Journal of the American Medical Informatics Association, 27(9), 2020, 1431–1436
doi: 10.1093/jamia/ocaa091
Advance Access Publication Date: 8 July 2020
Brief Communications

Brief Communications

TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19

Kirk Roberts [1], Tasmeer Alam,[2] Steven Bedrick,[3] Dina Demner-Fushman,[4] Kyle Lo,[5] Ian Soboroff,[2] Ellen Voorhees,[2] Lucy Lu Wang,[5] and William R. Hersh [3]

[1]University of Texas Health Science Center at Houston, Houston, Texas, USA, [2]National Institute of Standards and Technology, Gaithersburg, Maryland, USA, [3]Oregon Health & Science University, Portland, Oregon, USA, [4]US National Library of Medicine, Bethesda, Maryland, USA and [5]Allen Institute for AI, Seattle, Washington, USA

*Corresponding Author: Kirk Roberts, PhD, University of Texas Health Science Center at Houston, 7000 Fannin St, #600, Houston, TX 77030, USA (kirk.roberts@uth.tmc.edu)

Received 28 April 2020; Editorial Decision 29 April 2020; Accepted 1 May 2020

ABSTRACT

TREC-COVID is an information retrieval (IR) shared task initiated to support clinicians and clinical research during the COVID-19 pandemic. IR for pandemics breaks many normal assumptions, which can be seen by examining 9 important basic IR research questions related to pandemic situations. TREC-COVID differs from traditional IR shared task evaluations with special considerations for the expected users, IR modality considerations, topic development, participant requirements, assessment process, relevance criteria, evaluation metrics, iteration process, projected timeline, and the implications of data use as a post-task test collection. This article describes how all these were addressed for the particular requirements of developing IR systems under a pandemic situation. Finally, initial participation numbers are also provided, which demonstrate the tremendous interest the IR community has in this effort.

Key wordsKey words: information retrieval, shared task, COVID-19

MOTIVATION

During the last major global pandemic, the 1918–19 influenza ("Spanish Flu"), the information landscape was very different than today: flu viruses had not yet been discovered; worldwide literacy was considerably lower; information spread largely by word-of-mouth; and the digital content we depend on so greatly today for scientific advancement did not exist, from PubMed and preprints to social media. Medically, COVID-19 itself is different: rapidly spreading through many asymptomatic individuals but also having high morbidity and mortality, especially for certain groups, such as the elderly, infirm, and those facing existing health disparities.[1] However, another key difference in this pandemic is the quantity of information, including the use of preprints and rapid publication policies, which has resulted in a scientific corpus that grows by hundreds of COVID-19 articles per day.[2]

These changes in the conduct and dissemination of science all create challenges for information retrieval (IR), the scientific field behind search engines.[3] The technical goal of IR is to rapidly search through a large collection of documents (the "corpus") to find relevant information to address a particular information need. The biomedical and health goals of IR range from promoting scientific discovery,[4,5] to providing clinical decision support,[6,7] to addressing the health needs of consumers and combating misinformation.[8] All of these are, of course, highly relevant in a pandemic.

There are many important basic research questions surrounding the use of IR in a pandemic situation:

1. How does one identify the set of appropriate content (the corpus) over which to search?
2. How can a search engine be quickly deployed under these circumstances?

3. What are the appropriate IR modalities (ad hoc search, filtering, question-answering, etc.) for this kind of event?
4. What are effective methods for customizing the search engine to the specific needs of the situation?
5. Further, can existing data be leveraged (eg, via machine learning) to improve the search engine?
6. Further still, can event-specific training data be created fast enough to have an impact?
7. How does one *quantitatively* evaluate the search engine's performance (ranking)?
8. Further, how likely is it that different search engines have divergent enough performance to merit a quantitative comparison during a crisis?
9. How does one *qualitatively* evaluate the search engine?

For COVID-19, there are some initial resources to help answer these questions. The COVID-19 Open Research Dataset (CORD-19)[2] was created (and updated weekly) to provide a suitable corpus for retrieval (Question 1). Meanwhile, existing search engines were quickly repurposed for this dataset,[9] helping to answer Question 2. But while these more engineering-type questions have preliminary answers, the other questions, which dive deeper into the science of IR, still remain.

