On the Quality of the TREC-COVID IR Test Collections

Ellen M. Voorhees
ellen.voorhees@nist.gov
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Kirk Roberts
kirk.roberts@uth.tmc.edu
The University of Texas Health Science Center at Houston
Houston, Texas, USA

ABSTRACT

Shared text collections continue to be vital infrastructure for IR research. The COVID-19 pandemic offered an opportunity to create a test collection that captured the rapidly changing information space during a pandemic, and the TREC-COVID effort was created to build such a collection using the TREC framework. This paper examines the quality of the resulting TREC-COVID test collections, and in doing so, offers a critique of the state-of-the-art in building reusable IR test collections. The largest of the collections–called 'TREC-COVID Complete'–is found to be on par with previous TREC ad hoc collections with existing quality tests uncovering no apparent problems. Yet the lack of any way to definitively demonstrate the collection's quality and its violation of previously used quality heuristics suggest much work remains to be done to understand the factors affecting collection quality.

CCS CONCEPTS

• **Information systems** → **Test collections**.

KEYWORDS

datasets, test collections, TREC

ACM Reference Format:
Ellen M. Voorhees and Kirk Roberts. 2021. On the Quality of the TREC-COVID IR Test Collections. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463244

The COVID-19 pandemic offered an opportunity to create an IR test collection that captured the search needs of clinicians and biomedical researchers during a pandemic, thereby providing the infrastructure to improve search systems for future public health crises. TREC-COVID was a community evaluation that used the TREC framework to build a set of test collections in a series of iterations. Taken together, this set of collections forms a general-purpose ad hoc test collection that we call the TREC-COVID Complete collection. Earlier papers describe the motivation for and mechanics of building TREC-COVID Complete [5, 6, 11]. This paper focuses on examining the quality of the collection, and offers a critique of the state-of-the-art in building large, reusable IR test collections.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463244

We examine the quality of TREC-COVID Complete using two tests of an IR collection's quality. We assess reusability by adapting the leave-out-uniques test; this test assesses whether the collection's judgments are sufficiently complete to allow runs that did not contribute to the construction of the collection to be fairly evaluated. We then assess the quality of individual topics since the TREC-COVID Complete collection has an unusually high rate of relevant documents. To do so, we investigate whether a collection that omits the topics with the highest percentage of judged relevant documents would result in more stable rankings of systems.

The next section recaps how the TREC-COVID collections were constructed and introduces the notation used to designate precise segments of the total collection. The following section examines the quality of the collection by demonstrating its reusability and sensitivity to run differences. The TREC-COVID Complete collection is found to be on par with the TREC-8 ad hoc collection (generally considered a high-quality collection), with existing tests uncovering no apparent problems. The final section concludes by noting the need for new incremental, diagnostic tests that are able to definitively characterize a test collection's quality to support building high-quality collections more economically.

1 BUILDING TREC-COVID COMPLETE

The genesis of the TREC-COVID effort was the release of the COVID-19 Open Research Dataset (CORD-19) [13]. Begun in March 2020, this dataset is an open-access collection of biomedical literature articles on COVID-19 and other coronavirus research that is updated regularly. TREC-COVID used the CORD-19 datasets to build a series of test collections with the twin goals of evaluating how well existing search systems could cope with the rapidly growing corpus of scientific literature related to COVID-19, and discovering methods for improving the management of scientific information in future global health crises.

1.1 TREC-COVID rounds

TREC-COVID was based on the TREC model of building retrieval test collections through community evaluations of search systems. It consisted of a series of five rounds, with each round using a later version of CORD-19 and a expanded set of COVID-related topics. The TREC-COVID topics were developed based on search requests gathered from logs of medical library search systems and major trends on social media. The incremental building of the collection was designed to capture topics and documents as the pandemic progressed. The first round started with 30 topics and five new topics were added in each subsequent round.

Each TREC-COVID round functioned as a TREC track compressed into a few weeks. Organizers announced the version of the CORD-19 dataset and the topic set to be used in a given round, and

participants had approximately one week to submit ranked retrieval results (called *runs*) for that combination. The runs were used to create shallow pools and the documents in the pools were judged by biomedical experts for relevance on a three-point scale of Not Relevant, Partially Relevant or Fully Relevant. Runs were scored using the judgments (called *qrels*) and the scores returned to the participants. Finally, all data were archived on the TREC-COVID website[1]. Participation was excellent: approximately 90 distinct teams submitted a total of 556 runs across the five rounds.

