# CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview

Evangelos Kanoulas[1], Dan Li[1], Leif Azzopardi[2], and Rene Spijker[3]

[1] Informatics Institute, University of Amsterdam, Netherlands, E.Kanoulas@uva.nl, D.Li@uva.nl
[2] Computer and Information Sciences, University of Strathclyde, Glasgow, UK, leif.azzopardi@strath.ac.uk
[3] Cochrane Netherlands and UMC Utrecht, Julius Center for Health Sciences and Primary Care, Netherlands, R.Spijker-2@umcutrecht.nl

**Abstract.** Systematic reviews are a widely used method to provide an overview over the current scientific consensus, by bringing together multiple studies in a reliable, transparent way. The large and growing number of published studies, and their increasing rate of publication, makes the task of identifying all relevant studies in an unbiased way both complex and time consuming to the extent that jeopardizes the validity of their findings and the ability to inform policy and practice in a timely manner. The CLEF 2017 e-Health Lab Task 2 focuses on the efficient and effective ranking of studies during the abstract and title screening phase of conducting Diagnostic Test Accuracy systematic reviews. We constructed a benchmark collection of fifty such reviews and the corresponding relevant and irrelevant articles found by the original Boolean query. Fourteen teams participated in the task, submitting 68 automatic and semi-automatic runs, using information retrieval and machine learning algorithms over a variety of text representations, in a batch and iterative manner. This paper reports both the methodology used to construct the benchmark collection, and the results of the evaluation.

**Keywords:** Evaluation, Information Retrieval, Systematic Reviews, TAR, Text Classification, Active Learning

## 1 Introduction

Evidence-based medicine has become an important pillar in health care and policy making. In order to practice evidence-based medicine, it is important to have a clear overview over the current scientific consensus. These overviews are provided in systematic review articles, that summarize all available evidence regarding a certain topic (e.g., a treatment or diagnostic test). In order to write a systematic review, researchers have to conduct a search that will retrieve all the studies that are relevant. The large and growing number of published studies, and their increasing rate of publication, makes the task of identifying relevant studies in an unbiased way both complex and time consuming to the extent

that jeopardizes the validity of their findings and the ability to inform policy and practice in a timely manner. Hence, the need for automation in this process becomes of utmost importance. Finding all relevant studies in a corpus is a difficult task, known in the Information Retrieval (IR) domain as the total recall problem.

To this date, retrieval of evidence to inform systematic reviews is being conducted in multiple stages:

1. Boolean Search: At the first stage information specialists build a broad Boolean query expressing what constitutes relevant information. The query is then submitted to a medical database containing titles, abstracts, and indexing terms of a controlled vocabulary of medical studies. The result is a set, A, of potentially interesting studies.
2. Title and Abstract Screening: At a second stage experts are screening the titles and abstracts of the returned set and decide which one of those hold potential value for their systematic review, a set D. If screening an abstract has a cost $C_a$, screening all |A| abstracts has a cost of $C_a * |A|$.
3. Study Screening: At a third stage experts are downloading the full text of the potentially relevant abstracts, $D$, identified in the previous phase and examine the content to decide whether indeed these studies are relevant or not. Examining a document has typically a larger cost of $C_d > C_a$. The result of the second screening is a set of references to be included in the systematic review.

Unfortunately, the precision of the Boolean searches is typically low, hence reviewers often need to look manually through many thousands of irrelevant titles and abstracts in order to identify a small number of relevant ones. Furthermore, the recall of the searches is often assumed to be 100%, which may not be the case.

To overcome some of the limitations of the Boolean search, researchers have been testing the effectiveness of machine learning and information retrieval methods. O'Mara-Eves et al.[15] provide a systematic review of the use of text mining techniques for study identification in systematic reviews.

The goal of this lab is to bring together academic, commercial, and government researchers that will conduct experiments and share results on automatic methods to retrieve relevant studies with high precision and high recall, and release a reusable test collection that can be used as a reference for comparing different retrieval and mining approaches in the field of medical systematic reviews.

## 2 Benchmark Collection

To construct the benchmark collection, the organizers of the task considered 58 systematic reviews on Diagnostic Test Accuracy conducted by the Cochrane researchers. These reviews are publicly available through the Cochrane Library[4]

[4] http://www.cochranelibrary.com/

and can be identified by setting the topic filter in the library to "Diagnostic" and "Diagnostic Test Accuracy" and the stage fitler to "Review". At the date of the publication of this article 79 such studies are available, however the last 22 were performed after the organizers put the collection together. The 58 systematic reviews considered can be found in the Appendix of this articles at Table 6.

Participants were provided with two data sets: (a) a development set, and (b) a test set. The development set consists of 20 topics for Diagnostic Test Accuracy (DTA) systematic reviews, while the test set consists of 30 topics. For both sets, one *topic* file and two files of relevance judgments at abstract and document level respectively are constructed (*qrel's*).

The topic file is generated through the following procedure. For each systematic review, we reviewed the search strategy from the corresponding study in Cochrane Library. A search strategy, among others, consists of the exact Boolean query developed and submitted to a medical database, at the time the review was conducted, and typically can be found in the Appendix of the study. Rene Spijker, a co-author of this work and a Cochrane information specialist examined the grammatical correctness of the search query and specified the date range which dictated the valid dates for the articles to be included in this systematic review. The date range was necessary because a study published after the systematic review should not be included even though it might be relevant, since that would require manually examining its content to quantify its relevance. Important note: A number of medical databases, and search interfaces to these databases is available for search, and for each one information specialists construct a different variation of their query that better fits the data and meta-data of the database. For this task, we only considered the Boolean query constructed for the MEDLINE database, using the Wolters Kluwer Ovid interface. Then we submitted the constructed Boolean query to the OVID system[5] and collected all the returned PubMed document identification numbers (PMID's) which satisfied the date range constraint. This step was automated by a Python script we put together and through an interface available to the University of Amsterdam[6]. Out of the 58 reviews 8 were discarded since the provided Boolean query was not in the right format, which made it difficult if not impossible to reconstruct the set of PMID's, hence the 50 topics in the development and test set.

The topic file is in a text format and contains four sections, Topic, Title, Query, and PMID's, where Topic is the topic ID, a substring of DOI of the document (e.g. CD010438 for 10.1002/14651858.CD010438.pub2), and PMID's are the document IDs returned by the Boolean query. The PIDs can be used to access the corresponding document through the National Center for Biotechnology Information (NCBI)[7]. An example of a topic file can be viewed below.

[5] http://demo.ovid.com/demo/ovidsptools/launcher.htm
[6] https://github.com/dli1/tar_data_collection
[7] https://www.ncbi.nlm.nih.gov/books/NBK25497/

```
Topic: CD009551
Title: Polymerase chain reaction blood tests for the diagnosis of
       invasive aspergillosis in immunocompromised people

Query:
exp Aspergillosis/
exp Pulmonary Aspergillosis/
exp Aspergillus/
(aspergillosis or aspergillus or aspergilloma or "A.fumigatus" or
"A. flavus" or "A. clavatus" or "A. terreus" or "A. niger").ti,ab.
or/1-4
exp Nucleic Acid Amplification Techniques/
pcr.ti,ab.
"polymerase chain reaction*".ti,ab.
or/6-8
5 and 9
exp Animals/ not Humans/
10 not 11

Pmid's:
    25815649
    26065322
    ...
```

For the construction of the *qrel* files, we considered the reference section of the 50 systematic reviews. The references are split into three categories: Included, Exclude, and Additional. Included are the studies that are relevant to the systematic review. Excluded are the studies that in the abstract and title screening stage were considered relevant, but at the article screening phase were considered irrelevant to the study and hence excluded from it. Additional are additional references that do not impact the outcome of the study, and hence irrelevant to it. The included references were the relevant studies at the document-level qrels, while both the included and excluded references were considered relevant at the abstract-level qrels. The format of the qrels followed the standard TREC format:

Topic Iteration Document Relevance

where Topic is the topic ID of the systematic review, Iteration in our case is a dummy field always zero and not used, Document is the PMID, and Relevancy is a binary code of 0 for not relevant and 1 for relevant studies. The order of documents in the qrel files is not indicative of relevance. Studies that were returned by the Boolean query but were not relevant based on the above process, were considered irrelevant. Those are studies that were excluded at the abstract and title screening phase. All other documents in MEDLINE were also assumed to be irrelevant, given that they were not judged by the human assessor.