This article describes the rationale and preliminary structure of TREC-COVID, a shared task focused on analyzing Questions 3 through 8 above. The goals of the task are to galvanize the informatics community and provide the necessary data to help answer these important questions. The last concern about qualitative evaluation (Question 9) remains, but was added to the list above to acknowledge its well-established importance (eg,[9]) and to encourage other informatics experts to take up its banner.

This article provides a preliminary overview of TREC-COVID, which has just begun accepting submissions. Its purpose is to encourage further participation in this task as well as gather critical feedback from the informatics community, all with the goal of answering the above critical questions.

TASK STRUCTURE

The basic TREC (Text REtrieval Conference) ad hoc evaluation structure[10] provides participants with a corpus and set of topics (which they fashion into queries entered into their IR systems). Participants then submit "runs" of up to N results per topic (usually $N = 1000$). The results of all participants are pooled and the top-ranked results are manually assessed. Note that unlike natural language processing (NLP) evaluations,[11] IR evaluations generally perform annotation after system submission because the gold standard relevance data is unknown. Participant runs are then scored according to the assessed data. Evaluating a search engine for a pandemic, however, breaks many of these assumptions: new topics arise as the pandemic develops; new documents are published with updated information; and search engines are modified to keep pace. A new evaluation paradigm was thus warranted for TREC-COVID. Notably, the task is iterative, with new documents, new topics, and new system submissions every few weeks. Figure 1 provides an illustration of the task structure and the key aspects of this structure are described below.

Users

Given that the CORD-19 dataset (see Wang et al[2] for more details on CORD-19) is composed largely of scientific articles, the intended user of a TREC-COVID-compatible system is broadly defined as an "expert," including researchers, clinicians, policy makers, and journalists. The content of the articles in CORD-19 is likely beyond the understanding of many health consumers.

Modality

Three IR modalities were initially considered: (1) *ad hoc*, where a query is issued by a user and ranked documents are returned immediately–this is the most widely used IR modality; (2) *filtering*, where a standing query is issued, and then over time, as new batches of documents become available, they are filtered down to the relevant subset for the query; and (3) *question answering*, which is an extension of ad hoc with the notable differences that the query is a full natural language question and the answer is in the form of a passage, not an entire document. Given the large paradigm shift from the standard TREC evaluation, it was decided to start with an ad hoc evaluation, being the most familiar and likely the simplest modality. However, a question answering task that extends the ad hoc task has been proposed and will likely be announced soon. A filtering task is also being considered.

Topics

An initial set of 30 topics was created, with 5 new topics planned for each additional round. The inspiration for the topics came from a variety of sources: posts by high-profile researchers on Twitter, medical library searches, search logs of MedlinePlus, and suggestions on Twitter using #COVIDSearch. Due to the nature of the CORD-19 data, it is assumed that users will be willing to enter longer, clearer queries than normal. To account for this, each topic has 3 fields with increasing levels of expressiveness: (1) query, a few simple keywords (eg, "*coronavirus mortality*"), (2) question, which provides a more specific natural language version ("*what are the mortality rates overall and in specific populations?*"), and (3) narrative, which adds additional clarifications and suggestions of the user's intent ("*Seeking information on fatality rates in different countries and in different population groups based on gender, blood types, or other factors*"). Examples of a further 5 topics are provided in Table 1.

Participant requirements

Participants are given roughly 1 week from topic release to result submission. They submit up to 1000 documents (by CORD-19 id) for each topic in the standard "trec_eval" format. To reduce barriers to entry, participants are allowed to take part in any round, without any prior or subsequent round submission requirements. This means that teams are ranked on a by-round basis, instead of overall.

Assessment

Manual judgment of IR results is a time- and resource-intensive process but essential for a gold-standard test collection. It is estimated that it takes approximately 1 minute to judge a single article for a topic, and the goal is to assess several hundred results per topic, requiring hundreds of hours of assessment over the course of the task. The assessment is conducted with a custom platform. See Figure 2 for a screenshot.

