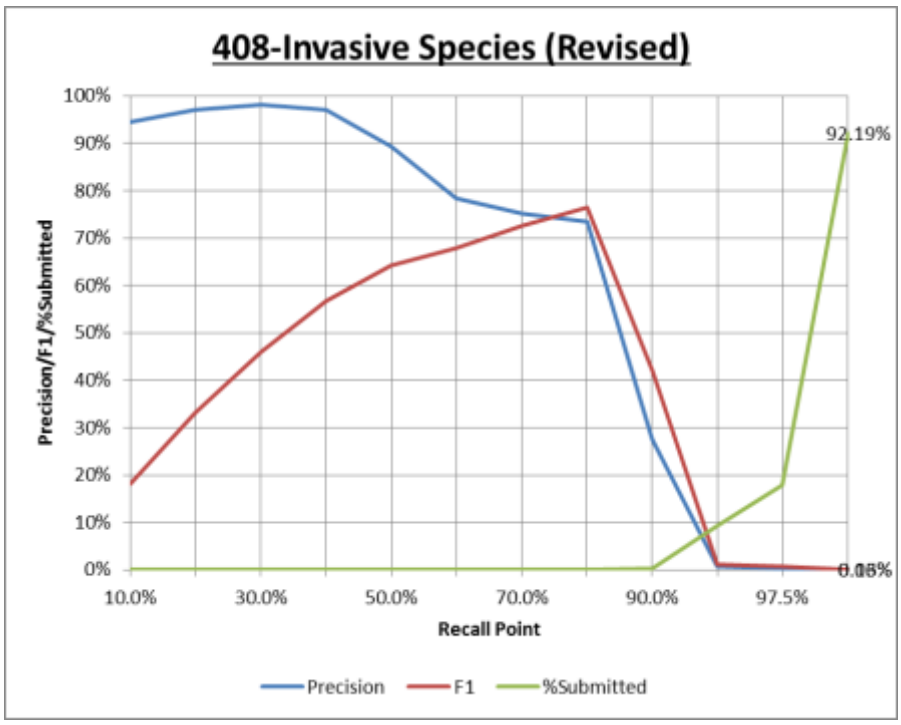
This topic was run by Tony Reichenberger. A google search of non-native species in Florida and the state Invasive Species webpage served as the basis for creating a list of keywords to search for relevant documents. It was apparent from the first submission that only select invasive species were considered relevant. Documents solely relating to species found irrelevant from the TREC feedback were coded irrelevant. Documents were submitted until the keywords were exhausted at which point the Reasonable call was made.

However, the standard was inconsistent in coding; for instance within the first submission was a document explicitly about Burmese python (a well-known invasive species to Florida causing a myriad of problems in the Everglades) which was returned from TREC as irrelevant. However, later submissions relating to Burmese Pythons were found relevant. Likewise, assessors seemed to confuse "endangered' species such as manatees, with "invasive" species on a number of calls. Assessors also made the mistake of confusing species that are nuisances, such as particular red algae blooms, with being invasive, even though they are native to the area.

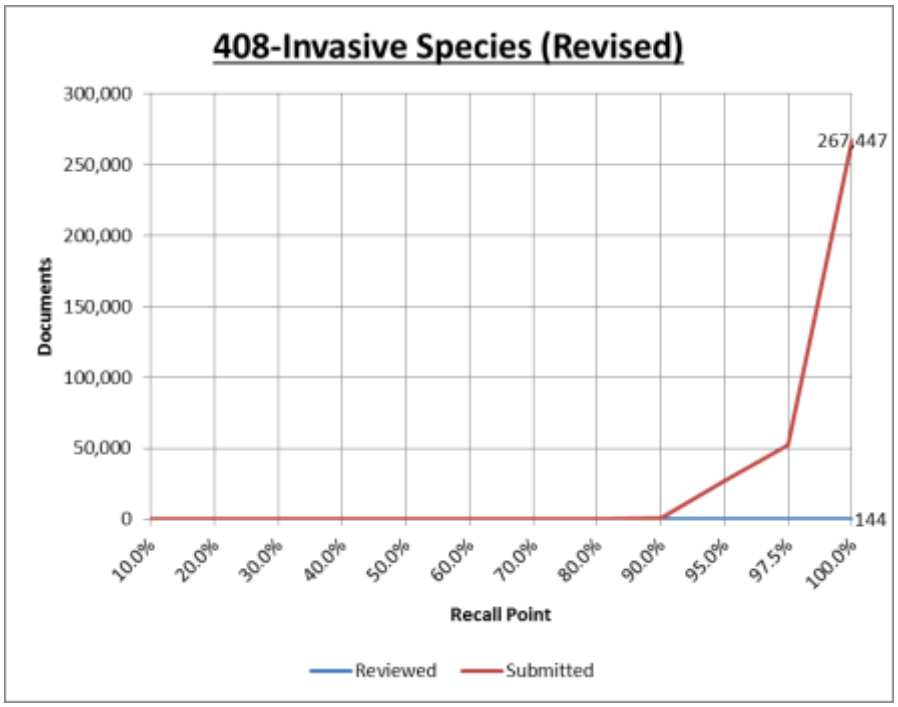
As such, this was an issue that the standard (particularly for lawyers) was inherently flawed, and not really indicative of the issue. Therefore, it is not representative of comparisons between human-only or hybrid reviewers and machine learning auto-runs.

Graphs

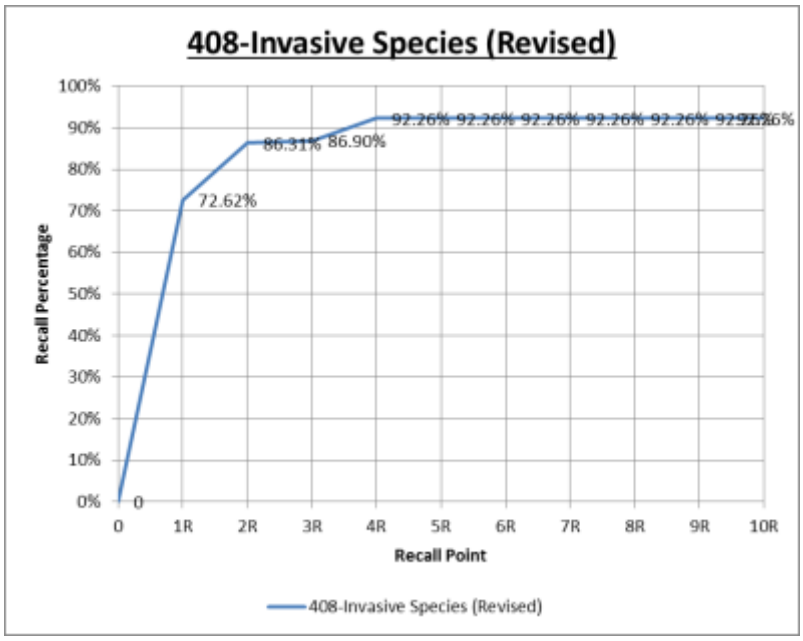
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Invasive Species topic, the 90% recall threshold had been attained by submitting only 0.19%% of the corpus, 553 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 409 - Climate Change
Total Documents: 290,099
Total Relevant: 224
Total Prevalence: 0.08%

Confusion Matrix - Climate Change

	@Reasonable	@90% Recall	@95% Recall
True Positives	*198*	*202*	*213*
True Negatives	*289,653*	*289,254*	*273,227*
False Positives	*222*	*621*	*16,648*
False Negatives	*26*	*22*	*11*
Recall	88.39%	90.18%	95.09%
Precision	47.14%	24.54%	1.26%
F1 Measure	61.49%	38.59%	2.49%
Accuracy	99.9145%	99.7784%	94.2575%
Error	0.0855%	0.2216%	5.7425%
Elusion	0.01%	0.01%	0.00%
Fallout	0.08%	0.21%	5.74%

Topic 409 - Climate Change - UNCORRECTED
Total Documents: 290,099
Total Relevant: 202
Total Prevalence: 0.07%

Confusion Matrix - Climate Change

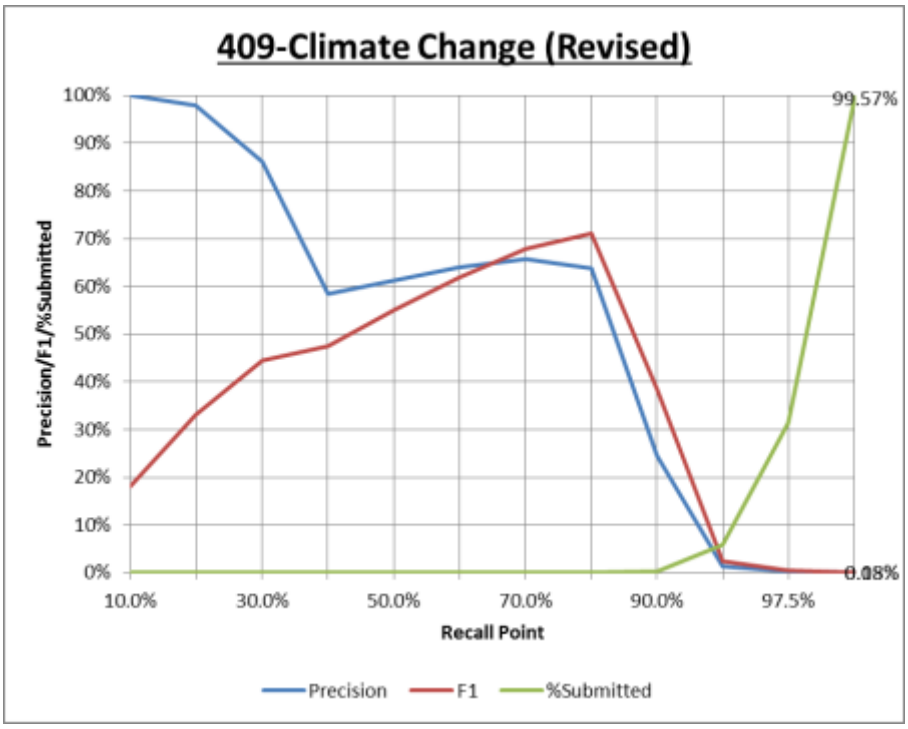
	@Reasonable	@90% Recall	@95% Recall
True Positives	*171*	*182*	*192*
True Negatives	*289,648*	*285,786*	*248,332*
False Positives	*249*	*4,111*	*41,565*
False Negatives	*31*	*20*	*10*
Recall	84.65%	90.10%	95.05%
Precision	40.71%	4.24%	0.46%
F1 Measure	54.98%	8.10%	0.92%
Accuracy	99.90%	98.58%	85.67%
Error	0.10%	1.42%	14.33%
Elusion	0.01%	0.01%	0.00%
Fallout	0.09%	1.42%	14.34%

Summary

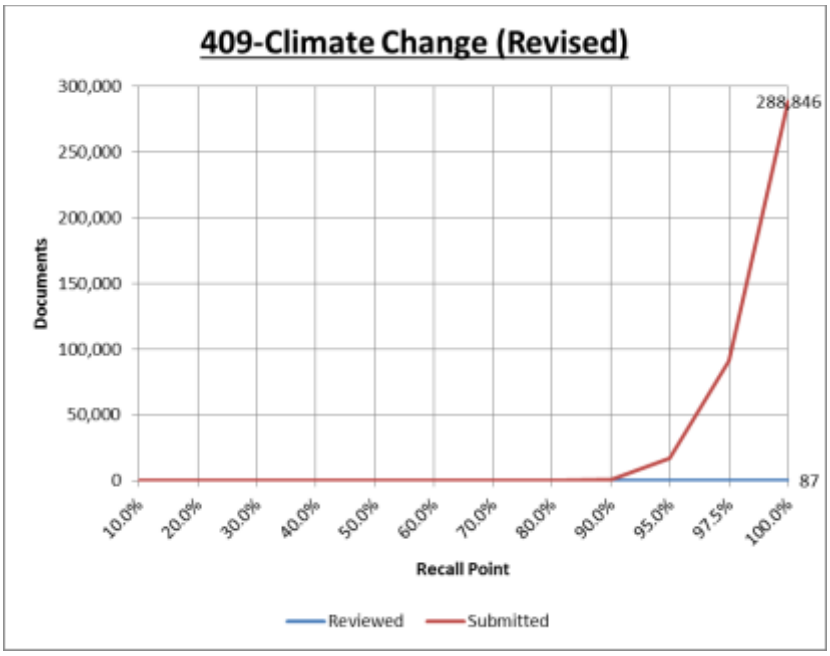
This topic was run by Levi Kuehn. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. When scores dropped to 5%, a final search was submitted, another learning session run, and documents were submitted in probability order.

Graphs

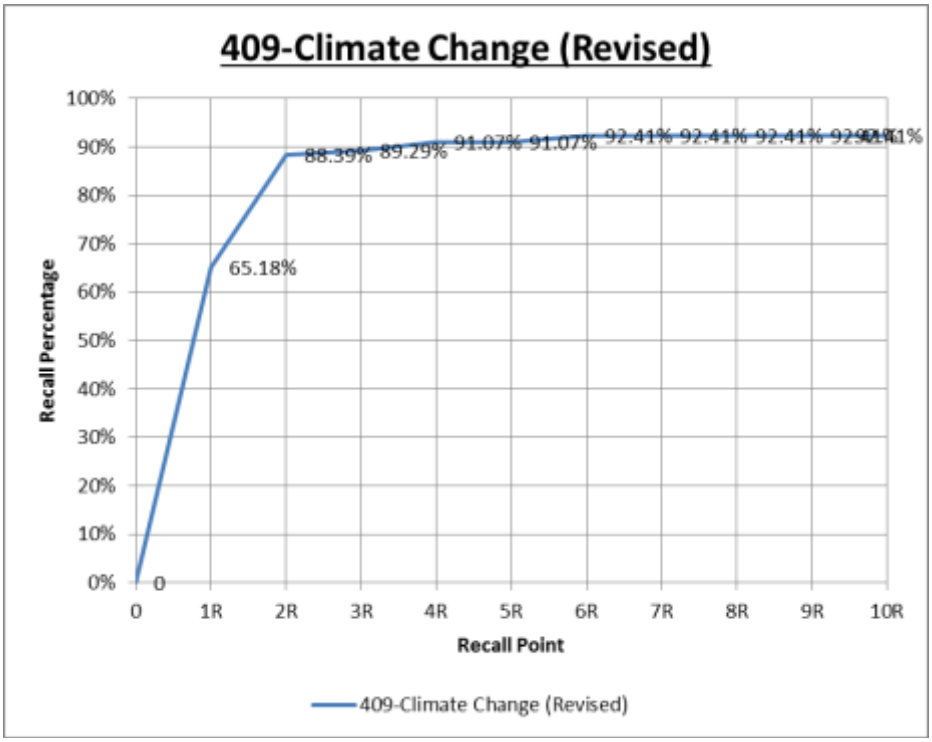
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Climate Change topic, the 90% recall threshold had been attained by submitting only 0.28%% of the corpus, 823 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 410 - Condominiums
Total Documents: 290,099
Total Relevant: 1,317
Total Prevalence: 0.45%

Confusion Matrix - Condominiums

	@Reasonable	@90% Recall	@95% Recall
True Positives	*1,314*	*1,186*	*1,252*
True Negatives	*287,321*	*287,583*	*287,497*
False Positives	*1,461*	*1,199*	*1,285*
False Negatives	*3*	*131*	*65*
Recall	99.77%	90.05%	95.06%
Precision	47.35%	49.73%	49.35%
F1 Measure	64.22%	64.07%	64.97%
Accuracy	99.4953%	99.5415%	99.5346%
Error	0.5047%	0.4585%	0.4654%
Elusion	0.00%	0.05%	0.02%
Fallout	0.51%	0.42%	0.44%

Topic 410 – Condominiums - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,346
Total Prevalence: 0.46%

Confusion Matrix - Condominiums

	@Reasonable	@90% Recall	@95% Recall
True Positives	*1,280*	*1,212*	*1,279*
True Negatives	*287,258*	*287,445*	*287,305*
False Positives	*1,495*	*1,308*	*1,448*
False Negatives	*66*	*134*	*67*
Recall	95.10%	90.04%	95.02%
Precision	46.13%	48.10%	46.90%
F1 Measure	62.12%	62.70%	62.80%
Accuracy	99.46%	99.50%	99.48%
Error	0.54%	0.50%	0.52%
Elusion	0.02%	0.05%	0.02%
Fallout	0.52%	0.45%	0.50%

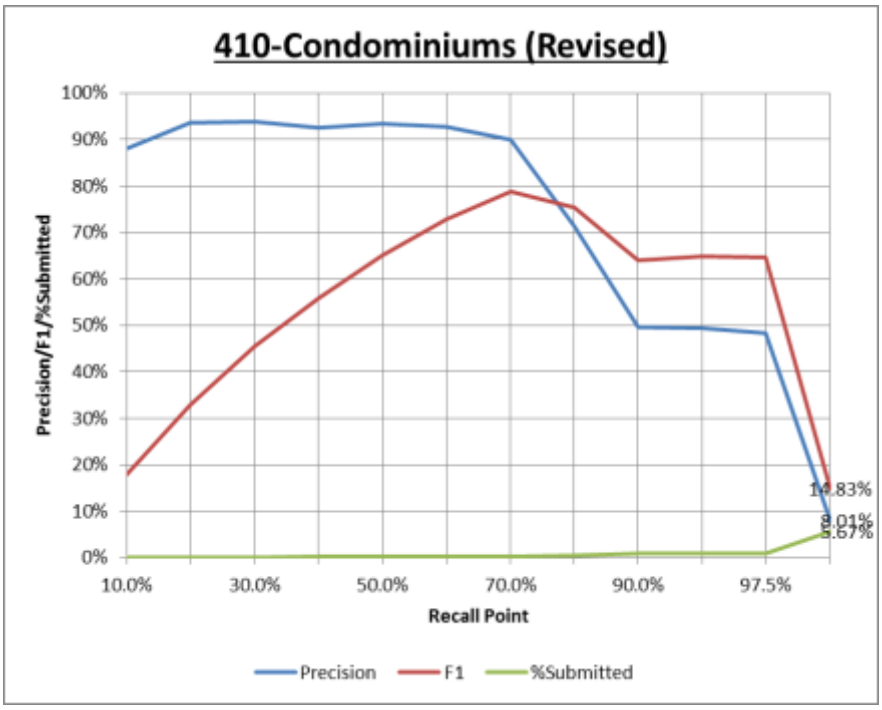
Summary

This topic was run by Tony Reichenberger. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (50-100 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing; when additional relevant materials were found, subsequent searches for similar documents were partaken.

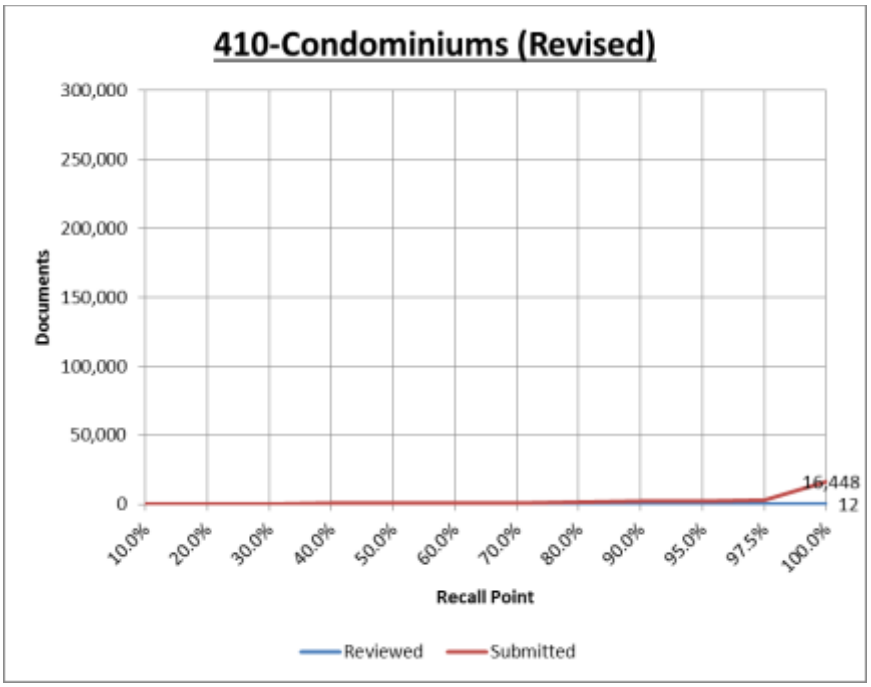
Reasonable was called when keywords were exhausted and the precision within the submission dropped to less than 5%.

Graphs

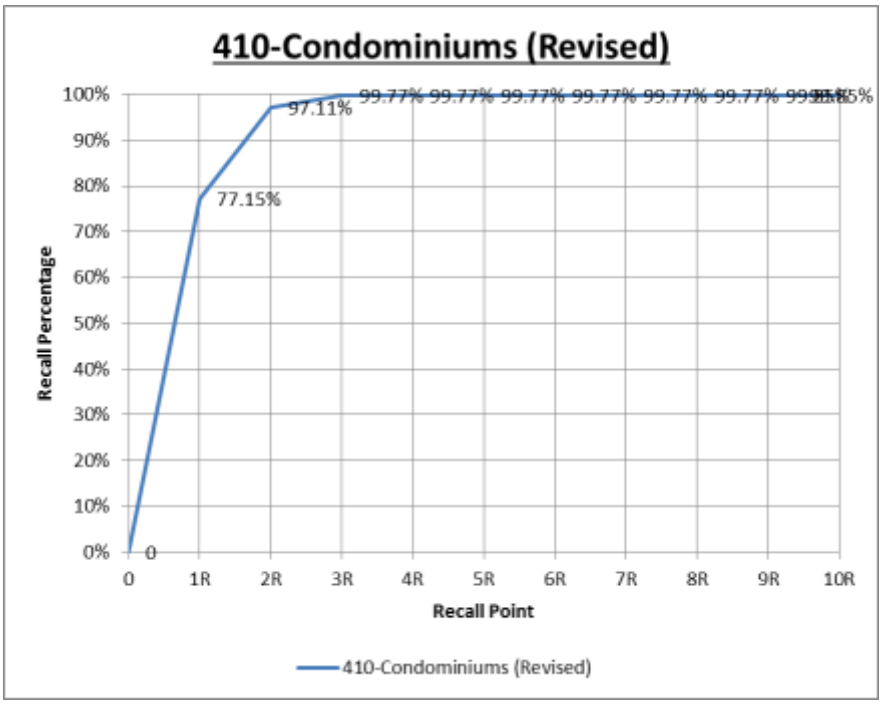
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Condominiums topic, the 90% recall threshold had been attained by submitting only 0.82%% of the corpus, 2,385 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 411 - Stand Your Ground
Total Documents: 290,099
Total Relevant: 59
Total Prevalence: 0.02%

Confusion Matrix - Stand Your Ground

	@Reasonable	@90% Recall	@95% Recall
True Positives	*59*	*54*	*57*
True Negatives	*290,011*	*290,027*	*290,019*
False Positives	*29*	*13*	*21*
False Negatives	*0*	*5*	*2*
Recall	100.00%	91.53%	96.61%
Precision	67.05%	80.60%	73.08%
F1 Measure	80.27%	85.71%	83.21%
Accuracy	99.9900%	99.9938%	99.9921%
Error	0.0100%	0.0062%	0.0079%
Elusion	0.00%	0.00%	0.00%
Fallout	0.01%	0.00%	0.01%

Topic 411 - Stand Your Ground - UNCORRECTED
Total Documents: 290,099
Total Relevant: 89
Total Prevalence: 0.03%

Confusion Matrix - Stand Your Ground

	@Reasonable	@90% Recall	@95% Recall
True Positives	*59*	*81*	*85*
True Negatives	*289,981*	*250,502*	*143,021*
False Positives	*29*	*39,508*	*146,989*
False Negatives	*30*	*8*	*4*
Recall	66.29%	91.01%	95.51%
Precision	67.05%	0.20%	0.06%
F1 Measure	66.67%	0.41%	0.12%
Accuracy	99.98%	86.38%	49.33%
Error	0.02%	13.62%	50.67%
Elusion	0.01%	0.00%	0.00%
Fallout	0.01%	13.62%	50.68%

Summary

This topic as run by Losey who worked on it from August 14th to August 16th 2016 for five hours. He reviewed 274 document and manually categorized 198.

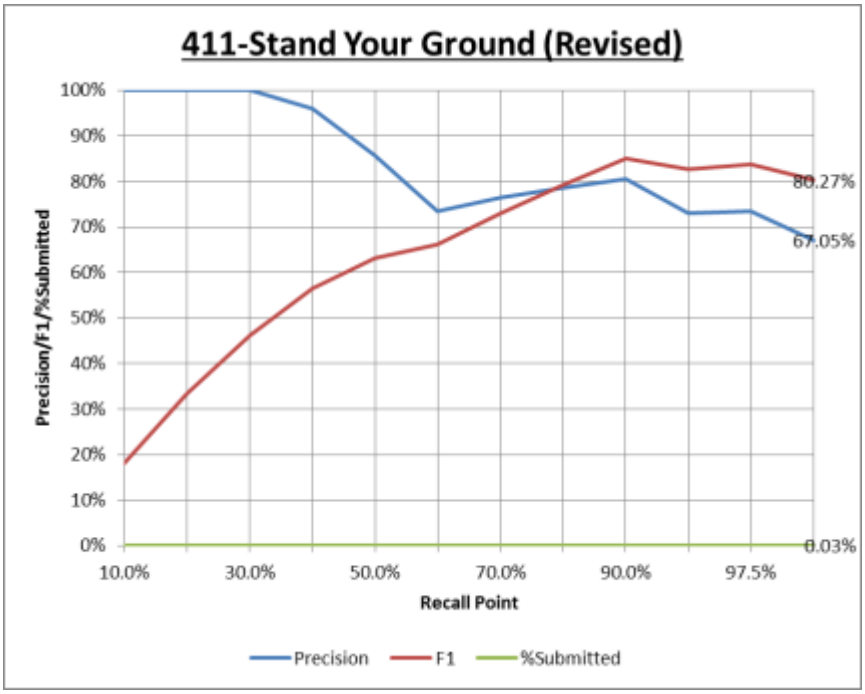
The full description of h topic is: **Stand Your Ground - All documents concerning a Florida bill permitting the use of deadly force to protect one's self or one's property.**

Of course most everyone in Florida with half a brain knows all about this controversial law. Losey did not find this a difficult assignment, especially because the scope of relevance was clear and so were the documents. As an experiment Losey called reasonable with his first submission. Before the submission Losey created 28 search folders. His review was entirely based on keyword search, and similarity type searches. Most of his five-hour time was spent doing these searches.

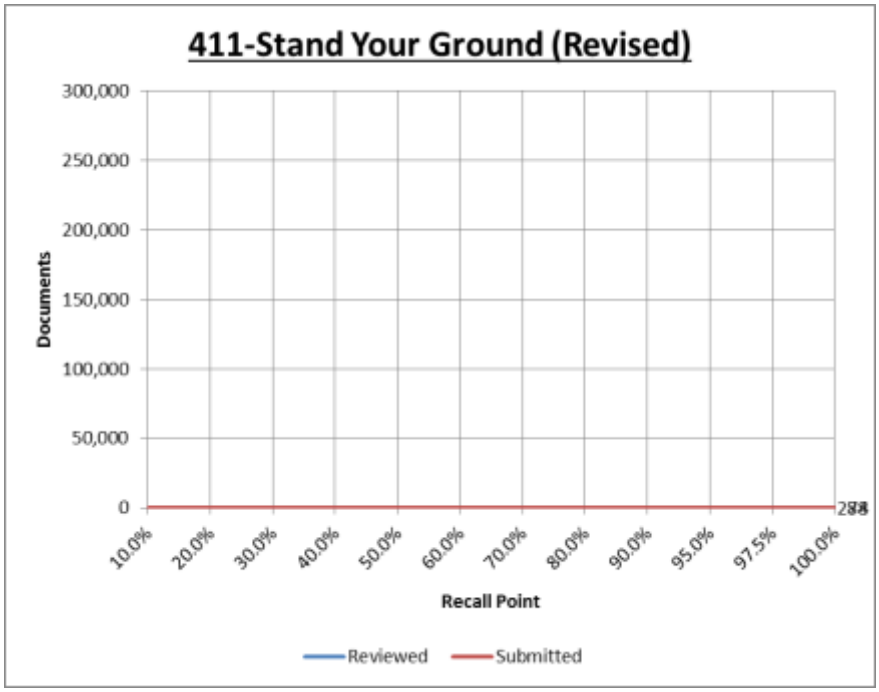
Losey then used TREC as a QC check to see if he had missed anything. Unfortunately the judging by TREC on this topic was poor. TREC found 58 additional documents, but they were all False Positives, iw - not relevant. Trec also missed 29 docs in my first submission of all relevant.