Relevance judgments from one round were available to participants before the start of each subsequent round. Participants were explicitly allowed to use the judgments to train systems for later rounds. To support the use of the relevance feedback search techniques, TREC-COVID used *residual collection* evaluation [8] for each round except the first. In residual collection evaluation, any document that has already received a judgment with respect to a particular topic is (conceptually) removed from the collection for that topic in all later rounds. One consequence of using residual collection evaluation was that runs submitted to a particular TREC-COVID round could only be scored using judgments made for that round, not the cumulative set of judgments across rounds.

1.2 Family of collections

Assessing time was the most critical resource for TREC-COVID—judgments had to be made by biomedical experts during a biomedical crisis. To have predictable assessor workloads, sometimes different assessors judged a given topic in different rounds, and occasionally more than one assessor judged a topic within a single round. This is a departure from standard TREC procedure. Since we know that different assessors can have different opinions about relevance, the TREC-COVID relevance judgment sets might be less self-consistent than is ideal. Investigating the extent to which this is a problem (if at all) would require yet more judgments.

To keep assessors occupied while participants were generating their runs and thus increase the total number of judgments that could be made, judgments were made in two sets per TREC-COVID round. The judgments used to evaluate Round $X+1$ runs consisted of the documents in the pools for Round $X+1$, called Round $X+1$ judgments, in union with previously unjudged documents selected from Round X runs, called Round $X.5$ judgments. Unanticipated changes in the various CORD-19 versions further complicated relevance judgment bookkeeping. Later versions of CORD-19 are not strict supersets of earlier versions. Instead, some documents from early versions are dropped in later versions and other documents receive new document ids, reflecting the messy reality of the churn in the literature. Document content might also have changed between CORD-19 versions. If a document that was judged in a previous round had a significant change in content (defined as the addition of an abstract or full-text or a change in title) in a version of CORD-19 used in a later round, the document was rejudged in the later round (the only time a document was judged more than once per topic).

The variety of different judgment sets, the document content changes, and the changes in the composition of the CORD-19 versions makes it difficult—but vital—to know exactly what test collection is used to evaluate a run. Proper evaluation requires the document set assumed in the qrels file to exactly match the document set used to produce the run. TREC-COVID converged on a notation to make the relationships explicit, and this convention is used for the qrels files in the TREC-COVID archive and to refer to judgment sets in this paper. A qrels designation contains two parts consisting of a document round and a sequence of contiguous judgment rounds: dX_jY-Z where X is an integer from 1 to 5 and refers to the CORD-19 version used in Round X, and Y and Z are beginning and ending judgment rounds. A dX qrels file contains only document ids that are valid in the TREC-COVID Round X document set. A jY-Z qrels file contains the judgments made in any judgment (half) round starting at Y up to and including Z.

Table 1: Datasets comprising the TREC-COVID Collections

	Round1	Round 2	Round 3	Round4	Round5
Release	April 10	May 1	May 19	June 19	July 16
# docs	51,103	59,851	128,492	157,817	191,175
# topics	30	35	40	45	50
Cumulative qrels					
Label	d1_j0.5-1	d2_j0.5-2	d3_j0.5-3	d4_j0.5-4	d5_j0.5-5
Size	8,691	20,727	33,068	46,203	69,318
Chronological qrels					
Label	d1_j0.5-5	d2_j0.5-5	d3_j0.5-5	d4_j0.5-5	d5_j0.5-5
Size	21,545	31,842	46,074	58,844	69,318

TREC-COVID organizers distinguished three broad types of qrels files derivable from the TREC-COVID judgments, and these are the qrels posted in the archive. The first type is the residual collection qrels used to evaluate official TREC-COVID submissions. For example, the qrels used to evaluate TREC-COVID Round 3 submissions is d3_j2.5-3. A second type is a Cumulative qrels, which includes all judgments up to and including the judgments for the corresponding round. For example, d3_j0.5-3 is the Round 3 Cumulative qrels that contains all judgments made before or during Round 3. The last type is a Chronological qrels. Taken together, the set of Chronological qrels across document rounds define test collections for different time periods as the pandemic unfolded (hence the "chronological" designation). The individual Chronological qrels for Round X contains a judgment for every [*topic,docid*] pair such that *topic* is one of the topics used in Round X, *docid* is a valid document in Round X, and *docid* was judged (in any judgment round) for *topic*. These are designated as dX_j0.5-5. Table 1 gives statistics for the primary test collections derived from TREC-COVID judgments.