Relevance

As is typically done in TREC, including its medical tracks,[6,7,12–15] each assessed document is judged as *relevant*, *partially relevant*, or *not relevant* to the topic. Details and clarifications on the relevance

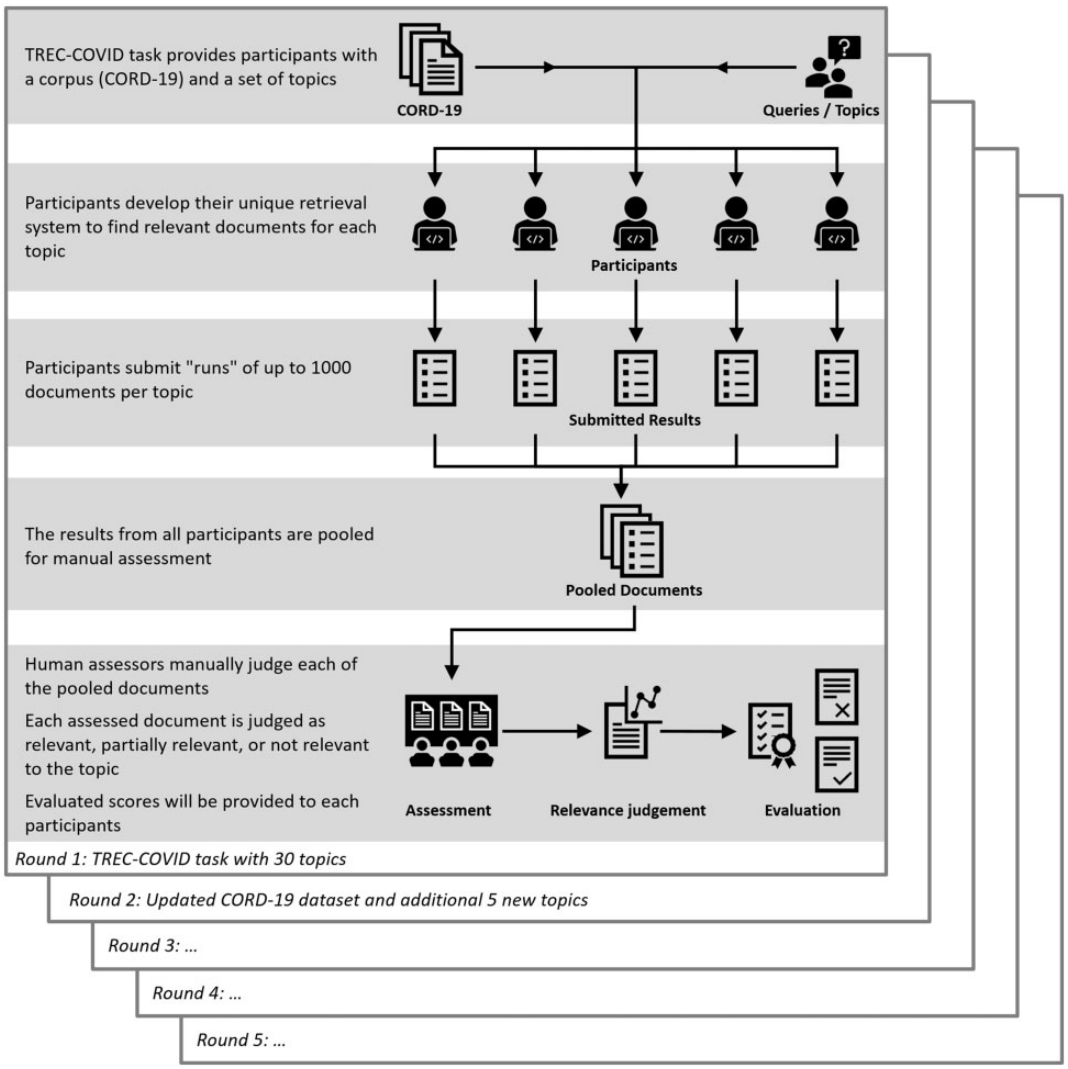

Figure 1. Graphical illustration of TREC-COVID task.

Table 1. Illustrative examples of topics for TREC-COVID task

Query	Question	Narrative
Coronavirus response to weather changes	How does the coronavirus respond to changes in the weather?	Seeking range of information about virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions
Coronavirus social distancing impact	Has social distancing had an impact on slowing the spread of COVID-19?	Seeking specific information on studies that have measured COVID-19's transmission in 1 or more social distancing (or non-social distancing) approaches
Coronavirus outside body	How long can the coronavirus live outside the body?	Seeking range of information on the virus's survival in different environments (surfaces, liquids, etc.) outside the human body while still being viable for transmission to another human
coronavirus asymptomatic	What is known about those infected with Covid-19 but are asymptomatic?	Studies of people who are known to be infected with Covid-19 but show no symptoms?
Coronavirus hydroxy-chloroquine	What evidence is there for the value of hydroxychloroquine in treating Covid-19?	Basic science or clinical studies assessing the benefit and harms of treating Covid-19 with hydroxychloroquine.