Graphs

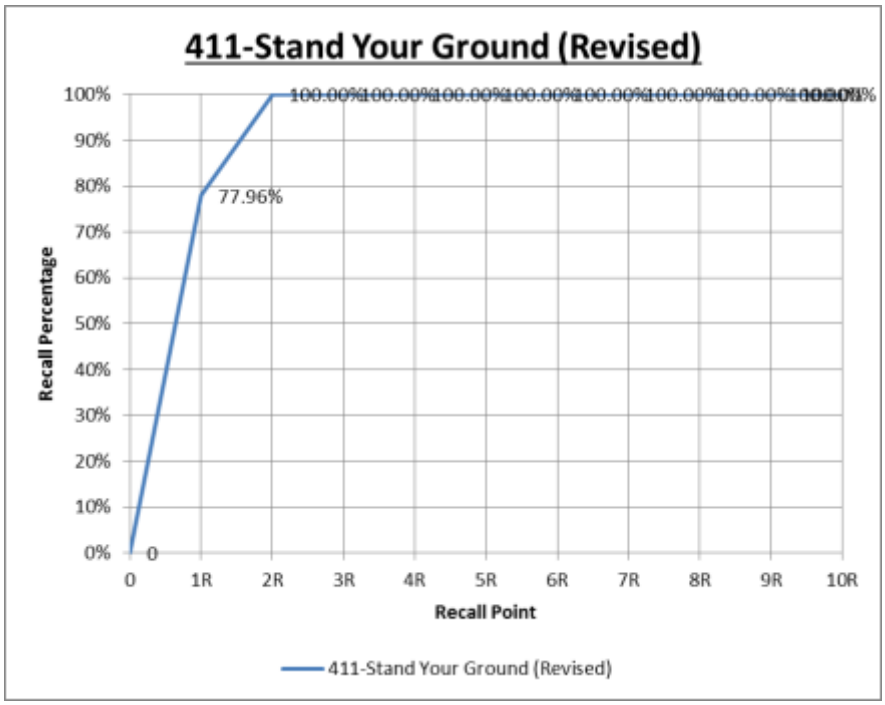
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Stand Your Ground topic, the 90% recall threshold had been attained by submitting only 0.02%% of the corpus, 67 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


411-Stand Your Ground (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
77.96%
0
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
411-Stand Your Ground (Revised)

Topic 412 - 2000 Recount
Total Documents: 290,099
Total Relevant: 850
Total Prevalence: 0.29%

Confusion Matrix - 2000 Recount

	@Reasonable	@90% Recall	@95% Recall
True Positives	*747*	*765*	*808*
True Negatives	*288,351*	*287,968*	*285,458*
False Positives	*898*	*1,281*	*3,791*
False Negatives	*103*	*85*	*42*
Recall	87.88%	90.00%	95.06%
Precision	45.41%	37.39%	17.57%
F1 Measure	59.88%	52.83%	29.66%
Accuracy	99.6549%	99.5291%	98.6787%
Error	0.3451%	0.4709%	1.3213%
Elusion	0.04%	0.03%	0.01%
Fallout	0.31%	0.44%	1.31%

Topic 412 - 2000 Recount - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,410
Total Prevalence: 0.49%

Confusion Matrix - 2000 Recount

	@Reasonable	@90% Recall	@95% Recall
True Positives	*809*	*1,269*	*1,340*
True Negatives	*287,853*	*276,191*	*215,249*
False Positives	*836*	*12,498*	*73,440*
False Negatives	*601*	*141*	*70*
Recall	57.38%	90.00%	95.04%
Precision	49.18%	9.22%	1.79%
F1 Measure	52.96%	16.72%	3.52%
Accuracy	99.50%	95.64%	74.66%
Error	0.50%	4.36%	25.34%
Elusion	0.21%	0.05%	0.03%
Fallout	0.29%	4.33%	25.44%

Summary

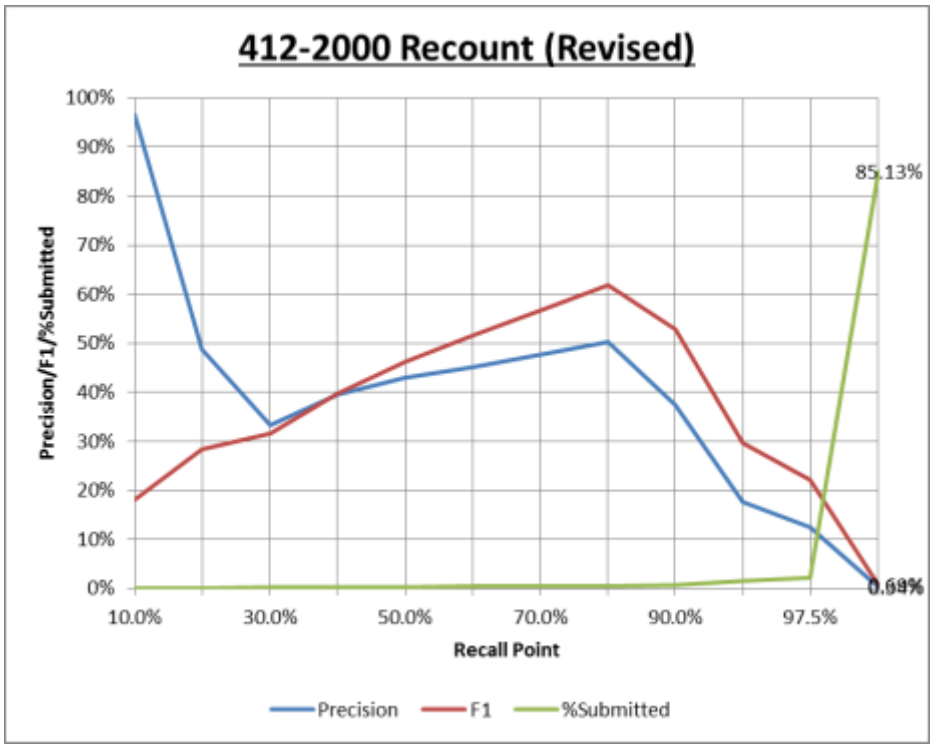
This project was run by Tony Reichenberger. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, within a date filter, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted within the initial date range, the filter was opened up and then finally, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. The fourth submission size was in error, far in excess of what was intended to be submitted; however, other submission sizes were as appropriate given their scoring and expectation.

Reasonable was called when all keywords were exhausted, there was no longer a date filter being applied and scores on documents remaining dropped to 10%.

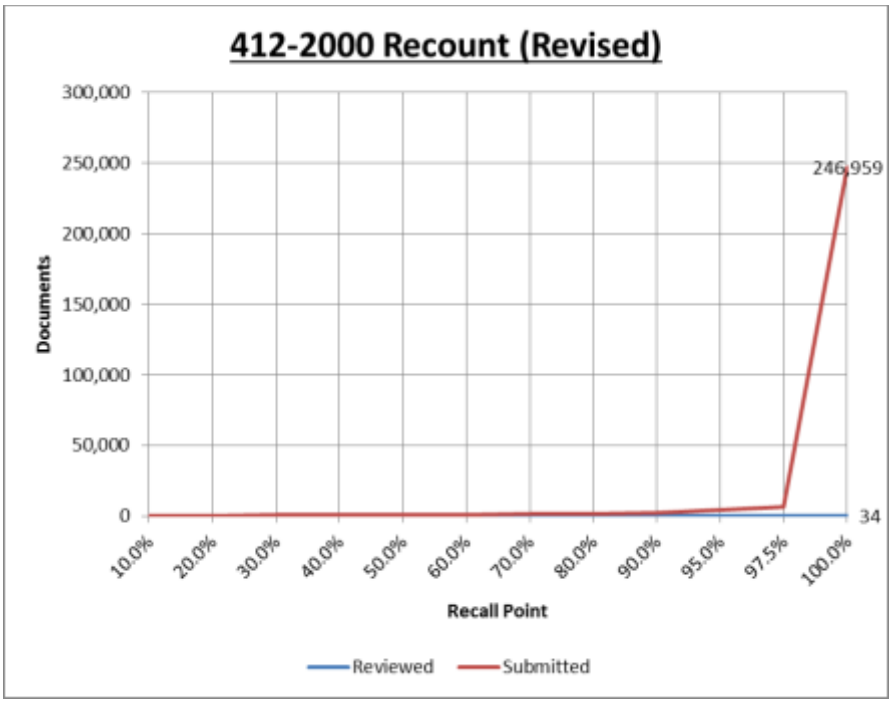
Common errors found in the TREC standard focused on issues for subsequent elections (2002-2008) that had similar problems as in 2000 (e.g. voter disenfranchisement, long lines at polling stations, etc.), but specifically referenced other elections (including a circuit court election, congressional elections, primaries for down ballot races, etc.). Without a reference to the 2000 election in these instances, they should be irrelevant.

Graphs

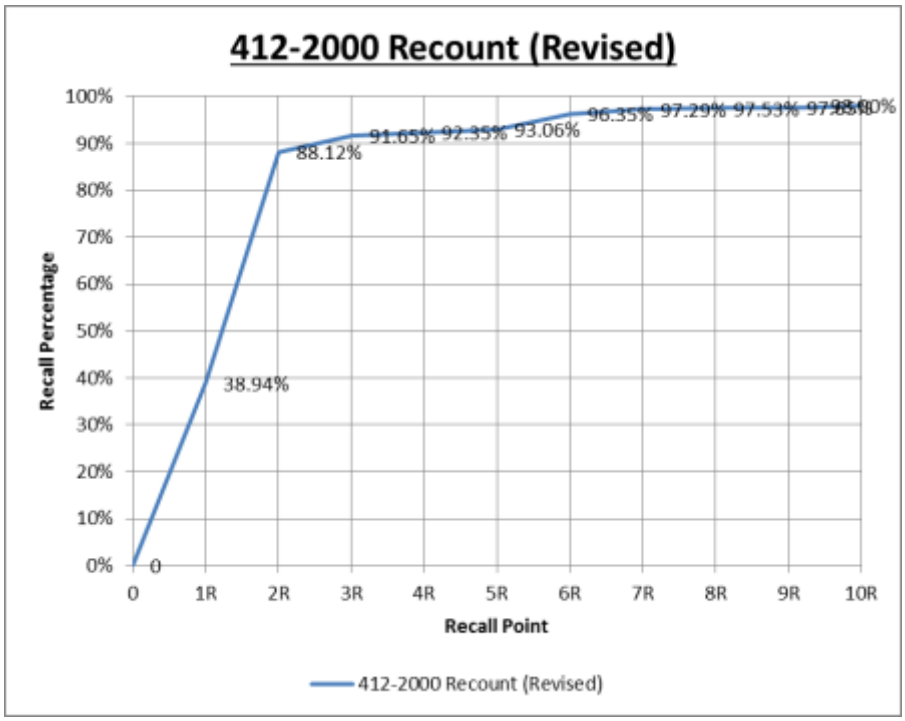
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the 2000 Recount topic, the 90% recall threshold had been attained by submitting only 0.71%% of the corpus, 2,046 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 413 - James V. Crosby
Total Documents: 290,099
Total Relevant: 600
Total Prevalence: 0.21%

Confusion Matrix - James V. Crosby

	@Reasonable	@90% Recall	@95% Recall
True Positives	*581*	*540*	*570*
True Negatives	*289,489*	*289,492*	*289,492*
False Positives	*10*	*7*	*7*
False Negatives	*19*	*60*	*30*
Recall	96.83%	90.00%	95.00%
Precision	98.31%	98.72%	98.79%
F1 Measure	97.57%	94.16%	96.86%
Accuracy	99.9900%	99.9769%	99.9872%
Error	0.0100%	0.0231%	0.0128%
Elusion	0.01%	0.02%	0.01%
Fallout	0.00%	0.00%	0.00%

Topic 413 - James V. Crosby - UNCORRECTED
Total Documents: 290,099
Total Relevant: 546
Total Prevalence: 0.19%

Confusion Matrix - James V. Crosby

	@Reasonable	@90% Recall	@95% Recall
True Positives	*526*	*492*	*519*
True Negatives	*289,488*	*289,495*	*289,493*
False Positives	*65*	*58*	*60*
False Negatives	*20*	*54*	*27*
Recall	96.34%	90.11%	95.05%
Precision	89.00%	89.45%	89.64%
F1 Measure	92.52%	89.78%	92.27%
Accuracy	99.97%	99.96%	99.97%
Error	0.03%	0.04%	0.03%
Elusion	0.01%	0.02%	0.01%
Fallout	0.02%	0.02%	0.02%

Summary

Topic 413 was run by Jim Sullivan, who started on August 12, 2016 and concluded on the same day.

Sullivan entered this topic with no prior knowledge of James V. Crosby. At first he thought it was a legal case, with James as the Plaintiff and Crosby as the Defendant. That was not accurate.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on variations of crosby subject line, and moved broader variations of the name anywhere in the document. He called 70% recall after submitting 422 documents, with 397 relevant. Almost all of the 25 false positives were obvious errors in the TREC standard. 500 random documents were trained Not Relevant and a learning session was initiated.

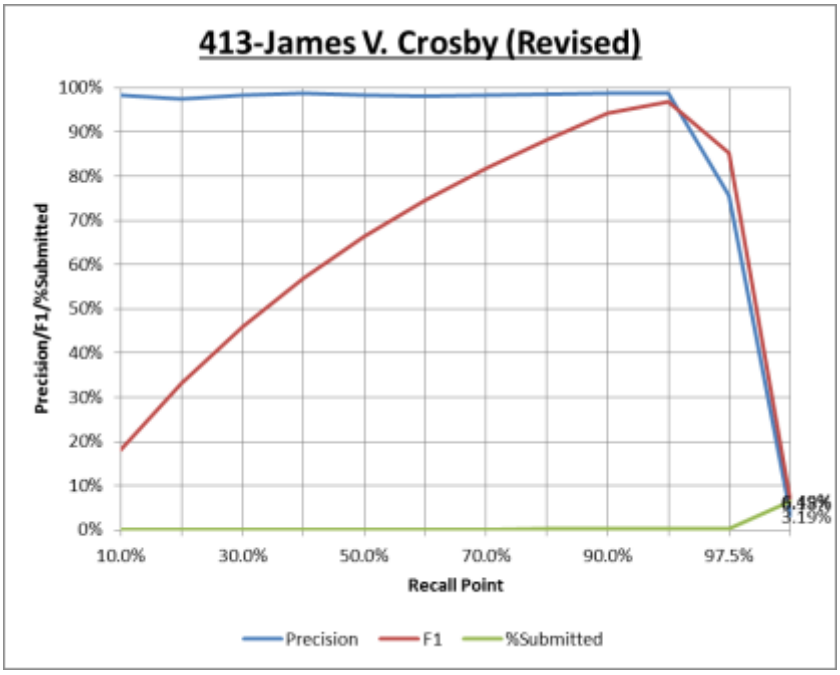
Sullivan continued with variations of keyword terms until he called Reasonable after 591 documents submitted, with 526 being returned Relevant. Most of the 65 documents returned Not Relevant were again clear errors.

He submitted all remaining documents that contained the term Crosby, followed by the rest with the highest scores being submitted first. A total of 546 documents were returned relevant by TREC. In total, 3.0 hours were spent reviewing this very easy topic. The use of predictive coding on this topic was unnecessary.

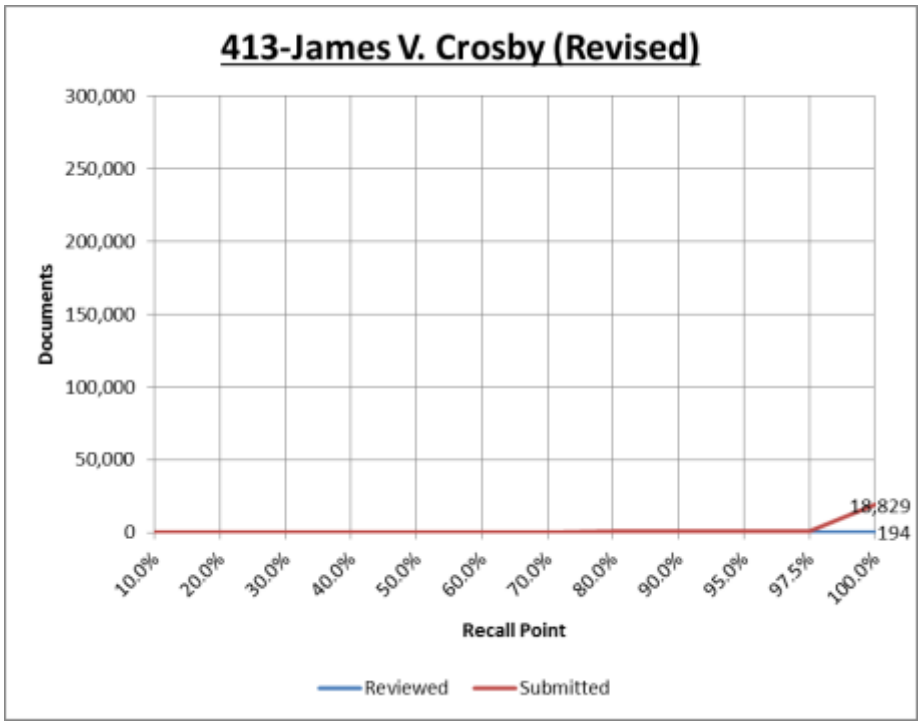
This topic had an average TREC standard. Though he identified 56 documents that were clearly erroneous, overall the standard was clear and the inconsistencies weren't widespread.

Graphs

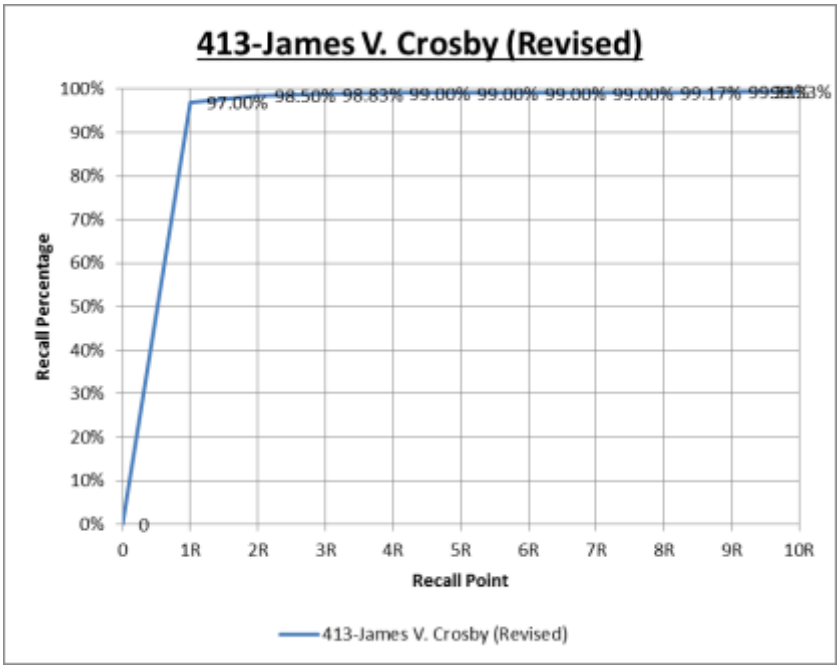
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the James V. Crosby topic, the 90% recall threshold had been attained by submitting only 0.19%% of the corpus, 547 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


413-James V. Crosby (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
97.00%
0
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
413-James V. Crosby (Revised)

Topic 414 - Medicaid Reform
Total Documents: 290,099
Total Relevant: 844
Total Prevalence: 0.29%

Confusion Matrix - Medicaid Reform

	@Reasonable	@90% Recall	@95% Recall
True Positives	*783*	*760*	*802*
True Negatives	*287,858*	*288,177*	*286,907*
False Positives	*1,397*	*1,078*	*2,348*
False Negatives	*61*	*84*	*42*
Recall	92.77%	90.05%	95.02%
Precision	35.92%	41.35%	25.46%
F1 Measure	51.79%	56.67%	40.16%
Accuracy	99.4974%	99.5994%	99.1761%
Error	0.5026%	0.4006%	0.8239%
Elusion	0.02%	0.03%	0.01%
Fallout	0.48%	0.37%	0.81%

Topic 414 - Medicaid Reform - UNCORRECTED
Total Documents: 290,099
Total Relevant: 839
Total Prevalence: 0.29%

Confusion Matrix - Medicaid Reform

	@Reasonable	@90% Recall	@95% Recall
True Positives	*0*	*756*	*798*
True Negatives	*287,115*	*288,111*	*286,515*
False Positives	*2,145*	*1,149*	*2,745*
False Negatives	*839*	*83*	*41*
Recall	0.00%	90.11%	95.11%
Precision	0.00%	39.69%	22.52%
F1 Measure	#DIV/0!	55.10%	36.42%
Accuracy	98.97%	99.58%	99.04%
Error	1.03%	0.42%	0.96%
Elusion	0.29%	0.03%	0.01%
Fallout	0.74%	0.40%	0.95%

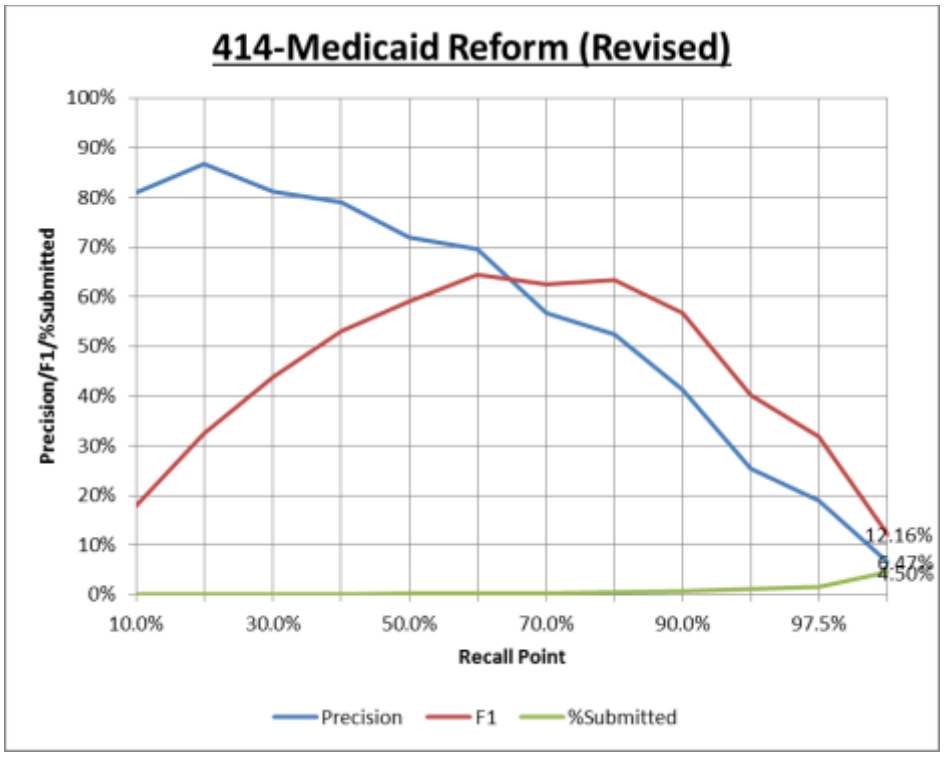
Summary

This topic was run by Tony Reichenberger. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken.

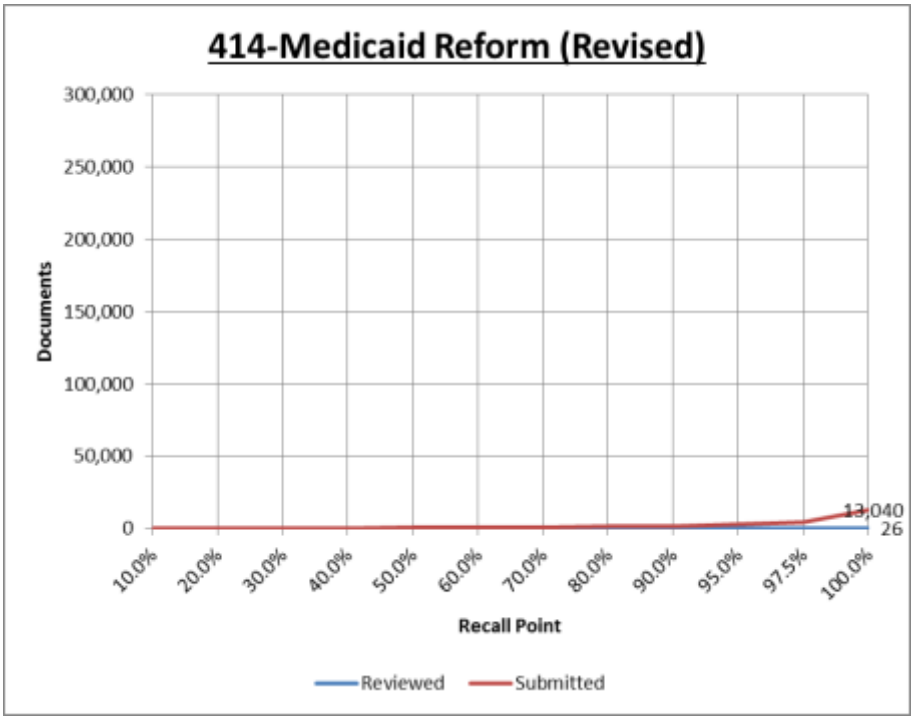
Reasonable was called when all scores dropped below 7.5% probability.

Graphs

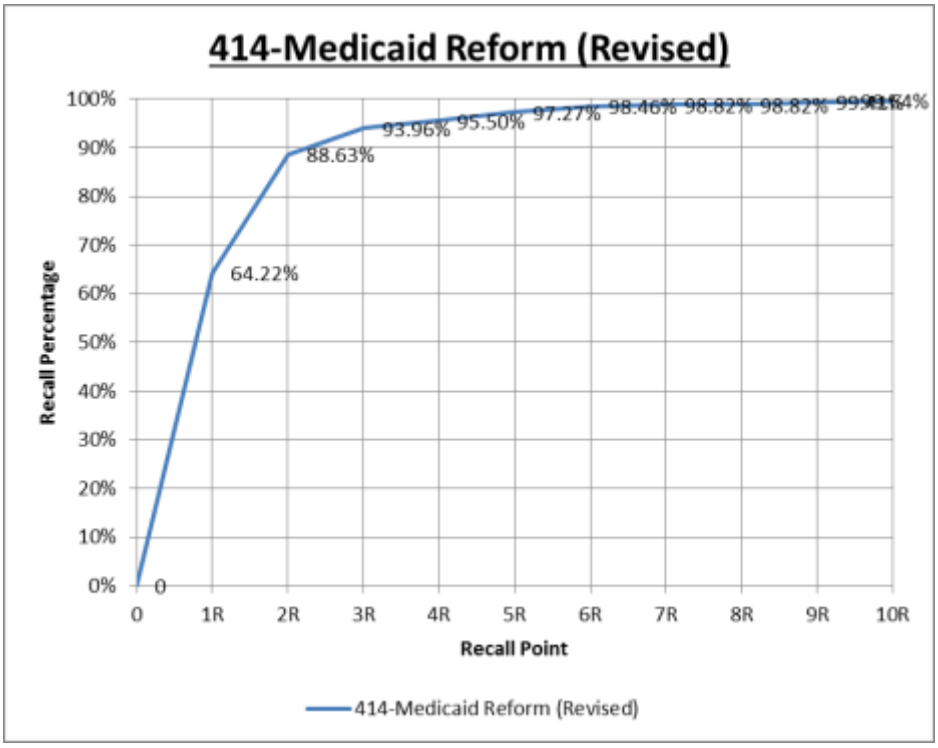
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Medicaid Reform topic, the 90% recall threshold had been attained by submitting only 0.63%% of the corpus, 1,838 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 415 - George W. Bush
Total Documents: 290,099
Total Relevant: 12,267
Total Prevalence: 4.23%

Confusion Matrix - George W. Bush

	@Reasonable	@90% Recall	@95% Recall
True Positives	*11,554*	*11,041*	*11,654*
True Negatives	*276,876*	*277,056*	*275,461*
False Positives	*956*	*776*	*2,371*
False Negatives	*713*	*1,226*	*613*
Recall	94.19%	90.01%	95.00%
Precision	92.36%	93.43%	83.09%
F1 Measure	93.26%	91.69%	88.65%
Accuracy	99.4247%	99.3099%	98.9714%
Error	0.5753%	0.6901%	1.0286%
Elusion	0.26%	0.44%	0.22%
Fallout	0.34%	0.28%	0.85%

Topic 415 - George W. Bush - UNCORRECTED
Total Documents: 290,099
Total Relevant: 12,106
Total Prevalence: 4.17%

Confusion Matrix - George W. Bush

	@Reasonable	@90% Recall	@95% Recall
True Positives	*11,389*	*10,896*	*11,501*
True Negatives	*276,872*	*277,056*	*275,265*
False Positives	*1,121*	*937*	*2,728*
False Negatives	*717*	*1,210*	*605*
Recall	94.08%	90.00%	95.00%
Precision	91.04%	92.08%	80.83%
F1 Measure	92.53%	91.03%	87.34%
Accuracy	99.37%	99.26%	98.85%
Error	0.63%	0.74%	1.15%
Elusion	0.26%	0.43%	0.22%
Fallout	0.40%	0.34%	0.98%

Summary

Topic 415 was run by Jim Sullivan, who started on August 22, 2016 and concluded on August 29, 2016.