Important Note: Note that, as mentioned earlier, the references of a systematic review were produced after a number of Boolean queries were submitted

to a number of medical databases, and their titles and abstracts were screened. The PMID's provided however were only those that came out of the MEDLINE query. Therefore, there was a number of abstract-level relevant studies (the gray area in the Venn diagram below) that were not part of the result set of the Boolean query provided to the participants. For the development set, the qrel file contained those additional PMIDs, for those participants that would decide to search the entire MEDLINE database, and not only consider the studies provided to them in the Topic files. To the best of our knowledge, no one submitted such a system, hence to avoid any bias we excluded those relevant studies from the test set.

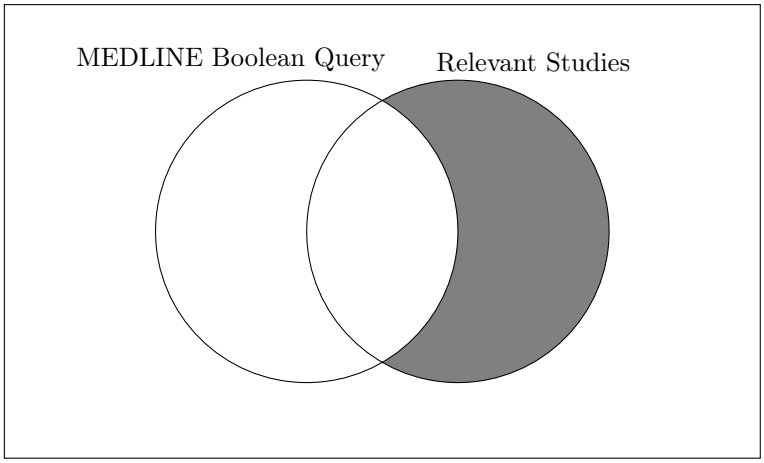


Table 1 shows the distribution of the relevant documents at abstract or document level for all the topics in the development set and the test set. The total number of unique PMID is 149,405 for the development set and 117,562 for the test set. Their percentages of relevant documents at abstract level are quite close, which is 1.88% for the development set and 1.58% for the test set. This is not true at document level, however, where the relevant documents in the test set is almost twice as large as in the development set, even though there are 0.52% and 0.33% of relevant studies, respectively. In [17], a test collection was developed based on a random selection of 93 Cochrane systematic reviews (not just DTAs), and reported a slightly higher rate of relevance ($\frac{14}{1159} = 1.2\%$). However, compared with the TREC campaign, the rate of relevant documents is 5.45%, 2.78% for the Adhoc track of TREC 8 and the Web track of TREC 2002. Overall, the number of relevant documents is not very high in this lab, making locating them quite a difficult task.

Important Note: As one can observe in Table 1, there are topics for which the output of the Boolean query is rather narrow, with as few as 64 studies to be reviewed for topic CD008760. Cochrane is conducting systematic reviews on a regular basis, in an attempt to update each review every two-three years. Some of the reviews considered for the construction of the benchmark collection, such

| file name | Topic | # total PMIDs | # abs rel | # doc rel | % abs rel | % doc rel |
|---|---|---|---|---|---|---|
| Development Set | | | | | | |
| 1 | CD010438 | 3250 | 39 | 3 | 1.20 | 0.09 |
| 11 | CD007427 | 1521 | 123 | 17 | 8.09 | 1.12 |
| 14 | CD009593 | 14922 | 78 | 24 | 0.52 | 0.16 |
| 19 | CD011549 | 12705 | 2 | 1 | 0.02 | 0.01 |
| 23 | CD011134 | 1953 | 215 | 49 | 11.01 | 2.51 |
| 28 | CD008686 | 3966 | 7 | 5 | 0.18 | 0.13 |
| 33 | CD011975 | 8201 | 619 | 60 | 7.55 | 0.73 |
| 35 | CD009323 | 3881 | 122 | 9 | 3.14 | 0.23 |
| 37 | CD009020 | 1584 | 162 | 12 | 10.23 | 0.76 |
| 38 | CD011548 | 12708 | 113 | 5 | 0.89 | 0.04 |
| 4 | CD011984 | 8192 | 454 | 28 | 5.54 | 0.34 |
| 43 | CD010409 | 43363 | 76 | 41 | 0.18 | 0.09 |
| 44 | CD008054 | 3217 | 274 | 41 | 8.52 | 1.27 |
| 45 | CD010771 | 322 | 48 | 1 | 14.91 | 0.31 |
| 50 | CD009591 | 7991 | 144 | 41 | 1.80 | 0.51 |
| 53 | CD008691 | 1316 | 73 | 20 | 5.55 | 1.52 |
| 54 | CD010632 | 1504 | 32 | 14 | 2.13 | 0.93 |
| 55 | CD007394 | 2545 | 95 | 47 | 3.73 | 1.85 |
| 6 | CD008643 | 15083 | 11 | 4 | 0.07 | 0.03 |
| 9 | CD009944 | 1181 | 117 | 64 | 9.91 | 5.42 |
| total | | 149405 | 2804 | 486 | 1.88 | 0.33 |
| Test Set | | | | | | |
| 10 | CD007431 | 2074 | 24 | 15 | 1.16 | 0.72 |
| 12 | CD008803 | 5220 | 99 | 99 | 1.90 | 1.90 |
| 15 | CD008782 | 10507 | 45 | 34 | 0.43 | 0.32 |
| 16 | CD009647 | 2785 | 56 | 17 | 2.01 | 0.61 |
| 17 | CD009135 | 791 | 77 | 19 | 9.73 | 2.40 |
| 18 | CD008760 | 64 | 12 | 9 | 18.75 | 14.06 |
| 2 | CD010775 | 241 | 11 | 4 | 4.56 | 1.66 |
| 21 | CD009519 | 5971 | 104 | 46 | 1.74 | 0.77 |
| 22 | CD009372 | 2248 | 25 | 10 | 1.11 | 0.44 |
| 25 | CD010276 | 5495 | 54 | 24 | 0.98 | 0.44 |
| 26 | CD009551 | 1911 | 46 | 16 | 2.41 | 0.84 |
| 27 | CD012019 | 10317 | 3 | 1 | 0.03 | 0.01 |
| 29 | CD008081 | 970 | 26 | 10 | 2.68 | 1.03 |
| 31 | CD009185 | 1615 | 92 | 23 | 5.70 | 1.42 |
| 32 | CD010339 | 12807 | 114 | 9 | 0.89 | 0.07 |
| 34 | CD010653 | 8002 | 45 | 0 | 0.56 | 0.00 |
| 36 | CD010542 | 348 | 20 | 8 | 5.75 | 2.30 |
| 39 | CD010896 | 169 | 6 | 3 | 3.55 | 1.78 |
| 40 | CD010023 | 981 | 52 | 14 | 5.30 | 1.43 |
| 41 | CD010772 | 316 | 47 | 11 | 14.87 | 3.48 |
| 42 | CD011145 | 10872 | 202 | 48 | 1.86 | 0.44 |
| 47 | CD010705 | 114 | 23 | 18 | 20.18 | 15.79 |
| 48 | CD010633 | 1573 | 4 | 3 | 0.25 | 0.19 |
| 49 | CD010173 | 5495 | 23 | 10 | 0.42 | 0.18 |
| 5 | CD009786 | 2065 | 10 | 6 | 0.48 | 0.29 |
| 51 | CD010386 | 626 | 2 | 1 | 0.32 | 0.16 |
| 56 | CD010783 | 10905 | 30 | 11 | 0.28 | 0.10 |
| 57 | CD010860 | 94 | 7 | 4 | 7.45 | 4.26 |
| 7 | CD009579 | 6455 | 138 | 79 | 2.14 | 1.22 |
| 8 | CD009925 | 6531 | 460 | 55 | 7.04 | 0.84 |
| total | | 117562 | 1857 | 607 | 1.58 | 0.52 |

