An Exploration of Total Recall with Multiple Manual Seedings

Final Version

Jeremy Pickens, Tom Gricks, Bayu Hardi, Mark Noel, John Tredennick
Catalyst Repository Systems
1860 Blake Street, 7th Floor
Denver, CO 80202
jpickens,tgricks,bhardi,mnoel,jtredennick@catalystsecure.com

ABSTRACT

The Catalyst participation in the manual at home Total Recall Track had one fundamental question at the core of its run: What effect various kinds of limited human effort have on a total recall process. Our two primary modes were one-shot (single query) and interactive (multiple queries).

1. OFFICIAL RUN

Our official run approach consisted of two main parts:

1. Manual selection of a minimal number of initial documents (seeding)
2. Continuous learning with active learning as relevance feedback, only

1.1 Manual seeding

Four different reviewers participated in every topic. For each topic, a reviewer was assigned to manually seed that topic either by doing a single query (one-shot) and flagging (reviewing) the first (i.e. not necessarily the best) 25 documents returned by the query, or run as many queries (interactive) as desired within a short time period, but stop after 25 documents had been reviewed. In the one-shot approach, the reviewer is not allowed to examine any documents before issuing the query, i.e. the single query is issued "blind" after only reading the topic title and description. The first 25 documents returned by that query are flagged as having been reviewed, but because there is no further interaction with the system, it does not matter whether or not the reviewer spends any time looking at those documents. In the interactive case, reviewers were free to read documents in as much or little depth as they wished, issue as many or as few queries as they wished, and use whatever affordances were available to them from the system to find documents (e.g. synonym expansion, timeline views, communication tracking views, etc.)

Every document that the reviewer laid eyeballs on during this interactive period had to be flagged as having been seen and submitted to the Total Recall server, whether or not the reviewer believed the document to be relevant. This was of course done in order to correctly assess total effort, and therefore correctly measure gain (recall as a function of effort). We also note that the software did not strictly enforce the 25 document guideline. As a result, sometimes the interactive reviewers went a few documents over their 25 document limit and sometimes they went a few documents under, as per natural human variance and mistake, but we do not consider this to be significant. Regardless, all documents reviewed, even with duplication, were noted and sent to the Total Recall server.

The reviewers working on each topic were randomized, assigned to run each topic either in one-shot or in interactive mode. Each topic had two one-shot and two interactive 25-document starting points. For our one allowed official manual run, these starting points were combined (unioned) into a separate starting point. Because we did not control for overlap or duplication of effort, the union of these reviewed documents is often smaller than the sum. Reviewers working asynchronously and without knowledge of each other often found (flagged as seen) the same exact documents.

In this paper, we augment the official run with a number of unofficial runs, four for each topic, two one-shot starting points and two interactive starting points. This will be discussed further in Section 3

1.2 Continuous Learning, Active Learning

In more consumer facing variations of our technology, multiple forms of active learning such as explicit diversification are used together in a continuous setting. As the purpose of those additional types of active learning are to balance out and augment human effort, we turned them off for this experiment in order to be able to more clearly assess the effects of just the manual seeding. Therefore, in this experiment, we do continuous learning with relevance feedback as the only active document selection mechanism [1].

2. OFFICIAL RESULTS

2.1 Gain Curves

Figure 1 is a plot of the official results in the form of gain curves over the relevant documents produced by averaging performance across all 34 athome4 topics. The plots on the left have the entire collection along the x-axis, and the plots on the right are narrowed down to the more interesting top few thousand. There are two BMI baselines: Topic title only (red) and title+description (blue), with our method given in black. Figure 2 is a plot of the official results in the form of gain curves over the *highly* relevant documents, similarly characterized.

There is not much to say about these results other than, on average, we come close but do not beat the baseline. The difference at 25% and 50% recall is small – only an extra 38

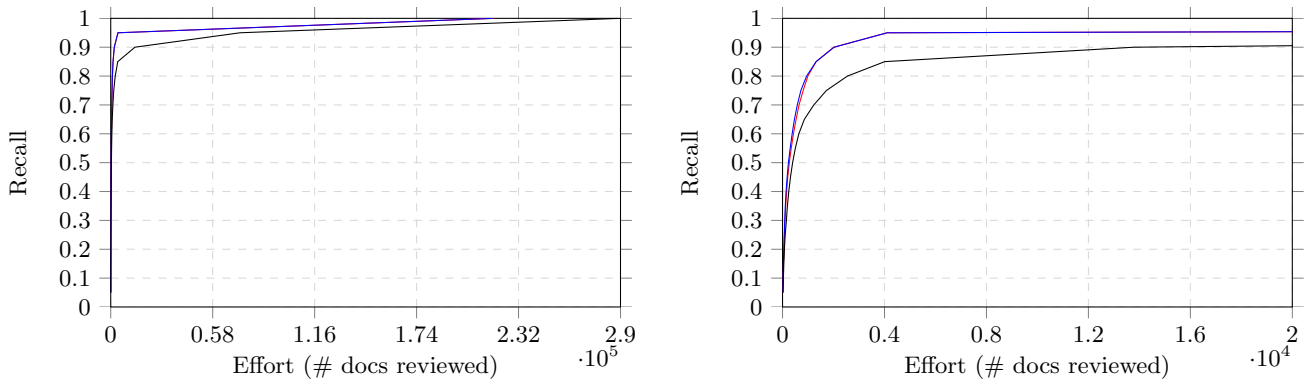

Figure 1: Relevant documents. Full collection gain curve (left), top 20k documents (right). BMI title-only = red, BMI description = blue, this paper = black.

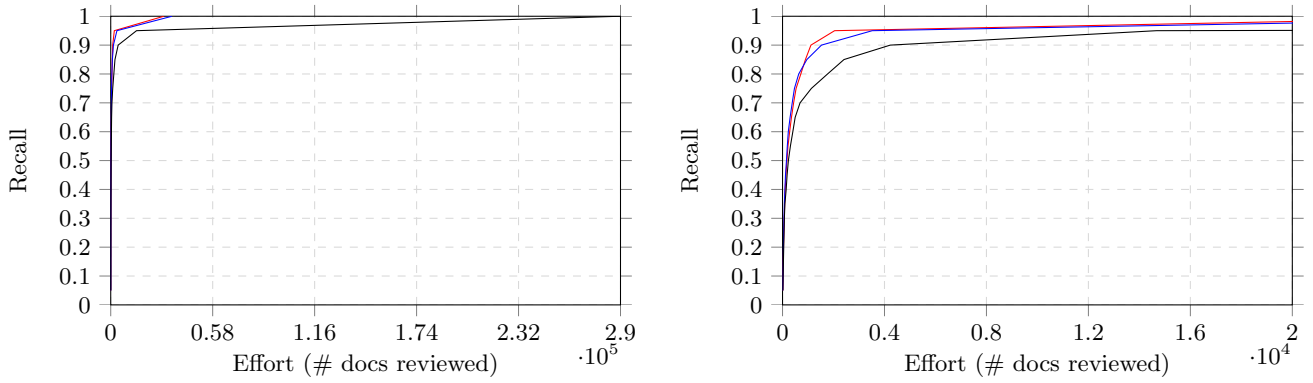

Figure 2: Highly relevant documents. Full collection gain curve (left), top 20k documents (right). BMI title-only = red, BMI description = blue, this paper = black.

documents are needed on average with our technique to get to 25% recall, and an extra 165 documents on average to get to 50% recall. The difference grows at 85% recall, with an extra 2692 documents needed. This is not a huge difference relative to all 290,000 documents in the collection, but is a non-zero difference nonetheless. We have some hypotheses related to this gap, but the purpose of this work is not to focus on comparisons to other methods, but on the effects of a variety of manual seeding approaches on our own methods.

However, we do note one interesting aspect of the results in Figures 1 and 2. And that's the fact that our technique seems to do better relative to the baseline on the highly relevant documents than it does relative to the baseline on the standard relevant documents. We have attempted to quantify that narrowing gap in the following manner: First, at all levels of recall in 5% increments, we calculate the difference in precision (d_{rel}) between our technique and each baseline for the relevant documents. Second, we calculate this difference in precision (d_{high}) between the techniques for the highly relevant documents. Then we calculate how much smaller d_{high} is than d_{rel} at each of these recall points, expressed as a percentage difference, using the following formula:

$$\frac{|d_{rel} - d_{high}|}{0.5 * (d_{rel} + d_{high})}$$

The result is shown in Figure 3. A negative value indicates that our highly relevant result is closer to the highly relevant baseline than our regular relevant result is to the relevant baseline. Across most recall points, the highly relevant result is between 100% and 300% closer to the baseline than is the regular relevant result.

However, while this is an interesting pattern in the data, it is difficult to know how to interpret this. Does the gap narrow on the highly relevant results because the highly relevant documents are also the more findable ones, and there is not a big difference between any reasonable technique? I.e., is the overall dynamic range on the highly relevant documents smaller? Or does the gap narrow because there is something specific about our technique that is doing a better job on highly relevant documents than on regular relevant documents?

Stepping back for a moment, the larger question that we are trying to answer is how one would compare two methods against each other in terms of their ability to find highly relevant documents, when what they are retrieving is relevant documents. The confounding factor is that one method may retrieve more highly relevant documents simply because it retrieves more total relevant documents for the same level of effort. So is that technique doing better than the other because it is better at highly relevant documents, or because it is better at all relevant documents?

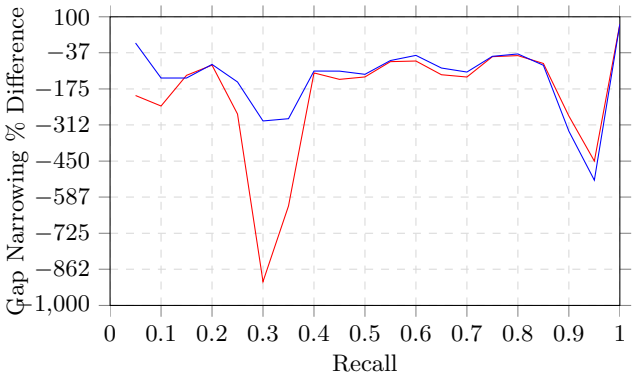

Figure 3: % Difference in the size of the gap between highly relevant and regular relevant runs, for our approach versus the BMI baseline title-only (red), and our approach versus BMI baseline title+description (blue)

At one level, this question doesn't matter, as a technique that retrieves fewer relevant documents requires the reviewer to expend more effort. And avoiding that additional effort is of primary concern. However, it would also be interesting to see from a purely exploratory standpoint how two highly relevant gain curves would appear if all non-relevant documents were removed and only relevant and highly relevant documents remained. In that manner, the highly relevant document ordering could be compared more directly. Again, this would be more of an exploratory comparison, but could yield interesting insights.

Another approach to the evaluation of highly relevant document finding would be to run systems in which training is done either purely on the highly relevant documents, or in which the highly relevant documents are given a larger weight than the regular relevant documents when they are encountered during review. This becomes especially pertinent to the main focus of this paper, which is the effect of manual seeding on the process. At some point we would like to be able to see not only whether reviewers are able to more quickly or easily find highly relevant documents, but whether that makes a difference to the underlying machine learning algorithms supporting the review.

2.2 Call Your Shot

Our main focus in this paper was not on calling our shot, but on exploring the effects of manual seeding. Nevertheless, "call your shot" was part of the track, so we borrowed with attribution the basic intuition of [2] and implemented a quick, naive version thereof. [2] had noted that a reasonable stopping point seems to be when batch richness (total relevant / total reviewed) drops to about 1/10th of the high water mark. We implemented this by calling the following function periodically throughout the continuous learning review.