Sullivan entered this topic with general knowledge of George W. Bush. Like most people, he is familiar with the former President of the United States, but he didn't have any special knowledge.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. This topic was especially tricky due to "Bush" appearing in every document in the database. He started by submitting documents that hit on obvious terms in the subject line, and moved broader variations anywhere in the document. By the end of the first day, he was comfortable that he had found most of the relevant material. He was way off. He called 70% recall after submitting 1,233 documents, with 1,207 returned Relevant. He disagreed with most returned Not Relevant, but the mistakes seemed reasonable given such a high prevalence.

On day two, he started submitting large batches of search term hits and found a very significant volume of new hits. He had previously missed a large collection of documents with nothing more than a reference to the "President." He trained 2,000 randomly selected documents as Not Relevant, and initiated a learning session. From there he decided to rely much more heavily on the predictive coding scores as to not miss another significant set of documents.

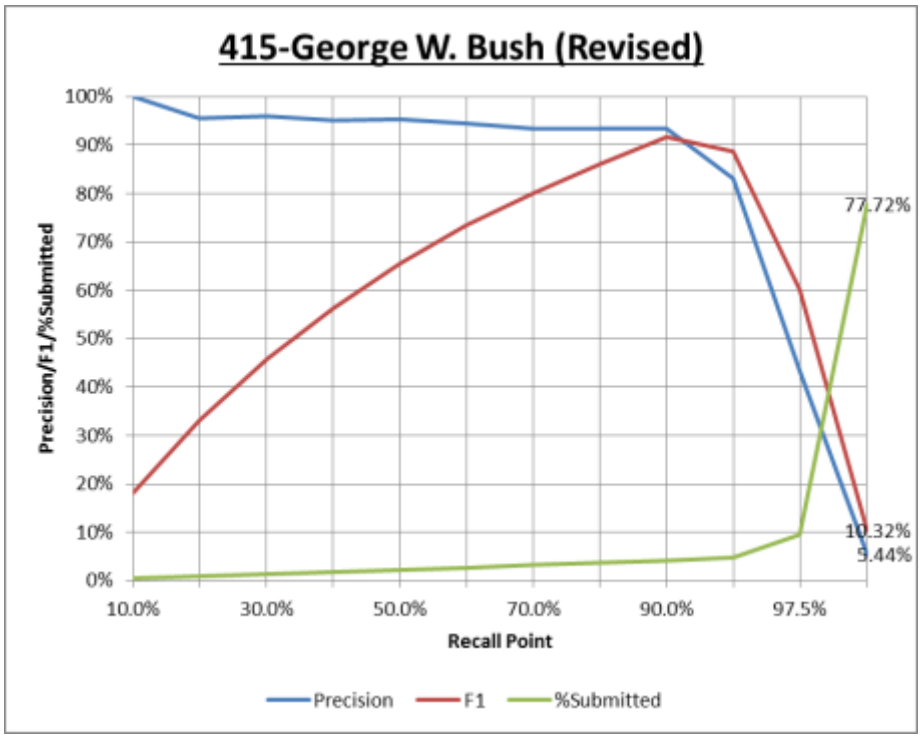
Relying on a combination of keywords and predictive coding scores, a large set of additional relevant documents were discovered. Reasonable recall wasn't called until 12,510 documents were submitted, with 11,389 being returned as Relevant.

To finish up, he submitted all remaining documents with the highest scores being submitted first. A total of 12,106 documents were returned relevant by TREC. In total, 3.5 hours were spent reviewing this high prevalence topic.

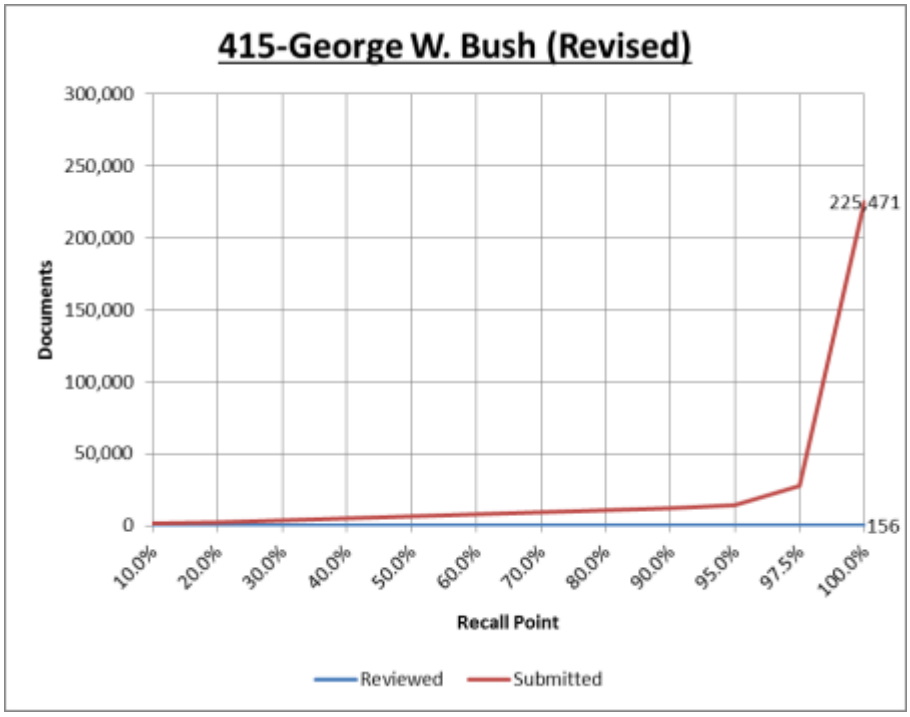
This topic had an above average TREC standard. Though he identified 169 documents that were clearly erroneous, overall the standard was clear and the inconsistencies weren't widespread. He was impressed how TREC properly returned vague references to George W. Bush without any relevant keywords present. The small number of errors is very reasonable for a topic with such high overall prevalence.

Graphs

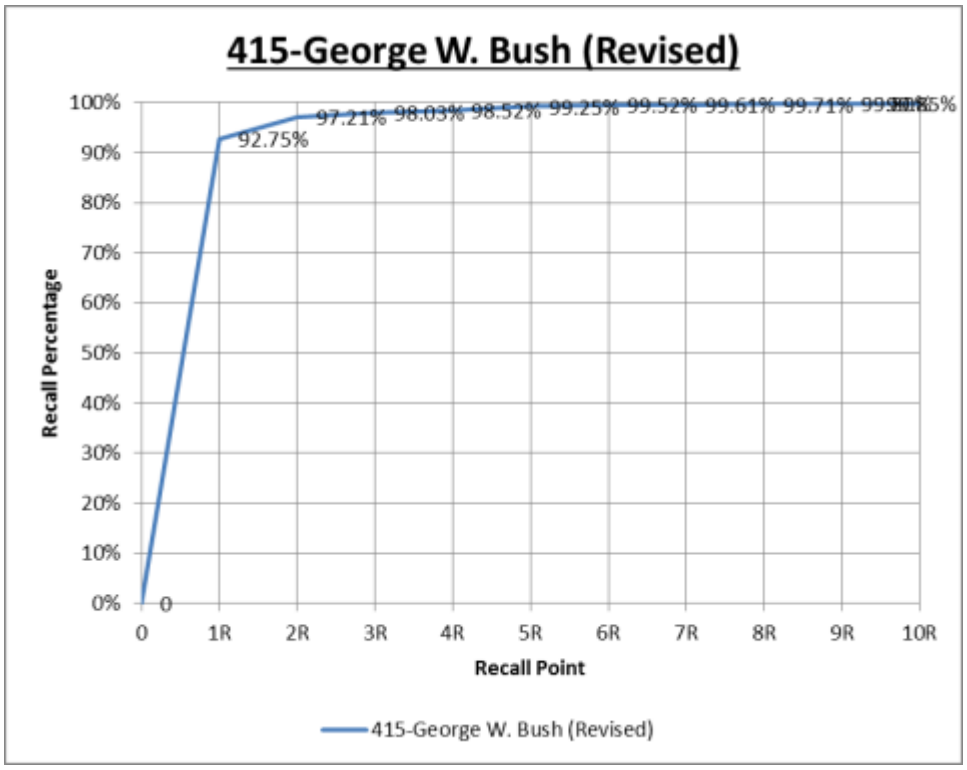
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the George W. Bush topic, the 90% recall threshold had been attained by submitting only 4.07%% of the corpus, 11,817 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 416 - Marketing
Total Documents: 290,099
Total Relevant: 1,485
Total Prevalence: 0.51%

Confusion Matrix - Marketing

	@Reasonable	@90% Recall	@95% Recall
True Positives	*911*	*1,337*	*1,411*
True Negatives	*287,453*	*269,283*	*263,314*
False Positives	*1,161*	*19,331*	*25,300*
False Negatives	*574*	*148*	*74*
Recall	61.35%	90.03%	95.02%
Precision	43.97%	6.47%	5.28%
F1 Measure	51.22%	12.07%	10.01%
Accuracy	99.4019%	93.2854%	91.2533%
Error	0.5981%	6.7146%	8.7467%
Elusion	0.20%	0.05%	0.03%
Fallout	0.40%	6.70%	8.77%

Topic 416 – Marketing - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,446
Total Prevalence: 0.50%

Confusion Matrix - Marketing

	@Reasonable	@90% Recall	@95% Recall
True Positives	*872*	*1,302*	*1,374*
True Negatives	*287,453*	*269,113*	*263,258*
False Positives	*1,200*	*19,540*	*25,395*
False Negatives	*574*	*144*	*72*
Recall	60.30%	90.04%	95.02%
Precision	42.08%	6.25%	5.13%
F1 Measure	49.57%	11.68%	9.74%
Accuracy	99.39%	93.21%	91.22%
Error	0.61%	6.79%	8.78%
Elusion	0.20%	0.05%	0.03%
Fallout	0.42%	6.77%	8.80%

Summary

Topic 416 was run by Jim Sullivan, who started on July 27, 2016 and concluded on August 26.

Sullivan entered this topic blind to what could be meant by Marketing in Florida. He was far from being an expert by any standard.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms, and moved to more generic lists. While he entered the topic blind, things only got more difficult once he began reviewing TREC's feedback on his initial submissions. Finding documents relating to "visit florida" or "marketing" were only returned relevant 1/3 of the time, and for seemingly indistinguishable reasons.

Though frustrated and confused by the TREC standard, 80% recall was called after 373 documents were submitted, with 130 relevant. He was only able to achieve 34.9% precision on his own.

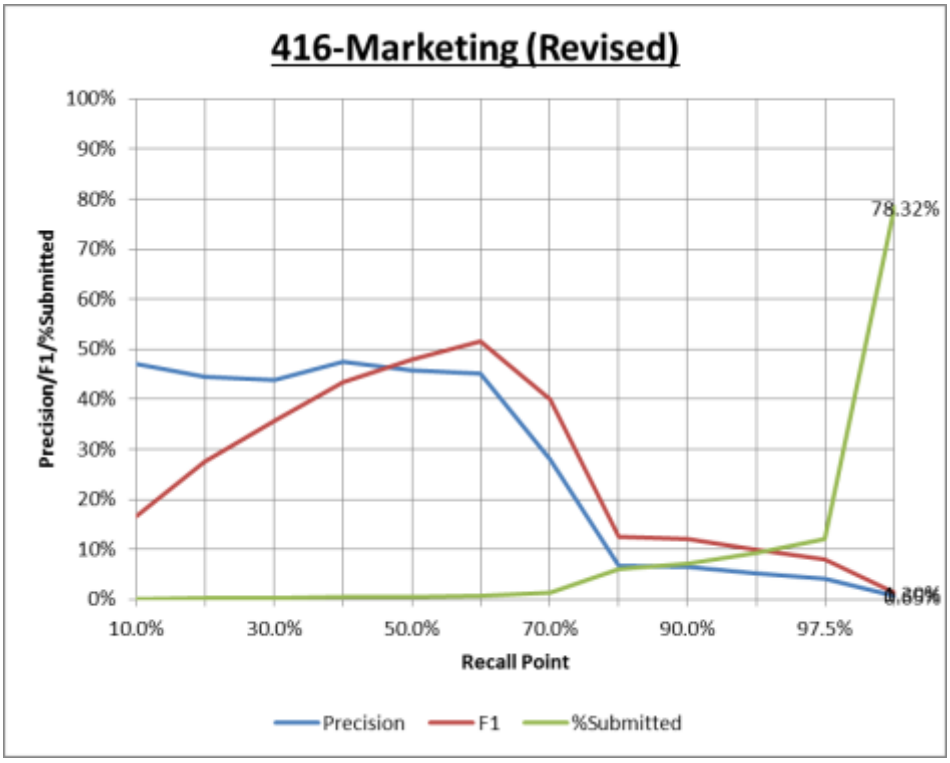
At this point he just started blindly submitting the highest scoring documents based on predictive coding, and got better results than he did by looking at anything. He continued iterations of submissions and learning sessions until calling reasonable after 2,072 submitted, with 872 relevant. Mr. EDR was able to get 43.7% precision without any input.

After the reasonable call, all remaining documents were submitted by predictive coding score with the highest scores being submitted first. A total of 1,446 documents were returned relevant by TREC. In total, 7.0 hours were spent reviewing this topic.

This topic was the poorest gold standard Sullivan faced of all his TREC topics. Though he could only identify 39 documents that were clearly erroneous, most of the errors were related to inconsistencies, where similar documents were classified differently. In the end, he was rarely able to understand what was supposed to be relevant well enough to determine what was a mistake.

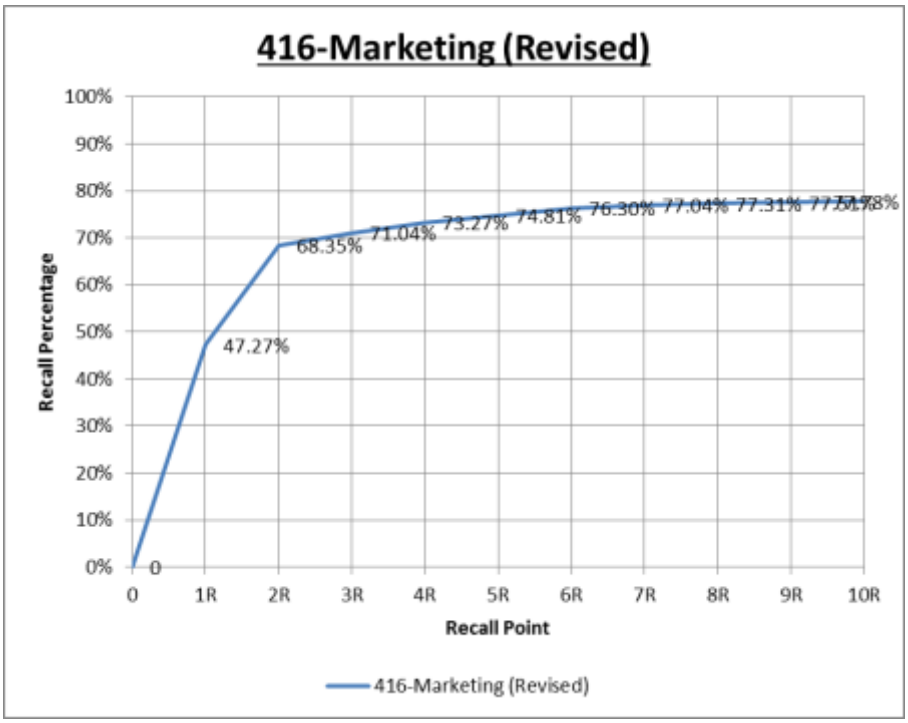
Graphs

The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Marketing topic, the 90% recall threshold had been attained by submitting only 7.12%% of the corpus, 20,668 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.

The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 417 - Movie Gallery
Total Documents: 290,099
Total Relevant: 5,945
Total Prevalence: 2.05%

Confusion Matrix - Movie Gallery

	@Reasonable	@90% Recall	@95% Recall
True Positives	*5,945*	*5,351*	*5,648*
True Negatives	*284,154*	*284,154*	*284,154*
False Positives	*0*	*0*	*0*
False Negatives	*0*	*594*	*297*
Recall	100.00%	90.01%	95.00%
Precision	100.00%	100.00%	100.00%
F1 Measure	100.00%	94.74%	97.44%
Accuracy	100.0000%	99.7952%	99.8976%
Error	0.0000%	0.2048%	0.1024%
Elusion	0.00%	0.21%	0.10%
Fallout	0.00%	0.00%	0.00%

Topic 417 - Movie Gallery - UNCORRECTED
Total Documents: 290,099
Total Relevant: 5,931
Total Prevalence: 2.04%

Confusion Matrix - Movie Gallery

	@Reasonable	@90% Recall	@95% Recall
True Positives	*5,908*	*5,338*	*5,635*
True Negatives	*284,131*	*284,146*	*284,141*
False Positives	*37*	*22*	*27*
False Negatives	*23*	*593*	*296*
Recall	99.61%	90.00%	95.01%
Precision	99.38%	99.59%	99.52%
F1 Measure	99.49%	94.55%	97.21%
Accuracy	99.98%	99.79%	99.89%
Error	0.02%	0.21%	0.11%
Elusion	0.01%	0.21%	0.10%
Fallout	0.01%	0.01%	0.01%

Summary

This topic was run by Losey from July 6th to 7th 2016. He took a total of five hours on this fairly simple project with most of the time doing keyword searches. He created 27 search fodlers, reviewed only 66 documents, but manually catergorized 5,966 documents (bulk coding).

The topic is defined as: **Movie Gallery-All documents concerning investments or divestments by the State of Florida in Movie Gallery.**

The Movie Gallery is a publically traded pornography company in which the great State of Florida decided to invest some of its employee pension funds. When this was eventually discovered by the public, and then a form email campaign was launched by citizens and employees both.

The work began in an unusual fashion. Losey did keyword search and then submitted all 5,932 documents that have the keyword phrase “movie gallery” in them. He only did a 15 minute judgmental sample review of this folder to see they all were relevant. They seemed to all be pretty much the same form email. So, as an experiment, he decided to just submit them all at once. They were in fact all relevant.

There were 5,945 Relevant documents on this issue out of the total of 290,099 (after correcting for the 58 obvious errors in coding made by the TREC assessor).

By use of one keyword search "movie gallery" Losey found 5,932 of them. That is 99.78% RECALL, 100% Precision from one search.

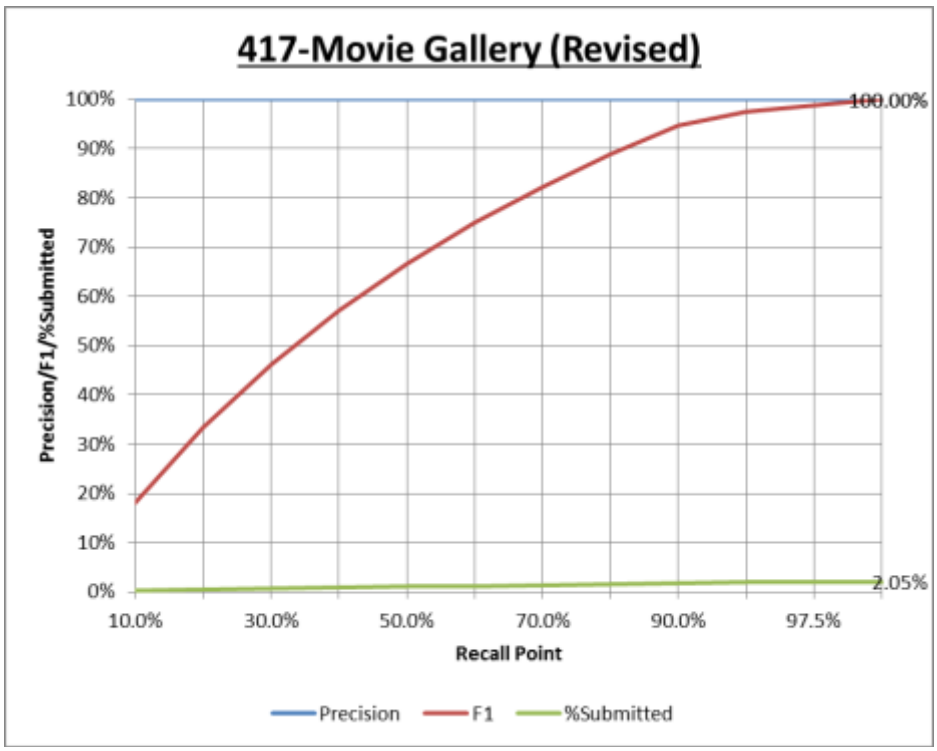
By use of a second series of keyword searches Losey found 7 more relevant documents, for a total of 5,939. That is 99.90% RECALL. 100% Precision.

By use of Mr. EDR – AI based ranking - he found 6 more relevant documents, for a total of 5,945, and called REASONABLE. That is 100% RECALL and 100% Precision.

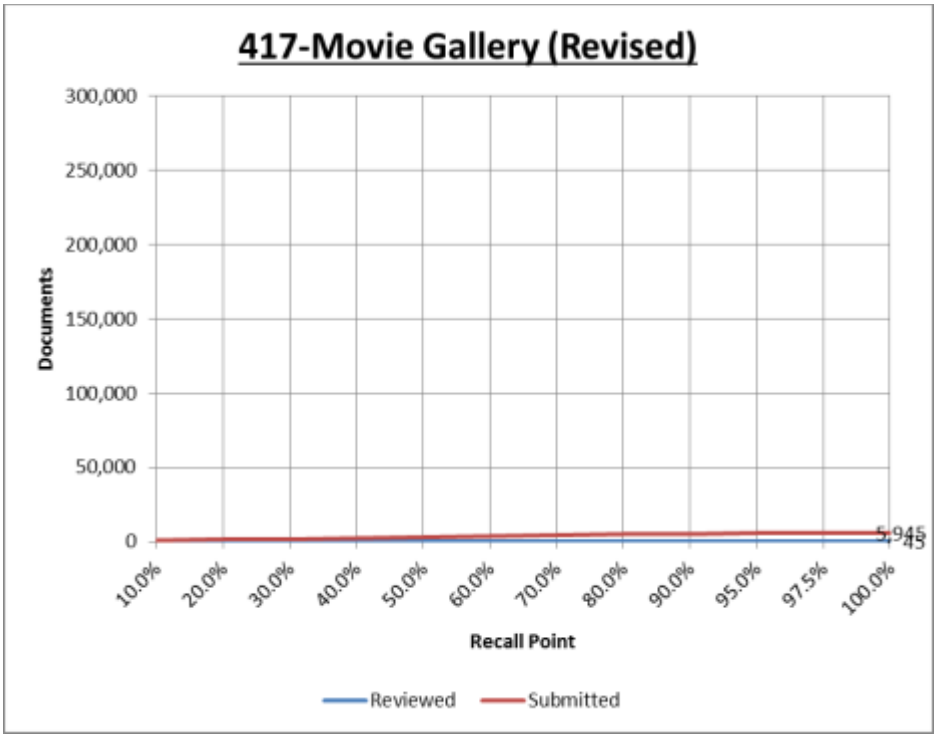
This topic was fairly easy, but did have some subtleties, including the selection of the right balance of irrelevant training docs and having the confidence to call reasonable early. The confidence was provided by Mr. EDR. Just before his perfect call, Losey looked all the way down to 3%, and only 8 new documents were seen, none even close to relevant). The document ranking served as an excellent quality assurance tool and made it easier to make the right Stop call.

Graphs

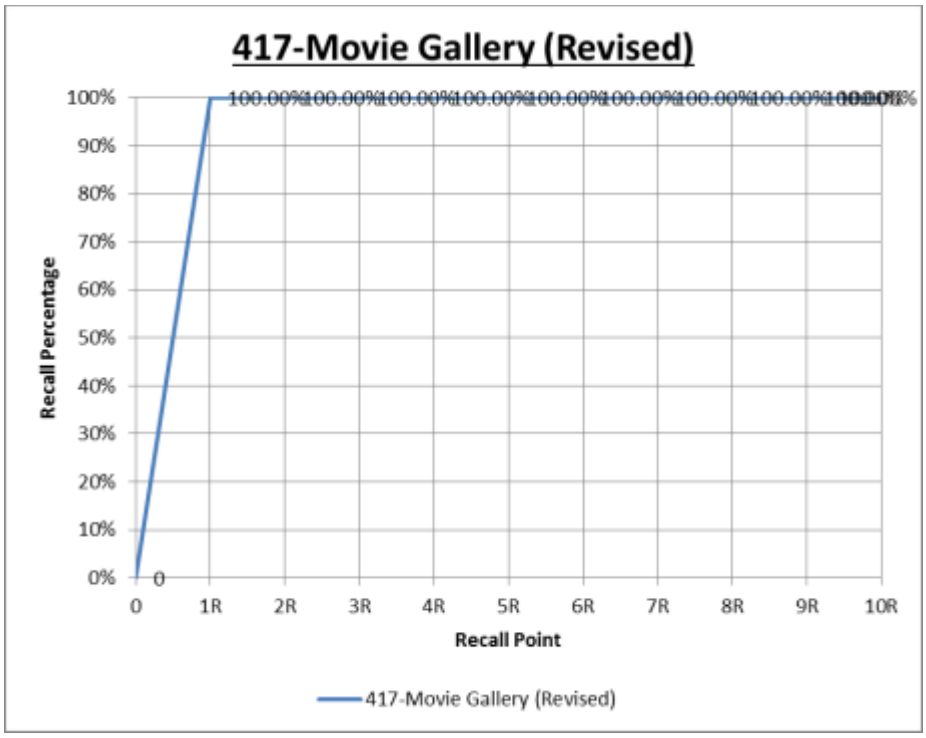
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Movie Gallery topic, the 90% recall threshold had been attained by submitting only 1.84%% of the corpus, 5,351 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 418 - War Preparations
Total Documents: 290,099
Total Relevant: 141
Total Prevalence: 0.05%

Confusion Matrix - War Preparations

	@Reasonable	@90% Recall	@95% Recall
True Positives	*114*	*127*	*134*
True Negatives	*289,925*	*287,707*	*286,196*
False Positives	*33*	*2,251*	*3,762*
False Negatives	*27*	*14*	*7*
Recall	80.85%	90.07%	95.04%
Precision	77.55%	5.34%	3.44%
F1 Measure	79.17%	10.08%	6.64%
Accuracy	99.9793%	99.2192%	98.7008%
Error	0.0207%	0.7808%	1.2992%
Elusion	0.01%	0.00%	0.00%
Fallout	0.01%	0.78%	1.30%

Topic 418 - War Preparations - UNCORRECTED
Total Documents: 290,099
Total Relevant: 187
Total Prevalence: 0.06%

Confusion Matrix - War Preparations

	@Reasonable	@90% Recall	@95% Recall
True Positives	*74*	*169*	*178*
True Negatives	*289,839*	*279,562*	*271,871*
False Positives	*73*	*10,350*	*18,041*
False Negatives	*113*	*18*	*9*
Recall	39.57%	90.37%	95.19%
Precision	50.34%	1.61%	0.98%
F1 Measure	44.31%	3.16%	1.93%
Accuracy	99.94%	96.43%	93.78%
Error	0.06%	3.57%	6.22%
Elusion	0.04%	0.01%	0.00%
Fallout	0.03%	3.57%	6.22%

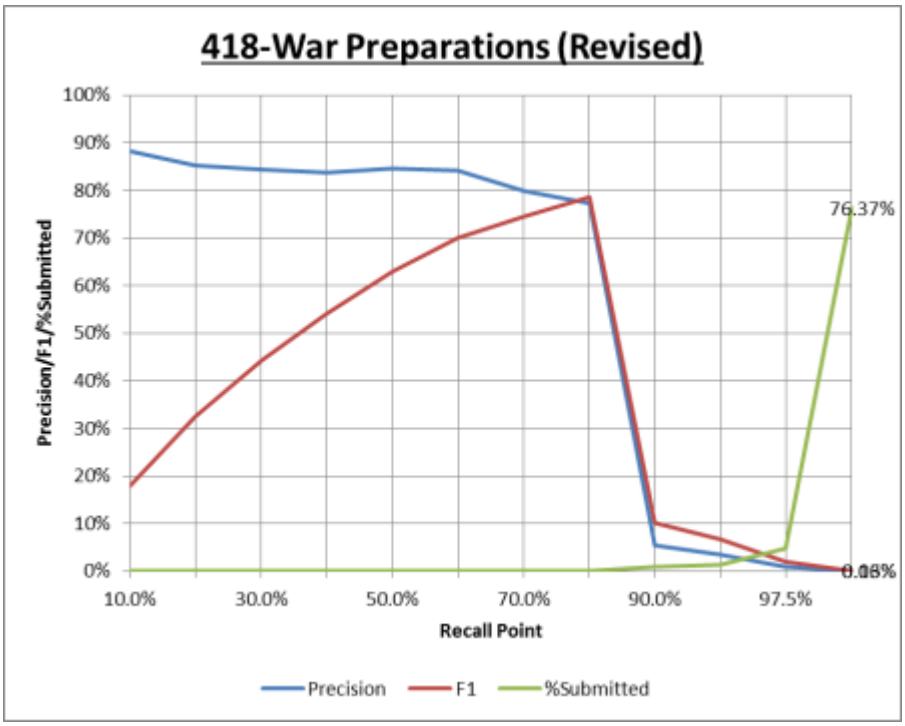
Summary

This project was run by Tony Reichenberger. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-20 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring; when additional relevant materials were found, subsequent searches for similar documents were partaken.

