Relativity®

White Paper

# The Impact of Judgmental Sampling on Assisted Review

# Contents

# Introduction

Computer-assisted review relies on human reviewers submitting coded example documents to a machine learning algorithm. The computer uses those examples to develop a model of responsiveness, which it can then apply to the entire document population.

In this white paper, our goal is to study the impact of sample size and sampling methodology on the effectiveness of a computer-assisted review project. We also compare the results of using keyword searches to judgmentally select the first set of exemplars to the results of identifying and reviewing a random sample.

The observations made in this paper are derived from an experiment in Relativity Assisted Review. The results suggest that randomly selected example documents can be a more reliable, cost-efficient methodology, while providing some statistical assurances about the project's results.

## Goals of the Study

Before we started running our experiments, we began with some overarching questions we sought to answer.

### Question 1: How large should your sample be?

This question pertains to the size of a random sample. Statistically speaking, the more exemplars we use in a computer-assisted review process, the better we can assume the project's overall effectiveness will be. Because there is an actual cost associated with every exemplar document subjectively coded by a human reviewer, is there an optimal sample size that offers a perfect trade-off between review investment and machine categorization performance? To help address this question, we will call the cost-efficient sample set volume the price performer size.

### Question 2: Is it better to use only random samples, or should you start with judgmental samples?

At the outset of such an experiment, it seems intuitive that keyword searches will help in finding good exemplars for a machine learning algorithm. Judgmental sampling is the process of selecting a relatively small set of documents via manually identified criteria,  as opposed to randomly selecting the documents without any bias. Keyword searches served as the judgmental sampling methodology used for this experiment.

# Background

Computer-assisted review begins when reviewers identify responsive and non-responsive documents and submit them to a document categorization algorithm. The system then labels the entire document collection based on these exemplars: it predicts which documents are responsive, which are non-responsive, and leaves undetermined documents uncategorized. In a typical workflow, the case team will then seek to quality control the system's results and then re-train based on the corrected exemplars.

The process continues until the case team determines that their project has achieved satisfactory results. One way to track the computer's progress is to use a control set of documents. The control set is selected via random sampling, and is subjectively reviewed and set aside. These documents then serve as the ground truth, and each round of training results will be evaluated against the control set. The comparison of the results is done using information retrieval measurements of precision, recall, and F1.

Different categorization algorithms may yield varied results. That said, more important than the algorithm itself is the size and quality of the training exemplars provided to the system. To test the variations of submitting exemplars to the system, multiple iterations of workflow are not needed—a single round is sufficient to determine the impact of categorization effectiveness. A single round tests the effect of the sample on the categorization's effectiveness without clouding the results with the impact of a multi-round workflow.

We tested different sizes of sample sets as well as different types—i.e., judgmental and random. Additionally, our experiment used a data set that has been 100 percent reviewed. The data for this experiment is derived from the Text Retrieval Conference (TREC)[1] data set. The data has been fully human reviewed, so we're able to calculate recall. We used roughly 20,000 reviewed documents from the TREC Legal Track Enron data set that include judgments for each identified issue. This subset is shown in Figure 1.

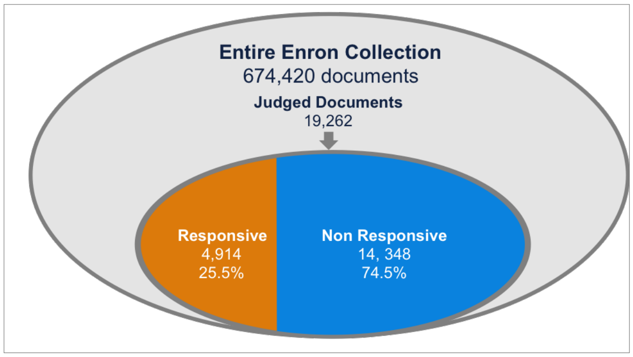

Figure 1

We then ran a single round of categorization for different exemplars and measured precision and recall. Figure 2 visualizes the difference between precision and recall, which are both explained in more detail below.