The test collection we have designated as TREC-COVID Complete is d5_j0.5-5, which is judgments from all rounds, using all 50 topics and using the Round 5 (July 16) version of the CORD-19 document set. Note that the TREC-COVID Complete qrels, the Round 5 Chronological qrels, and the Round 5 Cumulative qrels are all exactly the same judgment set. No TREC-COVID participant runs correspond to TREC-COVID Complete because the use of residual collection evaluation meant that previously-judged documents were removed from Round 5 runs.

1.3 Judgment pools

The Common Core and Deep Learning tracks in recent TRECs studied the use of dynamic assessment schemes to economically

[1] https://ir.nist.gov/trec-covid/

Table 2: Pool creation strategy for judgment round

0.5	Three coordinator-produced Anserini runs pooled to depth 40.
1.0	Top priority Round 1 run from each team (56 runs) pooled to depth 7.
1.5	Top priority Round 1 runs pooled to depth 15.
2.0	Top priority run from each Round 2 team plus two baseline runs (52 runs total) to depth 7 for topics 1–30 and depth 15 for topics 31–35.
2.5	Top 2 priority Round 2 runs (97 runs) pooled to depth 10 for topics 1–30 and depth 20 for topics 31–35.
3.0	Top priority Round 3 runs (31 runs) plus 7 Kaggle submissions pooled to depth 10. Roughly 250 documents mistakenly did not get judged, so were judged in round 3.5.
3.5	Top 2 priority Round 3 runs (58 runs) pooled to depth 11 for topics 1–30. All Round 3 runs (79 runs) pooled to depth 15 for topics 31–40.
4.0	Top 2 priority Round 4 runs (51 runs) pooled to depth 8 for topics 1–30 and depth 20 for topics 31–45.
4.5	All Round 4 runs (72 runs) pooled to depth 15 for topics 1-35 and depth 30 for topics 36-45.
5.0	For topics 1–35, we selected a topic-specific depth such that the pool size was as large as possible without exceeding 200 documents. These pools were created over the top 5 priority Round 5 runs (99 runs). For topics 36–45, we used depth 50 pools over the top 5 priority runs, and for topics 46–50, we used depth-50 pools over all Round 5 runs (126 runs).

build high-quality test collections [4, 12]. A dynamic protocol selects the next document to be judged based on the current set of assessments. A practical protocol needs to specify how to allocate a given total judgment budget among the set of topics. The protocols used in TREC each started with depth-10 pools and then used various heuristics regarding the ratio of relevant documents to judged documents to terminate the judgment process for a topic.

Unfortunately, none of this experience transferred to TREC-COVID. Assessing time was the most critical resource, and for some rounds even depth-10 pools created too large of a document set to be judged in the allotted time. Each topic in the test set had to be judged in every subsequent round both to make sure the new part of the document set was explored and because the use of residual collection evaluation meant only documents judged for that round could be used to evaluate submissions. As a result, judgment sets were simply traditional pools with small cut-offs, where the cut-off was determined empirically each round to be as large as possible without exceeding the overall budget for the round. Half-round judgment sets were formed using deeper cut-offs and/or including runs that had not previously contributed to the pools. Table 2 summarizes the pool formation strategy used in each round. Run *priority* refers to the priority order assigned to a run by its submitter. In all cases, documents that had already been judged were removed from the pools.

Table 3: Per-topic (T) counts of total number of judged documents (Tot J) and the fraction of judged documents that are some form of relevant (% Rel).