Reasonable was called too early on this topic, as precision and quality of documents preceding the call steeply diminished. Subsequent submissions post-call were confined to a date filter to enhance precision which resulted in additional relevant materials not previously considered being found. As additional relevant documents were found, additional searches and learning sessions were conducted as follow ups, with those documents being included in the next submission.

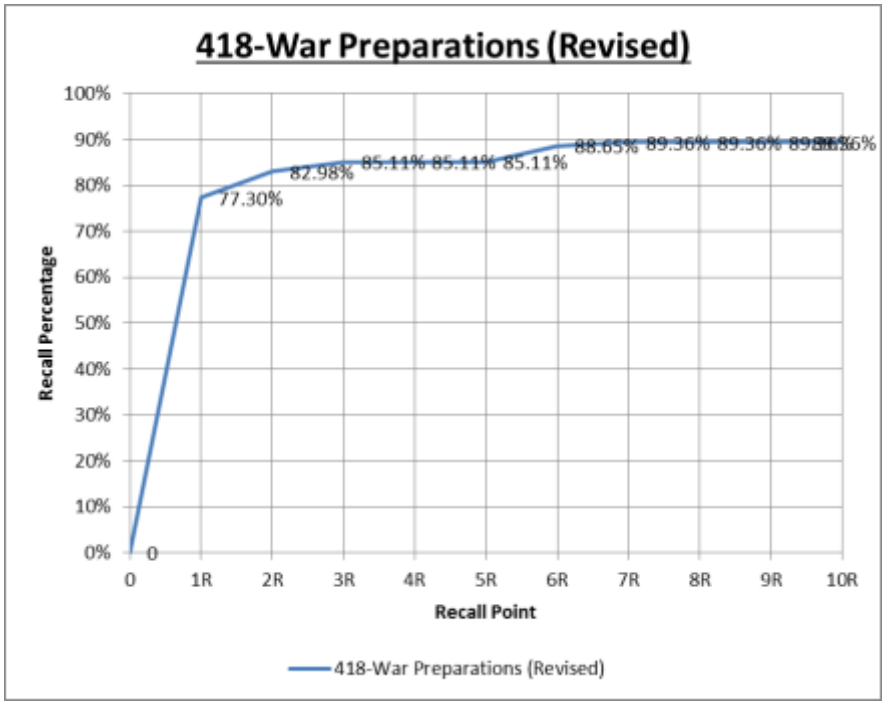
Graphs

The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the War Preparations topic, the 90% recall threshold had been attained by submitting only 0.82%% of the corpus, 2,378 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.

The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


418-War Preparations (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
77.30%
82.98%
85.11%
85.11%
85.11%
88.65%
89.36%
89.36%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
418-War Preparations (Revised)

Topic 419 - Lost Foster Child Rilya Wilson
Total Documents: 290,099
Total Relevant: 1,982
Total Prevalence: 0.68%

Confusion Matrix - Lost Foster Child Rilya Wilson

	@Reasonable	@90% Recall	@95% Recall
True Positives	*1,964*	*1,784*	*1,883*
True Negatives	*277,007*	*285,486*	*283,977*
False Positives	*11,110*	*2,631*	*4,140*
False Negatives	*18*	*198*	*99*
Recall	99.09%	90.01%	95.01%
Precision	15.02%	40.41%	31.26%
F1 Measure	26.09%	55.78%	47.05%
Accuracy	96.1641%	99.0248%	98.5388%
Error	3.8359%	0.9752%	1.4612%
Elusion	0.01%	0.07%	0.03%
Fallout	3.86%	0.91%	1.44%

Topic 419 - Lost Foster Child Rilya Wilson - UNCORRECTED
Total Documents: 290,099
Total Relevant: 1,989
Total Prevalence: 0.69%

Confusion Matrix - Lost Foster Child Rilya Wilson

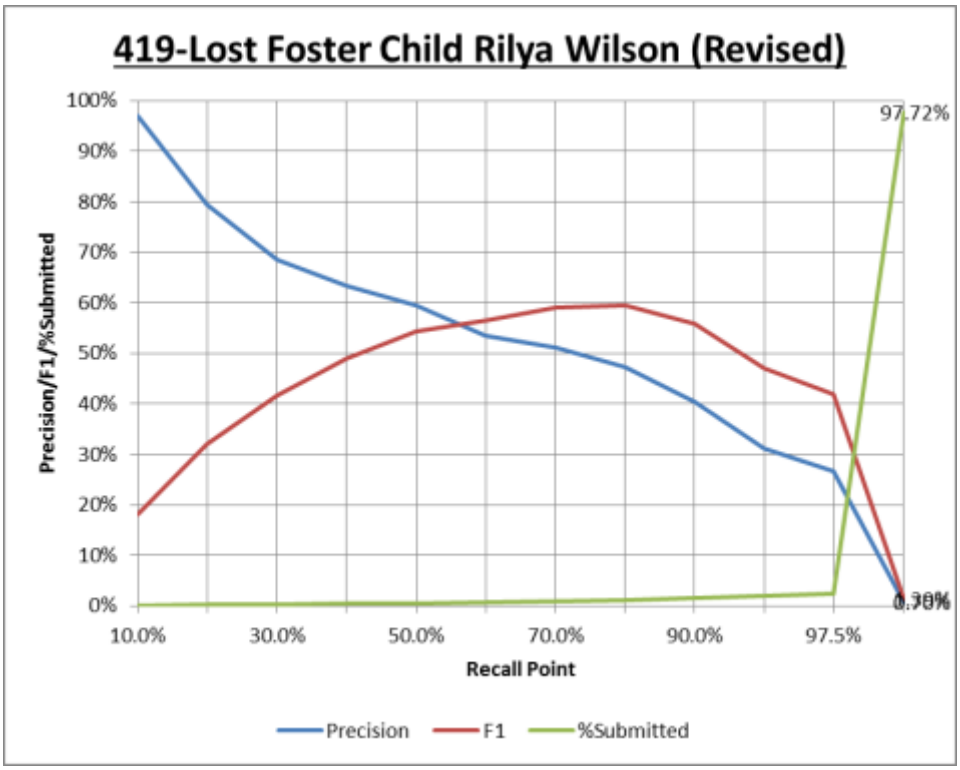
	@Reasonable	@90% Recall	@95% Recall
True Positives	*1,966*	*1,791*	*1,890*
True Negatives	*277,002*	*285,321*	*283,642*
False Positives	*11,108*	*2,789*	*4,468*
False Negatives	*23*	*198*	*99*
Recall	98.84%	90.05%	95.02%
Precision	15.04%	39.10%	29.73%
F1 Measure	26.10%	54.53%	45.29%
Accuracy	96.16%	98.97%	98.43%
Error	3.84%	1.03%	1.57%
Elusion	0.01%	0.07%	0.03%
Fallout	3.86%	0.97%	1.55%

Summary

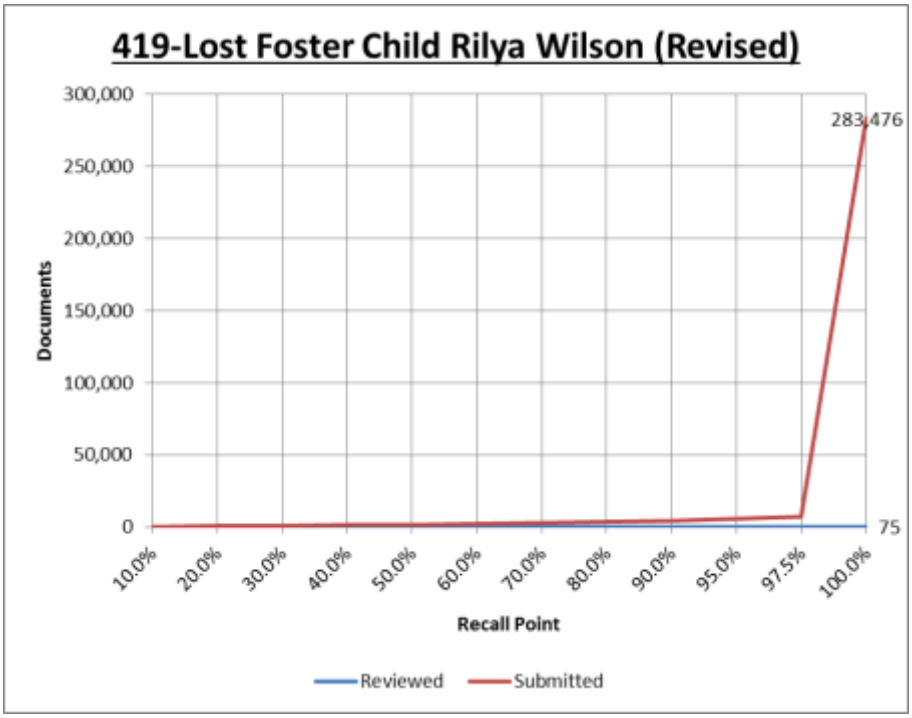
This topic was run by Levi Kuehn. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. When scores dropped to 5%, a final search was submitted, another learning session run, and documents were submitted in probability order.

Graphs

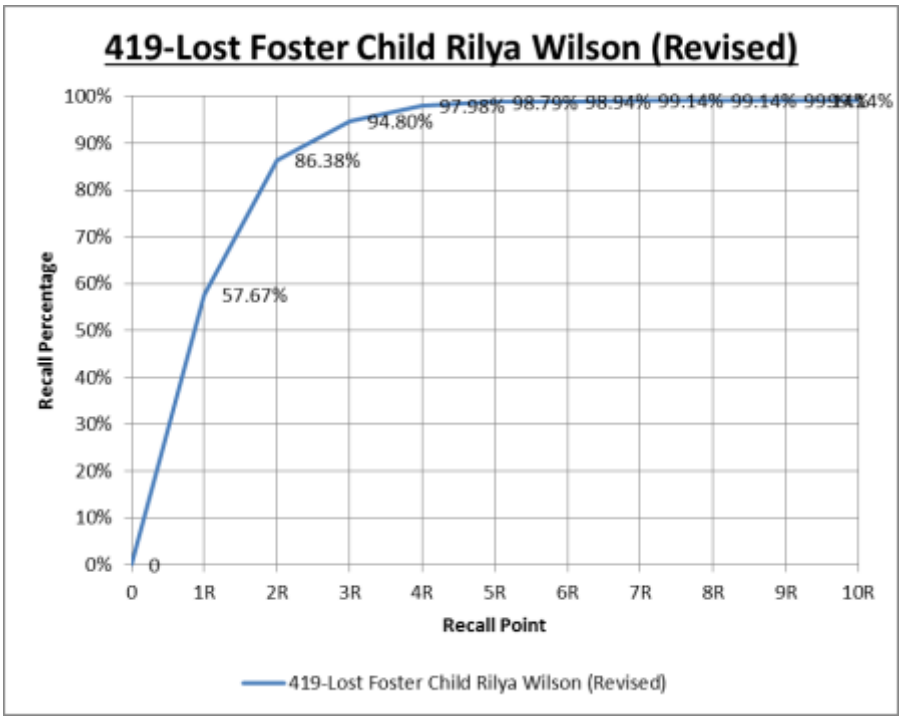
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Lost Foster Child Rilya Wilson topic, the 90% recall threshold had been attained by submitting only 1.52%% of the corpus, 4,415 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 420 - Billboards
Total Documents: 290,099
Total Relevant: 739
Total Prevalence: 0.25%

Confusion Matrix - Billboards

	@Reasonable	@90% Recall	@95% Recall
True Positives	*707*	*666*	*703*
True Negatives	*289,327*	*289,327*	*289,327*
False Positives	*33*	*33*	*33*
False Negatives	*32*	*73*	*36*
Recall	95.67%	90.12%	95.13%
Precision	95.54%	95.28%	95.52%
F1 Measure	95.61%	92.63%	95.32%
Accuracy	99.9776%	99.9635%	99.9762%
Error	0.0224%	0.0365%	0.0238%
Elusion	0.01%	0.03%	0.01%
Fallout	0.01%	0.01%	0.01%

Topic 420 – Billboards - UNCORRECTED
Total Documents: 290,099
Total Relevant: 737
Total Prevalence: 0.25%

Confusion Matrix - Billboards

	@Reasonable	@90% Recall	@95% Recall
True Positives	*682*	*664*	*701*
True Negatives	*289,304*	*289,304*	*289,224*
False Positives	*58*	*58*	*138*
False Negatives	*55*	*73*	*36*
Recall	92.54%	90.09%	95.12%
Precision	92.16%	91.97%	83.55%
F1 Measure	92.35%	91.02%	88.96%
Accuracy	99.96%	99.95%	99.94%
Error	0.04%	0.05%	0.06%
Elusion	0.02%	0.03%	0.01%
Fallout	0.02%	0.02%	0.05%

Summary

Topic 420 was run by Jim Sullivan, who started on August 22, 2016 and concluded on August 25, 2016.

Sullivan entered this topic with little knowledge of billboard and their legal status in Florida. While he certainly has driven by his share of billboards on the highway, that's as far as his prior knowledge extends.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms in the subject line, and moved broader variations anywhere in the document. By the end of the first day he had a very good understanding of what was relevant to the TREC standard for the topic. He called 70% recall after submitting 557 documents, with 516 returned Relevant.

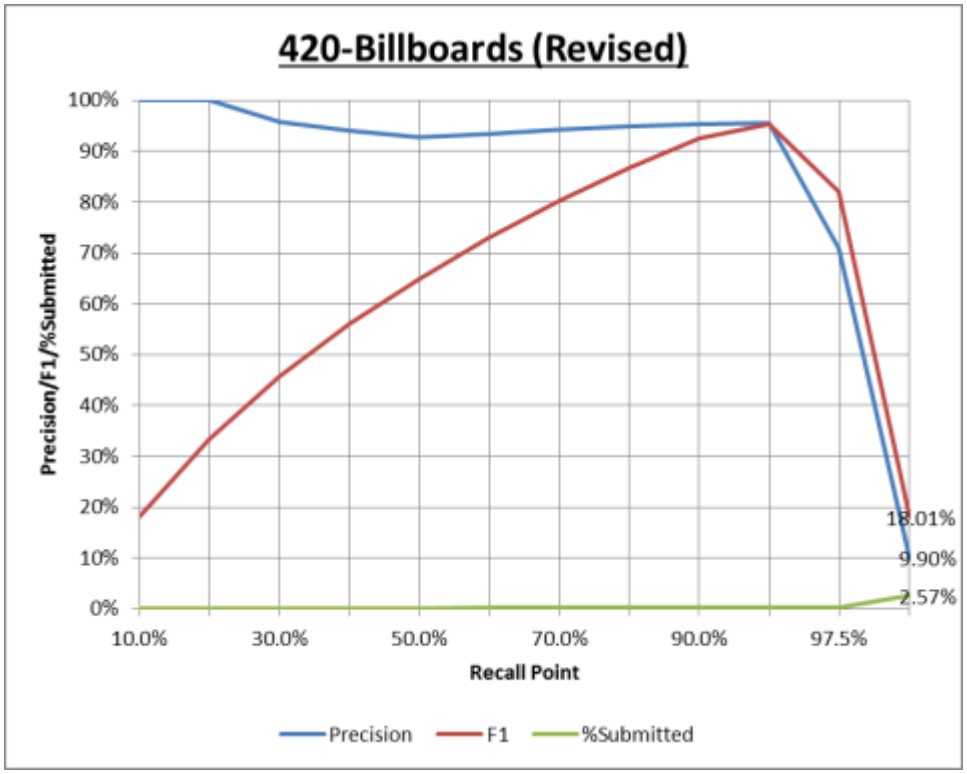
On day two, the final search results were submitted and 80% recall was called after 628 submitted, with 573 returned relevant. He trained 1,000 randomly selected documents as Not Relevant and initiated a learning session.

On the final day, he submitted the highest scoring documents, and quickly called Reasonable after 740 submitted. 682 were returned relevant. He submitted all remaining documents with the highest scores being submitted first. A total of 737 documents were returned relevant by TREC. In total, 4.0 hours were spent reviewing this topic.

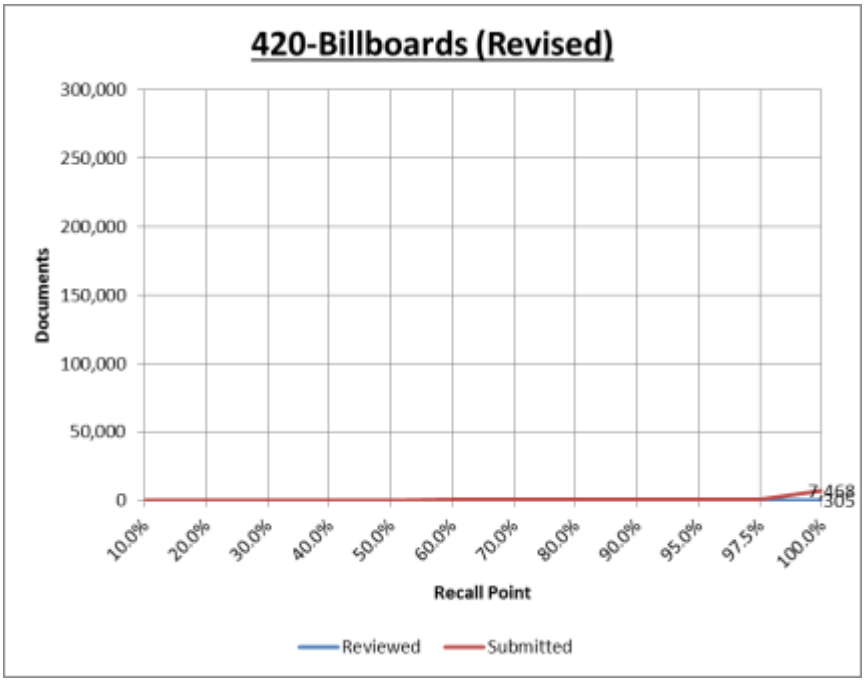
This topic had an above average TREC standard. Though he identified 48 documents that were clearly erroneous, overall the standard was clear and the inconsistencies weren't widespread.

Graphs

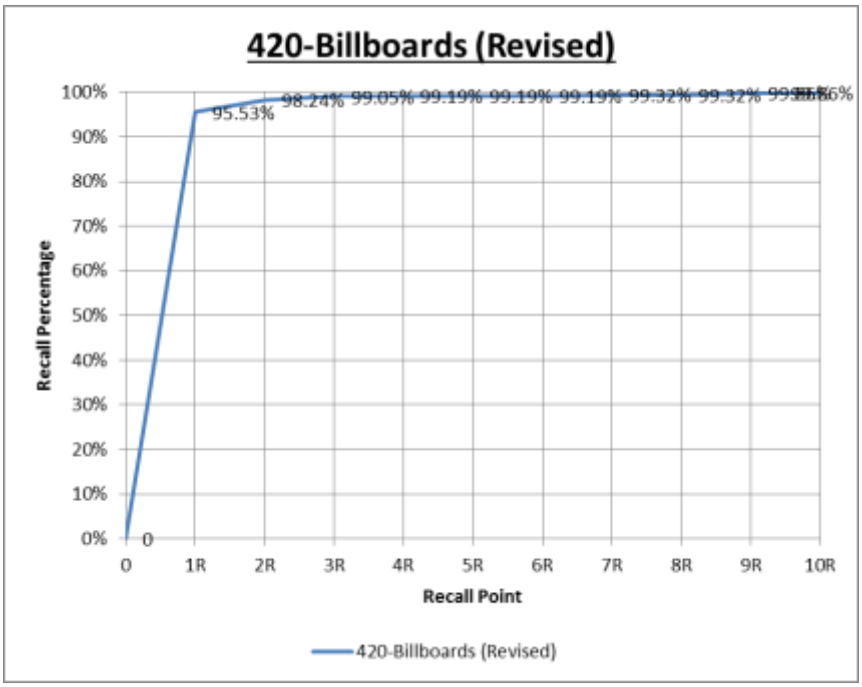
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Billboards topic, the 90% recall threshold had been attained by submitting only 0.24%% of the corpus, 699 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 421 - Traffic Cameras
Total Documents: 290,099
Total Relevant: 54
Total Prevalence: 0.02%

Confusion Matrix - Traffic Cameras

	@Reasonable	@90% Recall	@95% Recall
True Positives	*52*	*49*	*52*
True Negatives	*289,945*	*290,045*	*290,045*
False Positives	*100*	*0*	*0*
False Negatives	*2*	*5*	*2*
Recall	96.30%	90.74%	96.30%
Precision	34.21%	100.00%	100.00%
F1 Measure	50.49%	95.15%	98.11%
Accuracy	99.9648%	99.9983%	99.9993%
Error	0.0352%	0.0017%	0.0007%
Elusion	0.00%	0.00%	0.00%
Fallout	0.03%	0.00%	0.00%

Topic 421 - Traffic Cameras - UNCORRECTED
Total Documents: 290,099
Total Relevant: 21
Total Prevalence: 0.01%

Confusion Matrix - Traffic Cameras

	@Reasonable	@90% Recall	@95% Recall
True Positives	*19*	*19*	*20*
True Negatives	*289,945*	*290,047*	*281,036*
False Positives	*133*	*31*	*9,042*
False Negatives	*2*	*2*	*1*
Recall	90.48%	90.48%	95.24%
Precision	12.50%	38.00%	0.22%
F1 Measure	21.97%	53.52%	0.44%
Accuracy	99.95%	99.99%	96.88%
Error	0.05%	0.01%	3.12%
Elusion	0.00%	0.00%	0.00%
Fallout	0.05%	0.01%	3.12%

Summary

Topic 421 was run by Jim Sullivan, who started on August 20, 2016 and concluded on the same day.

Sullivan entered this topic with basic knowledge of traffic cameras and a solid understanding of related keywords. This knowledge was acquired by completing the traffic cameras topic in TREC 2015. This experience proved very helpful.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms in the subject line, and moved broader variations anywhere in the document. He quickly realized the low prevalence rate of this topic and called 70% recall after submitting 43 documents, with 17 relevant. He disagreed with TREC the classification on the remaining 26.

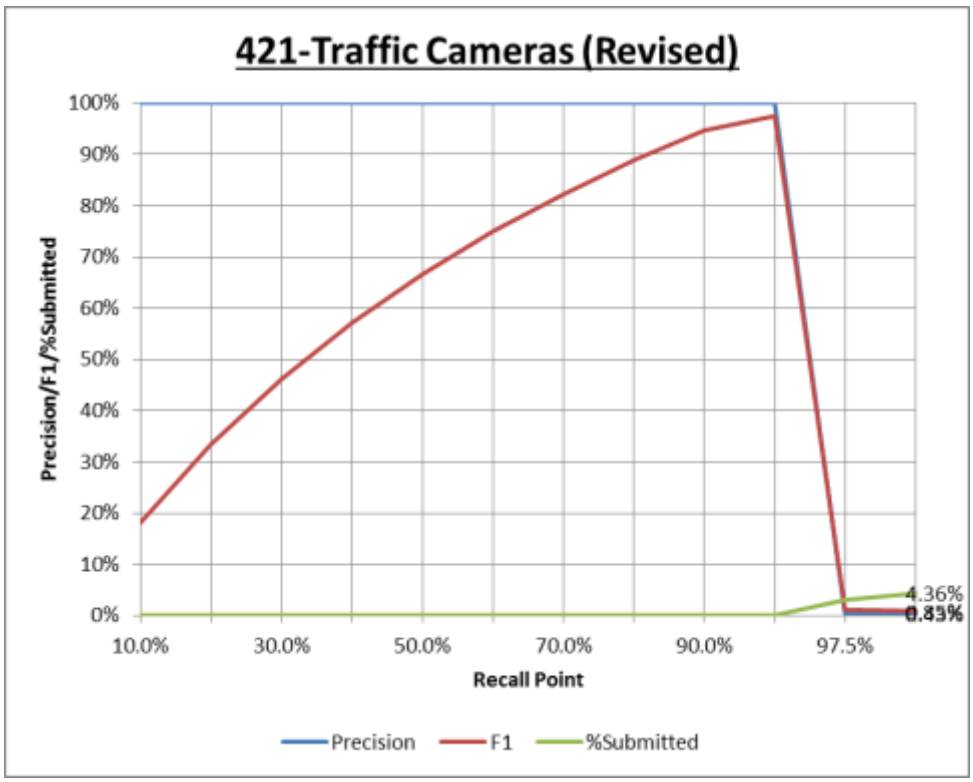
Sullivan continued with variations of keyword terms and high predictive coding scores to find a couple more Relevant documents until he called Reasonable after 152 documents submitted, with 19 being returned Relevant.

He submitted all remaining documents with the highest scores being submitted first. 2 more relevant documents were returned, in which he did not disagree. A total of 21 documents were returned relevant by TREC. In total, 2.0 hours were spent reviewing this very easy topic. The use of predictive coding on this topic was unnecessary.