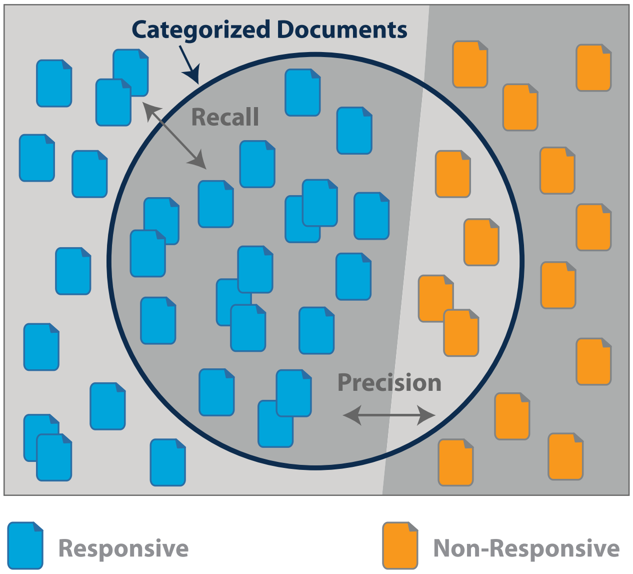

Figure 2

[1] http://trec.nist.gov/proceedings/proceedings.html

Precision is the ratio of documents correctly categorized as responsive by the computer to the total number of documents categorized as responsive. It's relatively easy to calculate, as precision is found once the team QCs the documents the computer categorized as responsive. For example, if we categorized 100 documents and found that 70 were correct—as demonstrated by the blue and orange documents in the inner circle of the graph—precision is 0.70.

Recall is more difficult to quantify because it's a measure of how many responsive documents we found compared to how many were possible to be found. To determine recall, therefore, we would have to read every document in the collection and consider whether or not it was responsive to each query.

Mathematically, recall is the ratio of the number correctly categorized as responsive divided by the total number to be found. In Figure 2, you'd find it by dividing the same 70 blue boxes inside of the circle we discussed by all of the blue boxes in the diagram. If we found 70 inside the circle but 700 existed in the collection, recall would be computed as 70 divided by 700, which equals 10 percent.

Finally, a measure called F1 is used to provide a single number for effectiveness. F1 is the harmonic mean of precision and recall, and is computed according to the following formula.

$$\mathbf{F1 = 2 * \frac{precision * recall}{precision + recall}}$$

# Methodology

## Question #1

To answer the question of optimal sample size, we used sample size variables of 96; 378; 1,431; 2,152; 3,000; 4,000; 5,000; and 6,559. In this data set, a sample size of 96 documents yields a 95 percent confidence level with a +/- 10 percent margin of error. A 6,559-document sample yields a 95 percent confidence level with a +/- 1 percent margin of error.

Figure 3 shows a graph with an x-axis of sample size and a y-axis of recall. Note the logarithmic curve that shows a slight increase in recall as we increase sample size. The orange square indicates the point at which recall increases less than 0.05 percent for a new 100 training samples. This happened at a sample size of 2,010, between 1,431 and 2,152, which is obtained with a 95 percent confidence level and +/- 2 or +/- 5 percent margin of error.

The graph shows an R2 value of 0.96. R2 quantifies variation in data, indicating the goodness-of-fit of a curve to the numbers collected. A value of 1.0 would indicate that the line fits perfectly—meaning there is very little variation—so 0.96 is strong. These results will vary depending on the data set, but, in general, diminishing returns can be expected as sample size increases. In this study, all random sample tests were run five times to ensure that no significant variation occurred. Figure 4 shows that little variation occured for sample sizes once we got beyond a small size of 96 documents.