T	Tot J	% Rel	T	Tot J	% Rel	T	Tot J	% Rel
1	1647	0.424	18	1325	0.503	35	1360	0.176
2	1287	0.260	19	1489	0.079	36	1233	0.549
3	1688	0.386	20	1234	0.613	37	1234	0.416
4	1849	0.307	21	1600	0.411	38	1920	0.720
5	1697	0.381	22	1325	0.449	39	1264	0.773
6	1607	0.619	23	1293	0.305	40	1230	0.478
7	1382	0.379	24	1248	0.361	41	1043	0.341
8	1869	0.347	25	1590	0.362	42	769	0.362
9	1664	0.126	26	1720	0.484	43	878	0.342
10	1141	0.436	27	1477	0.610	44	1238	0.438
11	1821	0.243	28	1103	0.559	45	1171	0.769
12	1626	0.399	29	1241	0.523	46	680	0.294
13	1893	0.486	30	1035	0.390	47	1064	0.438
14	1296	0.211	31	1701	0.218	48	747	0.644
15	1981	0.225	32	1571	0.146	49	1093	0.244
16	1640	0.250	33	1270	0.242	50	889	0.168
17	1353	0.530	34	1842	0.107			

Topics that were added to the test set in a later round had a relatively greater proportion of the judgment budget in the subsequent rounds because the organizers wanted the TREC-COVID Complete collection to be as balanced as it could be given the other constraints. The last set of judgments (d5_j5-5) was significantly bigger than previous sets. Table 3 shows counts of documents within the TREC-COVID Complete collection. The table gives the total number of documents judged for the topic (Tot J) and the fraction of judged documents that are any form of relevant (% Rel) for each topic (T). Compared to other TREC ad hoc collections, TREC-COVID Complete contains both more topics with large numbers of relevant documents and more topics with a large percentage of judged documents that are relevant.

2 COLLECTION QUALITY

While the primary goal of TREC-COVID was to capture the dynamics of the information environment during a pandemic, the organizers also hoped to build a general-purpose, reusable test collection for traditional ad hoc retrieval. In this section we support the claim that TREC-COVID Complete is such a collection. Reusability tests show that the collection is comparable to early TREC ad hoc collections that are generally regarded as being high-quality collections. This is despite the fact that the collection includes all the topics that would have been rejected by the heuristics used in recent TRECs to decide topic inclusion. A limitation of all the existing quality tests is that they depend on the available data—the runs, topics, and relevance judgments used to construct the collection. Thus problems the tests detect are truly problems, but the tests may not detect problems that are nonetheless present.

2.1 Reusability

We call an IR test collection reusable if it is able to fairly compare retrieval runs that did not contribute to the collection-building process. While in principle any fair set of relevance judgments

produces a reusable collection, in practice the only known way of constructing a fair set of relevance judgments for traditional evaluation measures is to obtain a 'sufficiently complete' set of judgments. That is, we need judgment sets that contain all of the different types of relevant documents for a topic, if not all instances of relevant documents [1].

The traditional way of testing the reusability of a collection built through pooling is with the Leave-Out-Uniques (LOU) test [1, 14]. The LOU test compares system rankings induced by two test collections, the original collection and a reduced collection in which the relevant documents that were contributed to the pools by a single participant team are removed. The rationale for the test is that uniquely retrieved relevant documents would not be known to be relevant if the team had not participated in the collection-building process. When these uniquely retrieved relevant documents are removed from the collection, the runs that contributed them can be treated as new runs (i.e., runs that were not used to build the collection) and small changes in the relative ranking of the new runs with respect to the remaining runs supports a claim of reusability.

We cannot perform a LOU test on TREC-COVID Complete. First, the concept of uniquely retrieved relevant documents is undefined because there is not a single set of runs that contributed to the pools. Second, none of the TREC-COVID submissions are retrieval results over the full TREC-COVID Complete document set because previously judged documents were removed from Round 5 submissions. Nonetheless, the main idea of the LOU test, examining how system comparisons fare when the collection's judgment set is restricted, can be investigated using the Chronological qrels.