Not allowing the possibility of a positive value for the stopping condition until at least 500 documents have been reviewed is a hack, and is based on the fact that we have noticed through prior experience on TREC 2015 data that when *only manual seeds* are used in the initial iteration, early richness can fluctuate unpredictably on some topics. Whether this is a function of the reviewers choosing seeds in a biased manner, or whether it is a broader reflection of the nature of certain topics, we are not sure. But because

```
Data: R is a list of documents, in reviewed order
chunksize ← 250;
stop ← False;
if len(R) > 500 then
    for i ← 1 to len(R)-chunksize do
        window ← R[i..(i+chunksize)];
        rich ← numrel(window) / chunksize;
        if rich > maxrich then
            maxrich ← rich
        end
    end
    for i ← 1 to len(R)-chunksize do
        window ← R[i..(i+chunksize)];
        rich ← numrel(window) / chunksize;
        if (rich / maxrich) < 0.1 then
            stop ← True
        end
    end
Result: stop
```

of our our prior experience that the fluctuation could exist, we made a blanket decision to never stop before the 500th reviewed document. This of course affected our shot-calling performance on some of the topics for which there were only a few hundred, sometimes a few dozen, relevant documents. We also note that the size of the window (chunk size) over which we calculated richness was arbitrary and unoptimized; we did not investigate other settings, either prior to the run nor since.

Finally, we note that the code to calculate the stopping condition was a rushed last minute endeavor, put together in about half an hour. As such, it contains (at least) one glaring logical hole: The lowest to highest richness ratio is unordered. What should have been done is that only the more recent windows should have been emphasized, i.e. the ratio to be tested should have been the more recent richness window to the highest richness window.

Normally, that shouldn't be a problem, because overall richness is generally monotonically decreasing. However, in at least one case the opposite was true: Topic 403. The initial review started off moderately rich, then flattened for a longer period of time before rising sharply again. A a result, about halfway up that sharp ascent, the lowest richness

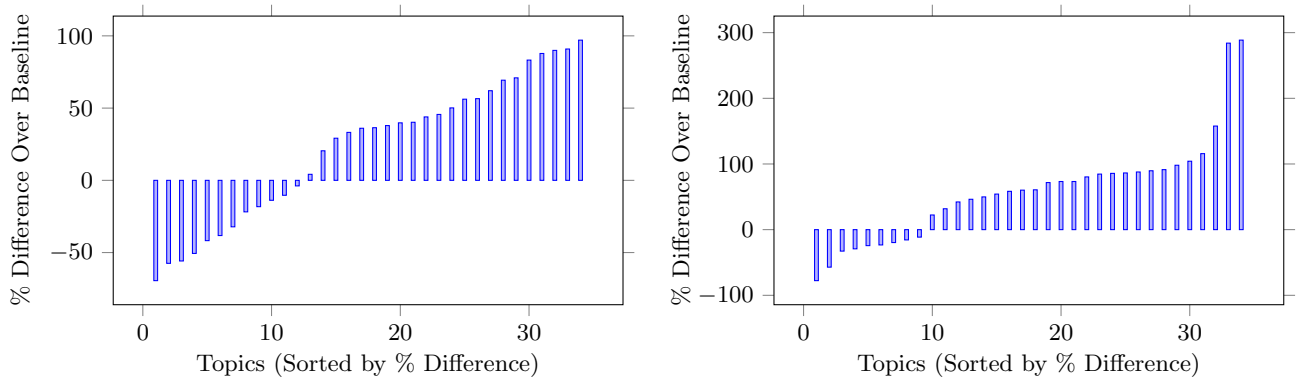

Figure 4: "Call Your Shot": Per topic histogram of F_1 score % change of our approach relative to the BMI baselines. Relevant documents (left), Highly relevant documents (right)

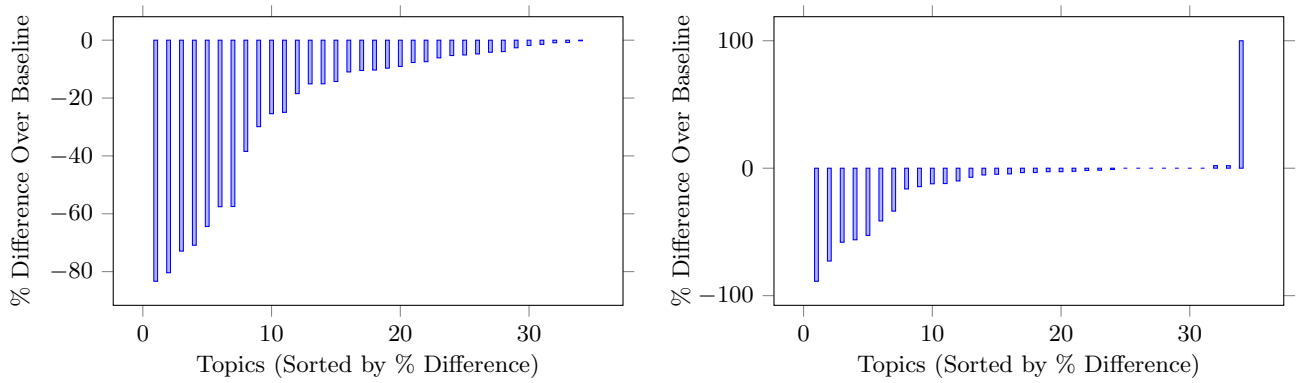

Figure 5: "Call Your Shot": Per topic histogram of *recall* score % change of our approach relative to the BMI baselines. Relevant documents (left), Highly relevant documents (right)

window (centered around document 400) became 1/10th as rich as the highest richness window (centered around document 900). Since the algorithm ignored window order, it then called the stopping point in the middle of that sharp rise, while documents were still being found at a very high rate (see Figure 8, Topic 403)

That boundary condition aside, Figure 4 shows that for most of the topics, and evaluating using F_1, our naive approach beat the baselines for both relevant and highly relevant documents. However, Figure 5 tells a different story: When it comes to pure recall (with no consideration of precision) the point at which our algorithm calls the shot is almost invariably at a lower recall point than the baseline. This of course raises the issue of what "reasonable" means. The task is a total recall task, so one would imagine that the higher recall point is the better evaluation metric. However, anyone can get higher recall, simply by continuing to review, i.e. calling a later stopping point. In the extreme, everyone – even those randomly reviewing documents – can always get 100% recall by reviewing the entire collection. This is not the intent of the task, because there is also a requirement to avoid wasted effort, i.e. to keep precision high.

F_1 is a traditionally common way to balance both precision and recall, and it was the one chosen by the Total Recall Track as an official metric, which is why we show it above. But it places equal weight on both recall and precision. So might it not be more reasonable, in a total recall task, to use F_2 or even F_3 instead? Or why not $F_{2.647}$? Is that not a more reasonable metric than $F_{2.883}$? Why or why not? The question still remains: What is reasonable, and how do we measure it?

Part of the difficulty is that the measurement of the quality of a stopping point algorithm is inextricably linked with the quality of the review ordering itself. Different review orderings are going to limit (or expand) the highest achievable stopping metric score of even the best stopping point algorithm. Perhaps future work from the community could separate out the review order quality issue from the stopping point issue, by creating stopping point test collections that standardize on fixed orderings.

3. MANUAL SEED EXPERIMENTS

The main subject of investigation for our TREC 2016 Total Recall run was what effect various manual initial selections (seedings) of documents has on overall task outcome. Section 1 explored the union of all four sets of seeds (two one-shot, two iterative), and in this section we explore each of the starting points, individually.

3.1 Manual Effort Statistics

Figure 6 contains statistics of the 34 athome4 topics. This data is presented in four parts: Query Count, Time Spent,

	Query Count				Time Spent (minutes)				Docs Reviewed				Review Overlap			
Topic	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	Unique	Sum	Ratio	TopicSize
athome401d	-	-	6	7	-	-	11	31	25	25	31	25	67	106	0.63	229
athome402d	3	-	5	-	16	-	7	-	23	24	33	25	85	105	0.81	638
athome403d	3	4	-	-	16	11	-	-	25	13	25	25	56	88	0.64	1090
athome404d	-	2	3	-	-	5	4	-	25	20	25	25	54	95	0.57	545
athome405d	-	-	5	5	-	-	7	26	25	25	29	25	62	104	0.60	122
athome406d	3	5	-	-	16	6	-	-	25	25	25	25	69	100	0.69	127
athome407d	4	3	-	-	16	7	-	-	25	16	25	25	42	91	0.46	1586
athome408d	-	5	-	7	-	11	-	32	19	25	25	25	60	94	0.64	116
athome409d	-	-	8	3	-	-	12	22	24	25	22	25	67	96	0.70	202
athome410d	4	-	3	-	15	-	6	-	25	25	35	25	91	110	0.83	1346
athome411d	4	2	-	-	16	4	-	-	24	22	22	25	41	93	0.44	89
athome412d	-	6	-	5	-	12	-	34	22	25	25	25	81	97	0.84	1410
athome413d	-	-	3	6	-	-	5	36	25	25	25	25	77	100	0.77	546
athome414d	1	-	3	-	16	-	5	-	25	25	27	25	59	102	0.58	839
athome415d	3	3	-	-	15	12	-	-	22	16	25	25	64	88	0.73	12106
athome416d	-	4	-	6	-	9	-	36	25	35	25	25	76	110	0.69	1446
athome417d	-	-	3	7	-	-	7	30	25	25	34	25	108	109	0.99	5931
athome418d	5	-	4	-	16	-	8	-	25	24	25	25	63	99	0.64	187
athome419d	2	3	-	-	16	9	-	-	26	15	25	25	50	91	0.55	1989
athome420d	-	3	-	4	-	9	-	25	25	33	25	25	66	108	0.61	737
athome421d	-	-	3	9	-	-	5	36	25	25	26	25	50	101	0.50	21
athome422d	-	-	3	-	-	-	5	-	25	25	32	25	75	107	0.70	31
athome423d	2	3	-	-	16	8	-	-	26	25	25	25	32	101	0.32	286
athome424d	2	2	-	4	16	7	-	28	25	20	25	25	60	95	0.63	497
athome425d	-	-	3	7	-	-	5	36	25	25	29	25	65	104	0.63	714
athome426d	3	-	5	-	15	-	6	-	25	25	32	25	54	107	0.50	120
athome427d	2	4	-	-	16	9	-	-	25	19	25	25	53	94	0.56	241
athome428d	-	1	-	6	-	7	-	32	25	24	25	25	72	99	0.73	464
athome429d	-	-	6	1	-	-	5	18	25	25	30	25	82	105	0.78	827
athome430d	3	-	4	-	16	-	5	-	25	24	28	25	88	102	0.86	991
athome431d	3	5	-	-	16	10	-	-	25	26	25	25	57	101	0.56	144
athome432d	-	2	-	4	-	5	-	27	25	23	25	25	51	98	0.52	140
athome433d	-	-	3	4	-	-	4	23	25	25	27	25	64	102	0.63	112
athome434d	4	-	7	-	16	-	11	-	25	25	29	25	45	104	0.43	38
average	3.0	3.4	4.3	5.3	15.8	8.3	6.6	29.5	23.6	24.6	26.9	25	64.3	100.2	0.64	1056

Figure 6: Manual Effort Statistics

Docs Reviewed, and Review Overlap. In the Query Count section, the number of queries that each reviewer (R1 through R4) issued for each topic is shown. For ease of reading, if the reviewer did a one-shot (single) query, that is indicated with a dash "-". If the reviewer did an interactive run, the actual number of queries is shown. The averages shown are averages of just the interactive runs; one-shot averages are 1.0, of course.

In the Time Spent section, the pattern is similar: dash indicates one-shot query, which likely look a minute or two but we did not record the exact amount of time each reviewer spent pondering his or her one query before issuing it. Actual numbers indicate time spent in an interactive run. The averages shown are averages of just the interactive runs.