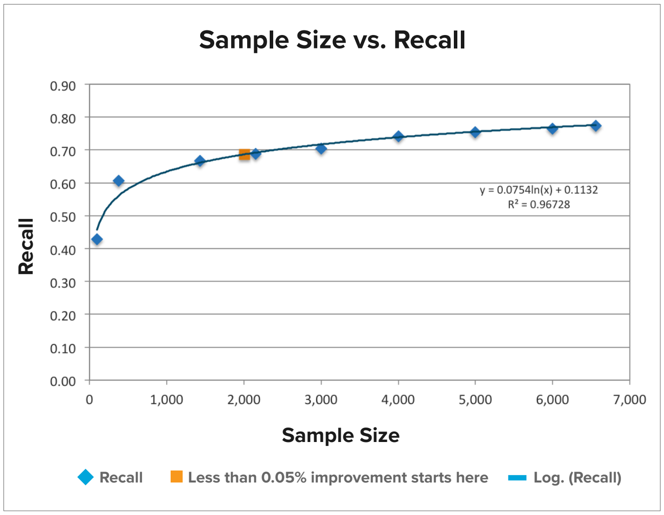

Figure 3

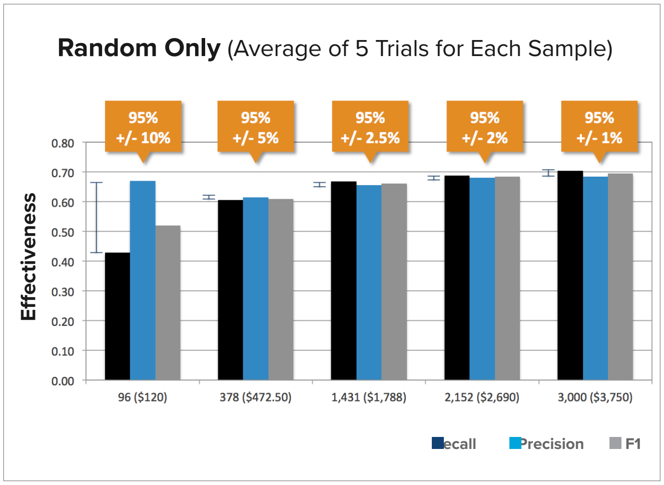

Figure 4

## Question #2

To evaluate the difference between judgmental sampling and random sampling, we took a random sample using a sample size of 2,152—essentially within the most efficient sample size range —and compared it to several different judgmental runs. To obtain results for judgmental sampling, we developed simple keyword searches that attempted to find documents based on the TREC descriptions. Figure

5 provides the TREC descriptions, and Figure 6 shows the keyword searches used to find documents of interest.

**First Set of Requests for Production:**

Plaintiffs request that Defendants produce all responsive documents on the following topics.

- All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
- All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
- All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
- All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.
- All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
- All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s), or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i)the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv)the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
- All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

**Important procedural note specific to Topic 207: solely for the purpose of the TREC 2009 Legal Track, any participant who chooses to submit results for Topic 207 must also submit results for at least one of the other topics (201-206) featured in the 2009 Interactive Task.**

Figure 5

| Difficulty | Topic No. | Keyword Search - Version 2 |
|---|---|---|
| easy | 201 | prepay transactionsor "structured commodity transactions" or "commodity transactions" |
| easy | 202 | FAS 140 or "FAS 125 |
| difficult | 203 | ("financial forecast" or "financial model" or "projection" or "projections" or "financial plans") and ("exceed" or "could meet" or "might meet" or "will meet" or "would meet" or "can meet") |
| difficult | 204 | destroy documents or "delete documents" or "shred documents" or "delete files" or "remove files" or alteration or destruction or retention or "lack of retention" or "deletion" or "shredding of documents" or "shred evidence" |
| medium | 205 | ("energy schedules" or "energy bids") and (characterization or analysis or estimate or estimates or forecast or forecasts or descriptions or characterizations or evaluations or plans or plan or report or reports or volume or "geographic location") and ("energy load" or "energy loads" or "load" or "loads") |
| difficult | 206 | "financial condition" or "analyst coverage" or "analyst rating" or "analyst coverage" or (impact w/20 relationship) |
| easy | 207 | Fantasy football or "football teams" or "football team" or "football statistics" or "football statistic" or "football performance" |
| easy | 207 | bears or lions or giants or packers or vikings or 49ers or falcons or rams or eagles or "Dallas Cowboys" or redskins or cardinals or rams or ravens or browns or steelers or bengals or patriots or dolphins or "buffalo bills" or "NY jets" or "N.Y. Jets" or "new york jets" or chargers or seahawks or chiefs or raiders or colts or "houston texans" or "jaguars" or "titans" or broncos |