We first need to define the set of runs to compare. Since Round 1 had no previously judged documents, the 143 Round 1 submissions are full runs with respect to the Round 1 version of CORD-19. We also have a set of 13 ad hoc (i.e., not feedback) runs created using the Round 5 topic and document sets but still containing all of the previously judged documents. This set of 13 runs, which we call the "Full Round 5" set, is comprised of nine baseline runs made available for Round 5 by the Anserini team[2] in union with four SMART runs provided by Chris Buckley. A Full Round 5 run can be converted into a valid Round 1 run by restricting the run to the first 30 topics and removing all documents that are not contained in the Round 1 document set from the ranked lists. We use the 156 runs in the union of the Round 1 submissions and the restricted Full Round 5 runs as the reusability test run set.

The LOU test generally measures the similarity of system rankings using the Kendall τ measure of association [10]. However, Kendall's τ is not ideal for this because its value depends on both the number of runs being ranked and the size of the differences in scores of the runs [9]. To remove these confounding effects, we instead directly compare all pairs of runs looking for *conflicts*: a pair of runs that 1) rank in different order on the original and reduced collection and 2) have a score difference that is statistically significant in at least one of the collections.

We consider a score difference to be statistically significant if the confidence intervals of the mean scores for the two runs do not overlap. Confidence intervals were computed using a bootstrap percentage method with 5000 iterations and $\alpha = 0.05$ [7]. Computing the confidence interval for a single run works as follows when there are T topics in the collection. For each iteration, T scores are drawn with replacement from the set of the run's individual topic scores and the mean of those T scores is computed. This produces a distribution of number-of-iterations (5000 in our case) means. The means at the 2.5 and 97.5 percentiles of the distribution are used as the lower and upper bounds of the confidence interval for that run.

We deem the system comparison induced by the entire set of relevance judgments (d1_j0.5-5) as the True order, and compare it to the order induced by removing the set of documents from a given judgment round from the qrels. That is, we compare the evaluation results scored using d1_j0.5-5 to those using d1_j0.5-4, then to those using d1_j0.5-3, and so on. We repeat the comparison of all run pairs for each of five measures: MAP, Precision@10, NDCG@10, Bpref, and RBP(p=0.5). Across all judgment rounds' qrels and all measures, we observe no conflicts at all. This means that the smallest of the collections, the Round 1 collection d1_j0.5-1 that has only 40% as many judgments as does d1_j0.5-5, is as good at detecting meaningful differences among this set of runs as d1_j0.5-5, even for measures such as MAP that depend on recall. Given the diverse make-up of the test set of runs, including the Round 5 runs as new runs, this is strong evidence that the TREC-COVID Complete collection is reusable.

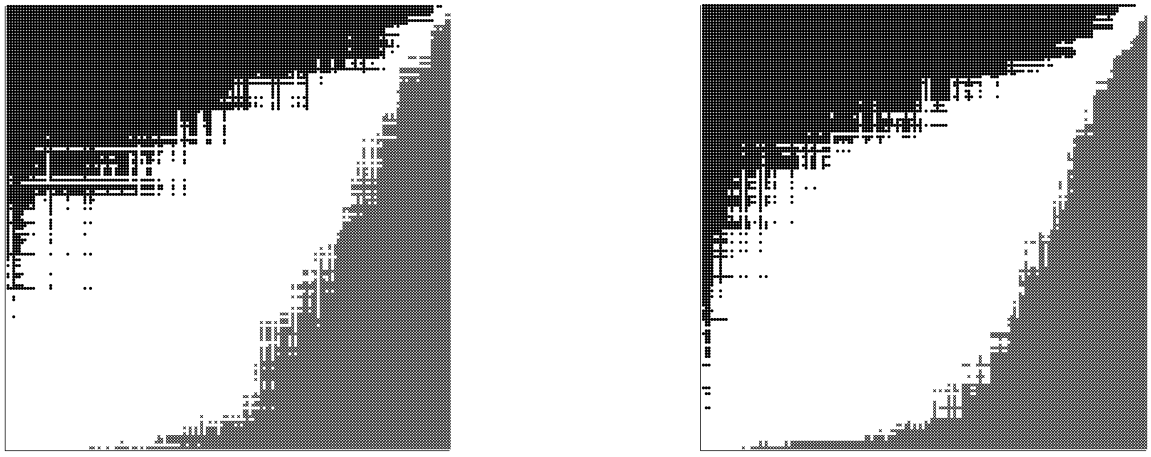
The full system rankings *are* different between the two collections. Comparing the system rankings produced for the d1_j0.5-1 and d1_j0.5-5 collections, the Kendall τ for MAP is 0.914 and for Precision@10 is 0.943, with a maximum change in rank of 35 and 25, respectively. There are no conflicts among these runs, however, because the changes all occur among runs that are statistically indistinguishable from one another.