In the third section (Docs Reviewed) the exact number of documents that the reviewer laid eyeballs on, both relevant and non-relevant, is indicated. The review software did not have explicit controls to stop a reviewer from going beyond 25 documents; this was left up to the individual reviewer. So as per normal human variance the count is sometimes a few docs over, sometimes a few docs under, but generally around the 25 document mark. The averages shown for Docs Reviewed are averages of all runs, both one-shot and iterative.

The final section is Review Overlap. Since the reviewers worked with no knowledge of each other, it was often the case that (even when issuing different queries) they reviewed some of the same documents. Therefore, we show statistics on not only the total number of documents reviewed, but the total number of unique documents reviewed. The ratio of unique to total is also shown, as is the log of the size of the topic (total number of available relevant documents for that topic). The averages shown are averages of all runs, both one-shot and iterative.

When examining these values, we noticed an interesting, though possibly spurious, relationship between uniqueness ratio and topic size (number of relevant documents available for that topic). Figure 7 is a plot of the relationship between the ratio of unique to total documents reviewed (x-axis) and the (log of the) size of the topic (y-axis). A line is fitted to this data, and seems to suggest that the more total available relevant documents there are to be found in the collection, the less overlap there was between what the four reviewers found.

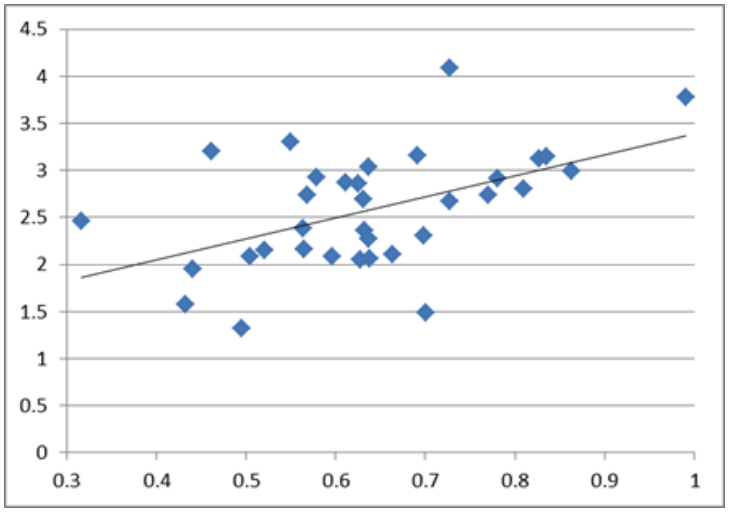

Figure 7: Log Topic Size (y-axis) against Ratio of Multi-Reviewer Unique to Total Seed Count (x-axis)

3.2 Individual Topic Runs

3.2.1 Individual Gain Curves

Gain curves for the 34 topics are shown in Figures 8 through 14 (left). The reviewer ID is indicated with a color: Red, blue, green, and brown for reviewers 1 through 4, respectively. If the reviewer ran that topic as a one-shot seeding process, the gain curve is a solid line. If the reviewer ran it iteratively, the line is dotted.

We acknowledge that there is an overabundance of graphs in this paper. The reason we chose to show them all in their entirety, rather than just show some sort of summary statistic such as precision@90% recall or average area under the gain curve is that this gives us a chance to observe the fine differences between topics and reviewers. These differences can be nuanced, which is both the strength and the weakness of presenting research results in this manner. It's a weakness, because it makes the results more difficult to summarize. It's a strength, because it lets one see where and how distinctions can arise. Furthermore, this is a TREC paper, rather than peer-reviewed scientific conference or journal paper, and we the authors feel it is more valuable and in keeping with the spirit of openness around TREC to show as much detail about our runs as possible. Displaying every graph allows us to do that.

That said, one general observation is that, no matter who the reviewer doing the initial seeding was, or what method they employed, high recall is achieved by almost every reviewer on almost every topic without having to review the vast majority of the collection. There remains a belief among many industry practitioners engaged in high recall tasks that only experts may select seed documents, that only experts have the capability to initiate a recall-oriented search by selecting the initial training documents. In our experiments, only one of our reviewers qualified as a practicing search expert: Reviewer 1 (red). Nevertheless, for 132 of the 136 starting points in Figures 8 through 14, all starting points hit high recall within a relatively similar amount of effort. For example, on Topic 405, Reviewer 2 (blue) gets to 90% recall a little faster, and Reviewer 1 (red) gets there a little later, with the other two reviewers somewhere in the middle. But the difference between the "best" and "worst" is literally 68 documents. Out of a collection of 290,000 documents, that is an insignificant difference. Other topics, such as Topic 411, show a bit of a back and forth as the review progresses between the various starting points. But all achieve high recall at about the same point.

The four starting points where there is a significant difference between best and worst were Reviewer 4 (brown) on Topic 415, Reviewers 2 and 3 (blue and green) on topic 418 and Reviewer 3 on Topic 419. Of those, 3 of the 4 were all one-shot queries. That is, the reviewer did not do any of the review of the collection before issuing his single query, did not receive any feedback on his or her single query, and did not issue more than one query. The final underperforming starting point was done by an iterative reviewer (4 queries in 8 minutes, as per Figure 6), but is still the exception rather than the rule.

There is one more data point worth noting in the gain curves on the left of each figure. The black curve is based on pooling all the unique seed documents from each of the four reviewers before running a continuous learning review. In the majority of the cases, the pooled seed starting point is at least as good as, if not better, than the best individual starting point. However, in a number of instances, the pooled seeds yield a result that is equal to the worst individual case, and in rare instances, even worse than the worst individual case. Given the casualness of this TREC paper, we don't have a formal metric, a concrete quantification of "better", "equal", "worse", etc. Rather, we did a casual eyeballing of the curves and looked at where the combined seeding came out generally, often, as per the individual seedings themselves, by the time the process hit high (85%-95%) recall, all the methods converged, anyway. So in this particular exploratory analysis, we are often looking at differences at lower levels of recall. Nevertheless, given that we're displaying all curves, the reader can see and judge for themselves whether or not significant differences exist.

The following table is a rough count of the number of times that the combined seeding approach was better than the best individual, approximately equal to the best individual, somewhere in the middle of all individuals, equal to the worst individual, or worse than the worse individual seeding run. For the most part, the pooled seeds tended to

be equal to or better than the best individual. However, it would be worth exploring those cases in which the pooled seeds do worse, and try to determine why. That analysis is beyond the scope of this paper. One more thing to keep in mind: Even when the pooled approach is better than the best, or worse than the worst, the absolute magnitude of the differences, especially relative to the size of the collection, are small.

Combined Seeding	Count
Better than the Best Individual	5
Equal to the Best Individual	16
In the Middle of the Individual	8
Equal to the Worst Individual	3
Lower than the Worst Individual	2

3.2.2 *Seeding Method-Averaged Gain Curves*

Lastly, for each topic in Figures 8 through 14, we show the average of the two one-shot seeding approaches, as well as the average of the two iterative seedings, in the charts on the right side. Perhaps the better approach would have been to pool the iterative and the one shot seeds, respectively, before running CAL on the combined seed pools. For now, however, we show the average of the gain curves of the individually seeded runs. There is variability among individual reviewers, so by averaging multiple runs and randomizing reviewers across topics, we hope to get a better sense of the the general approach, separate from the individual reviewer vagaries. Ideally we would have more than two reviewers doing each method, but resources are always limited.

Basically, we can see that for the most part, there is not much difference between the two approaches for most topics. The iterative approach may have a slight edge on Topics 415, 418, 419, 421, and 422. However, the one shot approach has a slight edge on topics 402, 411, 416, 428, and 433. On the remainder of the topics, where there are differences, those differences are slight.

4. GROUND TRUTH ANALYSIS

There is one final analysis of the data that we would like to present. It is perhaps a bit non-standard, and we do not offer any hard conclusions. But it was analysis that we found interesting so we would like to show it to provoke future thought.

4.1 Explanation of the Analysis

Our understanding of how the ground truth was created for this track was that the NIST assessors used a combination of methods, their own searches plus an algorithm that used the same underlying feature extraction mechanism as the baseline model implementation (BMI). The collection was not fully judged for each topic, but was judged to a depth proportionally deeper than the number of relevant documents that were found. So for example, for Topic 422, NIST assessors judged 31 relevant documents, and 317 non-relevant docs. No other documents were assessed. For Topic 423, 286 relevant documents and 1113 non-relevant documents were judged. No other documents were assessed.

As per common TREC practice, documents that are not assessed (non-judged) are presumed to be non-relevant, and treated as such for both training and evaluation purposes. In this last section, we wish to separate out, for the purpose of deeper analysis, judged non-relevant documents from non-judged documents. To this end we present Figures 15 through 21. For each topic, the figure on the left side shows recall on the x-axis, and the raw number of judged non-relevant documents on the y-axis. The blue line represents the number of judged non-relevant documents to that depth in the review unique to our pooled seeds method (see previous section), the red line shows the same information unique to the baseline (BMI) title+description method, and the dotted grey line is the number of judged non-relevant documents that both methods have in common. We only plot to 90% recall for all topics, as high non-judged counts can skew the visualization after that point.

So for example, see Topic 421 in Figure 19. Let's start with the figure on the left, which shows judged non-relevant documents on the y-axis. By the time that the baseline method (red) has hit 10% recall, it has seen 3 judged non-relevant documents, while our method (blue) has hit 0 judged non-relevant documents, and none of those documents are the same documents. At 75% recall, the baseline method has seen 28 judged non-relevant documents, while our method has seen 16 judged non-relevant documents. However, 13 of those documents are in common between the two methods. So at 75% recall, the baseline method has seen 15 judged non-relevant documents that our method has not (unique to BMI), and our method has seen about 3 judged non-relevant documents that the baseline method has not (unique to our method). In comparison, see the figure on the right, which shows non-judged documents on the y-axis. At 75% recall, there are only 5 documents that the baseline method has seen that are have not been judged, but 27 documents that our method has seen that have not been judged. None of these documents are in common, as the dotted grey line only starts to rise after about 82% recall. What this means is that, at 75% recall, the baseline method has only hit 15 judged + 5 non-judged = 20 non-relevant documents, but our method has hit 3 judged + 27 non-judged = 30 non-relevant documents.

From an overall evaluation standpoint, this means that (at 75% recall) the baseline method is better than our method, because it has hit fewer non-relevant documents. But when the majority of the documents in that comparison were judged for one method, and not judged for the other method, it raises questions about how things might be different if some of the non-judged documents had been judged. Would there be more relevant documents in those non-judged documents? Would there be more relevant documents in the non-judged documents unique to the baseline method, or unique to our method? And how would that affect overall recall and stopping points, not to mention training, especially when a half dozen newly found relevant documents could have significant effect in such low prevalence topics.

We did some casual, non-comprehensive spot checks on some of the topics by looking at the top 20 highest ranked documents that were unique to each method (i.e. 40 docs per topic in total). And we did find a fairly significant number of (what we thought were) relevant documents for some topics, almost none for other topics, at least within those first 20 documents.

However, we are not going to go into detail about how many additional relevant documents we believe that there were, for four reasons: (1) We did not do a full assessment of every topic, so any information we do present would be misleading, (2) We are not the same assessors as the NIST

assessors, and unless we were to go back and also review all the same documents that the NIST assessors reviewed, any assessment would be a skewed or biased by the disjointedness of the assessment, (3) Even if we did do a full reassessment of all judged and top-ranked non-judged documents, it would be unfair to the baseline method, because that method would not have had a chance to train on any newly-judged relevant documents, and finally (4) It is not within the spirit of TREC to publish research whose only goal it is to "beat" other systems. Rather, the purpose of TREC is to dive deep in to interesting questions, to challenge assumptions, to learn by trying crazy, unproven methods, to basically poke and prod a problem, and see what happens. This research hopefully accomplishes that by comparing multiple reviewers doing multiple approaches to seeding (one-shot query, iterative querying). To understand what ground truth data is being used and how that might affect things is a side goal, but not the primary one.