Figure 6

After running the keyword searches and comparing the results to ground truth, we noticed that one search (topic 205) found only four responsive documents out of 1,343 that could have been found.

We revised our queries to simulate a user who is learning more about the data set and can, therefore, input better keyword searches for the data. This step was rather subjective, as judgmental sampling tends to be, so it's worth keeping that in mind. Would a more experienced legal expert have come up with better search terms that would significantly improve the results? For the sake of this experiment, we ran several different sets of search terms to address the data in several ways, and to take a broad stroke in assessing effectiveness of keyword search.

Once we had the baseline of keyword responses, we ran categorization on the documents. Categorization was performed by Relativity Assisted Review, which uses latent semantic indexing (LSI). LSI categorization works by generating a query of the concepts found in the responsive documents—which are identified based on term co-occurrence—and running that query against the document collection. Documents that match the query above a given threshold are categorized as responsive. Similarly, a non-responsive query is run and documents that match above a threshold are categorized as non-responsive. Any document that does not exceed the threshold for either query is labeled as uncategorized. The threshold used for these experiments was 0.7 and the number of dimensions used in the concept space was 100. Dimensions drive the amount of semantic information kept by the LSI process as it works to reduce noise in the data.

These are the default settings for Assisted Review, and have been tested on various data sets. It's noteworthy that the best settings for these numbers may change from matter to matter.

As a result of the findings described above, we ran the following tests for our experiment:

1. Random sampling using a sample size of 2,152, followed by responsive and non-responsive queries for categorization.
2. Judgmental sampling using the search results of version two of the keyword queries on the entire document population coupled with responsive queries for document categorization. The sample size for this test was 1,401 documents.
3. Judgmental sampling using the search results for version two of the keyword queries coupled with only responsive queries for categorization. This was the only test in which we tried a responsive-only approach.
4. Judgmental sample with the 2,627 documents found as a result of the version two searches, plus 377 (95 percent confidence level and +/- 5 percent margin of error) additional, randomly selected documents to reach a sample size of 3,004. Responsive and non-responsive queries were used for document categorization in this test.

# Results

Figure 7 shows the results of running version one of the keyword searches with no categorization. The diamonds represent the number of documents to be found in each of the TREC queries: numbers 201–207. The results show recall levels were extremely low—below 10 percent—for most queries when using only a keyword search. Even when the queries were improved with version two, recall remained well below 50 percent. Quite a bit of noise was generated. We can see non-responsive documents—false positives—blossomed for topic 204. Essentially, by relying on keyword searching to pull the sample for this test, we added more wildcards to the queries. These queries did retrieve more responsive documents. However, the wildcards are not always as specific as we would like. Hence, these queries returned many non-responsive documents.