Figure 1 further illustrates the relationship between the d1_j0.5-1 and d1_j0.5-5 collections. The figure shows plots of the significantly different run pairs of the two collections using either MAP or Precision@10. Both axes plot the 156 runs in the run set ordered by decreasing mean score using d1_j0.5-5. A mark is placed at point (x,y) if the corresponding runs are statistically different from one another. The lower triangular half of a plot shows the results for the d1_j0.5-5 collection and the upper triangular for the d1_j0.5-1 collection. The empty swath along the diagonal demonstrates the (typical) result that most runs are statistically indistinguishable from their neighbors. The near symmetry of the two triangular halves shows that the Round 1 collection makes almost the same set of meaningful distinctions among runs as the full collection.

2.2 Discarding topics

When building a collection using a dynamic protocol, one of the main decisions to be made is when to stop judging a topic. A topic can be stopped because the builder believes a sufficient number of relevant documents has already been found, with such topics being included in the final collection, or because the builder believes the budget will not support finding a sufficient number of relevant documents so the topic should be abandoned before yet more judging resources are spent on it. The when-to-stop-assessing problem is closely related to the problem of when to terminate an interactive search; Cormack and Grossman suggest terminating a search after

[2]See https://github.com/castorini/anserini/blob/master/docs/experiments-covid.md.



a) MAP b) Precision@10

Figure 1: Significantly different run pairs for MAP (left) and Precision@10 (right). A mark is plotted at (x,y) if the confidence intervals around the mean scores for runs x and y do not overlap. The upper triangular half of the matrix compares runs as evaluated using collection d1_j0.5-1 and the lower triangular using collection d1_j0.5-5.

examining $2R + 100$ documents where R is the number of relevant documents found so far [3]. The high-quality TREC ad hoc collections' pools had a small percentage of relevant documents. For example, for the TREC-8 ad hoc collection fewer than 10% of the documents in the pool were relevant [1]. This experience led to using heuristics of accepting a topic to be part of a test collection only if the number of relevant found was less than a third [12] or at most a half [4] of the number of documents judged.

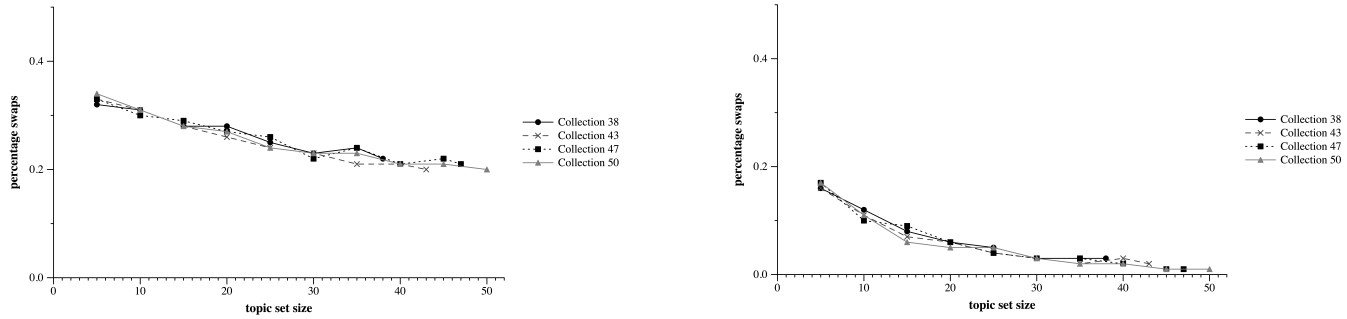
As shown in Table 3, TREC-COVID Complete contains topics with a large percentage of judged documents that are relevant. For twelve topics—about a quarter of the test set—more than half of the judged documents are relevant, and for three topics more than 70% of the judged documents are relevant. As TREC-COVID drew to a close, the question was whether to exclude some number of these topics from the final test set since, according to the heuristics, the large percentage of relevant indicate the judgments for these topics are too incomplete.