4.2 Discussion

Nevertheless, to understand the full context of this work, we felt it necessary to break out the analysis of the results into these two components: judged non-relevant and non-judged. Without even knowing whether the non-judged documents were truly relevant or non-relevant, some interesting patterns emerge. The first patterns is one exemplified by the topic that we already discussed above, Topic 421. In this pattern, our method has higher precision (lower number of non-relevant documents) on the judged set (left graph), but lower precision (higher number of non-relevant documents) on the non-judged set. This is a pattern seen in 10 other (11 total) instances: Topics 405, 410, 415, 422, 423, 425, 431, 432, 433, and 434. The next pattern is topics for which both methods find judged non-relevant documents at about the same rate, but our method hits a lot more non-judged documents. This pattern is found in 7 instances: Topics 401, 402, 406, 412, 418, 426, and 429. The remaining 16 topics are ones for which our method hits more non-relevant documents, both judged and non-judged, than does the baseline method.

How does one interpret this? Why is it that the two methods are finding, at times, vastly different non-relevant (judged or non-judged documents, while finding the same number of relevant documents. Or more specifically: For a large number of topics, why does our method find more *non-judged* documents, even as it is finding fewer judged non-relevant documents, at the same level of recall. By itself, finding more non-judged documents is not difficult: One can simply select documents at random. But this isn't a random selection of documents, because our method is finding relevant documents at a reasonably fast clip, while sometimes simultaneously finding fewer judged non-relevant documents at the same level of effort.

It's also interesting to note that for a fairly large number of topics, the baseline method finds almost no non-judged documents at all. Almost all non-relevant documents that it finds through the course of the review are ones that have already been judged. For example, see Topics 401, 402, 403, 404, 406, 407, 408, 409, 416, 417, 418, 419, 427, 428, 433, and 434, which are almost half the topics in the track. We are not certain why the unique documents found by one method are almost thoroughly judged while the unique documents found by the other are not. There may be something interesting in the way in which our method is working that is more naturally diverse (even without explicit diversification activated as explained in Section 1.2) relative to the baseline. It may mean that there was an aspect or facet of relevance that was found by our method that was not found during the ground truth assessment. Different doesn't necessarily mean better, however, as the non-judged documents are not necessarily going to be relevant, were they to be judged. We cannot answer this question now; we simply wish to show that there does seem to be consistent patterns of judged versus non-judged documents in the non-relevant set. This is a good opening into future work on Total Recall.

5. CONCLUSION

In conclusion, this paper examined the effect of multiple manual approaches to seeding using four different reviewers applying one of two different manual seeding strategies (one-shot vs iterative). For over 97% (132 of the 136) seedings, by the time high recall was hit, there was relatively little difference between the starting points no matter the method. When averaged across strategy (one-shot vs iterative), five topics slightly favored the iterative approach, five slightly favored the one-shot approach, and the remainder of the topics came out about the same.

Perhaps one of the challenges is that the iteration was brief; reviewers were only allowed to work until they had marked up to 25 documents. With more time or more queries, perhaps a larger difference could have been observable. On the other hand, many of these topics were relatively straightforward, and perhaps no differences and improvements via manual efforts may be possible. Nevertheless, we note that all topics achieved high recall without having to review the vast majority of the collection, no matter if an expert or a non-expert was used to manually seed each topic.

We also noted a possible relationship between reviewer overlap and topic size. Where different reviewers manually find many of the same documents, the topic may have a smaller number of documents, and vice versa when different reviewers manually find many different documents. Whether such an approach could be formalized enough to be broadly predictive remains an open question.

Additionally, we did some analysis of the ground truth itself, and examined the relative difference between our method and the baseline method in terms of how many judged non-relevant versus non-judged documents each found over the course of each topic's review. The visualization of these differences are interesting, but anything conclusive at this point would be pure speculation.

6. REFERENCES

[1] G. V. Cormack and M. R. Grossman. Evaluation of machine learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the ACM SIGIR Conference, Gold Coast, Australia, 6-11 July 2014*, Gold Coast, Australia, 2014.

[2] G. V. Cormack and M. R. Grossman. Waterloo (cormack) participation in the trec 2015 total recall track. In *Proceedings of Text REtrieval Conference (TREC)*, Gaithersburg, MD, 2015.