**Table 1.** Statistics of development and test set.

as the CD008760 review, are updates to previous reviews. These updates, only specify a query for a time range that starts after the last review on the topic was conducted. Hence, the 64 studies, are the output of the Boolean query for this short time range, hence its small number. If the Boolean query were to run against the entire MEDLINE database, the number of studies would be in the range of tens of thousands, as is the case for some other reviews considered, e.g. CD008782.

## 3 Task Description

The CLEF 2017 e-Health Lab [8], task 2, focused on retrieving studies for conducting Diagnostic Test Accuracy (DTA) systematic reviews. Retrieval in this area is generally considered very difficult, where sensitive searches result in large quantities of references to be screened manually, and a breakthrough in this field would likely be applicable to other areas as well. The task has a focus on the second stage of the process, i.e. given the results of a Boolean search how to make abstract and title screening more effective and efficient. Currently a typical number needed to read (NNR), the number of studies to screen to identify 1 eligible study, for DTA systematic reviews is approximately 80 when applied to potential abstracts that need further full text assessment. With an average of 7000 results to be screened, which would take approximately 120 hours to screen (1 minute per abstract [18]), a huge benefit can be made in reducing the workload in this process.

Given the results of the Boolean search from stage 1 as the starting point, participants were asked to rank the set of the provided abstracts. The task had two goals: (i) to produce an efficient ordering of the documents, such that all the relevant abstracts are retrieved above the irrelevant ones, and (ii) to identify the relevant subset of abstracts to be shown to a user, that is a stopping point in the ranked list of abstract, where a researcher could confidently stop screening abstracts and titles. Therefore, we solicited two types of submissions: (i) ranking submission: automatic or manual methods that rank all abstracts, with the goal of retrieving relevant abstracts as early in the ranking as possible, and (ii) thresholding submission: thresholding can be performed in a batch, or iterative manner as well.

We also considered two evaluation frameworks, (a) a simple evaluation, and (b) a cost-effective evaluation. The assumption behind the simple evaluation framework is the following: The user of your system is the researcher that performs the abstract and title screening of the retrieved articles. Every time an abstract is returned (i.e. ranked) there is an incurred cost/effort of CA, while the abstract is either irrelevant (in which case no further action will be taken) or relevant (and hence passed to the next stage of document screening) to the topic under review. The assumption behind the cost-effective evaluation is the following: The user that performs the screening is not the end-user. The user can interchangeably perform abstract and title screening, or document screening, and decide what PMIDs to pass to the end-user. Every time an abstract

is returned the user can either (a) read the abstract (with an incurred cost of CA) and decide whether to pass this PMID to the end-user, or (b) read the full document (with an incurred cost of CA+CD) and decide whether to pass this PMID to the end-user, or (c) directly pass the PMID to the end user (with an incurred cost of 0), or (d) directly discard the PMID and not pass it to the end user (with an incurred cost of 0). For every PMID passed to the end-user there is a cost of attached to it: CA if the abstract passed on is not relevant, and CA + CD if the abstract passed on is relevant (that is, we assume that the end-user completes a two-round abstract and document screening, as usual, but only for the PMIDs the algorithm+feedback user decided to be relevant). Although a small number of teams participated in the cost-effective sub-task, the lab focused on the simple evaluation sub-task, and this is what is described in the remaining of this report.

## 4 Evaluation

Evaluation within the context of using technology to assist in the reviewing process is very much dependent on how the user(s) interact with the system - and the goal of the technology assistance. For example, is the goal of the assistance to automate the screening process - where the system assess all the abstracts and returns a subset of the initial set to be screened by the end-user (i.e. screened in batch mode). Or, it could be used to identify all the relevant documents as soon as possible, in an iterative manner - where the system asks for feedback from the end-user to help improve the ranking. Of course, then the an open problem is decide when to stop requesting feedback, and when to stop assessing abstracts. In which case a subset of abstracts is identified, which consist of abstracts have been screened during the feedback cycles and the remainder that are screened but are not used for feedback (i.e. in batch mode). There are, of course, many other possible variations. For the purposes of this initial track/task, we consider the problem as a ranking task - that is to rank the set of documents associated with the topic in decreasing order of relevance. We consider a document relevant if the abstract passed the abstract screening phase (regardless of whether it was included or excluded from the study).

For this task we employ a number of standard measures, typically used in IR ranking evaluations, along with other measures from related tracks and some new measures we have developed.

- Standard Measures
  - Average Precision (AP)
  - Normalized cumulative gain @ 0% to 100% of documents shown; for the simple case that judgments are binary, normalized cumulative gain @ % is simply Recall @ % of shown documents[10]
  - Number of Relevant Found (nr)
  - Recall $r = nr/R$, where $R$ the total number of relevant documents
  - Number of documents returned/shown ($n$)

– Related Measures (from [6,5]
  - LOSS-R $loss_r = (1-r)^2$
  - LOSS-E $loss_e = (n/(R+100) * 100/N)^2$, where $N$ is the size of the collection
  - Reliability = $loss_r + loss_e$ [6]
  - Work Saved over Sampling at $r$, $WSS@Recall = (TN+FN)/N(1-r)$[5]
– Proposed Measures
  - Last Rel Found: Minimum number of documents returned to retrieve all $nr$ relevant documents
  - Total Cost (TC);
  - Total Cost with Uniform penalty (TCU)
  - Total Cost with Weighted Penalty (TCW)

To calculate the cost based measured, we considered three possible interactions to support a range of different ways to screen the items and to utilize feedback when ranking. We consider the follow possibilities:

1. suppose we have an ranking algorithm, which uses no feedback from the user, simply ranks the list of abstracts. The list is then presented to the end-user, who evaluates them in a batch. In this case, no feedback is requested, and abstracted are marked, NF.
2. suppose we have a ranking algorithm which uses feedback (i.e. abstract(s) are presented to the user, feedback on their relevance is obtained, which is then used by the algorithm, thus simulating online feedback from the user). In this case, for each document where feedback from the users is requested, abstracts are marked AF, but if no feedback is requested it is marked NF. Abstracts marked NF, are then presented to the end-user to evaluate in a final batch.
3. for either above option, the algorithm may decided that an abstract is not relevant, and thus it does not need to be shown to a user, and so are marked NS.