We used a bootstrapped test of collection sensitivity to examine the quality of the collection with and without topics with a large percentage of relevant documents. In the test we compare the evaluation scores for run pairs each evaluated on two different topic sets of size S. A run pair evaluating in a different order on the two different topic sets is called a *swap*. We repeat the comparison using all pairs of runs for each pair of topic sets. We generate many different topic set pairs where each set is generated from the set of available topics (the topic *universe*) by randomly drawing S topics with replacement. The percentage of the total number of comparisons that results in a swap is a measure of how reliably the collection can distinguish among runs using S topics, with smaller percentages indicating higher reliability. The entire process is repeated for increasing topic set size. All swap percentages are computed over bins representing a range of differences in scores. For example, in the implementation of the test used here, Bin 4 contains the set of run pairs for which the score difference is greater than 0.04 and less than or equal to 0.05.

Figure 2 shows the results of the test using MAP as the measure. Each plot contains a line for four different topic sets used as the universe, containing 38, 43, 47, and 50 topics respectively. The 50-topic universe contains all TREC-COVID topics, the 47 universe removes the three topics with more than 70% of the judged documents relevant, the 43 universe also removes the topics with more than 60% relevant, and the final universe contains none of the 12 topics with more than 50% relevant. For each topic set size S, 500 different topic set pairs are generated. Topic set sizes start at 5 topics and increase by 5 topics up to the universe size. The set of runs is the set of 126 runs submitted to TREC-COVID Round 5, so there are 7,875 run pairs[3]. There are 21 bins for score differences, where each bin is a range of size 0.01, differences start at 0, and the final bin contains all differences greater than 0.2. Figure 2 shows two representative bins. As expected, the swap percentage decreases with both larger mean differences and greater topic set size [2]. The results for each of the four topic set universes are very similar to one another, though the 50-topic universe has a lower swap rate slightly more often. This implies the heuristics do not hold in this case: the collection that includes all topics, including topics with very large percentage relevant, is at least as good and perhaps better at detecting differences among runs as the collections with fewer but more typical topics. Since larger topic sets are generally preferred, we kept all 50 topics in the TREC-COVID Complete collection.

We also looked at the distribution of evaluation scores across Round 5 submissions for each topic to ensure there were no obvious artifacts caused by large percentage of relevant documents. Figure 3 shows a box-and-whisker plot of average precision (AP) scores, where the heavy horizontal line is the median AP score across runs, the boxes range from the first to the third quartile of scores, and outlier scores are plotted as a circle. The topic set exhibits the typical variability in difficulty across topics with no apparent correlation to percentage relevant.

[3]We are examining the total topic set, so need to use runs produced using the Round 5 topic set.



a) Bin 1 with score differences between 0.01 and 0.02

b) Bin 4 with score differences between 0.04 and 0.05

Figure 2: Percentage of run pairs that evaluated in different orders on two different topic sets of a given size using MAP. Topic sets are drawn with replacement from the universe of topics, where there are either 38, 43, 47, or 50 topics in the universe.

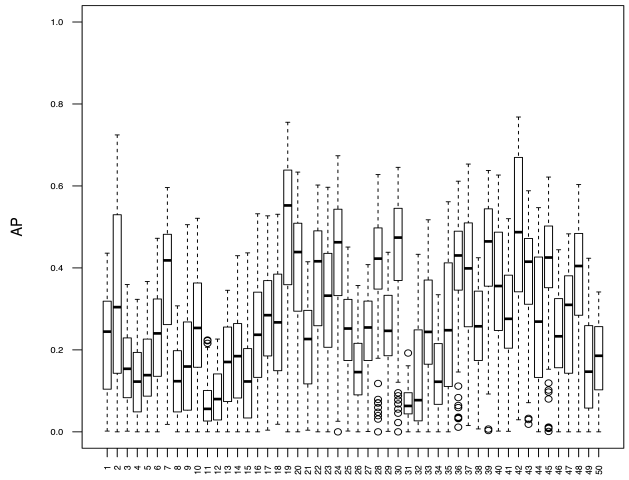


Figure 3: Distribution of per-topic AP scores across TREC-COVID Round 5 submissions. Topics with a large percentage of judged documents that are relevant do not have distinctive distributions.