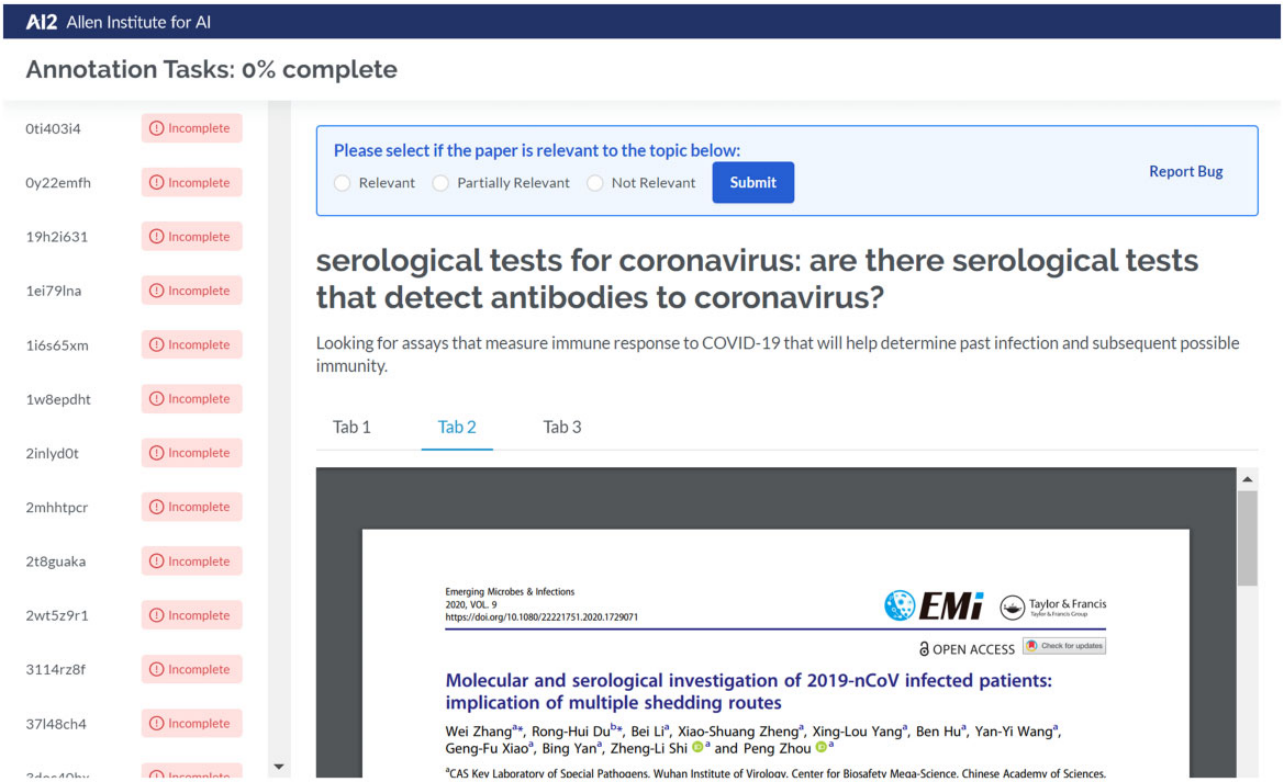

Figure 2. Screenshot of TREC-COVID assessment platform.

definition can be found at the TREC-COVID site (https://ir.nist.gov/covidSubmit/).

Evaluation

Traditional measures of retrieval effectiveness such as precision and recall assume the relevance judgments are complete. However, modern document sets are too large to have a human look at every document for every topic. TREC pioneered the use of pooling to create a smaller subset of documents to judge for a topic. The main assumption underlying pooling is that judging only the top-ranked documents from a wide variety of different retrieval results uncovers sufficiently many of the relevant documents that any unjudged document can be assumed to be not relevant. For TREC-COVID, the short time between rounds means that the subset of documents that can be judged for a topic will likely be too small to contain most of the relevant documents. Single-round scores will therefore be noisy, (ie, contain a large amount of uncertainty). One measure that does not rely on complete judgments is bpref (binary preference measure),[16] which is a function of the number of times a known irrelevant document is retrieved before a known relevant document, and thus disregards unjudged documents. TREC-COVID will score submissions using trec_eval (https://trec.nist.gov/trec_eval/index.html) that reports traditional measures as well as bpref scores.