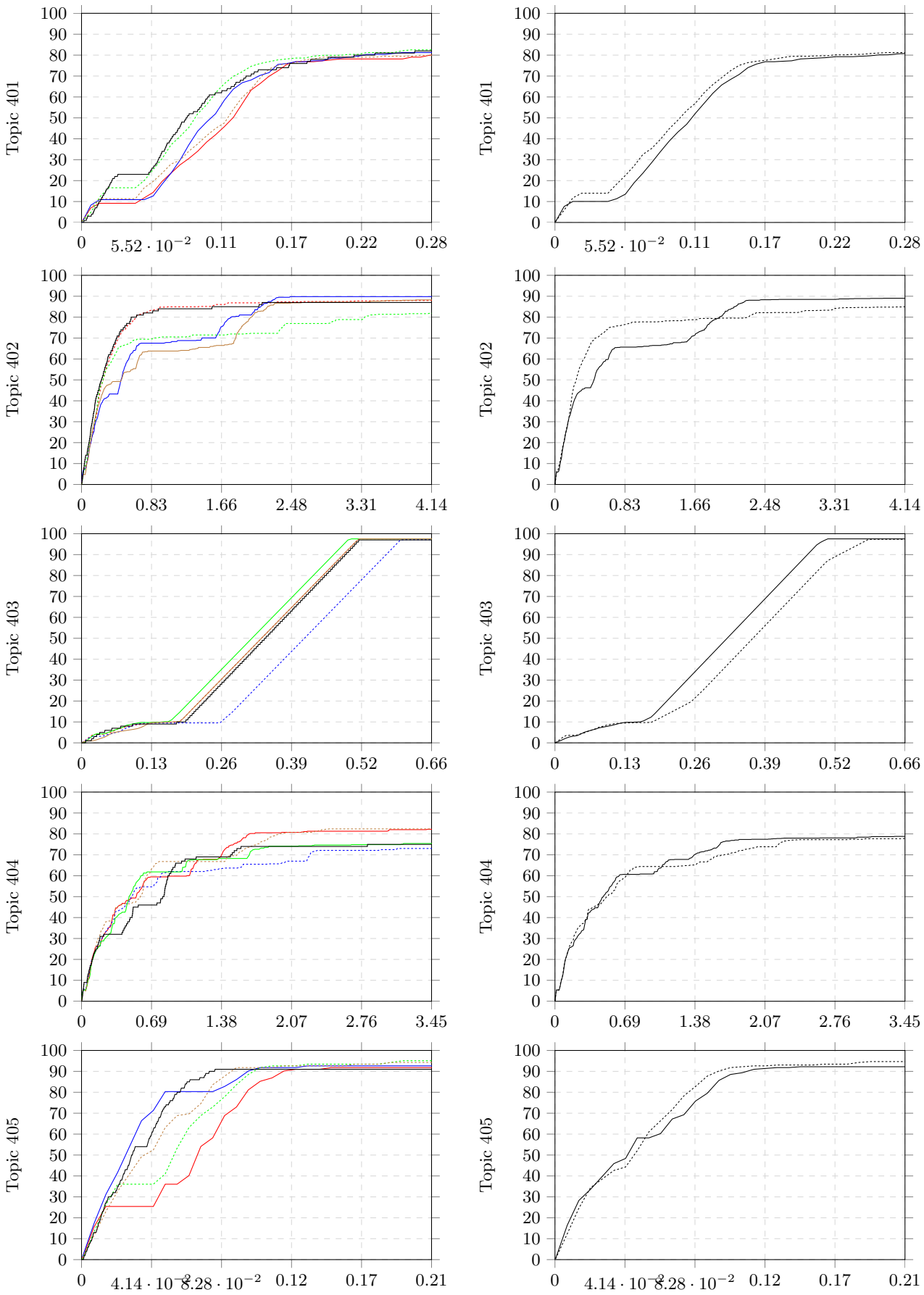

Figure 8: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

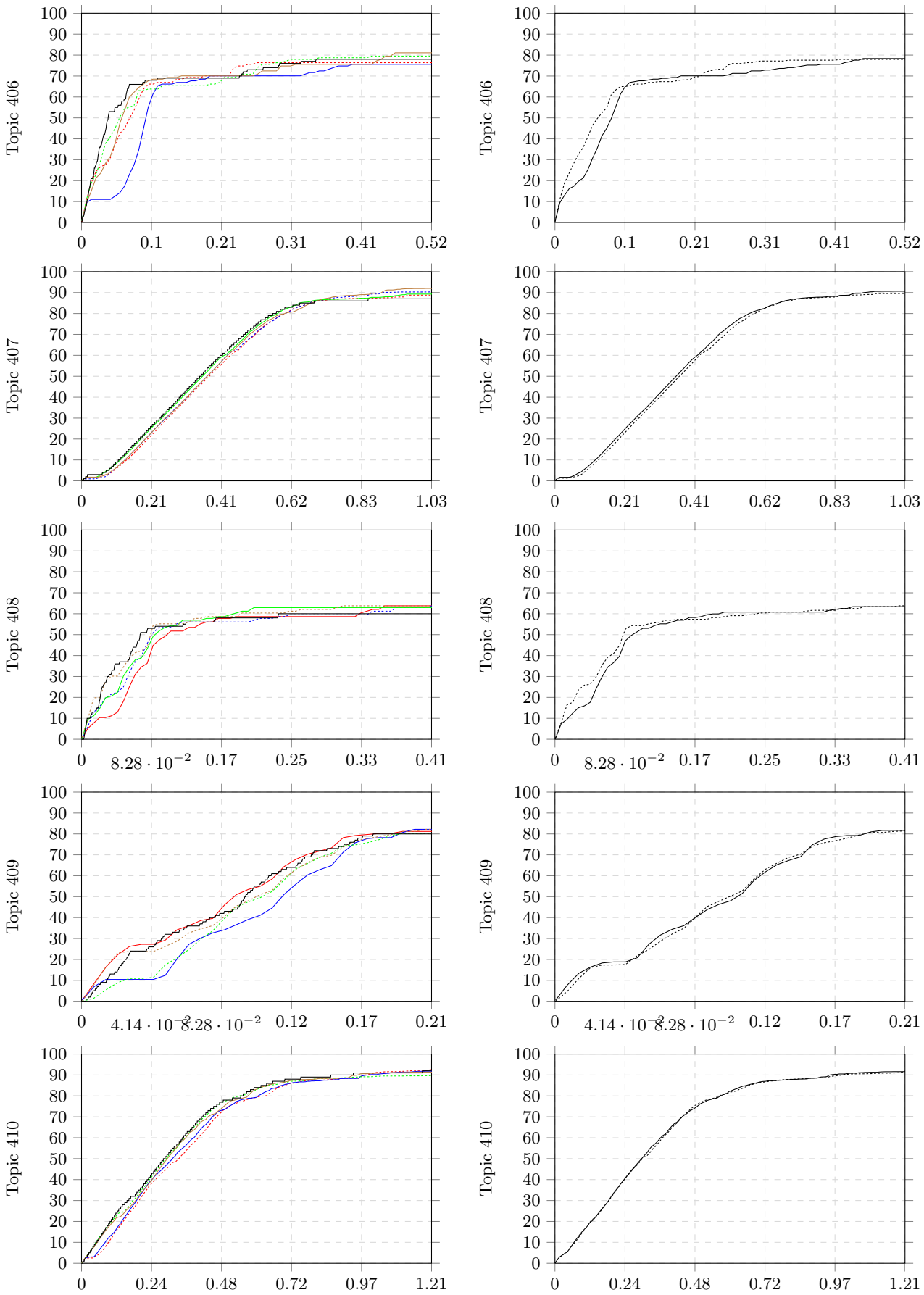

Figure 9: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

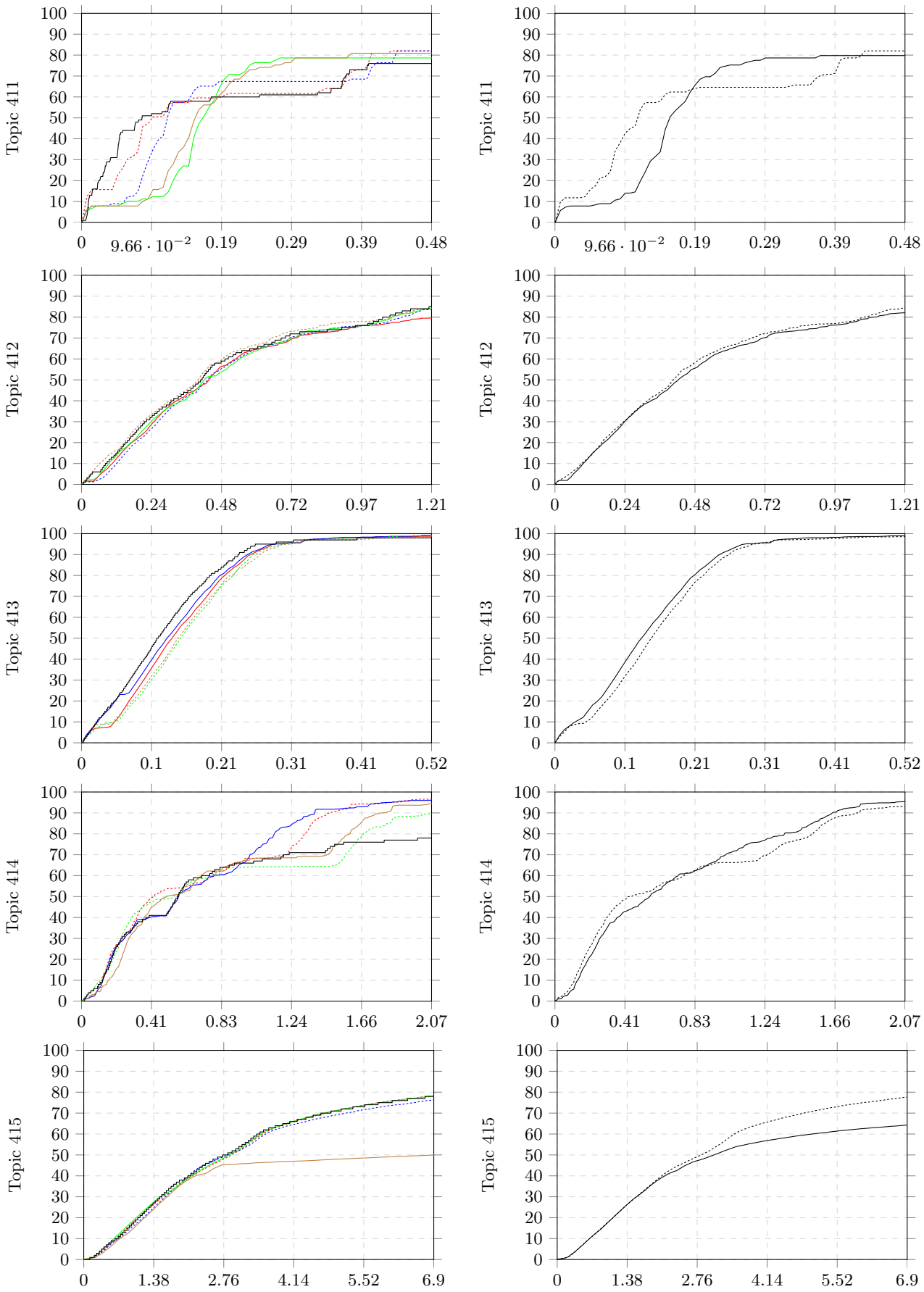

Figure 10: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

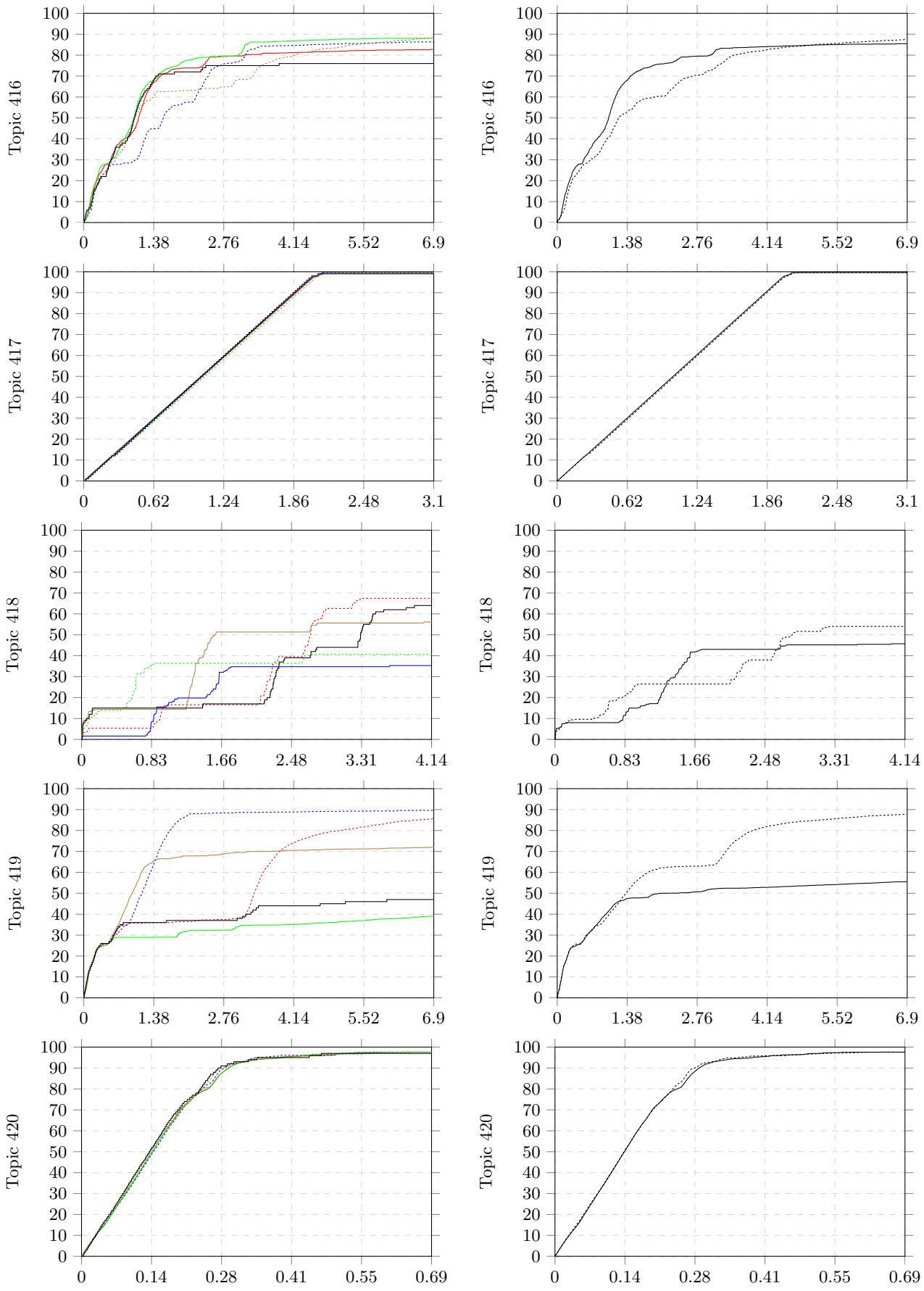

Figure 11: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

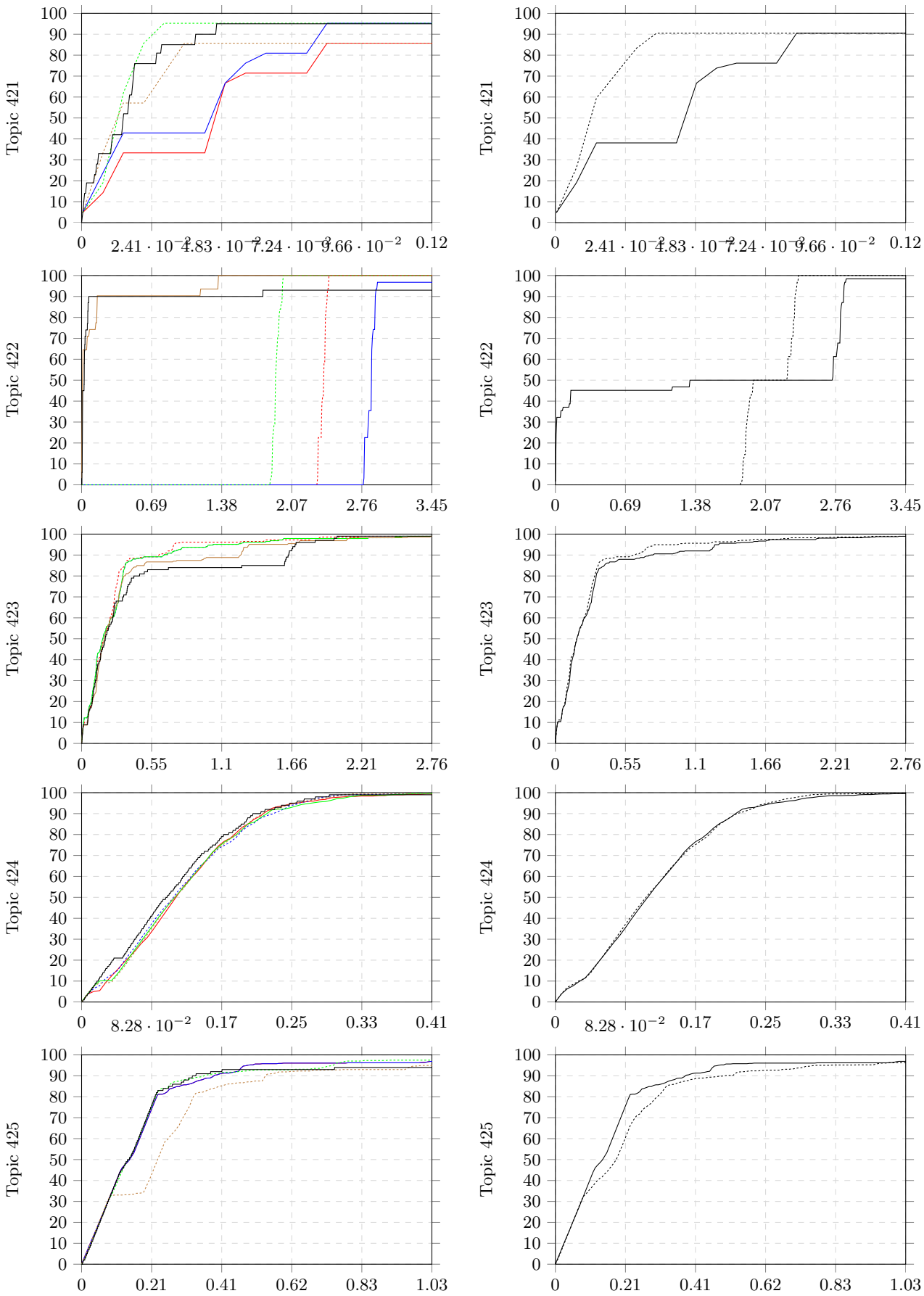

Figure 12: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

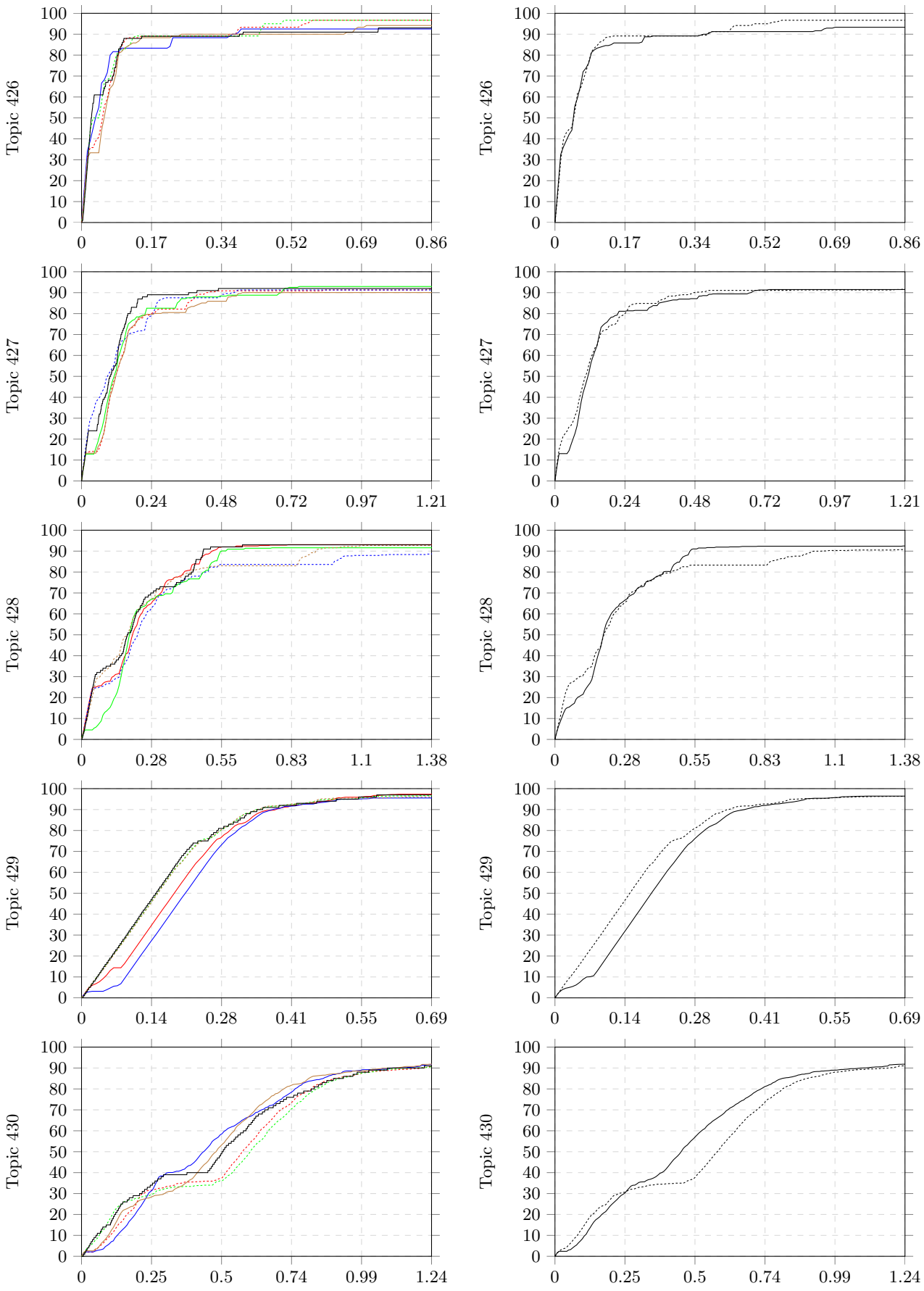