To calculate the total cost (TC), we calculated:

$$TC = \#NF.C_a + \#AF.(C_a + C_f) \tag{1}$$

where $C_a$ is the cost of assessing the abstract, $C_f$ is the cost of asking for feedback $\#NF$ is the number of NF items, $\#AF$ is the number of AF items.

We also created two additional cost measures which included a penalty for missing relevant abstracts (a) with a uniform penalty and (b) a weighted penalty. The uniform penalty was calculated as follows:

$$TCU = TC + (R - r/R) * (N - n) * C_p \tag{2}$$

where $C_p$ is the cost of the penalty of missing a relevant abstract, $N$ is the total number of documents in the set for the topic. The assumption behind this penalty is that the end-user would need to continue examining abstracts before they would from the remaining $(R - r)$ relevant items, and encounters them

at a uniform rate in the remaining $N - n$ abstracts which were not shown. So if half the relevant items were missing, then the penalty component would be $(N - n)C_p/2$. If no relevant items were missing the penalty component would be zero.

The weighted penalty was calculated as follows:

$$TCW = TC + \sum_{i=1}^{(R-r)} (1/2^i)(N - n) * CP \tag{3}$$

where the assumption is that the end user would been to examine half of the remaining documents to find the next relevant abstract, per missing relevant abstract. So if all relevant items were missing, then the summation would tend to one, and the penalty component tends to $(N - n) * C_p$, while if only one relevant item is missing then, the penalty component is $(N - n) * C_p/2$.

To compute these measures we set $C_a = 1$,$C_f = 2$ and $C_p = 2$, to represent the relative costs of the different actions. Note that these are not based on any empirical data and used as a way to regulate penalize feedback and no shows.

## 5 Participants

Fourteen groups from eleven countries submitted a total of 68 runs for this task:

1. Amsterdam Medical Center, The Netherlands (AMC)
2. Aristotle University of Thessaloniki, Greece (AUTH)
3. Centre Nationnal de la Recherche Scientifique, France & Amsterdam Medical Center, The Netherlands (CNRS)
4. East China Normal University, China (ECNU)
5. Eidgenoessische Technische Hochschule Zurich, Switzerland (ETH)
6. International Institute of Information Technology, Hyderabad, India (IIIT)
7. North Carolina State University, United States (NCSU)
8. Nanyang Technological University, Singapore (NTU)
9. University of Padua, Italy (Padua)
10. University of Sheffield, United Kingdom (Sheffield)
11. University College London, United Kingdom & Northeastern University, USA (UCL)
12. University of Waterloo, Canada (Waterloo)
13. Queensland University of Technology & CSIRO, Australia (QUT)
14. University of Strathclyde, United Kingdom (UOS)

Table 2 categorizes the participating runs along five dimensions: (a) automatic vs manual runs; (b) use of the development set; (c) use of supervised and semi-supervised learning algorithms, (d) use of relevance feedback; and (e) thresholding the ranked list of articles. The categorization has been performed by the lab coordinators – not by the participants – based on the submitted participants description of their algorithms. Hence, there is always a chance of mis-classifying some run. Out of the 68 runs submitted, 52 focused on the simple

| Team | Run | Auto | Develop-ment | Supervised | Feedback | Threshold |
|---|---|---|---|---|---|---|
| AMC | amc.run.res | ✓ | ✓ | ✓ | x | x |
| AUTH | simple.run1/run2/run3/run4 | ✓ | ✓ | ✓ | ✓ | x |
| BASELINE | BM25 | ✓ | x | x | x | x |
| BASELINE | random.pubmed | ✓ | x | x | x | x |
| CNRS | cnrs.abrupt.all | ✓ | ✓ | ✓ | ✓ | x |
| CNRS | cnrs.gradual.all | ✓ | ✓ | ✓ | ✓ | x |
| CNRS | cnrs.noaf.all | ✓ | ✓ | ✓ | x | x |
| CNRS | cnrs.noaffull.all | ✓ | ✓ | ✓ | x | x |
| ECNU | run1 | ✓ | x | x | x | x |
| ECNU | run2 | ✓ | ✓ | ✓ | x | ✓ |
| ECNU | run3 | ✓ | ✓ | ✓ | x | ✓ |
| ETH | m1 | ✓ | ✓ | ✓ | x | ✓ |
| ETH | m2 | ✓ | ✓ | ✓ | ✓ | ✓ |
| ETH | m4 | ✓ | ✓ | ✓ | x | ✓ |
| IIIT | run1/run2/run3/run4 | ✓ | x | x | ✓ | ✓ |
| NCSU | simple | ✓ | x | ✓ | ✓ | ✓ |
| NCSU | abs | ✓ | x | ✓ | ✓ | ✓ |
| NTU | run1/run2/run3 | ✓ | ✓ | ✓ | x | x |
| Padua | iafa_m10k150f0m10 | x | ✓ | ✓ | x | x |
| Padua | iafap_m10p2f0m10 | x | ✓ | ✓ | x | x |
| Padua | iafap_m10p5f0m10 | x | ✓ | ✓ | x | x |
| Padua | iafas_m10k50f0m10 | x | ✓ | ✓ | x | x |
| QUT | ca_bool_ltr | ✓ | ✓ | ✓ | x | x |
| QUT | ca_pico_ltr | x | ✓ | ✓ | x | x |
| QUT | rf_bool_ltr | ✓ | ✓ | ✓ | x | x |
| QUT | rf_pico_ltr | x | ✓ | ✓ | x | x |
| QUT | bool_es | ✓ | x | x | x | x |
| QUT | pico_es | x | x | x | x | x |
| Sheffield | run1/run2/run3/run4 | ✓ | x | x | x | x |
| UCL | abstract | ✓ | ✓ | ✓ | x | x |
| UCL | fulltext | ✓ | ✓ | ✓ | x | x |
| UOS | sis.AL30Q_BM25 | ✓ | x | x | ✓ | ✓ |
| UOS | sis.TMBEST_BM25 | ✓ | x | x | x | x |
| UOS | sis.TMAL30Q_BM25 | ✓ | x | x | ✓ | x |
| UOS | sis.bm25_t1.5 | ✓ | x | x | x | ✓ |
| UOS | sis.bm25_t1 | ✓ | x | x | x | ✓ |
| UOS | sis.bm25_t2.5 | ✓ | x | x | x | ✓ |
| UOS | sis.bm25_t2 | ✓ | x | x | x | ✓ |
| Waterloo | A-rank-normal.txt | ✓ | x | ✓ | ✓ | x |
| Waterloo | A-thresh-normal.txt | ✓ | x | ✓ | ✓ | ✓ |
| Waterloo | B-rank-normal.txt | ✓ | x | ✓ | ✓ | x |
| Waterloo | B-thresh-normal.txt | ✓ | x | ✓ | ✓ | ✓ |

**Table 2.** Categorization of participant's runs in the simple evaluation framework along five dimensions.

evaluation framework, while 16 on the cost-effective one. Out of the 52 submitted runs for the simple sub-task, 35 ranked all the PMIDs that were returned by the Boolean query, while 17 tested different stopping criteria over the ranking. Participants employed both supervised and unsupervised methods, for ranking articles. A large number of runs were trained over the provided development set, and their generalization was tested against the test topics. 26 runs used the development set in some fashion, while 26 made no explicit use of it; it may be the case that participants tried different models and algorithms over the development set, and selected to submit the best performing ones, hence there may be a flavor of model selection, however we did not consider this as use of the development set. Participants represented the textual data in a variety of ways, including document-topic features, bag-of-words, topic model distributions, embeddings, metadata. In the remainder of section, by article we mean the abstract and the title of an article. We are not aware of any participant that worked on the full text of these articles.