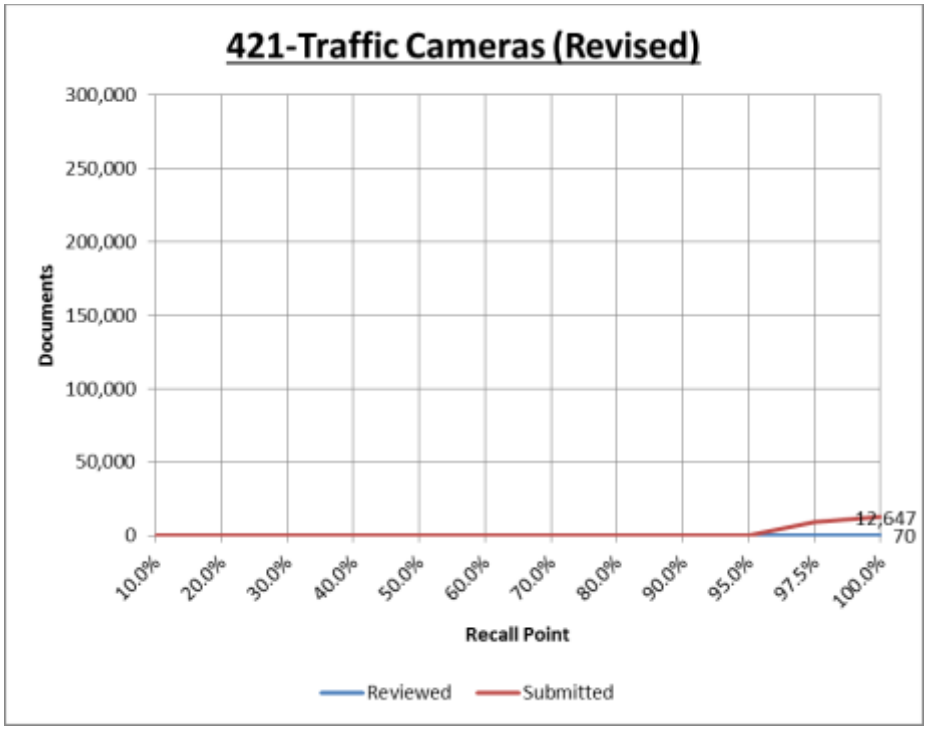
This topic had an average TREC standard. Though he identified 33 documents that were clearly erroneous, overall the standard was clear and the inconsistencies weren't widespread. Almost all errors were in situations where TREC had improperly classified a document as Not Relevant.

Graphs

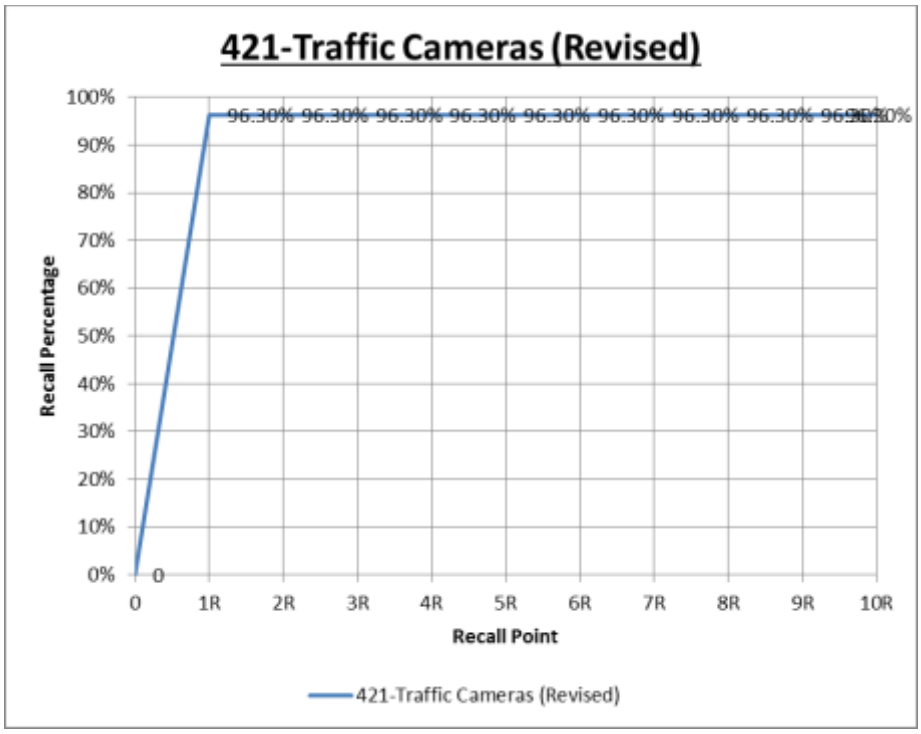
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Traffic Cameras topic, the 90% recall threshold had been attained by submitting only 0.02%% of the corpus, 49 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 422 - Non Resident Aliens
Total Documents: 290,099
Total Relevant: 48
Total Prevalence: 0.02%

Confusion Matrix - Non Resident Aliens

	@Reasonable	@90% Recall	@95% Recall
True Positives	*48*	*44*	*46*
True Negatives	*286,883*	*289,852*	*289,828*
False Positives	*3,168*	*199*	*223*
False Negatives	*0*	*4*	*2*
Recall	100.00%	91.67%	95.83%
Precision	1.49%	18.11%	17.10%
F1 Measure	2.94%	30.24%	29.02%
Accuracy	98.9080%	99.9300%	99.9224%
Error	1.0920%	0.0700%	0.0776%
Elusion	0.00%	0.00%	0.00%
Fallout	1.09%	0.07%	0.08%

Topic 422 - Non Resident Aliens - UNCORRECTED
Total Documents: 290,099
Total Relevant: 31
Total Prevalence: 0.01%

Confusion Matrix - Non Resident Aliens

	@Reasonable	@90% Recall	@95% Recall
True Positives	*29*	*28*	*30*
True Negatives	*286,881*	*289,814*	*286,003*
False Positives	*3,187*	*254*	*4,065*
False Negatives	*2*	*3*	*1*
Recall	93.55%	90.32%	96.77%
Precision	0.90%	9.93%	0.73%
F1 Measure	1.79%	17.89%	1.45%
Accuracy	98.90%	99.91%	98.60%
Error	1.10%	0.09%	1.40%
Elusion	0.00%	0.00%	0.00%
Fallout	1.10%	0.09%	1.40%

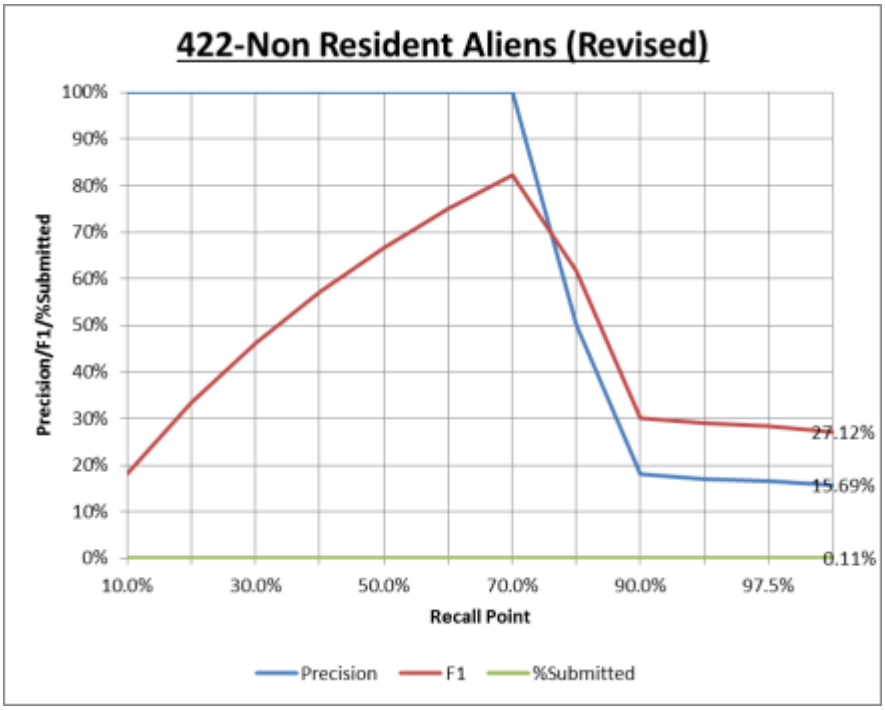
Summary

This project was run by Tony Reichenberger. Documents were submitted on this topic sparingly, based only on keywords initially. Feedback from TREC on the most documents relating to the topic came back as not relevant. Very few documents were being suggested by the machine learning as relevant, and those that were submitted were returned as not relevant. On the 10th submission, all remaining documents hitting on search terms were submitted (accidentally; it was only meant to be a subset of the remaining, but it was not realized until after the feedback from TREC that the whole set was submitted) and only 7 returned as relevant. With such low precision, reasonable was called.

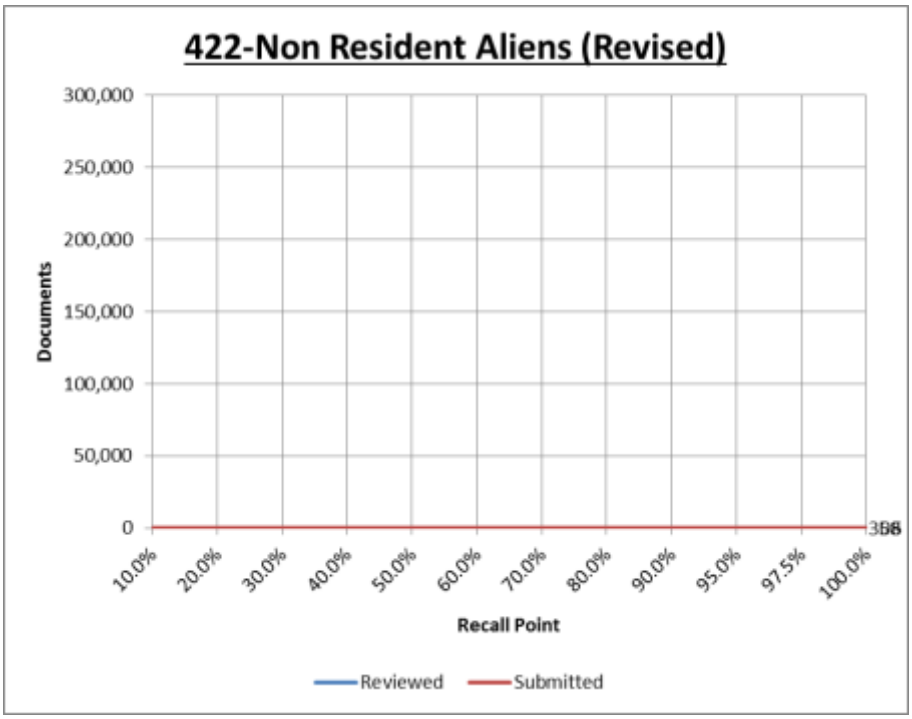
The TREC judgments here were poor, missing many obviously relevant documents. The accessors did not seem to understand the topic, despite the fact that the definition of relevance here was fairly clear: **Non-Resident Aliens (NRA) - All documents involving discussions of the non-resident alien issue. Documents concerning the National Rifle Association are not relevant.**

Graphs

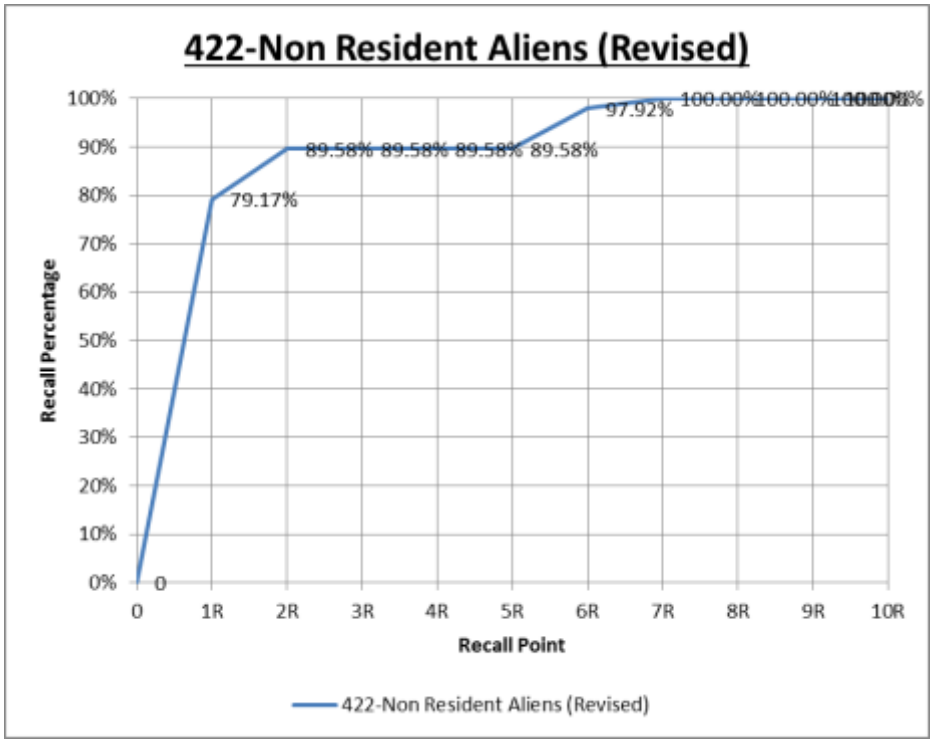
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Non Resident Aliens topic, the 90% recall threshold had been attained by submitting only 0.08%% of the corpus, 243 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


422-Non Resident Aliens (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
79.17%
89.58%
89.58%
89.58%
89.58%
97.92%
100.00%
100.00%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
422-Non Resident Aliens (Revised)

Topic 423 - National Rifle Association
Total Documents: 290,099
Total Relevant: 190
Total Prevalence: 0.07%

Confusion Matrix - National Rifle Association

	@Reasonable	@90% Recall	@95% Recall
True Positives	*147*	*171*	*181*
True Negatives	*289,616*	*289,072*	*288,856*
False Positives	*293*	*837*	*1,053*
False Negatives	*43*	*19*	*9*
Recall	77.37%	90.00%	95.26%
Precision	33.41%	16.96%	14.67%
F1 Measure	46.67%	28.55%	25.42%
Accuracy	99.8842%	99.7049%	99.6339%
Error	0.1158%	0.2951%	0.3661%
Elusion	0.01%	0.01%	0.00%
Fallout	0.10%	0.29%	0.36%

Topic 423 - National Rifle Association - UNCORRECTED
Total Documents: 290,099
Total Relevant: 286
Total Prevalence: 0.10%

Confusion Matrix - National Rifle Association

	@Reasonable	@90% Recall	@95% Recall
True Positives	*146*	*258*	*272*
True Negatives	*289,519*	*285,282*	*277,814*
False Positives	*294*	*4,531*	*11,999*
False Negatives	*140*	*28*	*14*
Recall	51.05%	90.21%	95.10%
Precision	33.18%	5.39%	2.22%
F1 Measure	40.22%	10.17%	4.33%
Accuracy	99.85%	98.43%	95.86%
Error	0.15%	1.57%	4.14%
Elusion	0.05%	0.01%	0.01%
Fallout	0.10%	1.56%	4.14%

Summary

This project was run by Tony Reichenberger. It is the "other NRA" topic specifically defined as: **National Rifle Association (NRA) - All documents concerning the National Rifle Association, its members, and its influences. Documents concerning the non-resident alien issue are not relevant.**

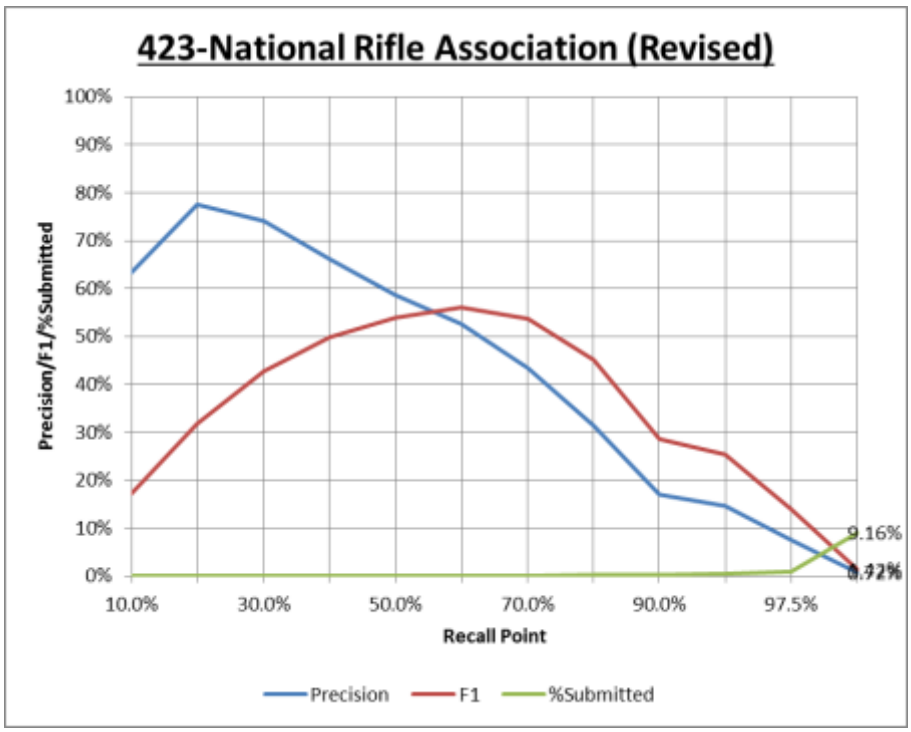
The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-20 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring; when additional relevant materials were found, subsequent searches for similar documents were partaken.

An inconsistent standard resulted in poor and conflicting results. Documents containing the exact same text were often found with contradictory coding, and likewise there were scores of missed relevant documents and documents coded relevant for little or no reason. The result was confusion based on TREC feedback for both the human reviewer and the machine learning.

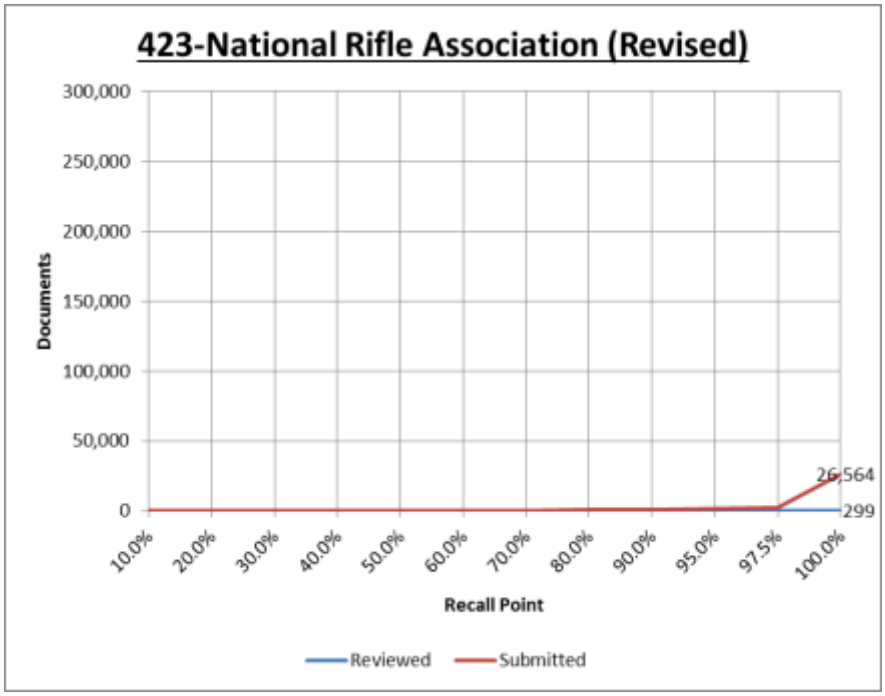
With the conflicting issues, Reasonable was called too early on this topic, as questions of what was irrelevant misled the human assessor. Submissions post-call of similar materials and keyword hits resulted in relevant materials that altered the Reasonable assessment. As additional relevant documents were found, additional searches and learning sessions were conducted as follow ups, with those documents being included in subsequent submissions.

Graphs

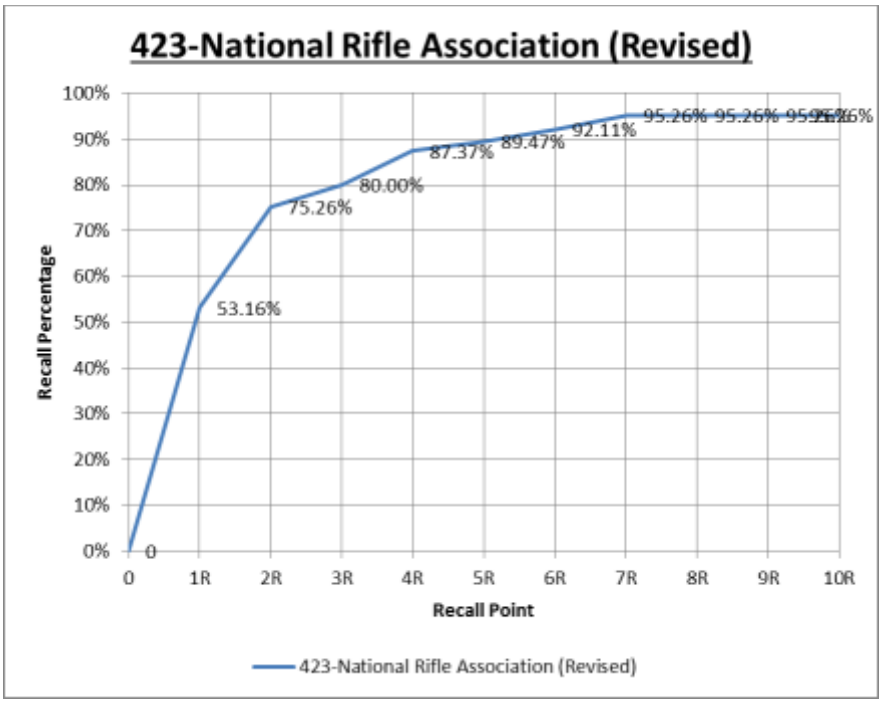
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the National Rifle Association topic, the 90% recall threshold had been attained by submitting only 0.35%% of the corpus, 1,008 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 424 - Gulf Drilling
Total Documents: 290,099
Total Relevant: 495
Total Prevalence: 0.17%

Confusion Matrix - Gulf Drilling

	@Reasonable	@90% Recall	@95% Recall
True Positives	*493*	*446*	*471*
True Negatives	*287,922*	*289,209*	*288,888*
False Positives	*1,682*	*395*	*716*
False Negatives	*2*	*49*	*24*
Recall	99.60%	90.10%	95.15%
Precision	22.67%	53.03%	39.68%
F1 Measure	36.93%	66.77%	56.00%
Accuracy	99.4195%	99.8469%	99.7449%
Error	0.5805%	0.1531%	0.2551%
Elusion	0.00%	0.02%	0.01%
Fallout	0.58%	0.14%	0.25%

Topic 424 - Gulf Drilling - UNCORRECTED
Total Documents: 290,099
Total Relevant: 497
Total Prevalence: 0.17%

Confusion Matrix - Gulf Drilling

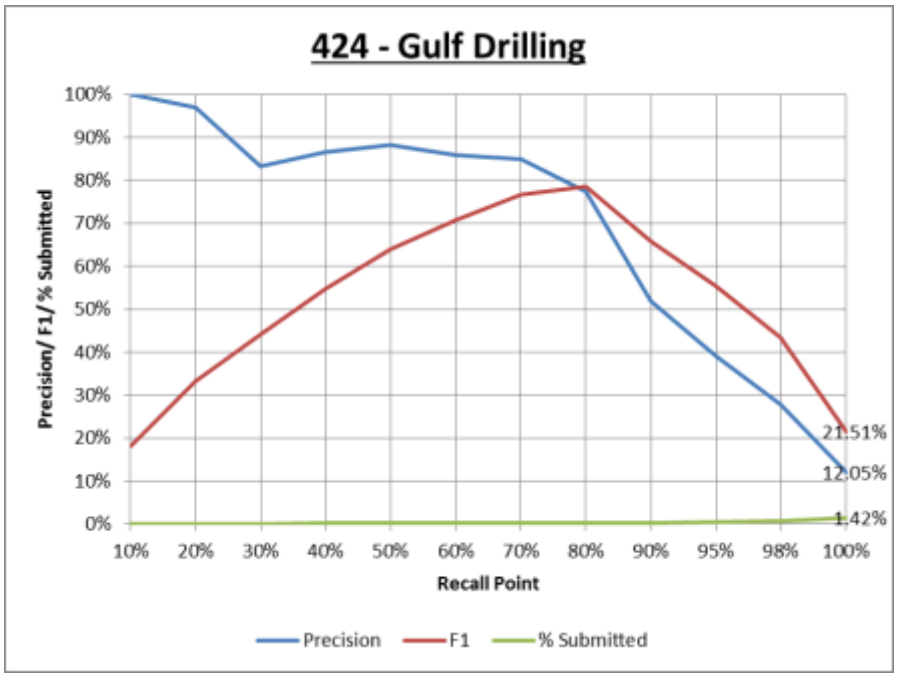
	@Reasonable	@90% Recall	@95% Recall
True Positives	*495*	*448*	*473*
True Negatives	*287,922*	*289,186*	*288,869*
False Positives	*1,680*	*416*	*733*
False Negatives	*2*	*49*	*24*
Recall	99.60%	90.14%	95.17%
Precision	22.76%	51.85%	39.22%
F1 Measure	37.05%	65.83%	55.55%
Accuracy	99.42%	99.84%	99.74%
Error	0.58%	0.16%	0.26%
Elusion	0.00%	0.02%	0.01%
Fallout	0.58%	0.14%	0.25%

Summary

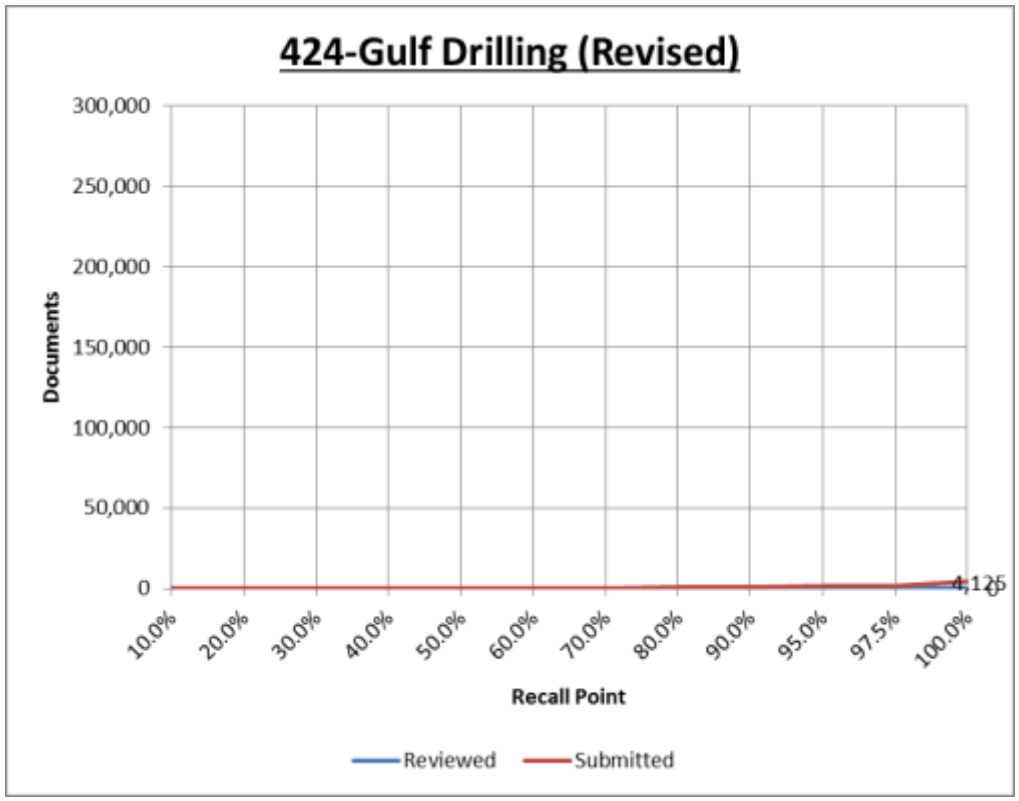
This topic was run by Levi Kuehn. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. When scores dropped to 5%, a final search was submitted, another learning session run, and documents were submitted in probability order.