Figure 13: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

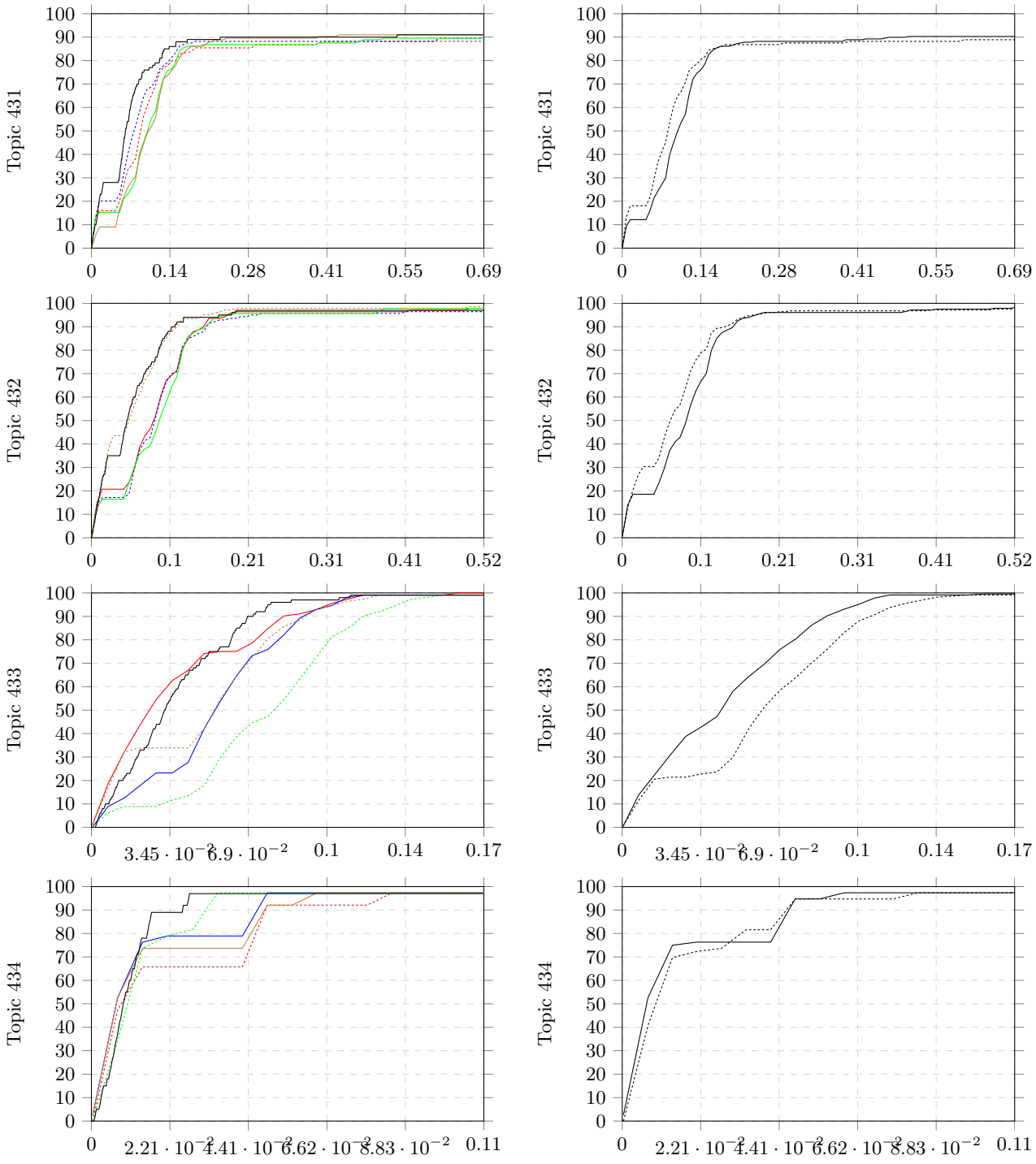

Figure 14: Gain Curves. x-axis = reviewed documents (in order), as a percentage of the entire collection. y-axis = recall. [Left] Red = Reviewer 1, Blue = Reviewer 2, Green = Reviewer 3, Brown = Reviewer 4, Black = Pooled seeds from all reviewers. For individual reviewers, solid line indicates one-shot query; dashed line indicates iterative searching. [Right] Solid line is the average of the one-shot reviewers; dashed is the average of iterative reviewers