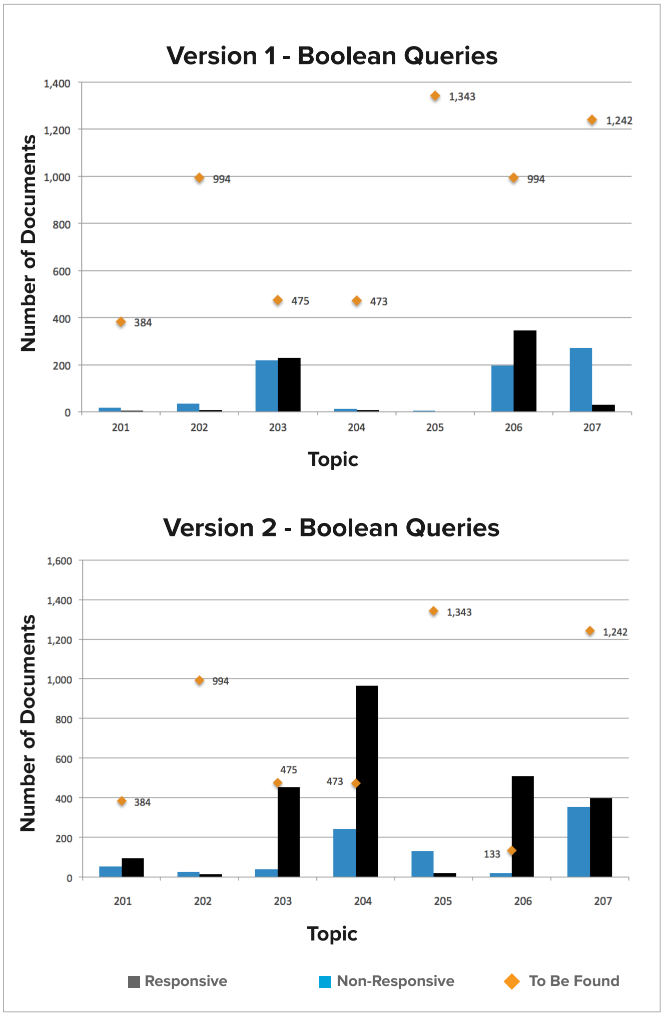

Figure 7

Topic-by-topic results for the four tests described in the previous section are shown in Figure 8. For each topic, the first two bars show results for random sampling, and the second two bars show samples derived from keyword queries. The chart indicates that random samples perform reasonably well across the set of queries. It's noteworthy that this graph depicts only the number of responsive documents, so it does not show how much noise appeared in each query.

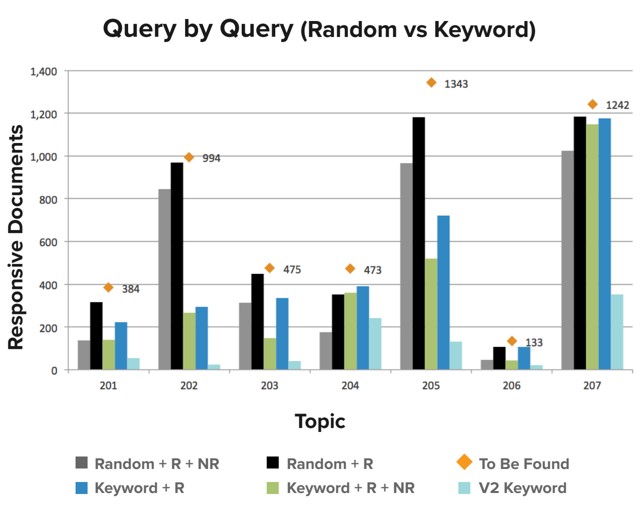

Figure 8

For any type of categorization, the number of responsive documents returned by the random approach significantly increases over the number found using only a keyword search.

The responsive-only topics did show significantly increased recall over categorization using responsive and non-responsive queries. Figure 9 shows the precision for each topic. It can be seen that precision drops significantly for responsive-only, so it may only be an ideal approach for projects that mainly require high recall, such as government investigations or production reviews.