Graphs

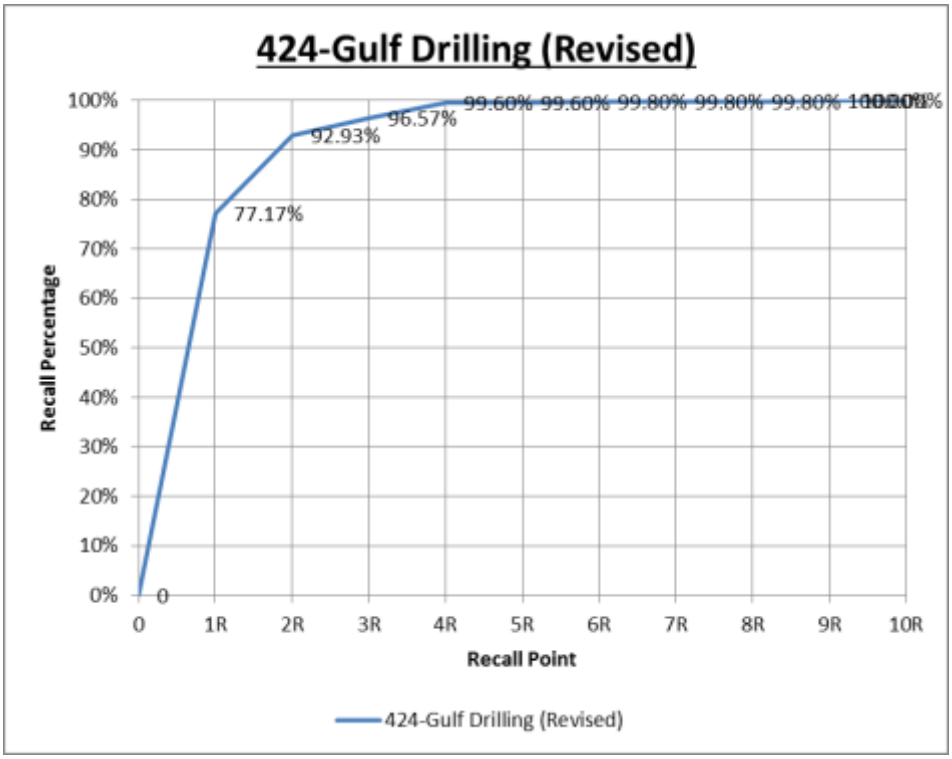
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Gulf Drilling topic, the 90% recall threshold had been attained by submitting only 0.29%% of the corpus, 841 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


424-Gulf Drilling (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
77.17%
92.93%
96.57%
99.60% 99.60% 99.80% 99.80% 99.80%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
424-Gulf Drilling (Revised)

Topic 425 - Civil Rights Act of 2003
Total Documents: 290,099
Total Relevant: 718
Total Prevalence: 0.25%

Confusion Matrix - Civil Rights Act of 2003

	@Reasonable	@90% Recall	@95% Recall
True Positives	*653*	*623*	*658*
True Negatives	*289,355*	*289,371*	*286,331*
False Positives	*26*	*10*	*3,050*
False Negatives	*65*	*95*	*60*
Recall	90.95%	86.77%	91.64%
Precision	96.17%	98.42%	17.75%
F1 Measure	93.49%	92.23%	29.73%
Accuracy	99.9686%	99.9638%	98.9280%
Error	0.0314%	0.0362%	1.0720%
Elusion	0.02%	0.03%	0.02%
Fallout	0.01%	0.00%	1.05%

Topic 425 - Civil Rights Act of 2003 - UNCORRECTED
Total Documents: 290,099
Total Relevant: 714
Total Prevalence: 0.25%

Confusion Matrix - Civil Rights Act of 2003

	@Reasonable	@90% Recall	@95% Recall
True Positives	*652*	*643*	*679*
True Negatives	*289,362*	*289,365*	*286,345*
False Positives	*23*	*20*	*3,040*
False Negatives	*62*	*71*	*35*
Recall	91.32%	90.06%	95.10%
Precision	96.59%	96.98%	18.26%
F1 Measure	93.88%	93.39%	30.63%
Accuracy	99.97%	99.97%	98.94%
Error	0.03%	0.03%	1.06%
Elusion	0.02%	0.02%	0.01%
Fallout	0.01%	0.01%	1.05%

Summary

This topic was run by Losey who put a substantial eight-hour effort into this search from June 15th to 22nd 2016. He reviewed 291 documents, created 35 different search folder and manually categorized 739 documents.

The topic was further defined as: **Civil Rights Act of 2003 - All documents involving discussions of the Florida Civil Rights Act of 2003.**

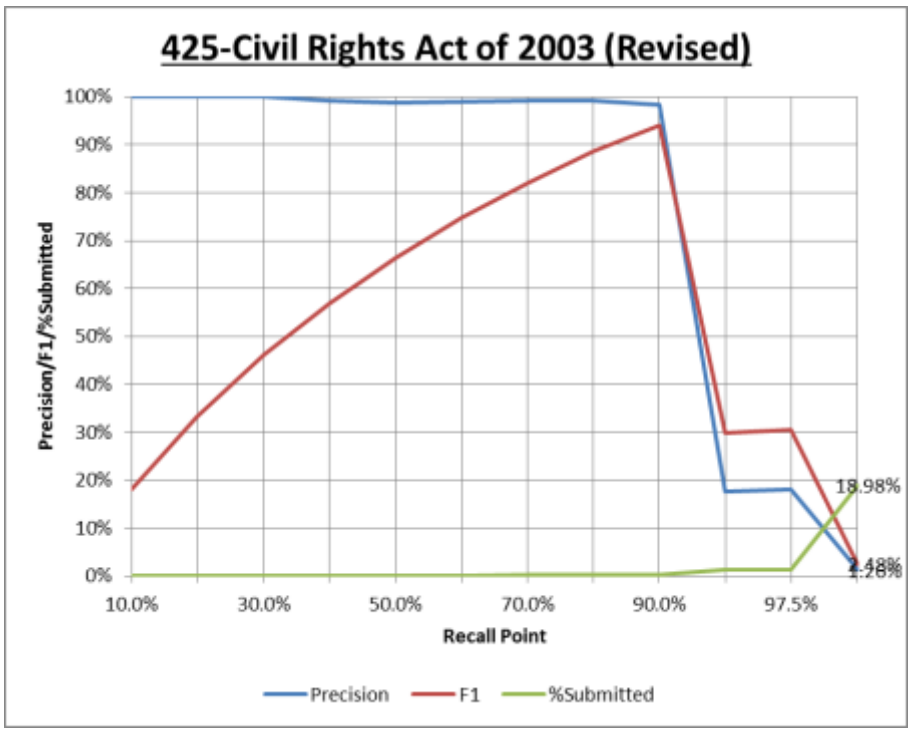
Losey began with a Google search to obtain detailed facts for the search beyond the obvious. He learned, among other things, that the legislation was called the "Dr. Marvin Davies Florida Civil Rights Act" and was signed into law by Governor Bush on June 18, 2003. Marvin Davies was a Florida civil rights leader who died cancer April 25, 2003.He also read the final law, and noted from its legislative history the various numbers associated with bill during the legislative process. The law supplemented to the original Florida Civil Rights Act of 1992. There was not much civil rights legislation during the Bush years so the relevant emails stuck out easily.

This was, fortunately, a topic with a well-judged TREC standard, one that required some legal acumen to do properly.

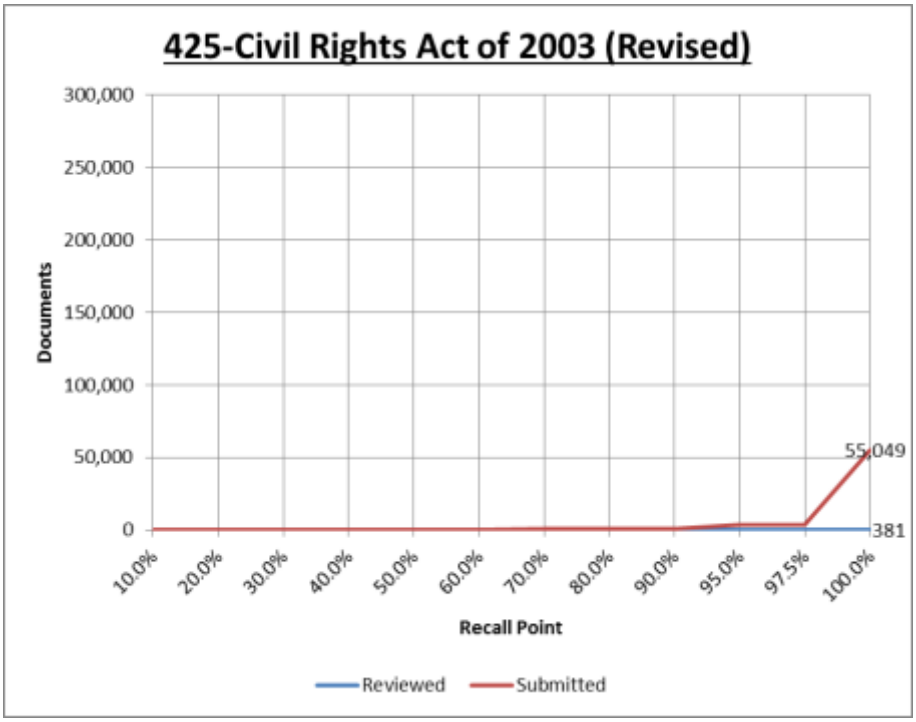
Losey would have cored even higher on this topic but for the fact he accidentally did not submit a set of documents he had identified as probable relevant until after the reasonable call. This is no doubt derived from rushing ad not using our usual quality controls. Such a mistake would not be possible under normal legal search conditions, or if the mistake was made, could be easily cured by a supplemental production.

Graphs

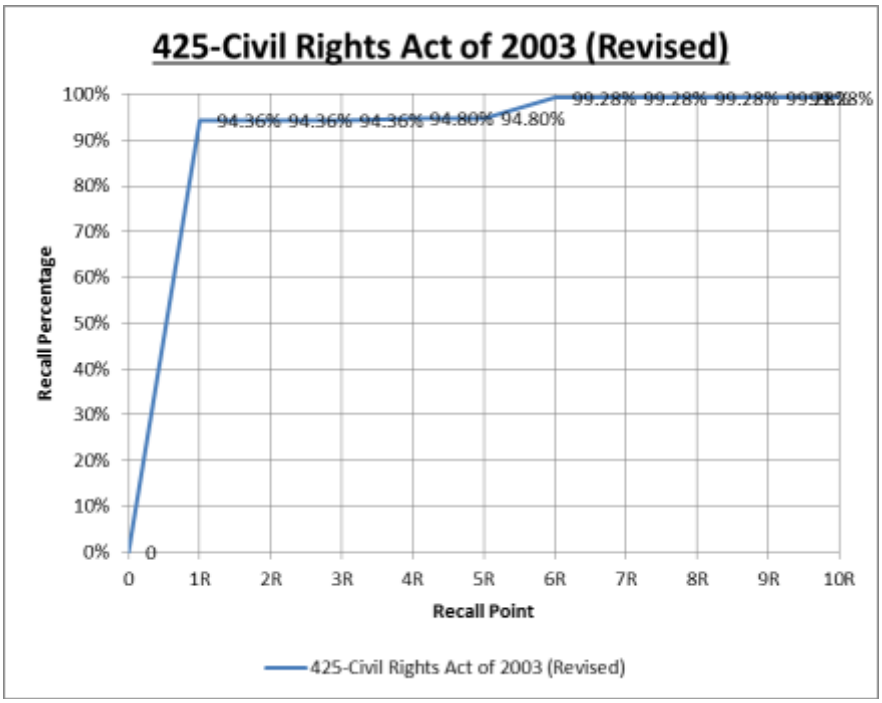
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Civil Rights Act of 2003 topic, the 90% recall threshold had been attained by submitting only 0.22%% of the corpus, 633 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 426 - Jeffrey Goldhagen
Total Documents: 290,099
Total Relevant: 98
Total Prevalence: 0.03%

Confusion Matrix - Jeffrey Goldhagen

	@Reasonable	@90% Recall	@95% Recall
True Positives	*91*	*89*	*94*
True Negatives	*289,996*	*289,996*	*288,587*
False Positives	*5*	*5*	*1,414*
False Negatives	*7*	*9*	*4*
Recall	92.86%	90.82%	95.92%
Precision	94.79%	94.68%	6.23%
F1 Measure	93.81%	92.71%	11.71%
Accuracy	99.9959%	99.9952%	99.5112%
Error	0.0041%	0.0048%	0.4888%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	0.49%

Topic 426 - Jeffrey Goldhagen - UNCORRECTED
Total Documents: 290,099
Total Relevant: 120
Total Prevalence: 0.04%

Confusion Matrix - Jeffrey Goldhagen

	@Reasonable	@90% Recall	@95% Recall
True Positives	*84*	*108*	*114*
True Negatives	*289,967*	*289,613*	*287,627*
False Positives	*12*	*366*	*2,352*
False Negatives	*36*	*12*	*6*
Recall	70.00%	90.00%	95.00%
Precision	87.50%	22.78%	4.62%
F1 Measure	77.78%	36.36%	8.82%
Accuracy	99.98%	99.87%	99.19%
Error	0.02%	0.13%	0.81%
Elusion	0.01%	0.00%	0.00%
Fallout	0.00%	0.13%	0.81%

Summary

This project was run by Losey from August 8th to 11th 2016. He spent five hours, made 22 submissions, called reasonable after 11, and created 18 search folders. He reviewed a total of 112 documents and manually categorized 141.

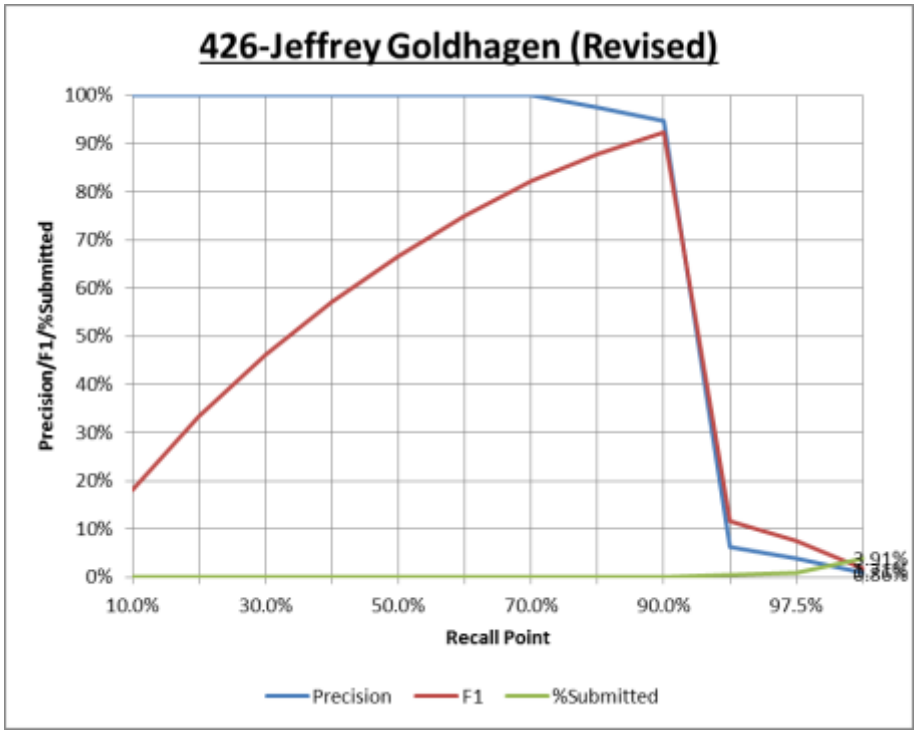
The full description of this topic is: **Jeffrey Goldhagen - All documents related to Jeffrey Goldhagen's role in the Bush administration, his firing, and reinstatement.'** Losey had never heard of this man but a Google search quickly provided the background. He was a doctor and medical director for Jacksonville that was fired by Bush, and then rehired.

This topic had a number of obvious errors in TREC judging, including such things as a tendency to call relevant any email about a physician in trouble, even if it was not Dr. Goldhagen. Also, the TREC classifier often seemed incapable of knowing when an email by Dr. Goldhagen's enemy, Holly Kartsonis, to Bush pertained to issues other than Dr. Goldhagen. She often wrote to Jeb on a number of topics, usually personal and flattering. She also asked for Jeb's help to get another job with the State. Kartsonis' husband was a doctor and Bush seemed to like to chat with her (part of his online nice guy persona, which is pretty much forced, but not entirely bogus). She appeared to think that creating an online relationship with the governor would help her, and it did to a point. In fact, it was amazing to see how the online relationship developed with Jeb. They had many emails over the years. There was no indication in this cleaned collection they ever met. Still, in the end, Jeb never intervened in the final decision by the State not to employ her. These emails have nothing to do with this topic, which is Dr. Goldhagen, not talkative Holly, although Losey found it interesting to read the many emails between them.

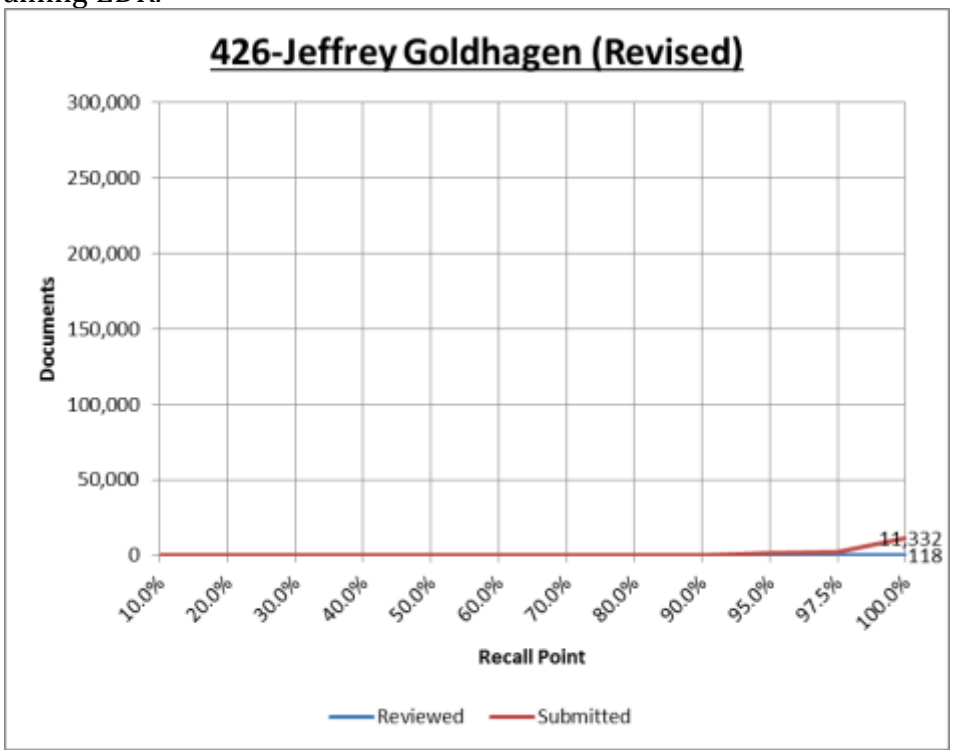
This was a topic that was once again driven primarily by keyword search. Losey used Mr. EDR primarily for QC. He also used use both Data Index and Concept based searches to look for misspelling and other words, and did find one useful variation, namely that Goldhagen was once referred to as "Dr. G." It turns out that there were two Dr. G's, and a few other false hits, but this abbreviation did allow location of two Relevant emails that otherwise would not have been found. They were found by concept search and manual review. This once again shows the power of using all search features - multimodal - and not just predictive coding, or keyword.

Graphs

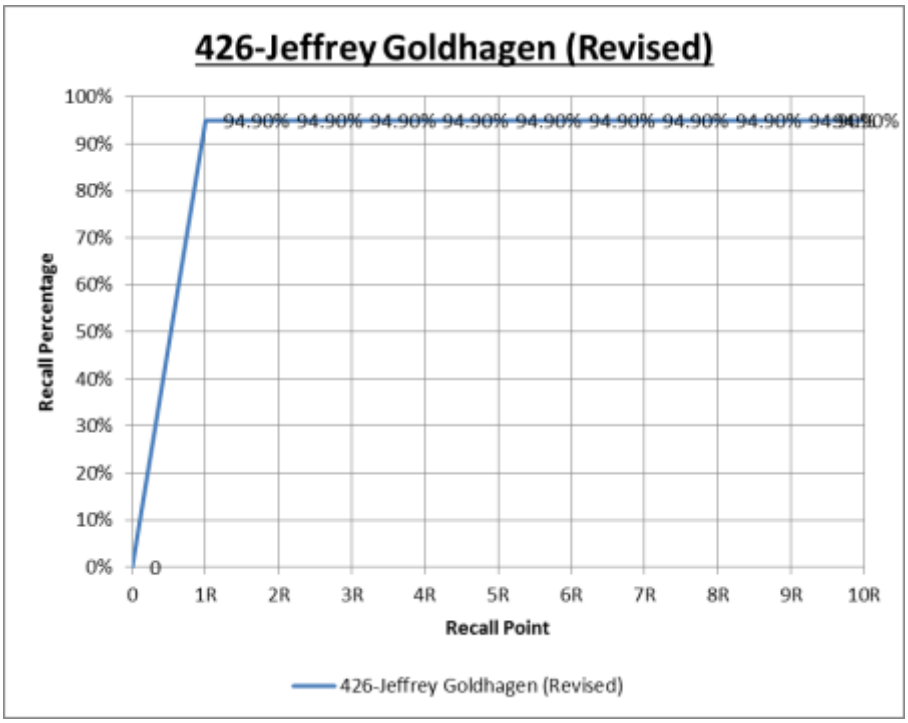
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Jeffrey Goldhagen topic, the 90% recall threshold had been attained by submitting only 0.03%% of the corpus, 94 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 427 - Slot Machines
Total Documents: 290,099
Total Relevant: 263
Total Prevalence: 0.09%

Confusion Matrix - Slot Machines

	@Reasonable	@90% Recall	@95% Recall
True Positives	*249*	*237*	*250*
True Negatives	*289,484*	*289,727*	*289,351*
False Positives	*352*	*109*	*485*
False Negatives	*14*	*26*	*13*
Recall	94.68%	90.11%	95.06%
Precision	41.43%	68.50%	34.01%
F1 Measure	57.64%	77.83%	50.10%
Accuracy	99.8738%	99.9535%	99.8283%
Error	0.1262%	0.0465%	0.1717%
Elusion	0.00%	0.01%	0.00%
Fallout	0.12%	0.04%	0.17%

Topic 427 - Slot Machines - UNCORRECTED
Total Documents: 290,099
Total Relevant: 241
Total Prevalence: 0.08%

Confusion Matrix - Slot Machines

	@Reasonable	@90% Recall	@95% Recall
True Positives	*215*	*217*	*229*
True Negatives	*289,472*	*289,178*	*275,153*
False Positives	*386*	*680*	*14,705*
False Negatives	*26*	*24*	*12*
Recall	89.21%	90.04%	95.02%
Precision	35.77%	24.19%	1.53%
F1 Measure	51.07%	38.14%	3.02%
Accuracy	99.86%	99.76%	94.93%
Error	0.14%	0.24%	5.07%
Elusion	0.01%	0.01%	0.00%
Fallout	0.13%	0.23%	5.07%

Summary

Topic 427 was run by Jim Sullivan, who started on July 21, 2016 and concluded on August 12, with four short days of review in that time period.

Sullivan has a long history with slot machines, both on the winning side and losing side. While he is no bona fide subject matter expert on the topic, he knows his way around the one-armed bandit.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms in the subject line, and moved to more generic terms in broader fields. At the end of the first day he had submitted 204 documents, with 165 relevant. To end the day, he kicked off a learning session after training 500 randomly selected documents as Not Responsive.

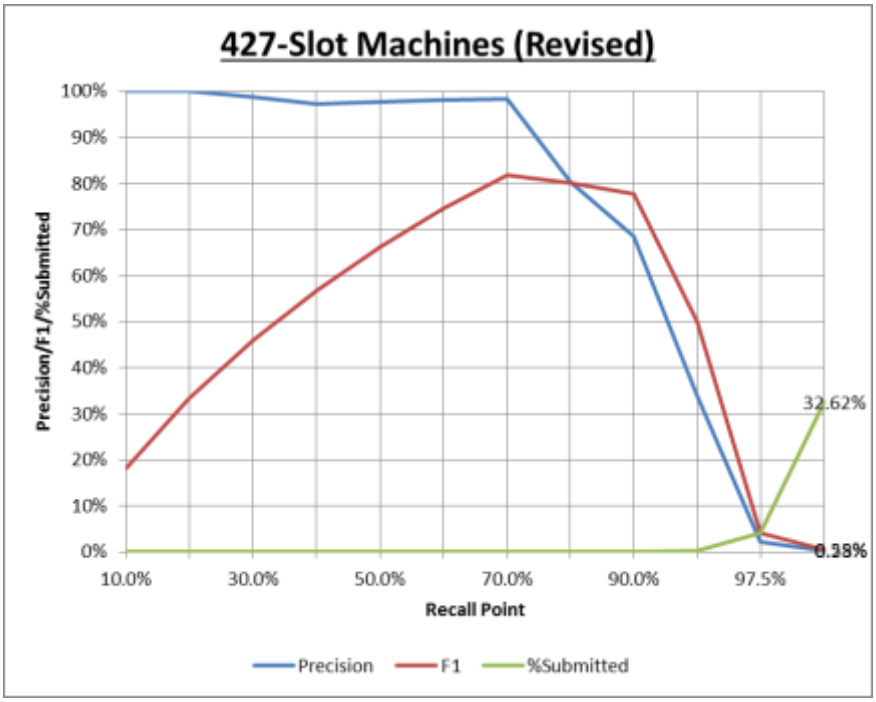
Day two was quick and consisted of submitting the last few docs that hit on "slot machin*" in the document or "slots*" in the subject line. Called 70% recall after 258 docs submitted with 172 relevant and called it a day. Day three was just as short, where the last docs that hit on "slots*" anywhere in the text were submitted.

80% recall was called early on day four, and escalated reliance was placed on the predictive coding scores. Once the predictive coding scores stopped yielding valuable results, Reasonable recall was called. After the reasonable call, all remaining documents were submitted by predictive coding score with the highest scores being submitted first. A total of 241 documents were returned relevant by TREC. In total, 4.25 hours were spent reviewing this topic.

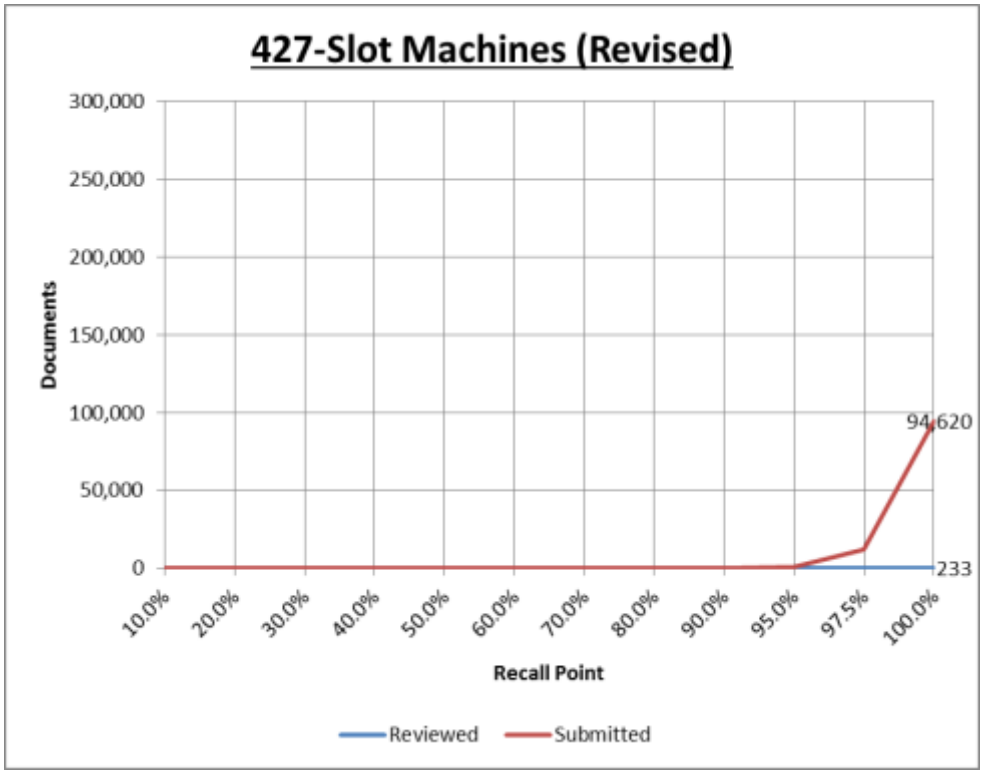
This topic was graded fairly and had a below average number of inconsistencies. There were only 46 documents where TREC had returned inconsistent or incorrect classifications.