Why does TREC-COVID Complete behave differently from other TREC collections despite the topics with large percentage relevant? We argue it is a combination of five factors:

- The CORD-19 document set is smaller than TREC ad hoc collections, and the total number of documents judged is a large percentage of the entire collection. A historically enormous 1% of the collection was judged for some topics.
- Most of the submitted runs were very effective. Many of the submissions were feedback runs that made good use of the available judgments. Since the runs were effective, there were even more known relevant documents to train on in later rounds and a large part of the information space was explored. There were also fewer mistake runs than in a typical TREC track, especially after the first round. Thus judgments weren't wasted on questionable runs.
- We counted partially relevant documents as part of the relevant set. Using fully relevant judgments only, the majority of topics have a relevant percentage less than 33%, and all topics have less than 50%. The very high percentage relevant when counting partially relevant documents may indicate that those topics have many instances of relevant documents, but fewer types of relevant documents, and the qrels contains each of the types.
- CORD-19 contains a relatively large number of duplicate and near-duplicate documents. With feedback, when one instance of a relevant document is found, the systems are likely to return all instances in the next round. This increases the density of relevant documents returned by the systems.
- The domain of CORD-19 is a single disease, and thus should not be expected to behave as a newswire collection. For example, a pandemic collection has no natural upper bound for the number of overlapping articles on a topic (e.g., there were more than 1200 clinical trials conducted on hydroxychloroquine for COVID-19).

3 CONCLUSION

The COVID pandemic offered an opportunity to create a test collection that captured the information needs of biomedical researchers as the pandemic unfolded, thereby providing infrastructure to improve search systems for future medical crises. The result of constructing the infrastructure is actually a series of separate, but interrelated, IR test collections. Existing tests of collection quality indicate that the largest of these collections, TREC-COVID Complete, is a reusable IR test collection. Having a set of diverse, high-quality runs facilitated good collection building as expected. But previous intuitions about the number of judgments needed for robust collections were shown to be (quite) pessimistic. These heuristics were developed within the traditional TREC framework with deep pools, one-shot retrieval, and generally some failed experiments. They did not generalize to the TREC-COVID experience where relevance feedback produced high-quality runs from experienced participants.

Existing quality tests have the major limitation that they cannot definitively demonstrate that a collection is of high-quality, just note the absence of any known problems with the run set used to build the collection after its construction. The availability of incremental, diagnostic quality tests would greatly facilitate collection-building since they could be used to determine when to stop judging a topic in dynamic assessment scenarios. This would prevent wasting as many (expensive) judgments on topics that are too big to adequately assess so never make it into a test collection and eliminate the need for assessments for post hoc evaluation of the collection.

REFERENCES

[1] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Information Retrieval* 10 (2007), 491–508.

[2] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, N. Belkin, P. Ingwersen, and M.K. Leong (Eds.). 33–40.

[3] Gordon V. Cormack and Maura R. Grossman. 2016. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, 75–84.

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track. In *Proceedings of Twenty-Eighth Text REtrieval Conference (TREC 2019)*.

[5] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association* 27, 9 (2020), 1431–1436. https://doi.org/10.1093/jamia/ocaa091

[6] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. Searching for Scientific Evidence in a Pandemic: An Overview of TREC-COVID. *arXiv* 2104.09632 (2021).

[7] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 525–532. https://doi.org/10.1145/1148170.1148261

[8] Gerard Salton and Chris Buckley. 1997. Improving retrieval performance by relevance feedback. *Journal of the Association for Information Science and Technology* 41 (1997), 355–364.

[9] Mark Sanderson and Ian Soboroff. 2007. Problems with Kendall's Tau. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 839–840.

[10] Alan Stuart. 1983. Kendall's Tau. In *Encyclopedia of Statistical Sciences*, Samuel Kotz and Norman L. Johnson (Eds.). Vol. 4. John Wiley & Sons, 367–369.

[11] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *ACM SIGIR Forum* 54, 1 (June 2020).

[12] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 407–416. https://doi.org/10.1145/3269206.3271766

[13] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* abs/2004.10706 (2020).

[14] Justin Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM Press, New York, Melbourne, Australia, 307–314.