In particular, **AMC** took a batch supervised approach, training a Random Forest over a topic model representation of the articles. A 75-topic model was fitted over all articles in the collection, and the Topic-to-Document matrix was used to extract features [2].

**AUTH** took a learning-to-rank approach, using both batch and active learning. Their model, HybridRankSVM, consists of two parts: an inter-topic model which utilizes XGBoost and is trained over the entire development corpus and an intra-topic model, an iteratively-built SVM, trained over relevance feedback provided partially in the test topics. For the inter-topic model a total of 24 topic-document (or solely topic) features were computed over the title, abstract and mesh terms of the articles and the query. For the intra-topic model a TF-IDF vectorization of the articles was used [3].

**CNRS** trained a logistic regression model on n-gram features from the titles and abstracts and structured data from the Medline citations. One of their models was trained using stochastic gradient descent on the majority of the features, and one on the principal components of a subset of the features. Class imbalance was handled by reweighting and undersampling, while two approaches for relevance feedback were investigated [13].

**ECNU** took a learning-to-rank approach, using BM25, PL2, and BB2 as features. The trained model was also combined with a vector space model [4].

**ETH** used a LAMBDA-Mart model trained on features, such as BM25, Fuzzy search, Vector content representation, publishing data. This model was used to experiment with different stopping criteria. One of the approaches taken was to use minimal relevance feedback to estimate the distribution of positive samples by score. This was done by sampling from the articles, preferring articles with higher score. A Gaussian distribution was fitted on the positive samples and the resulting biased distribution was corrected. The correction worked by first adapting the mean and then iteratively finding the standard deviation matching the sampled data the best. For more details the reader can refer to [9].

**NCSU** adopted a continuous active learning framework for this task. An SVM classifier was trained on the relevance feedback labels and undersampling of the negatively labeled articles removing those furthest from the SVM decision hyperplane was employed. Different runs made use of different weights on the labels depending on whether the abstract or the full text was considered relevant [20].

**NTU** examined the role of convolutional neural networks for classifying medical articles for systematic reviews [12].

**Padua** used a two-dimensional probabilistic version of BM25 to rank articles. The parameters were tuned using the development set. Further, the top abstract returned by BM25 was provided to two non-experts who generated one additional query each. The tree queries were then used to re-rank articles. Different approaches for relevance feedback and thresholding were investigated [14].

**QUT** trained a learning-to-rank model using domain specific features. As domain specific features, PICO annotations (Population, Intervention, Control, Outcome) were used; these were extracted automatically from articles and manually from the Boolean queries [16].

**Sheffield** automatically parsed the Boolean queries to extract both the terms and MeSH heading,s and used TF-IDF cosine similarity to calculate the similarity score between document title and abstracts [1].

**UOS** explored two methods: (i) topic models, where they used Latent Dirichlet Allocation to identify topics within the set of retrieved articles, and then ranking articles by the topic most likely to be relevant to the query, and (ii) relevance feedback, where they used Rocchio's algorithm to update the query model for subsequent rounds of interaction. A third approach combined the topic model and relevance feedback approaches to quickly identify the relevant articles. For the thresholding task, they applied a score threshold over BM25 [11].

**UCL** took a supervised approach and trained a deep model architecture to identify studies pertaining to a given review topic [19].

**Waterloo** applied the Baseline Model Implementation (BMI) from the TREC Total Recall Track (2015-2016). They further applied their "knee-method" stopping criterion to BMI to determine how many abstracts should be examined for each topic [7].

## 6 Results

Tables 7, 8, 9, 10 provide the results of a selection of the evaluation measures for all participating runs, both against the abstract and the document level relevance judgments, for the simple evaluation scenario. Figure 1 shows the corresponding box plots for Average Precision, with the Mean Average Precision against the abstract and document level judgments respectively denoted with a blue rectangle over the box plot.

In the following subsections we present results separately for ranking and thresholding runs, so that comparisons can be more meaningful.

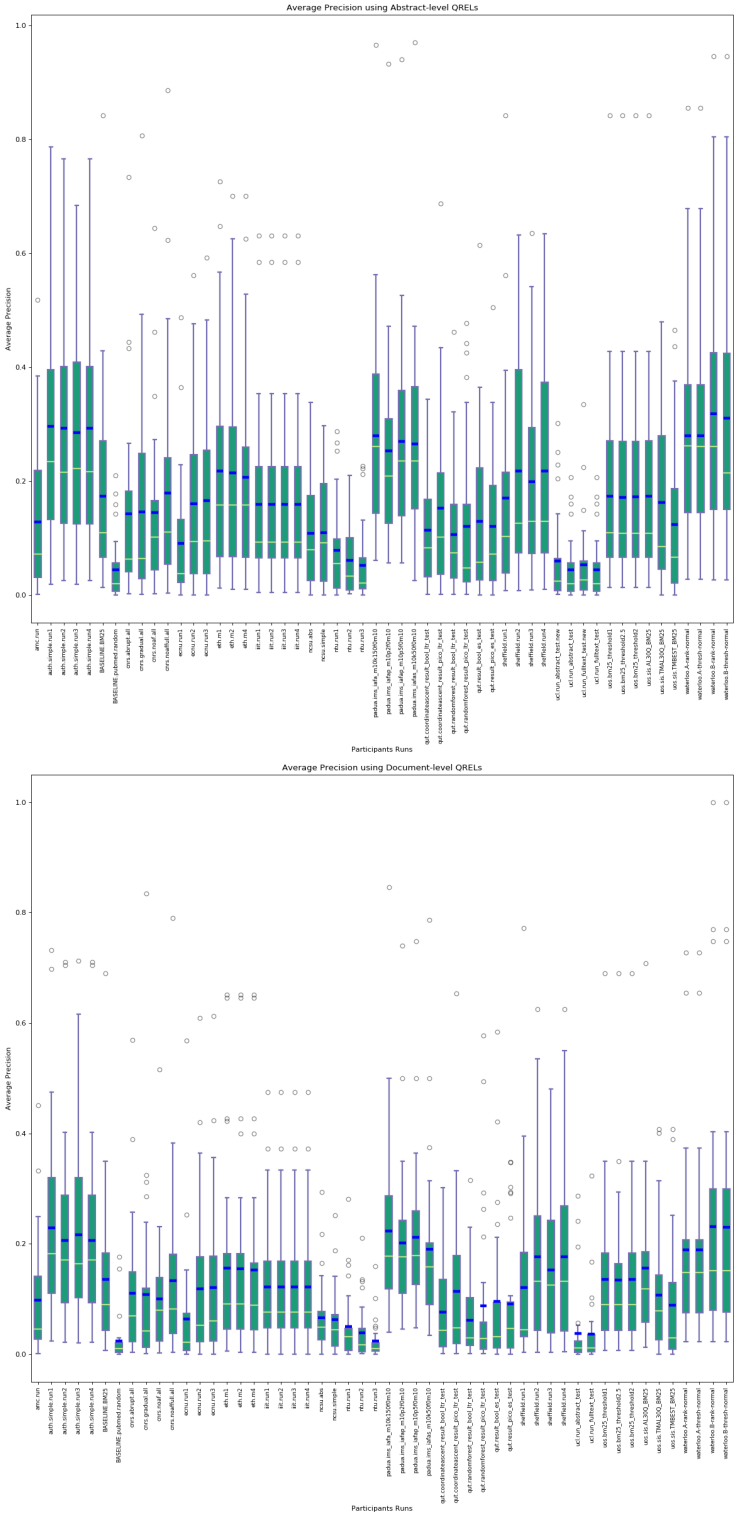


**Fig. 1.** Average precision using the abstract (top) and document (bottom) level relevance judgments.

### 6.1 Ranking Abstracts

Table 3 presents a number of evaluation measures for those runs that ranked the entire set of articles provided by the original Boolean queries; no thresholding has been applied. Some runs, as it may appear from Tables 7, 8, 9, 10, even though they applied no stopping criterion, still missed a number of documents. There may be multiple reasons for that, e.g. missing some topic, or not being able to download the abstract text, since participants were provided by PIDs only. The number of documents for which feedback was requested appears in the second column of the table, while the remaining of the columns report different measures of performance.