Graphs

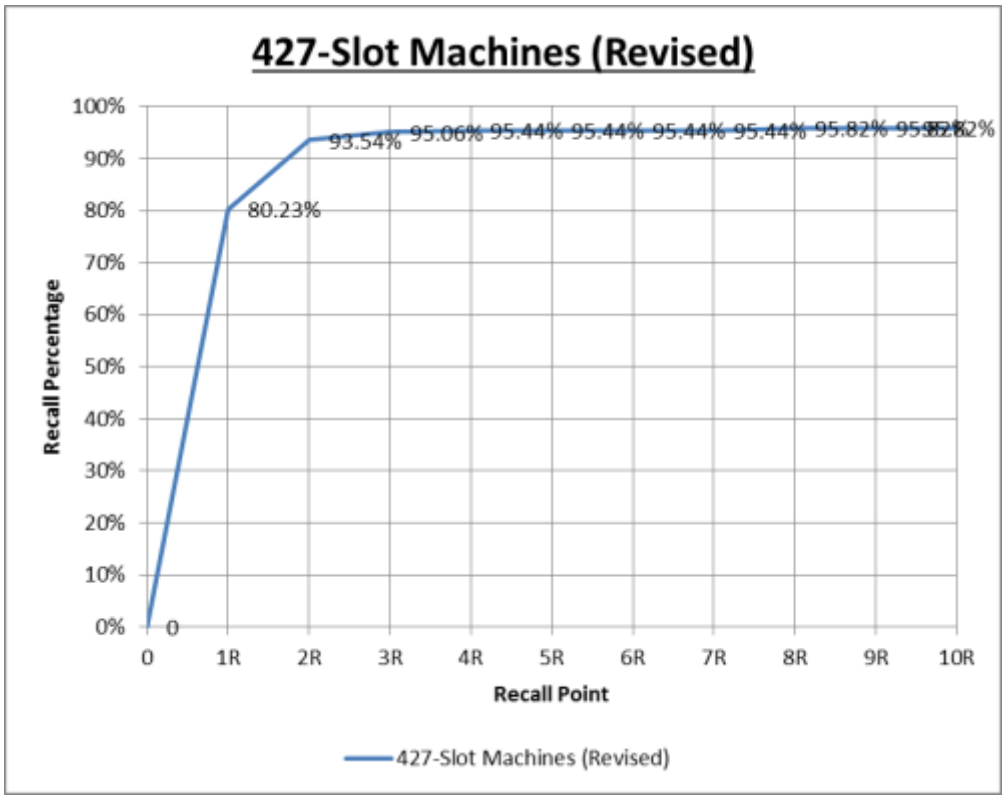
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Slot Machines topic, the 90% recall threshold had been attained by submitting only 0.12%% of the corpus, 346 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 428 - New Stadiums and Arenas
Total Documents: 290,099
Total Relevant: 476
Total Prevalence: 0.16%

Confusion Matrix - New Stadiums and Arenas

	@Reasonable	@90% Recall	@95% Recall
True Positives	*447*	*429*	*453*
True Negatives	*287,645*	*288,628*	*280,685*
False Positives	*1,978*	*995*	*8,938*
False Negatives	*29*	*47*	*23*
Recall	93.91%	90.13%	95.17%
Precision	18.43%	30.13%	4.82%
F1 Measure	30.82%	45.16%	9.18%
Accuracy	99.3082%	99.6408%	96.9111%
Error	0.6918%	0.3592%	3.0889%
Elusion	0.01%	0.02%	0.01%
Fallout	0.68%	0.34%	3.09%

Topic 428 - New Stadiums and Arenas - UNCORRECTED
Total Documents: 290,099
Total Relevant: 464
Total Prevalence: 0.16%

Confusion Matrix - New Stadiums and Arenas

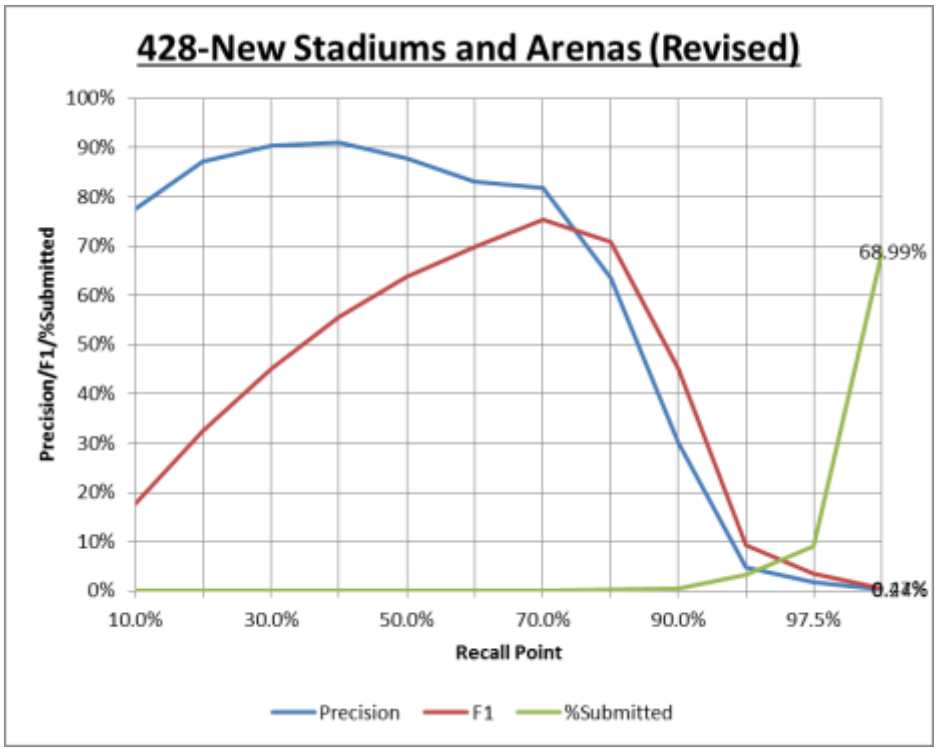
	@Reasonable	@90% Recall	@95% Recall
True Positives	*432*	*418*	*441*
True Negatives	*287,642*	*288,549*	*280,554*
False Positives	*1,993*	*1,086*	*9,081*
False Negatives	*32*	*46*	*23*
Recall	93.10%	90.09%	95.04%
Precision	17.81%	27.79%	4.63%
F1 Measure	29.91%	42.48%	8.83%
Accuracy	99.30%	99.61%	96.86%
Error	0.70%	0.39%	3.14%
Elusion	0.01%	0.02%	0.01%
Fallout	0.69%	0.37%	3.14%

Summary

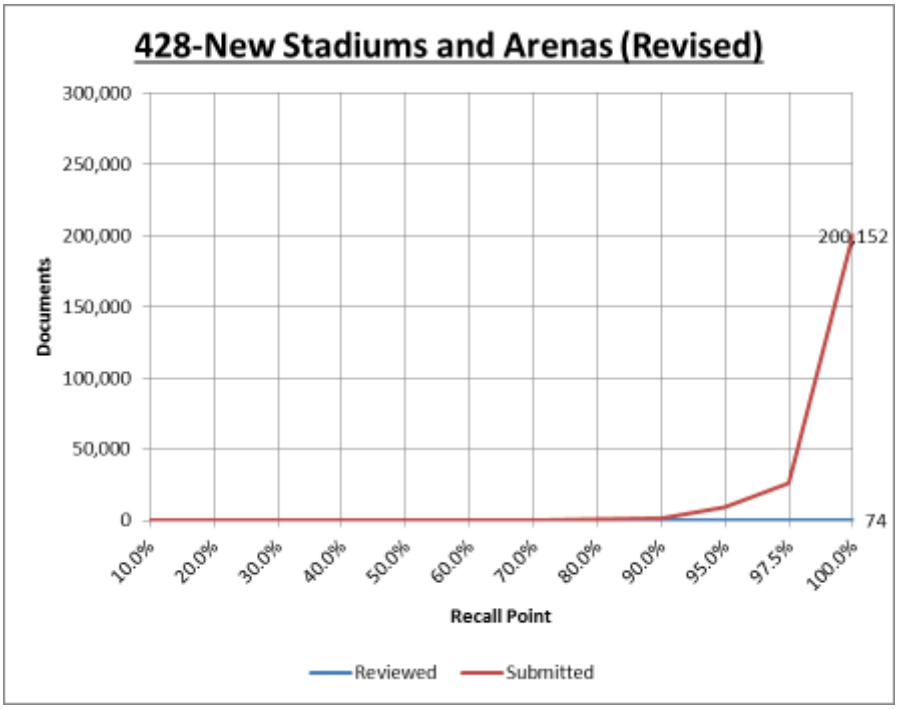
This topic was run by Levi Kuehn. The hybrid multimodal review was conducted by initially submitting keyword hits to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-50 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring, with the size of the submissions increasing (up to 100 documents); when additional relevant materials were found, subsequent searches for similar documents were partaken. When scores dropped to 5%, a final search was submitted, another learning session run, and documents were submitted in probability order.

Graphs

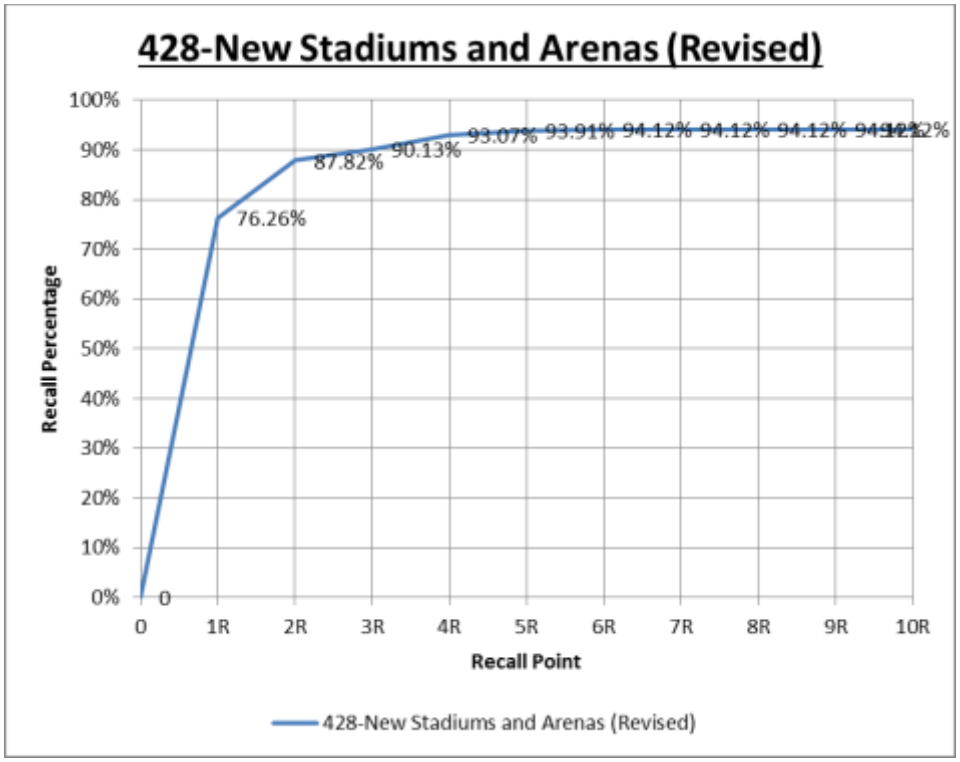
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the New Stadiums and Arenas topic, the 90% recall threshold had been attained by submitting only 0.49%% of the corpus, 1,424 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 429 - Elian Gonzalez
Total Documents: 290,099
Total Relevant: 844
Total Prevalence: 0.29%

Confusion Matrix - Elian Gonzalez

	@Reasonable	@90% Recall	@95% Recall
True Positives	*819*	*760*	*802*
True Negatives	*289,231*	*289,240*	*289,231*
False Positives	*24*	*15*	*24*
False Negatives	*25*	*84*	*42*
Recall	97.04%	90.05%	95.02%
Precision	97.15%	98.06%	97.09%
F1 Measure	97.10%	93.89%	96.05%
Accuracy	99.9831%	99.9659%	99.9772%
Error	0.0169%	0.0341%	0.0228%
Elusion	0.01%	0.03%	0.01%
Fallout	0.01%	0.01%	0.01%

Topic 429 - Elian Gonzalez - UNCORRECTED
Total Documents: 290,099
Total Relevant: 827
Total Prevalence: 0.29%

Confusion Matrix - Elian Gonzalez

	@Reasonable	@90% Recall	@95% Recall
True Positives	*779*	*745*	*786*
True Negatives	*289,208*	*289,226*	*289,006*
False Positives	*64*	*46*	*266*
False Negatives	*48*	*82*	*41*
Recall	94.20%	90.08%	95.04%
Precision	92.41%	94.18%	74.71%
F1 Measure	93.29%	92.09%	83.66%
Accuracy	99.96%	99.96%	99.89%
Error	0.04%	0.04%	0.11%
Elusion	0.02%	0.03%	0.01%
Fallout	0.02%	0.02%	0.09%

Summary

Topic 429 was run by Jim Sullivan, who started on June 3, 2016 and concluded on June 9. Being his first attempted topic on the year, he spent more time understanding the dataset than was necessary on later topics.

While Sullivan had heard the name Elian Gonzalez in the past, he had not read any of the news about him prior to this exercise.

Sullivan started by testing terms and creating a keyword highlight list, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms, and moved to more generic lists. At the end of the first day he had submitted 409 documents, with 404 relevant. At this point, he predicted 700 total relevant documents and kicked off a learning session after training 500 randomly selected documents as Not Responsive.

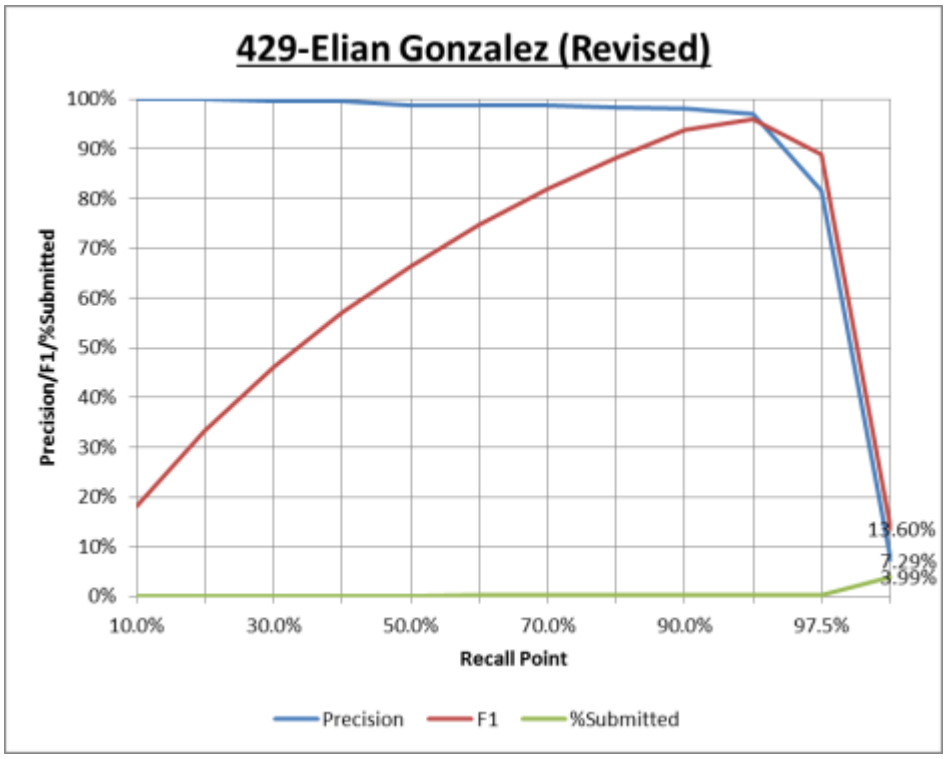
The second day of review was spent combining predictive coding scores with date searches. This was one of the few topics that had a very relevant time period. High scoring documents within the date range were submitted. He called 80% recall after 731 total documents submitted, with 699 relevant.

Day three was spent digging through any remaining search terms and high scoring documents. Exhausting all options, he called reasonable after finding 779 relevant documents. After the reasonable call, all remaining documents were submitted by predictive coding score with the highest scores being submitted first. A total of 827 documents were returned relevant by TREC. In total, 6.25 hours were spent reviewing this topic.

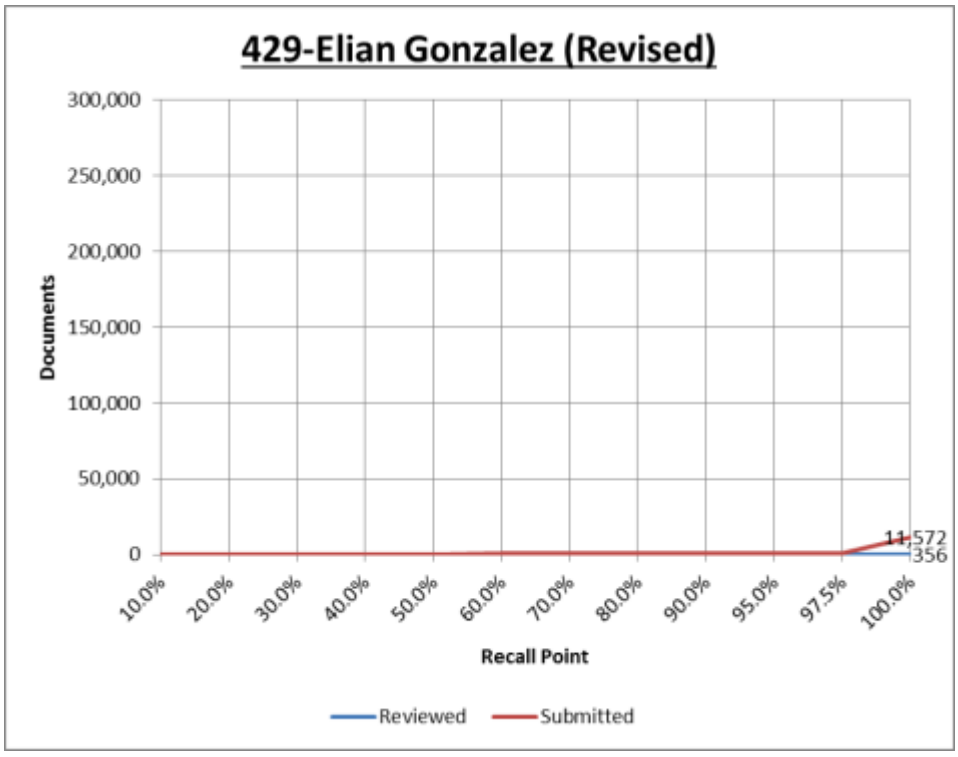
This topic was graded fairly and had a below average number of inconsistencies. There were only 63 documents where TREC had returned inconsistent or incorrect classifications. He was especially impressed by TREC's ability to identify misspellings of Elian and documents within the date range that referenced the event without any meaningful keywords.

Graphs

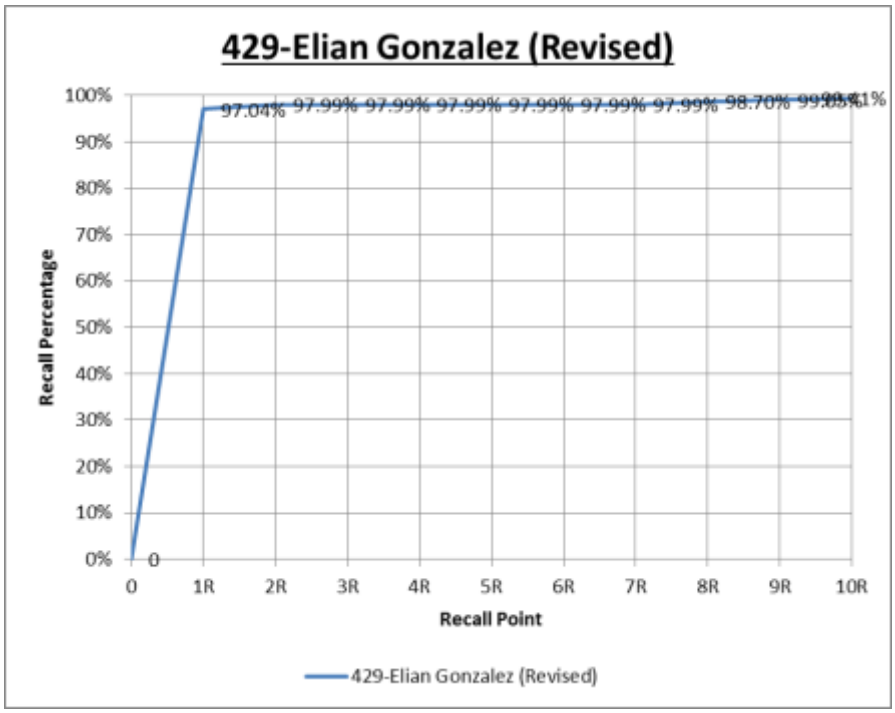
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Elian Gonzalez topic, the 90% recall threshold had been attained by submitting only 0.27%% of the corpus, 775 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 430 - Restraints and Helmets
Total Documents: 290,099
Total Relevant: 1,013
Total Prevalence: 0.35%

Confusion Matrix - Restraints and Helmets

	@Reasonable	@90% Recall	@95% Recall
True Positives	*735*	*912*	*963*
True Negatives	*288,724*	*281,357*	*279,080*
False Positives	*362*	*7,729*	*10,006*
False Negatives	*278*	*101*	*50*
Recall	72.56%	90.03%	95.06%
Precision	67.00%	10.55%	8.78%
F1 Measure	69.67%	18.89%	16.07%
Accuracy	99.7794%	97.3009%	96.5336%
Error	0.2206%	2.6991%	3.4664%
Elusion	0.10%	0.04%	0.02%
Fallout	0.13%	2.67%	3.46%

Topic 430 - Restraints and Helmets - UNCORRECTED
Total Documents: 290,099
Total Relevant: 991
Total Prevalence: 0.34%

Confusion Matrix - Restraints and Helmets

	@Reasonable	@90% Recall	@95% Recall
True Positives	*713*	*892*	*942*
True Negatives	*288,724*	*281,318*	*278,884*
False Positives	*384*	*7,790*	*10,224*
False Negatives	*278*	*99*	*49*
Recall	71.95%	90.01%	95.06%
Precision	65.00%	10.27%	8.44%
F1 Measure	68.30%	18.44%	15.50%
Accuracy	99.77%	97.28%	96.46%
Error	0.23%	2.72%	3.54%
Elusion	0.10%	0.04%	0.02%
Fallout	0.13%	2.69%	3.54%

Summary

This topic was run by Jani Grantz. This was her first attempted topic.

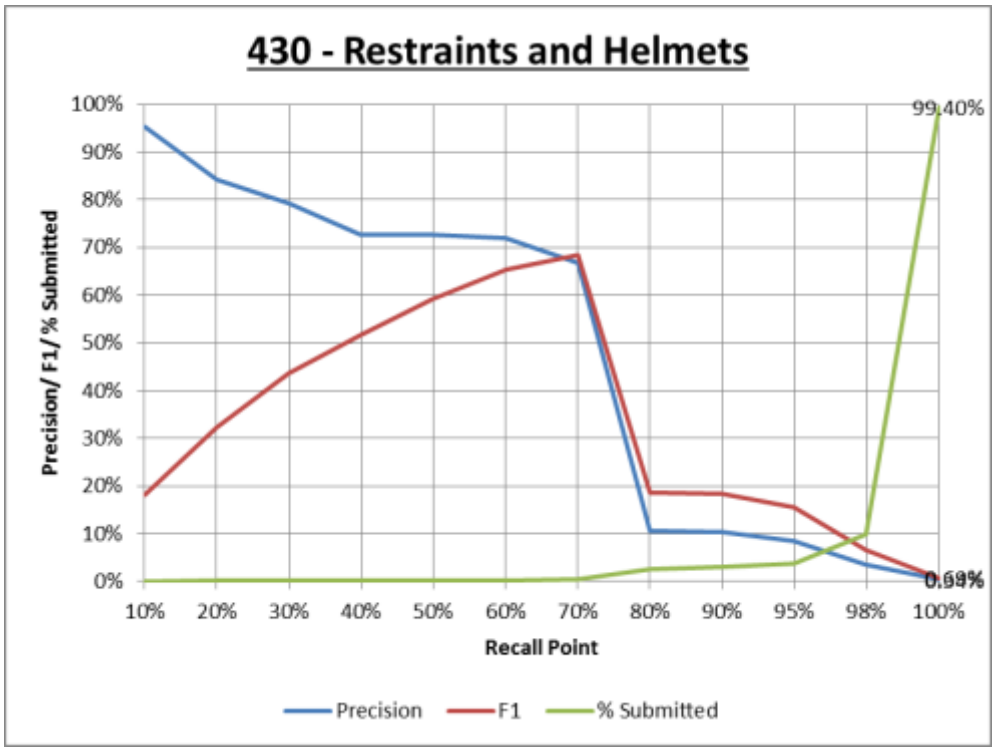
She began the process by running keywords that seemed logical to the topic and set up highlighting with those words. She split this topic up into its two parts 1) restraints and 2) helmets. Then she did some informal Doc Review on the docs that hit on multiple terms/most important terms for responsiveness. She started with small submissions of documents that she marked responsive for Restraints and found that almost every doc that hit on a term was Relevant, so this topic seemed easy and complete quickly.

However for the Helmets topic she did the same thing but found little rhyme or reason to docs that were relevant versus not relevant. She tried people outside of Florida as not relevant. Some were not relevant, but some were, She tried generic form responses to be Relevant at first, but they returned as both evenly. After that I gave up on trying to determine which docs were relevant and ran learning sessions and just submitted by probability since she had nothing else to go on. She started with the highest probability from ones that hit on some terms and went from there. She called reasonable when she got below a certain threshold percent where no more docs seemed like they should be relevant.

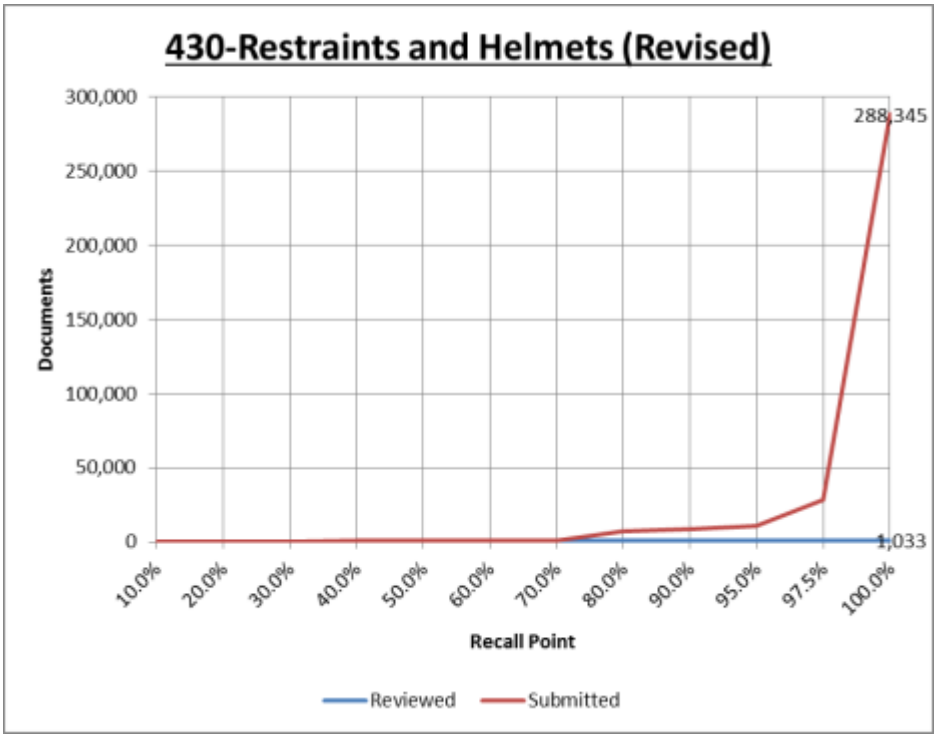
There is not much work placed into determining a corrected gold standard for this topic.