Iteration

The Round 1 topics were issued April 15, 2020 concurrently with the official press release (https://www.nist.gov/news-events/news/2020/04/nist-and-ostp-launch-effort-improve-search-engines-covid-19-research), with the initial runs due April 23. Round 1 judgment should be finished by May 3. Round 2 will start soon thereafter. To test the assessment process, there was also a "Round 0" based on runs from 3 baseline systems using Anserini.[17] Each subsequent round will have 5 new topics, while retaining the prior topics. An evaluation side effect of this is that participants will have access to gold standard data for the very topics on which they are retrieving results. This "feedback" scenario is seen as a feature instead of a bug: new documents will continue to be added to the collection, and many of the topics will still be important to the pandemic. So having a set of known relevant results for a topic is a legitimate use case. However, this requires "residual" evaluation: only the results assessed in the current round (not prior rounds) are considered for pooling and scoring.

Projected timeline

Allowing for roughly 1 week from a round's topic release to result submission, and around 1 week for result assessment, it is expected that each round takes between 2 and 3 weeks. New rounds will continue to be offered so long as there is interest, new topics worth issuing, and resources for assessment.

Post-task test collection

The final set of judgments will be useful beyond the life of the task. Each set of judgments will be associated with a snapshot of CORD-19, allowing future systems to simulate the streaming nature of the document collection and issuance of topics. The data could alternatively be used as a benchmark for a simple, standard ad hoc evaluation as well. The goal is to enable studying how IR systems are developed so as to improve search engines for the next major health outbreak, not just COVID-19.

Figure 3. Assessment results for Round 0.

Table 2. Baseline (Anserini) results on the Round 0 assessments using just the Question field of the topics.

	P@10	MAP	NDCG	bpref
Title/abstract	**0.5167**	0.3563	0.6061	0.3584
Full text	0.4233	0.2991	0.5249	0.3220
Paragraph*	0.5033	**0.3946**	**0.6594**	**0.3872**

Abbreviations: bpref, binary preference measure; MAP, Mean Average Precision; NDCG, Normalized Discounted Cumulative Gain; P@10, Precision of top 10 results.

*Note that the "paragraph" baseline indexes the title and abstract with each paragraph.

RESULTS

In Round 1, 56 teams submitted 143 runs, which is an extremely high level of participation and interest from the community. For perspective, only 1 task in the 28-year history of TREC (including 193 separate tasks) had more participants,[18] and TREC-COVID had a submission deadline less than 1 month after it was unofficially announced and 1 week after it was officially announced.

As of the time of writing, the assessments for Round 1 are unavailable, but the assessment results from Round 0 are shown in Figure 3 and the baseline run results are shown in Table 2. As can be seen, most topics have at least some relevant articles in CORD-19, though the distribution is uneven. While planned, no double-assessments have yet occurred, so there are no interrater agreement numbers to report. As the focus of this brief communication is the rationale and structure of the task, a detailed analysis of the results is left to a future publication.

DISCUSSION

The TREC-COVID task serves several purposes: (1) immediate support for researchers and clinicians fighting the pandemic caused by SARS-CoV-2 virus; (2) development of a new IR evaluation process as the document collection, state of knowledge, and users' interests rapidly evolve; and (3) a collection and approach to standing up systems capable of satisfying information needs during pandemics.

Limitations

While based on decades of IR evaluation experience, TREC-COVID is still a new evaluation paradigm being developed with unprecedented speed, which contributes to several limitations. The most important limitation is the incomplete judgments. Due to the pace of the evaluation, the growth of the collection (which doubled within a month), and limited availability of qualified annotators, the depth of the above-described judgment pools is fairly shallow, and some relevant documents will remain unjudged and therefore be considered not relevant. The second limitation is the nature of the collection that combines peer-reviewed and preprint work that is judged solely for topical relevance, which might lead to some less rigorous and potentially erroneous publications judged as relevant. When using this collection in the future, some of the errors will be mitigated by corrections in the subsequent versions, but some will remain. Finally, the collection does not cover the interests of health consumers. This limitation will be alleviated in an upcoming QA task, which will combine the CORD-19 collection with a collection of consumer-friendly COVID-related documents published by the WHO, CDC, and other government sites.