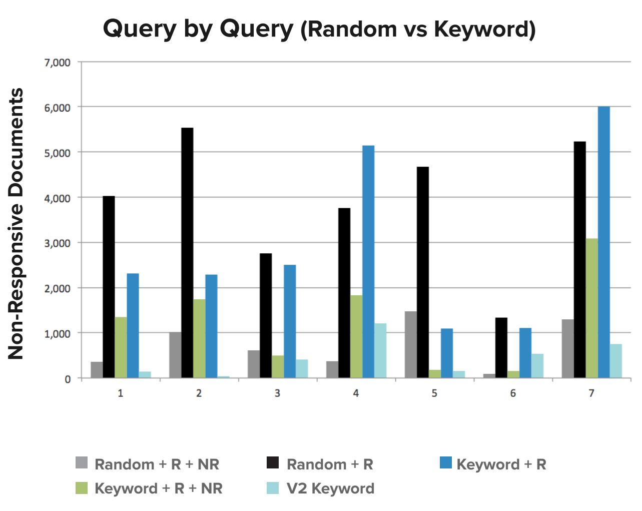

Figure 9

To summarize, we averaged the seven topics and showed precision, recall, and F1 in Figure 10. This figure also includes a cost per responsive document, which is computed as $1.25 to read each sample document and $1.25 to read each categorized document. The cost per responsive document is computed as the cost of reading the exemplars and the categorized documents, and dividing by the number of responsive documents that are returned. This shows that random sampling achieved the highest precision, reasonable recall, and the highest F1. It also obtained the lowest cost per responsive document.

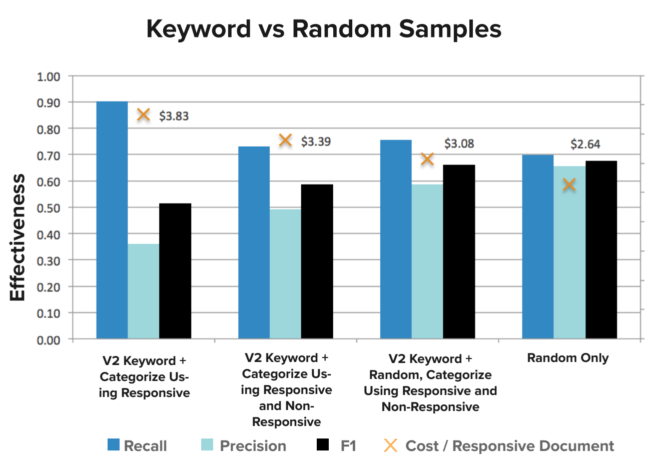

Figure 10

# Conclusions

This study reasonably suggests that a random sample is a sound approach in computer-assisted review, and may outperform a keyword-based sample. Given the statistical foundation provided by a random sample and the inherent biases of keyword searches, the results suggest that random samples work well for the task of seeding a categorization engine with a representative population of exemplars for training. That said, there may be some cases where a random sample may not be practical—particularly cases where very few responsive documents exist and the number of exemplars needed to ensure a statistically significant result require a larger budget than a client is willing or able to spend.

Additionally, the results did not show any real benefit from running a set of exemplars obtained from a keyword search and randomly obtained documents. It should be noted that such an approach would lack statistical meaning to derive conclusions about properties of the whole population.

Finally, the results did show that a responsive-only categorization can lead to high levels of recall (90 percent), but at a cost of extremely low precision (36 percent).

To summarize the answers to our key questions:

### Question 1: How large should your sample be?

We found that a sample size range between 1,431 and 2,152 documents was the best price-performer size and yielded good recall. This number is determined by the statistical parameters we set for the project, so individual projects will vary accordingly.

### Question 2: Is it better to use only random samples, or should you start with judgmental samples?

This study has shown that random sampling is an efficient methodology for submitting to a categorization engine. We suspect this result will be true for a number of different document collections. However, we do have concerns about using random samples when the number of responsive documents is very small, as the amount of training exemplars the case team must find to ensure a robust result may exceed the funds available for a case.

kCura®

231 South LaSalle Street, 8th Floor, Chicago, IL 60604
T: 312.263.1177 • F: 312.263.4351
sales@kcura.com • www.relativity.com