Graphs

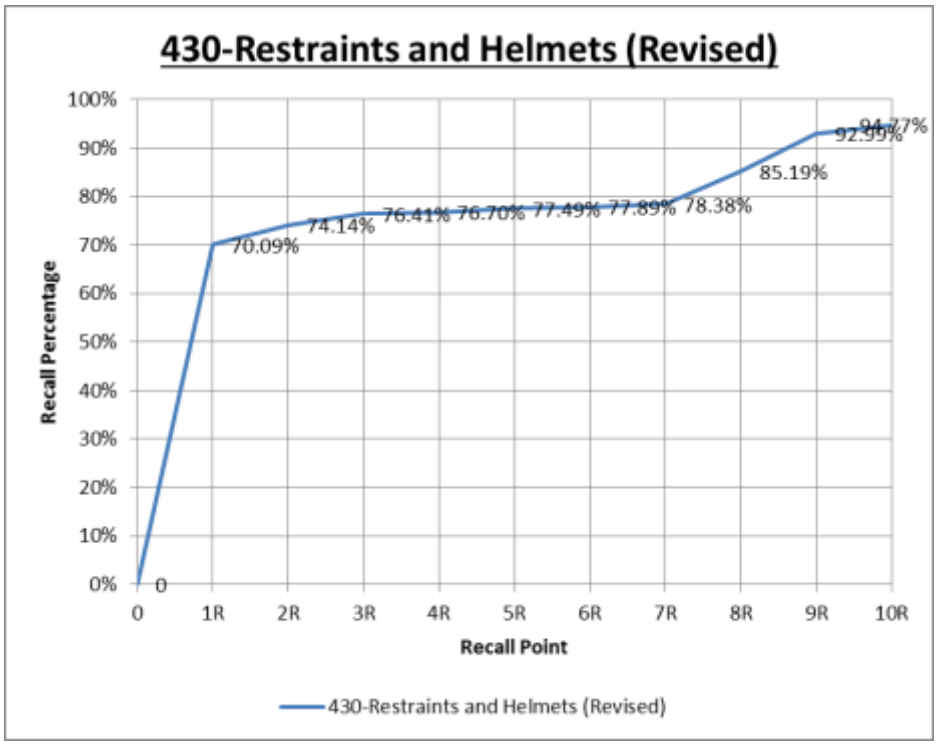
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Restraints and Helmets topic, the 90% recall threshold had been attained by submitting only 2.98%% of the corpus, 8,641 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).



Topic 431 - Agency Credit Ratings
Total Documents: 290,099
Total Relevant: 149
Total Prevalence: 0.05%

Confusion Matrix - Agency Credit Ratings

	@Reasonable	@90% Recall	@95% Recall
True Positives	*120*	*135*	*142*
True Negatives	*289,841*	*289,268*	*289,109*
False Positives	*109*	*682*	*841*
False Negatives	*29*	*14*	*7*
Recall	80.54%	90.60%	95.30%
Precision	52.40%	16.52%	14.45%
F1 Measure	63.49%	27.95%	25.09%
Accuracy	99.9524%	99.7601%	99.7077%
Error	0.0476%	0.2399%	0.2923%
Elusion	0.01%	0.00%	0.00%
Fallout	0.04%	0.24%	0.29%

Topic 431 - Agency Credit Ratings - UNCORRECTED
Total Documents: 290,099
Total Relevant: 144
Total Prevalence: 0.05%

Confusion Matrix - Agency Credit Ratings

	@Reasonable	@90% Recall	@95% Recall
True Positives	*109*	*130*	*137*
True Negatives	*289,835*	*289,242*	*277,498*
False Positives	*120*	*713*	*12,457*
False Negatives	*35*	*14*	*7*
Recall	75.69%	90.28%	95.14%
Precision	47.60%	15.42%	1.09%
F1 Measure	58.45%	26.34%	2.15%
Accuracy	99.95%	99.75%	95.70%
Error	0.05%	0.25%	4.30%
Elusion	0.01%	0.00%	0.00%
Fallout	0.04%	0.25%	4.30%

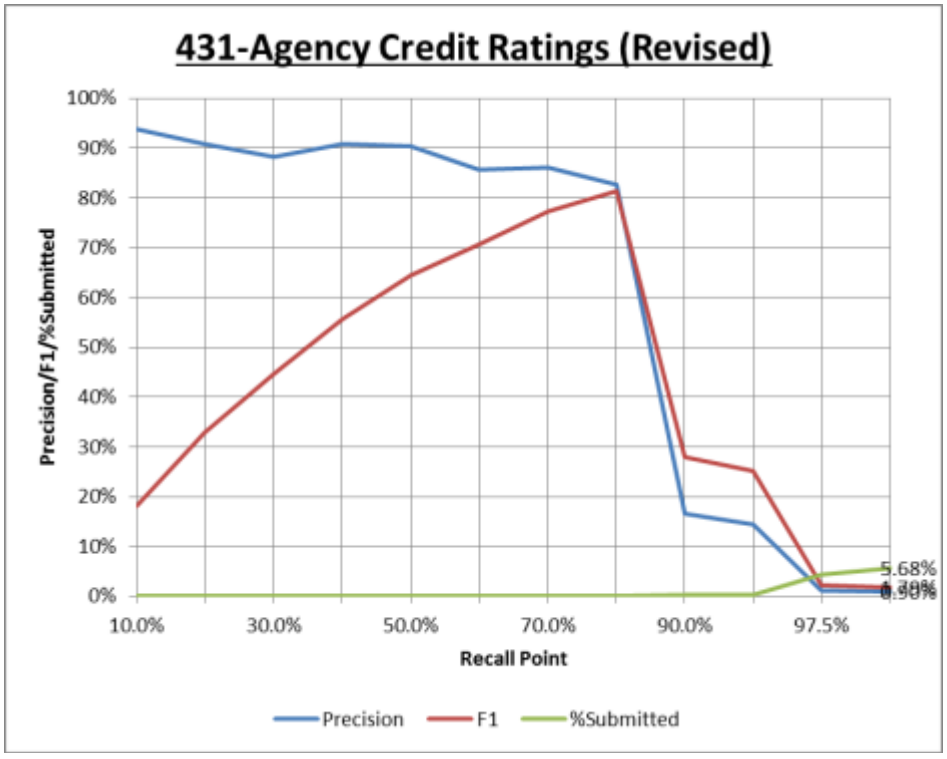
Summary

This topic was run by Tony Reichenberger. The hybrid multimodal review was conducted by initially submitting keyword hits (initially just the ratings agencies and various bond ratings) to train the machine learning, then letting the system suggest documents at various thresholds. Keyword hits were submitted in descending probability score order followed by learning sessions for the system, with submission sizes kept relatively small (10-20 documents each). Periodically, documents not hitting on keywords with high scores were submitted to ensure inclusiveness. Once all keyword hit documents were submitted, documents were submitted based solely on probability scoring; when additional relevant materials were found, subsequent searches for similar documents were partaken.

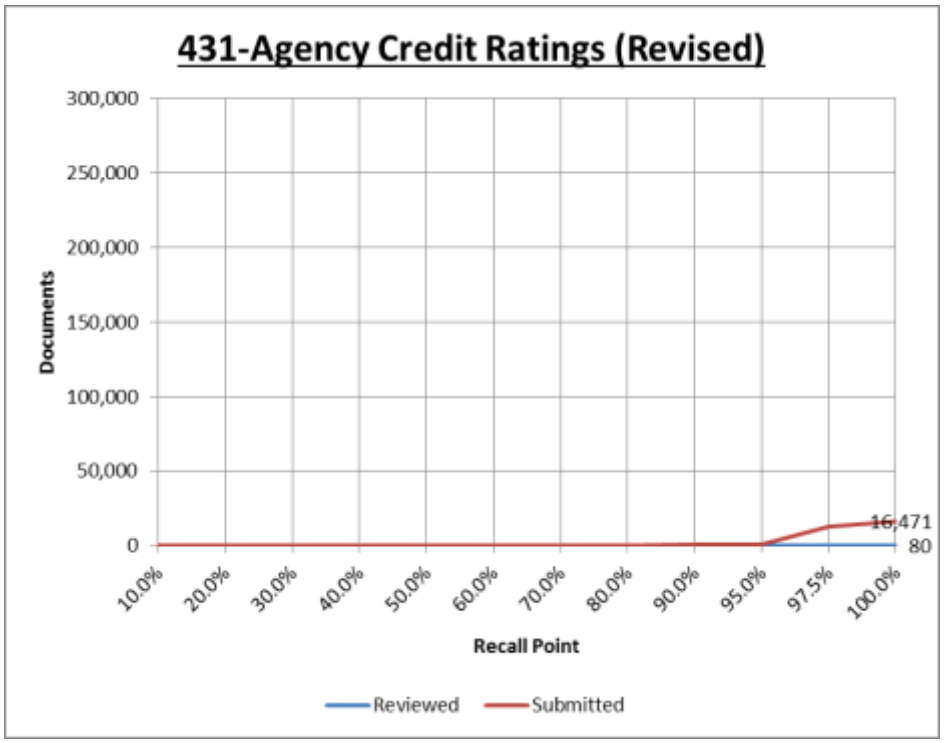
Reasonable was called when scores on keywords remaining were less than 25% and scores on all documents were less than 75%. Samples of keywords remaining at the time hit on only bond ratings but in a different context (AAA, B-, etc).

Graphs

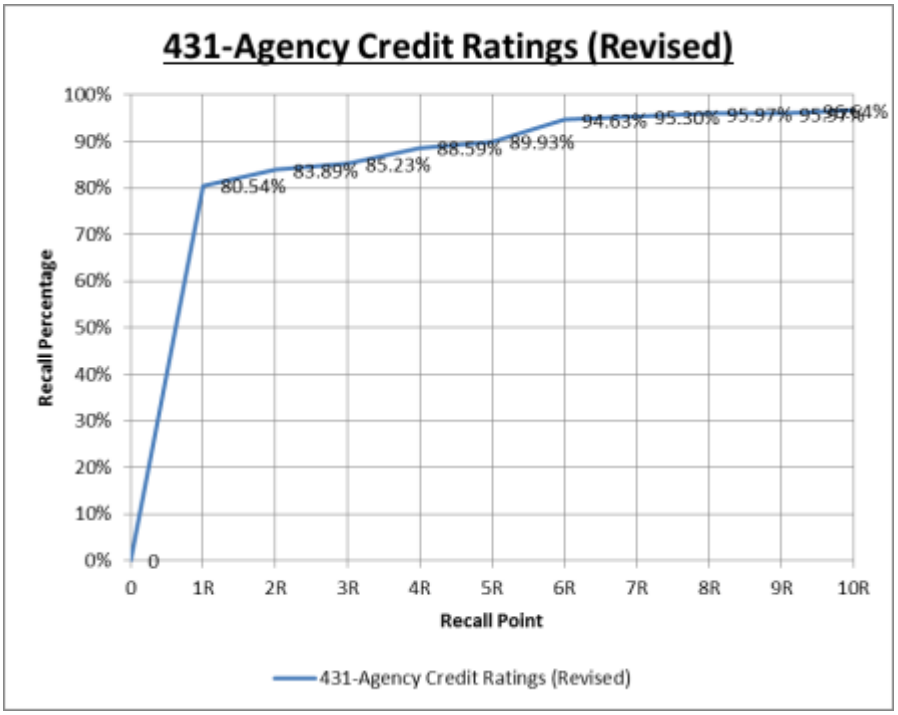
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Agency Credit Ratings topic, the 90% recall threshold had been attained by submitting only 0.28%% of the corpus, 817 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


431-Agency Credit Ratings (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
80.54%
83.89%
85.23%
88.59%
89.93%
94.63%
95.30%
95.97%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
431-Agency Credit Ratings (Revised)

Topic 432 - Gay Adoption
Total Documents: 290,099
Total Relevant: 137
Total Prevalence: 0.05%

Confusion Matrix - Gay Adoption

	@Reasonable	@90% Recall	@95% Recall
True Positives	*125*	*124*	*131*
True Negatives	*289,949*	*289,949*	*267,375*
False Positives	*13*	*13*	*22,587*
False Negatives	*12*	*13*	*6*
Recall	91.24%	90.51%	95.62%
Precision	90.58%	90.51%	0.58%
F1 Measure	90.91%	90.51%	1.15%
Accuracy	99.9914%	99.9910%	92.2120%
Error	0.0086%	0.0090%	7.7880%
Elusion	0.00%	0.00%	0.00%
Fallout	0.00%	0.00%	7.79%

Topic 432 - Gay Adoption - UNCORRECTED
Total Documents: 290,099
Total Relevant: 140
Total Prevalence: 0.05%

Confusion Matrix - Gay Adoption

	@Reasonable	@90% Recall	@95% Recall
True Positives	*119*	*126*	*133*
True Negatives	*289,940*	*279,621*	*245,846*
False Positives	*19*	*10,338*	*44,113*
False Negatives	*21*	*14*	*7*
Recall	85.00%	90.00%	95.00%
Precision	86.23%	1.20%	0.30%
F1 Measure	85.61%	2.38%	0.60%
Accuracy	99.99%	96.43%	84.79%
Error	0.01%	3.57%	15.21%
Elusion	0.01%	0.01%	0.00%
Fallout	0.01%	3.57%	15.21%

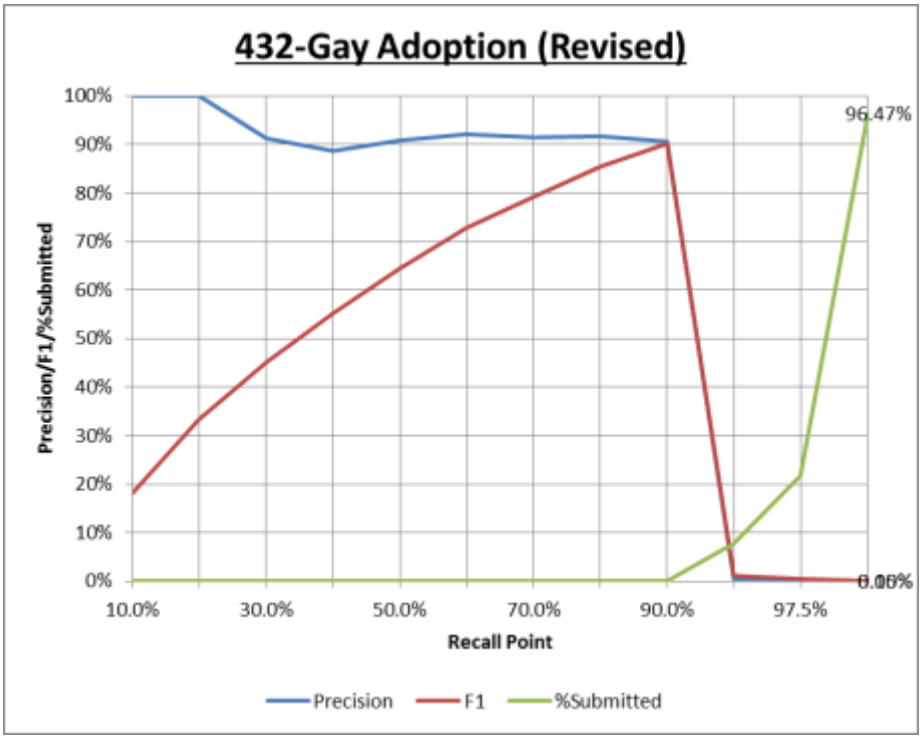
Summary

This topic was run by Jani Grantz. This was her second attempted topic. She began the process by running keywords that seemed logical to the topic and set up highlighting with those words. Then she did some informal Doc Review on the docs that hit on multiple terms/most important terms for responsiveness. She started with one moderately sized submission based on docs she found relevant. Then from the results that came back relevant she used Find Similar to find others that should be relevant. She did that to find additional keywords and relevant docs and then did a couple more submissions until she felt like she was out of clearly relevant docs. Then she ran learning sessions and submitted a few more that had a high percentage of likelihood to be relevant. When she felt like I exhausted those (reached a certain percentage) she called reasonable and submitted the rest.

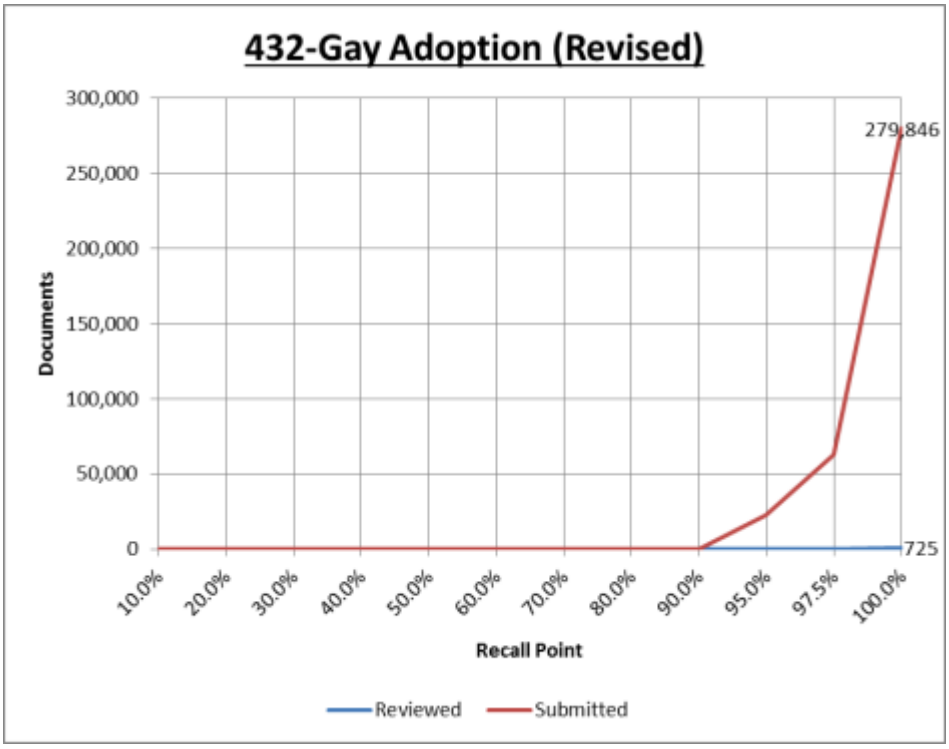
There is not much work placed into determining a corrected gold standard for this topic.

Graphs

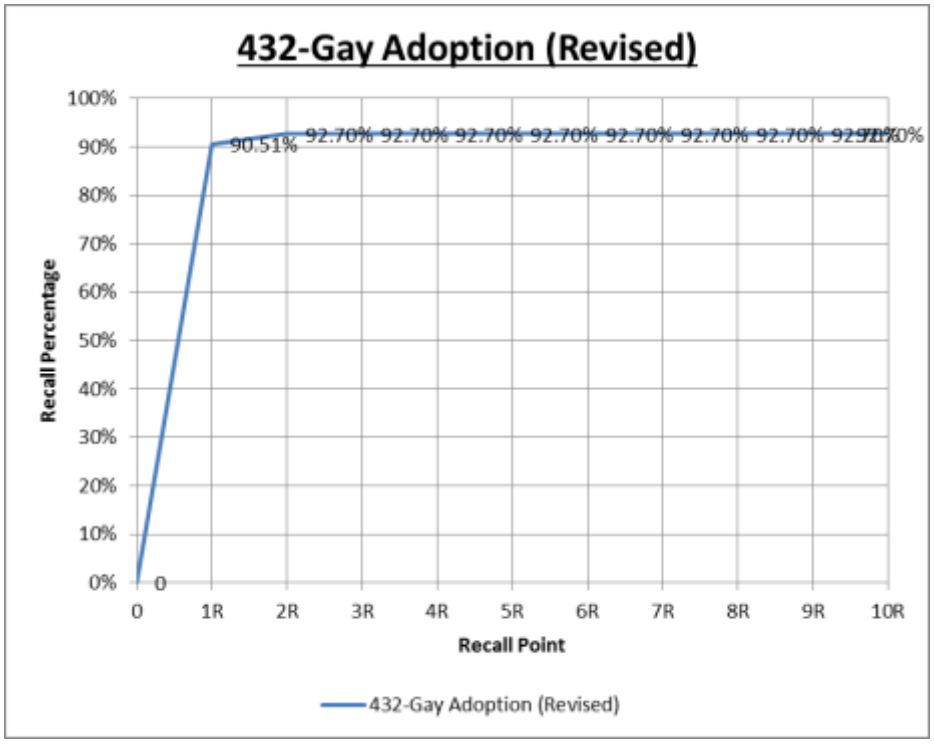
The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Gay Adoption topic, the 90% recall threshold had been attained by submitting only 0.05%% of the corpus, 137 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).


432-Gay Adoption (Revised)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recall Percentage
0
90.51%
92.70% 92.70% 92.70% 92.70% 92.70% 92.70% 92.70%
0
1R
2R
3R
4R
5R
6R
7R
8R
9R
10R
Recall Point
432-Gay Adoption (Revised)

Topic 433 - Abstinence
Total Documents: 290,099
Total Relevant: 141
Total Prevalence: 0.05%

Confusion Matrix - Abstinence

	@Reasonable	@90% Recall	@95% Recall
True Positives	*141*	*127*	*134*
True Negatives	*289,931*	*289,950*	*289,950*
False Positives	*27*	*8*	*8*
False Negatives	*0*	*14*	*7*
Recall	100.00%	90.07%	95.04%
Precision	83.93%	94.07%	94.37%
F1 Measure	91.26%	92.03%	94.70%
Accuracy	99.9907%	99.9924%	99.9948%
Error	0.0093%	0.0076%	0.0052%
Elusion	0.00%	0.00%	0.00%
Fallout	0.01%	0.00%	0.00%

Topic 433 – Abstinence - UNCORRECTED
Total Documents: 290,099
Total Relevant: 112
Total Prevalence: 0.04%

Confusion Matrix - Abstinence

	@Reasonable	@90% Recall	@95% Recall
True Positives	*111*	*101*	*107*
True Negatives	*289,930*	*289,957*	*289,956*
False Positives	*57*	*30*	*31*
False Negatives	*1*	*11*	*5*
Recall	99.11%	90.18%	95.54%
Precision	66.07%	77.10%	77.54%
F1 Measure	79.29%	83.13%	85.60%
Accuracy	99.98%	99.99%	99.99%
Error	0.02%	0.01%	0.01%
Elusion	0.00%	0.00%	0.00%
Fallout	0.02%	0.01%	0.01%

Summary

Topic 433 was run by Jim Sullivan, who started on June 14, 2016 and concluded on June 16, with two days of review.

Sullivan is not an expert in abstinence, neither in practice nor in theory.

Sullivan started by testing terms and creating a keyword highlight list for term hits and common variations, as was done on all topics reviewed. He started by submitting documents that hit on obvious terms in the subject line, and moved to more generic keywords in broader fields. At the end of the first day he had submitted 67 documents, with 57 relevant. He disagreed with the TREC categorization on the remaining 10. He initiated a learning session after training 500 randomly selected documents as Not Responsive.

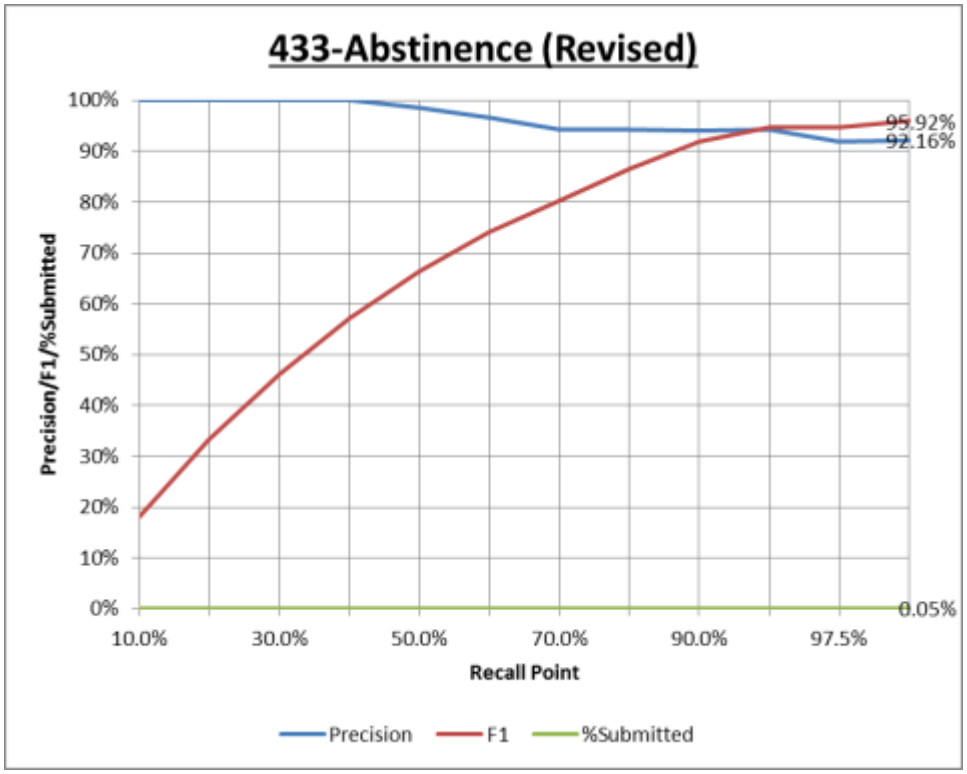
The second day of review was spent submitting documents with the highest predictive coding scores. He called 80% and reasonable recall after 168 total documents submitted, with 111 relevant. In total, 112 documents were returned Relevant by the TREC standard.

After the reasonable call, all remaining documents were submitted by predictive coding score with the highest scores being submitted first. Only 3.5 hours were spent reviewing this topic, considering 111 of the 112 TREC relevant documents hit on the term "abstinence," with only 40 documents in the entire database containing abstinence being returned as Not Relevant, and most of those being errors in the TREC standard.

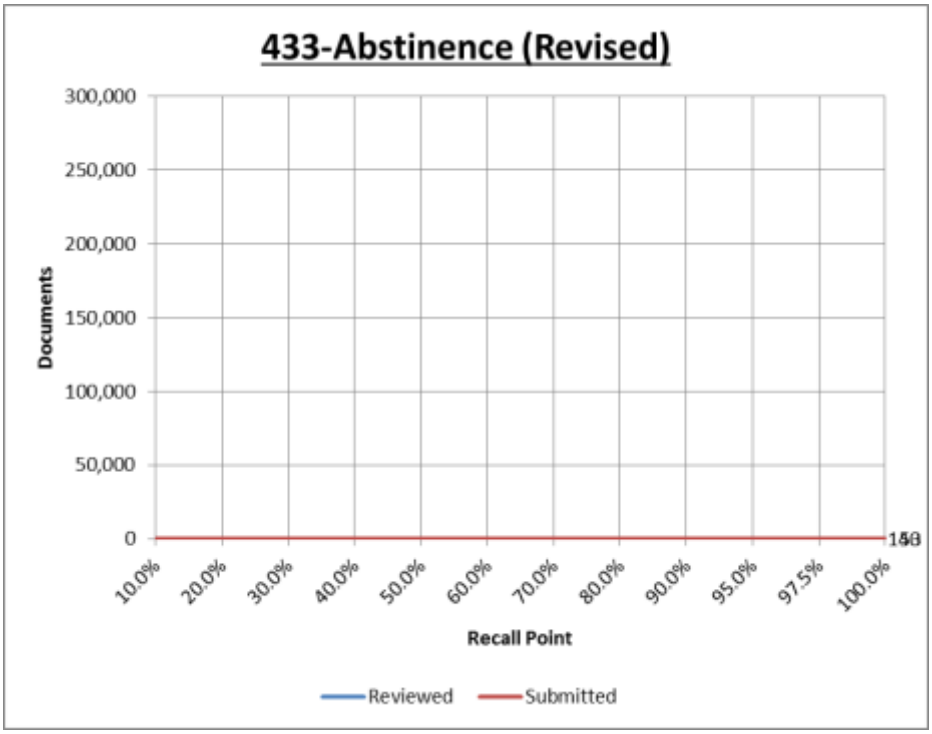
This topic was graded poorly for such an easy topic. While there were only 31 documents where TREC had returned inconsistent or incorrect classifications, the scope of documents containing the word abstinence was so small the high error rate was surprising. There were 2 documents that contained a misspelling of abstinence that were clearly missed (TRECID 285286 and 285292), and one document not containing the term abstinence marked Relevant by TREC for no apparent reason (TRECID 267623).

Graphs

The following chart shows Precision (blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Abstinence topic, the 90% recall threshold had been attained by submitting only 0.05%% of the corpus, 135 documents for adjudication.



The next chart below represents the amount of effort (documents actually reviewed eyes on) versus how many were submitted to attain 100% recall using the multi-modal hybrid model of training EDR.



The last chart shows the progression through the database submissions based on attained recall at various recall points throughout the database (2x # of recall documents, 3x Recall documents, etc).