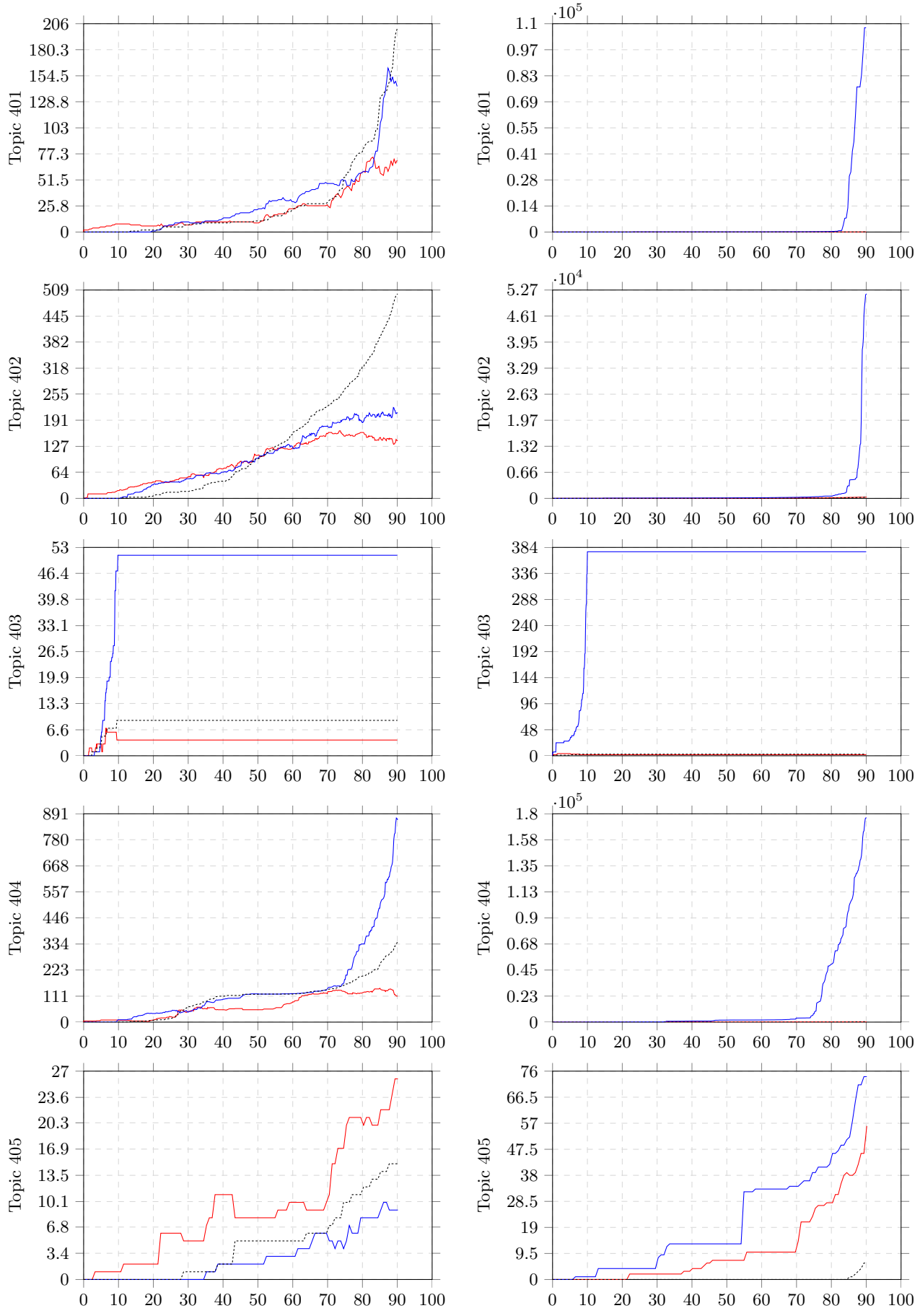

Figure 15: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

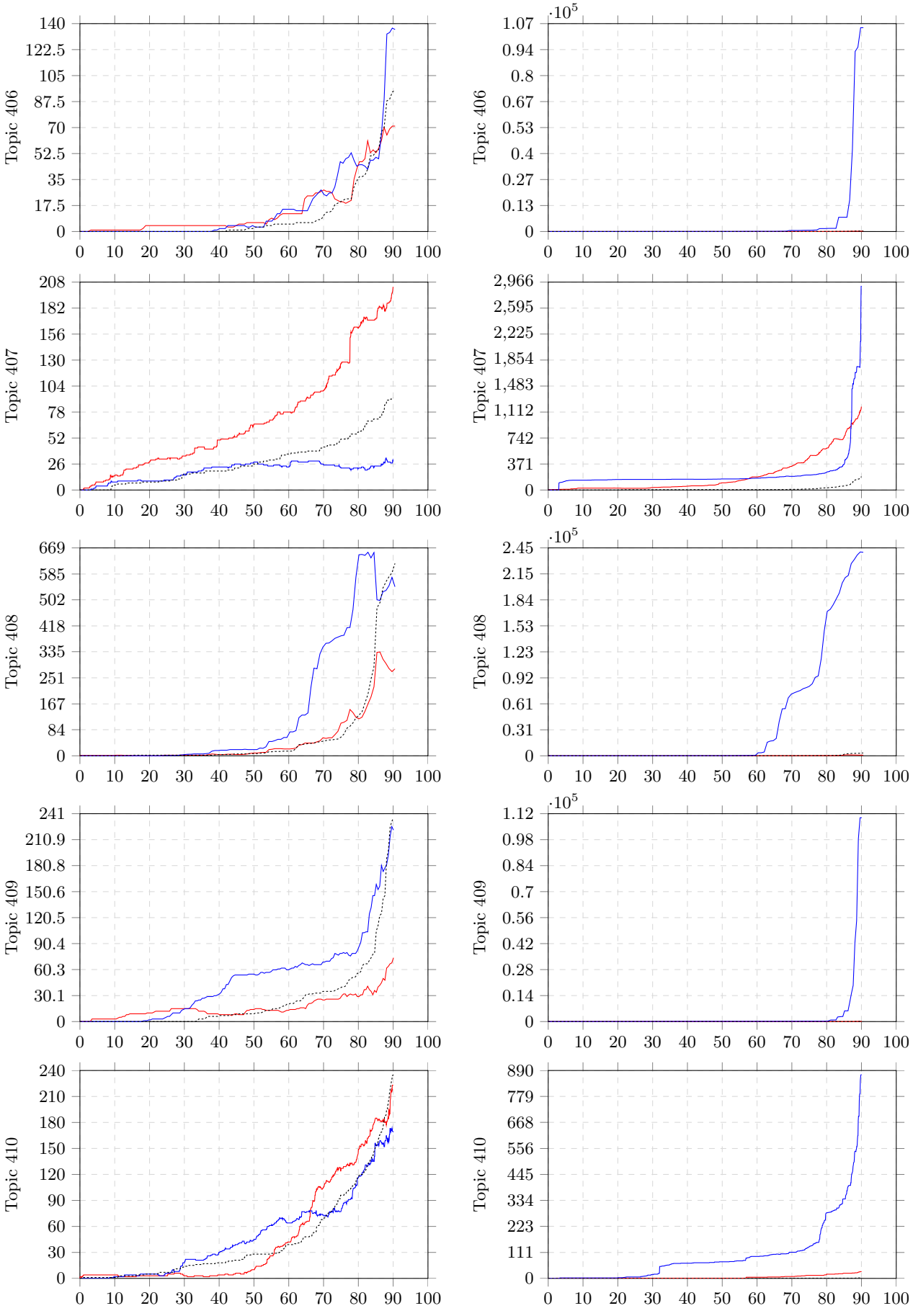

Figure 16: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

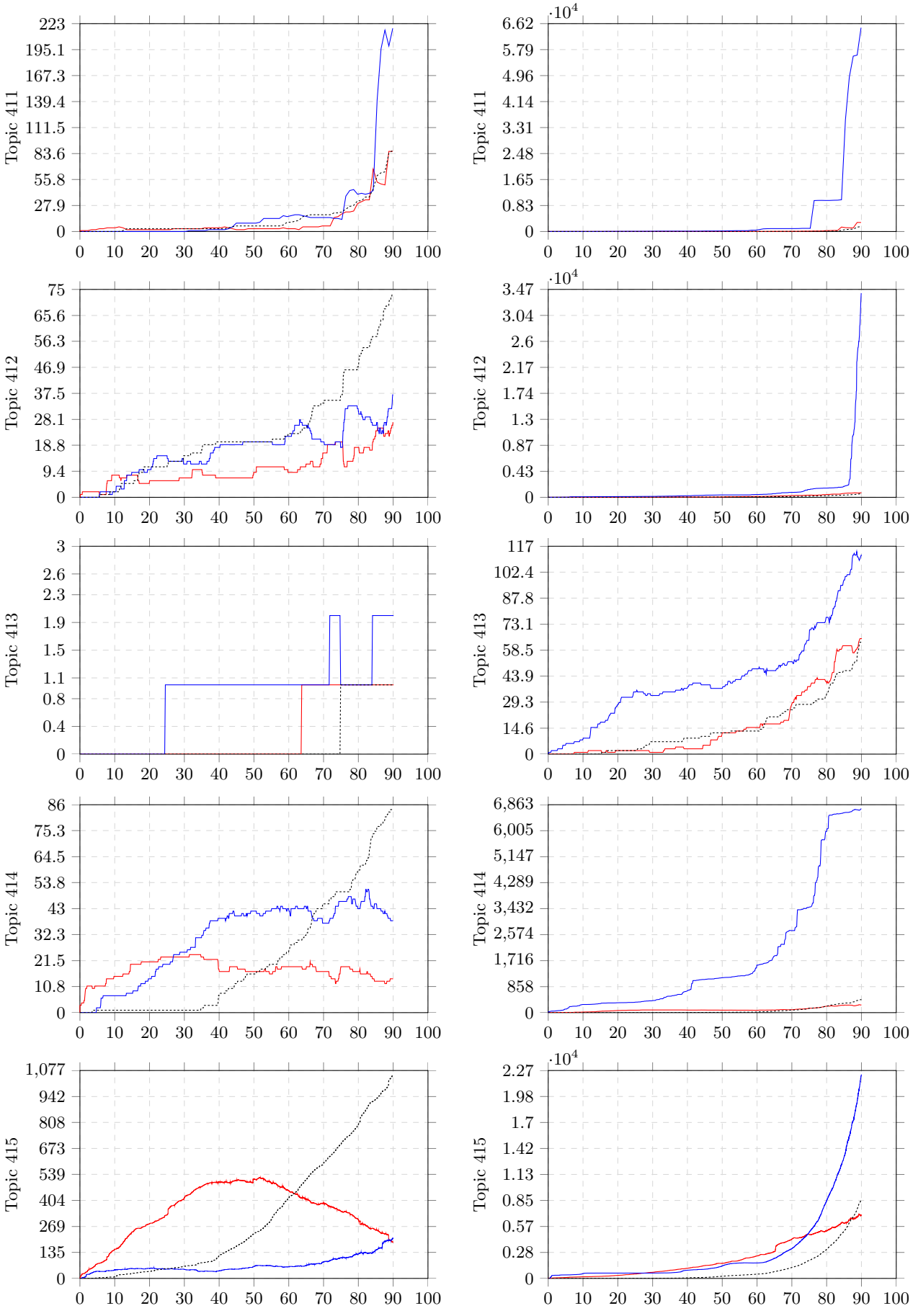

Figure 17: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

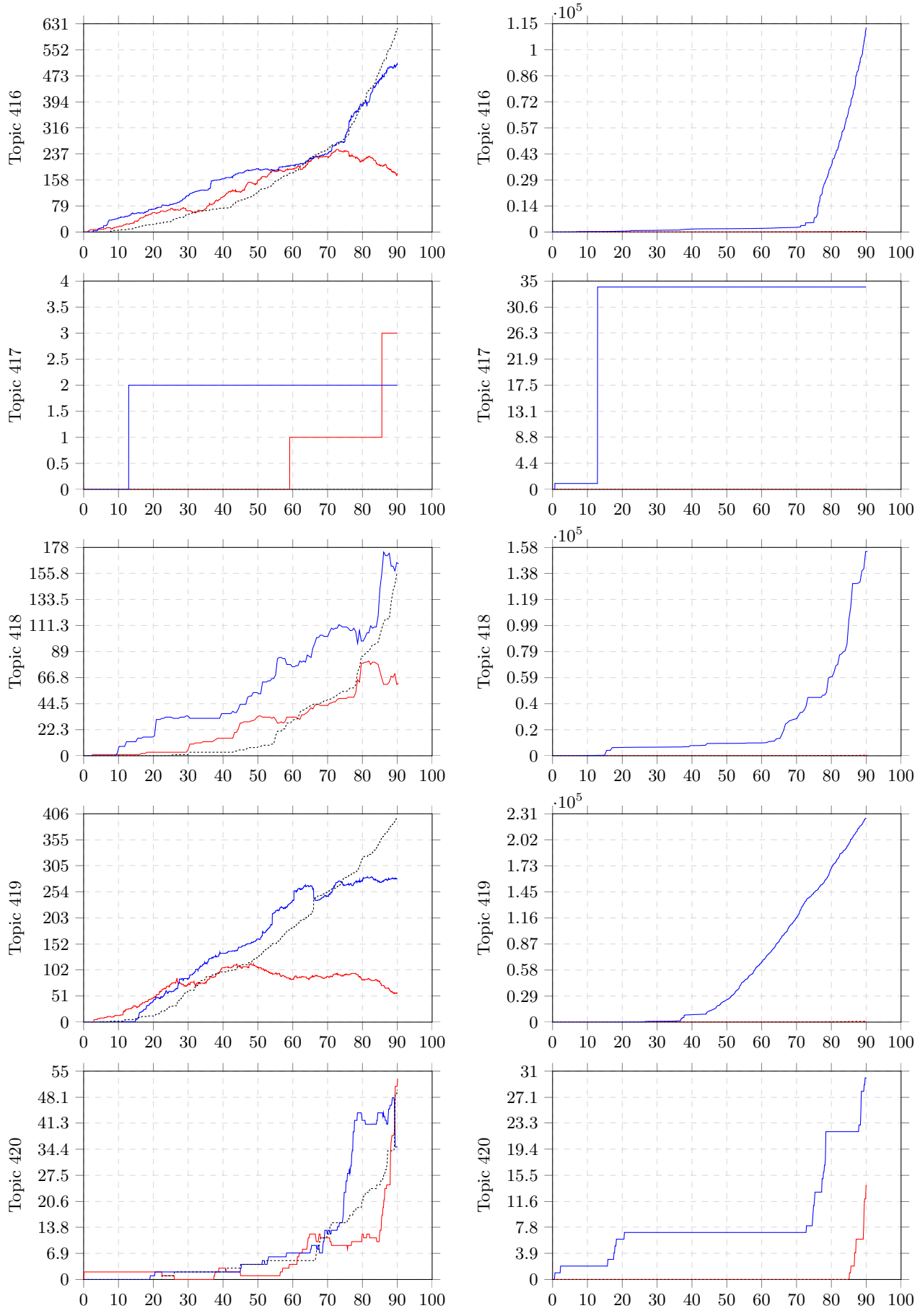

Figure 18: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

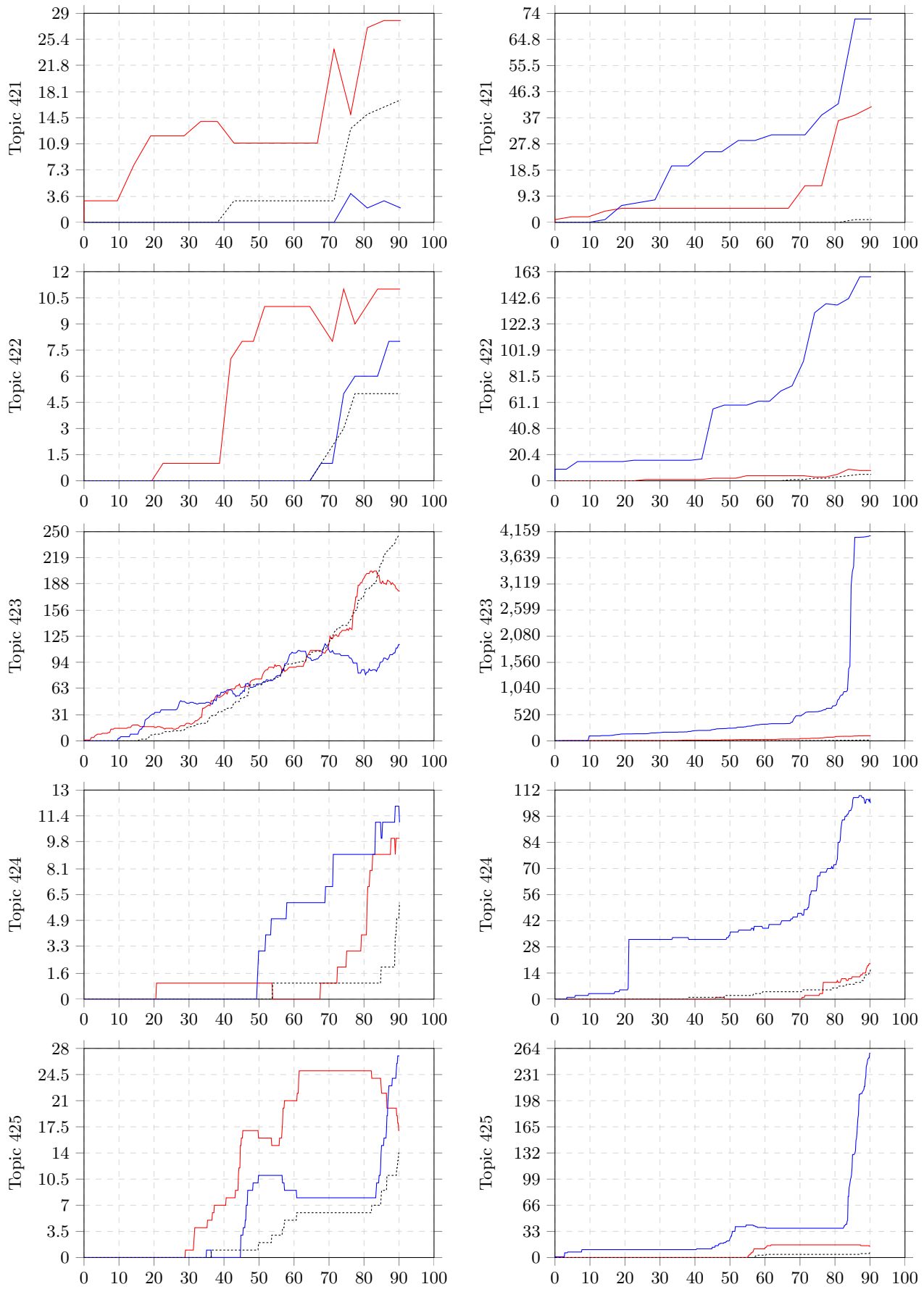