Figure 2 shows the recall-effort curves for the participating runs, that is the recall value at different percentage of documents shown to the user. The straight pink line with the triangular markers on x=y is the results of the Boolean query randomly shuffled, and it serves as a naive baseline, provided by the UOS team. The brown curve with the triangular markers is the BM25 retrieval function, also provided by the UOS team as a baseline; it ranks abstracts by BM25 over the Boolean query terms, with the default BM25 parameters setting.

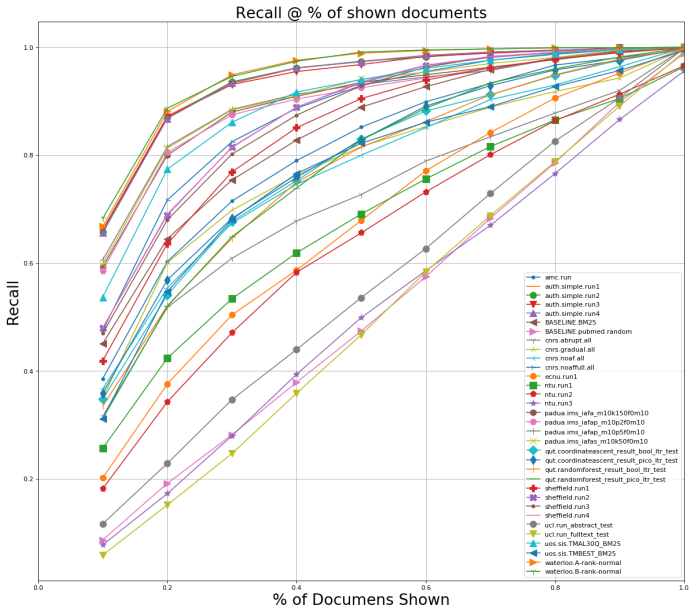


**Fig. 2.** Recall at different percentage of shown documents.

Figure 3 presents the box-plots of Mean Average Precision values for runs that do not make use of relevance feedback (left) and runs that make use of relevance feedback (right) respectively. On average relevance feedback boosts

| Run | Feedback | Last Rank Rel | wss@100 | wss@95 | Area Under Recall | AP |
|---|---|---|---|---|---|---|
| amc.run | 0 | 2913 | 0.249 | 0.333 | 0.761 | 0.129 |
| auth.simple.run1 | 41337 | 2143 | 0.519 | 0.693 | 0.928 | 0.297 |
| auth.simple.run2 | 41377 | 2124 | 0.521 | 0.697 | 0.920 | 0.293 |
| auth.simple.run3 | 23337 | 2183 | 0.511 | 0.678 | 0.924 | 0.285 |
| auth.simple.run4 | 41537 | 2119 | 0.519 | 0.690 | 0.920 | 0.293 |
| BASELINE.BM25 | 0 | 2851 | 0.285 | 0.400 | 0.809 | 0.174 |
| BASELINE.pubmed.random | 0 | 3722 | 0.040 | 0.034 | 0.484 | 0.045 |
| cnrs.abrupt.all | 19980 | 3414 | 0.173 | 0.243 | 0.735 | 0.143 |
| cnrs.gradual.all | 23683 | 3406 | 0.195 | 0.288 | 0.708 | 0.146 |
| cnrs.noaf.all | 0 | 2993 | 0.261 | 0.362 | 0.780 | 0.145 |
| cnrs.noaffull.all | 0 | 2250 | 0.412 | 0.497 | 0.839 | 0.179 |
| ecnu.run1 | 0 | 3633 | 0.099 | 0.121 | 0.627 | 0.091 |
| ntu.run1 | 0 | 3403 | 0.089 | 0.108 | 0.612 | 0.078 |
| ntu.run2 | 0 | 3204 | 0.117 | 0.131 | 0.595 | 0.060 |
| ntu.run3 | 0 | 3570 | 0.091 | 0.075 | 0.538 | 0.052 |
| padua.iafa_m10k150f0m10 | 2350 | 2269 | 0.415 | 0.508 | 0.896 | 0.280 |
| padua.iafap_m10p2f0m10 | 2367 | 2395 | 0.366 | 0.476 | 0.875 | 0.253 |
| padua.iafap_m10p5f0m10 | 5893 | 2260 | 0.398 | 0.496 | 0.885 | 0.269 |
| padua.iafas_m10k50f0m10 | 4320 | 2304 | 0.410 | 0.517 | 0.892 | 0.266 |
| qut.ca_bool_ltr | 0 | 3142 | 0.201 | 0.288 | 0.733 | 0.114 |
| qut.ca_pico_ltr | 0 | 3344 | 0.212 | 0.294 | 0.751 | 0.153 |
| qut.rf_bool_ltr | 0 | 3099 | 0.194 | 0.267 | 0.705 | 0.106 |
| qut.rf_pico_ltr | 0 | 3155 | 0.235 | 0.293 | 0.727 | 0.121 |
| sheffield.run1 | 0 | 2678 | 0.310 | 0.422 | 0.818 | 0.170 |
| sheffield.run2 | 0 | 2441 | 0.385 | 0.493 | 0.845 | 0.218 |
| sheffield.run3 | 0 | 2404 | 0.384 | 0.473 | 0.841 | 0.199 |
| sheffield.run4 | 0 | 2382 | 0.395 | 0.488 | 0.847 | 0.218 |
| ucl.run_abstract | 0 | 3801 | 0.072 | 0.064 | 0.507 | 0.060 |
| ucl.run_fulltext | 0 | 3755 | 0.077 | 0.076 | 0.522 | 0.053 |
| uos.sis.TMAL30Q_BM25 | 35432 | 2305 | 0.398 | 0.530 | 0.837 | 0.162 |
| uos.sis.TMBEST_BM25 | 0 | 3124 | 0.274 | 0.324 | 0.727 | 0.124 |
| waterloo.A-rank-normal | 117558 | 1464 | 0.601 | 0.700 | 0.927 | 0.279 |
| waterloo.B-rank-normal | 117558 | 1469 | 0.611 | 0.701 | 0.933 | 0.318 |

**Table 3.** Evaluation results for submitted runs ranking the entire set of articles provided by the Boolean query.

the effectiveness of the ranking algorithms, as expected, however it may come with additional cost in terms of assessing the relevance of abstract (based on the screening setup considered).