CONCLUSION

This article presented a brief description of the rationale and structure of TREC-COVID, a still-ongoing IR evaluation. TREC-COVID is creating a new paradigm for search evaluation in rapidly evolving crisis scenarios. Future publications will provide additional details about the results of the task.

FUNDING

The organizers would like to thank the Allen Institute for AI and Microsoft Research for funding support.

AUTHOR CONTRIBUTIONS

The authors are organizers of TREC-COVID. KER initially drafted the manuscript. All authors reviewed and approved the manuscript.

ACKNOWLEDGMENTS

The organizers would like to thank numerous individuals for their help in organizing this track: Aaron Cohen for task discussions; Sarvesh Soni for work on the baseline systems; Sam Skjonsberg, Paul Sayre, and Robert Gale for work on the assessment platform; and Julia Barton, Hannah Kim, Evan Mitchell, Isabelle Nguyen, Magdalena Hecht,

Adam Betcher, Miles Fletcher, Phu Nguyen, Meenakshi Vanka, Austen Yeager, Annemieke van der Sluijs, Brian Huth, Carol Fisher, Cathleen Coss, Cathy Smith, Deborah Whitman, Denise Hunt, Dorothy Trinh, Funmi Akhigbe, Janice Ward, Keiko Sekiya, Nick Miliaras, Oleg Rodionov, Olga Printseva, Preeti Kochar, Rob Guzman, Susan Schmidt, and Melanie Huston for work on manual assessments.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Oberfeld B, Achanta A, Carpenter K, *et al*. SnapShot: COVID-19. Cell 2020; 181 (4): 954.
2. Wang LL, Lo K, Chandrasekhar Y, *et al*. CORD-19: The COVID-19 open research dataset. *arXiv* abs/2004.10706, 2020.
3. Hersh WR. *Information Retrieval: A Biomedical and Health Perspective*. 4th ed. Springer; 2020.
4. Hersh W, Bhupatiraju RT. TREC Genomics Track Overview. In: Proceedings of the Twelfth Text REtrieval Conference; 2003.
5. Roberts K, Gururaj AE, Chen X, *et al*. Information retrieval for biomedical datasets: the 2016 bioCADDIE dataset retrieval challenge. *Database* 2017; 2017: bax068.
6. Roberts K, Simpson MS, Demner-Fushman D, Voorhees E, Hersh WR. State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS Track. *Information Retrieval* 2016; 19 (1): 113–48.
7. Roberts K, Demner-Fushman D, Voorhees, EM, *et al*. Overview of the TREC 2017 Precision Medicine Track. In: Proceedings of the Twenty-Sixth Text REtrieval Conference; 2017.
8. Jimmy, Zuccon G. UQ IElab at TREC 2019 Decision Track. In: Proceedings of the Twenty-Eighth Text REtrieval Conference, 2019.
9. Zhang E, Gupta N, Nogueira R, Cho K, Lin J. Rapidly deploying a neural search engine for the COVID-19 open research dataset: preliminary thoughts and lessons learned. *arXiv* 2004; 05125, 2020.
10. Voorhees EM, Harman DK. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 2005.
11. Uzuner Ö, South B, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
12. Roberts K, Simpson MS, Voorhees E, Hersh W. Overview of the TREC 2015 Clinical Decision Support Track. In: Proceedings of the Twenty-Fourth Text REtrieval Conference, 2015.
13. Roberts K, Demner-Fushman D, Voorhees E, Hersh W. Overview of the TREC 2016 clinical decision support track. In: Proceedings of the Twenty-Fifth Text REtrieval Conference, 2016.
14. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A. Overview of the TREC 2018 Precision Medicine Track. In: Proceedings of the Twenty-Seventh Text REtrieval Conference, 2018.
15. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A. Overview of the TREC 2019 Precision Medicine Track. In: Proceedings of the Twenty-Eighth Text REtrieval Conference, 2019.
16. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004.
17. Yang P, Fang H, Lin J. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017: 1253–1256.
18. Ounis I, Macdonald C, Lin J, Soboroff I. Overview of the TREC-2011 Microblog Track. In: Proceedings of the Twentieth Text REtrieval Conference, 2011.