Figure 19: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

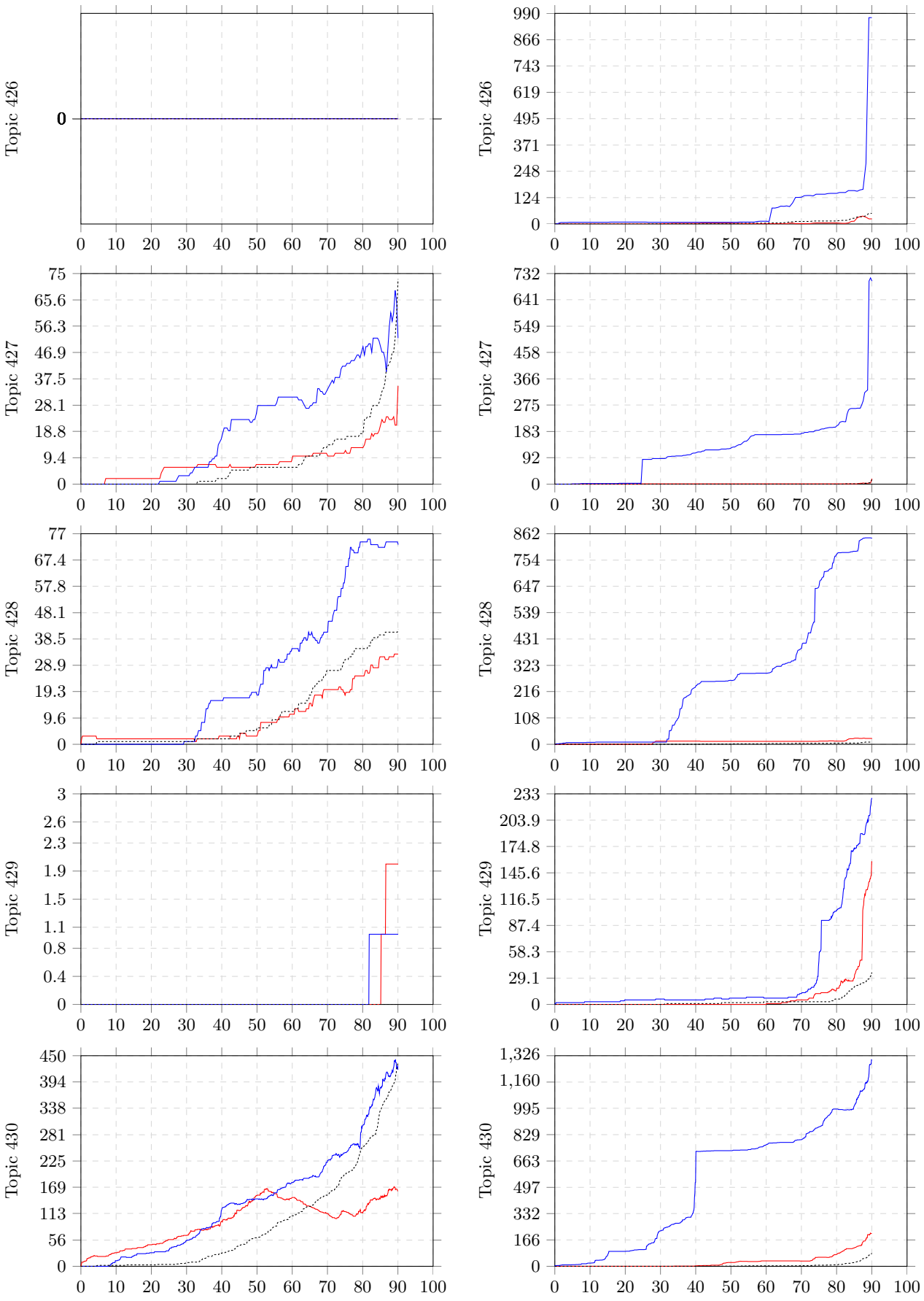

Figure 20: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.

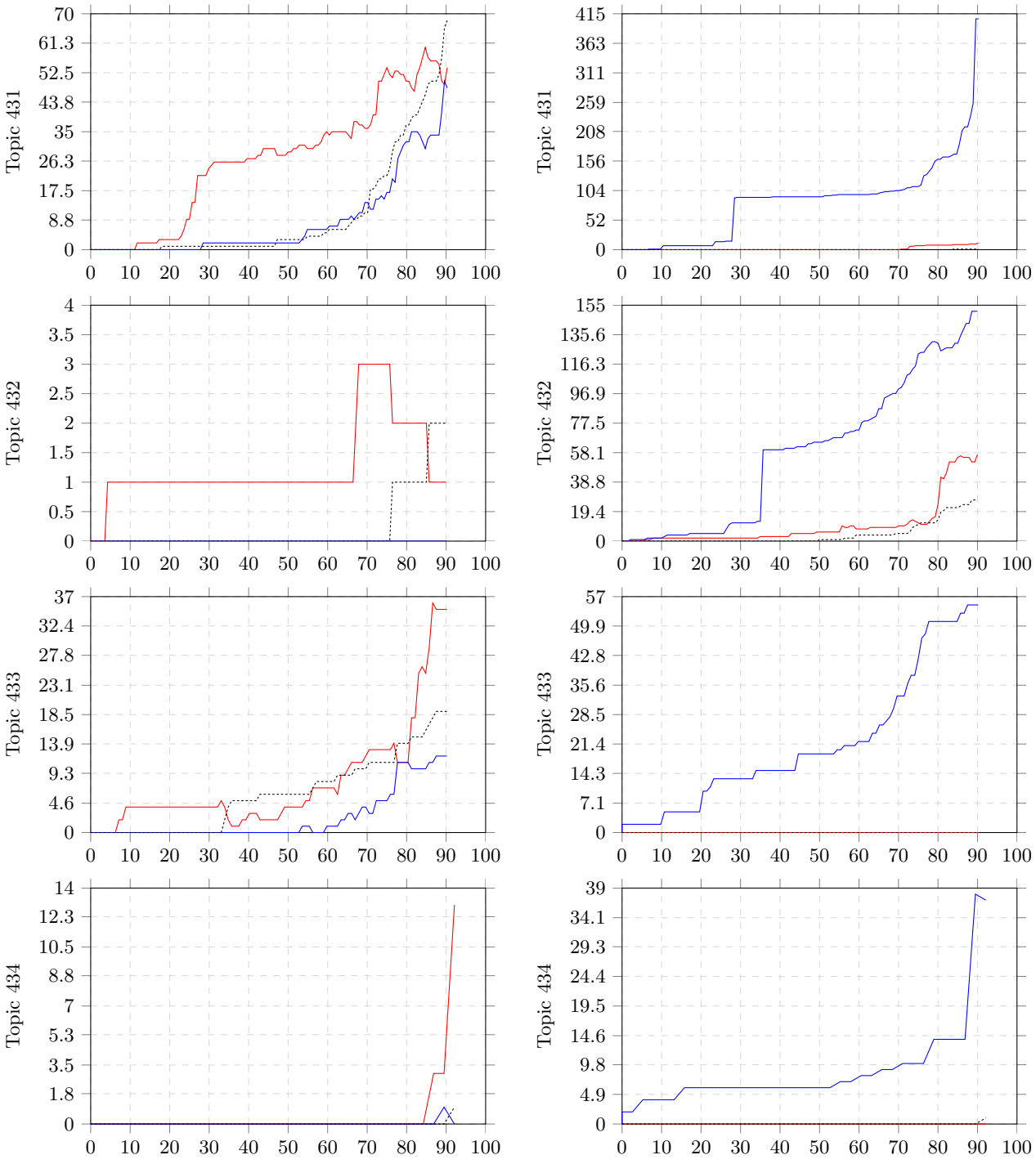

Figure 21: Analysis of Judged and Non-Judged Non-Relevant Documents. On both graphs, x-axis is recall level. y-axis is Judged Non-Relevant (left) and Non-Judged Non-Relevant (right). Red = documents unique to baseline method, blue = documents unique to our method, grey = documents common to both methods.