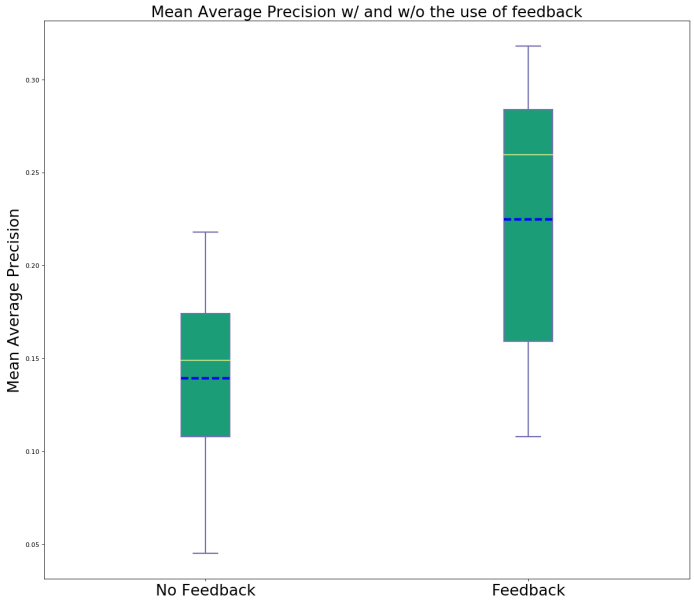


**Fig. 3.** Box-plots of Mean Average Precision for runs that do not make use of relevance feedback and those that do make use.

### 6.2 Drawing a Threshold

Table 4 presents a number of evaluation measures for those runs that applied a threshold criterion. The total number of shown to the user abstracts can be found in the second column of the table, the number of documents for which feedback was requested in the third, while the remaining of the columns report different measures of performance. The cost measures account both for the cost of presenting a document to the user and for the additional cost of requesting feedback for a document, while they also account for the cost one would need to pay to reach 100% recall, under certain assumptions. Reliability considers the cost of not finding all relevant documents but makes no discrimination between the documents returned to the user and those for which feedback is requested. Average precision is well defined under the stopping criterion but hard to be used for comparing runs that use different thresholds. An easy to understand measure is the achieved recall at the rank of the threshold.

Figure 5 presents recall at the point of the threshold as a function of the number of documents presented to the user; that is at different stopping criteria,

| Run | Docs Shown | Feedback | Rel Docs Found | Cost w/ Uniform Penalty | Cost w/ Weighted Penalty | Area Under Recall | AP | Recall@ Thresh | Relia- -bility |
|---|---|---|---|---|---|---|---|---|---|
| ecnu.run2 | 30000 | 0 | 1191 | 4003 | 6641 | 0.64 | 0.16 | 0.71 | 0.44 |
| ecnu.run3 | 30000 | 0 | 1197 | 4016 | 6717 | 0.65 | 0.17 | 0.72 | 0.44 |
| eth.m1 | 51640 | 0 | 1686 | 2306 | 4740 | 0.81 | 0.22 | 0.93 | 0.20 |
| eth.m2 | 51604 | 5063 | 1702 | 2676 | 4720 | 0.80 | 0.21 | 0.90 | 0.14 |
| eth.m4 | 27046 | 0 | 1406 | 2527 | 5590 | 0.74 | 0.21 | 0.82 | 0.14 |
| iiit.run1 | 15354 | 15354 | 1006 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run2 | 15354 | 15354 | 1006 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run3 | 15354 | 15354 | 1006 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run4 | 15354 | 15354 | 1006 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| ncsu.abs | 12942 | 12942 | 1073 | 4409 | 7695 | 0.61 | 0.11 | 0.71 | 0.33 |
| ncsu.simple | 27950 | 27950 | 1611 | 4145 | 6964 | 0.68 | 0.11 | 0.83 | 0.18 |
| qut.bool_es | 69951 | 0 | 1475 | 3480 | 4976 | 0.64 | 0.13 | 0.76 | 0.36 |
| qut.pico_es | 63018 | 0 | 1414 | 3527 | 5168 | 0.62 | 0.12 | 0.74 | 0.34 |
| uos.bm25_1 | 103051 | 0 | 1828 | 3454 | 3786 | 0.81 | 0.17 | 0.99 | 0.45 |
| uos.bm25_2.5 | 76104 | 0 | 1758 | 2905 | 3902 | 0.79 | 0.17 | 0.94 | 0.27 |
| uos.bm25_2 | 84740 | 0 | 1784 | 3117 | 3748 | 0.80 | 0.17 | 0.95 | 0.33 |
| uos.sis.AL30Q | 94967 | 0 | 1809 | 3280 | 3865 | 0.80 | 0.17 | 0.97 | 0.38 |
| waterloo.A-thresh | 87767 | 87767 | 1842 | 8809 | 9543 | 0.93 | 0.28 | 1.00 | 0.50 |
| waterloo.B-thresh | 60936 | 60936 | 1548 | 6470 | 7150 | 0.91 | 0.31 | 0.97 | 0.43 |

**Table 4.** Evaluation results for submitted runs using different threshold criteria; measures are computed using abstract-level relevance judgments.

but also with different ranking and thresholding algorithms. As expected the more documents presented to the user (the lower the threshold criterion) the higher the achieved recall. Nevertheless, there are still algorithms that dominate others. The figure present the Pareto frontier. Figure 5 presents recall at the point of the threshold as a function of the feedback documents requested. As it can be viewed, although feedback documents, are in principle helpful towards achieving a high recall, there are algorithms that used no relevance feedback and still achieved high recall at a threshold.

### 6.3 Topic Difficulty

Table 5 provides statistics on the topics used in the test set, along with the average Average Precision (AAP) for each topic, a measure that can be seen as a proxy of the difficult of each topic. The Pearson correlation coefficient between AAP and the percentage of relevant documents, the total number of documents, and the total number of relevant documents is -0.4868 (p-value = 0.006), 0.1295 (p-value = 0.495), and 0.8994 (p-value = 0). Figures 6 and 7 visually demonstrate this correlation.

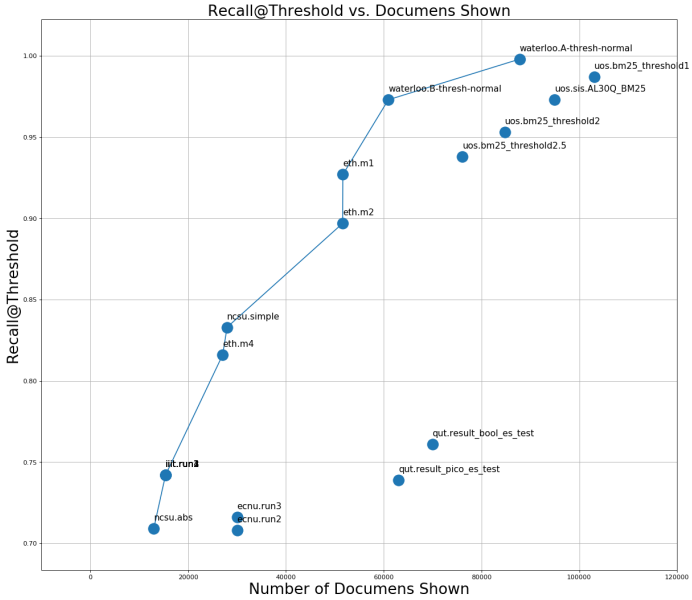


**Fig. 4.** Recall at the threshold rank as a function of the number of documents shown to the user.

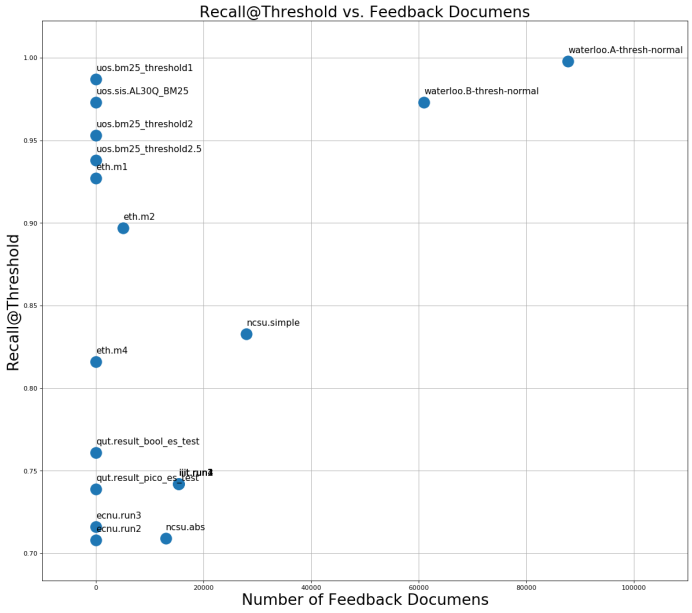


**Fig. 5.** Recall at the threshold rank as a function of the number of documents for which feedback is requested.

| Topic | Average AP | % of Relevant | Documents | Relevant |
|---|---|---|---|---|
| CD010173 | 0.035 | 0.42 | 5495 | 23 |
| CD010783 | 0.036 | 0.28 | 10905 | 30 |
| CD010386 | 0.040 | 0.32 | 626 | 2 |
| CD012019 | 0.042 | 0.03 | 10317 | 3 |
| CD010339 | 0.051 | 0.89 | 12807 | 114 |
| CD008081 | 0.076 | 2.68 | 970 | 26 |
| CD007431 | 0.077 | 1.16 | 2074 | 24 |
| CD009786 | 0.078 | 0.48 | 2065 | 10 |
| CD010653 | 0.079 | 0.56 | 8002 | 45 |
| CD010276 | 0.094 | 0.98 | 5495 | 54 |
| CD008782 | 0.096 | 0.43 | 10507 | 45 |
| CD009647 | 0.096 | 2.01 | 2785 | 56 |
| CD009372 | 0.102 | 1.11 | 2248 | 25 |
| CD011145 | 0.107 | 1.86 | 10872 | 202 |
| CD010896 | 0.119 | 3.55 | 169 | 6 |
| CD008803 | 0.132 | 1.90 | 5220 | 99 |
| CD010633 | 0.146 | 0.25 | 1573 | 4 |
| CD010542 | 0.149 | 5.75 | 348 | 20 |
| CD009551 | 0.156 | 2.41 | 1911 | 46 |
| CD009519 | 0.158 | 1.74 | 5971 | 104 |
| CD009185 | 0.254 | 5.70 | 1615 | 92 |
| CD010775 | 0.266 | 4.56 | 241 | 11 |
| CD009925 | 0.269 | 7.04 | 6531 | 460 |
| CD010023 | 0.290 | 5.30 | 981 | 52 |
| CD010860 | 0.310 | 7.45 | 94 | 7 |
| CD009579 | 0.317 | 2.14 | 6455 | 138 |
| CD009135 | 0.351 | 9.73 | 791 | 77 |
| CD010772 | 0.395 | 14.87 | 316 | 47 |
| CD008760 | 0.423 | 18.75 | 64 | 12 |
| CD010705 | 0.524 | 20.18 | 114 | 23 |

**Table 5.** Average Average Precision (AAP) per topic as a measure of topic difficulty, along with statistics about relevant documents.

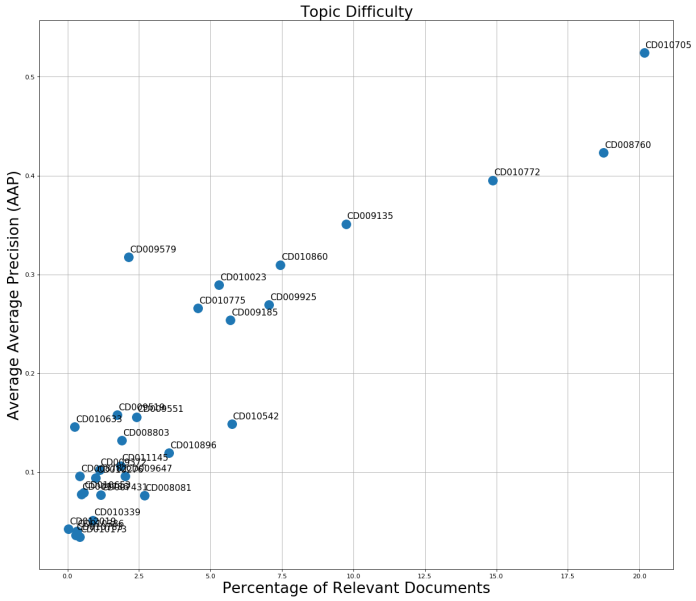


**Fig. 6.** Average Average Precision (AAP) as a function of the percentage of relevant documents.

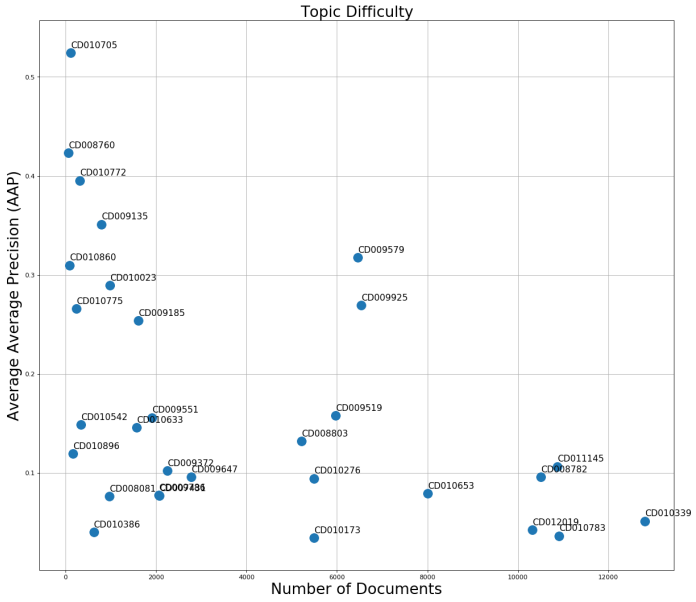


**Fig. 7.** Average Average Precision (AAP) as a function of the total number of documents.

## 7 Conclusions

The CLEF 2017 e-Health Lab Task 2 constructed a benchmark collection of 50 Diagnostic Test Accuracy systematic reviews to study the effectiveness and efficiency of information retrieval and machine learning algorithms in prioritizing the studies to be screened at the abstract and title screening stage, and providing a stopping criterion over the ranked list. The results demonstrate that automatic methods can be trusted for finding most, if not all, relevant studies in a fraction of the time manual screening can do the same. Given that across different runs many parameters change simultaneously it is not easy to come to certain conclusions about the relative performance of automatic methods.

Regarding the benchmark collection itself, there is a number of limitations to be considered: (a) Pivoting on the results of the the OVID MEDLINE Boolean query limits our ability to identify all relevant studies, i.e. relevant studies that are outputted by Boolean queries over different databases, and relevant studies that are actually not found by these Boolean queries. The former can be overcome by considering all the different queries submitted; for the latter extra manual judgments would be required. (b) Pivoting on abstract and title only we miss the opportunity to study the effect of automatic methods when applied to the full text of the studies, that would present an opportunity to completely overcome the multi-stage process of systematic reviews. However, most of the full text articles are protected under copyright laws that do not give all participants access to those. (c) The evaluation setup of ranking does not allows us to consider the cost of the process, since given a ranking a researcher would have to still go over all studies ranked. A more realistic setup, e.g. a double-screening setup, could be considered. (d) In the construction of relevant judgments we considered the included and excluded references of the systematic reviews under study, which prevented us to study the noise and disagreement between reviewers. (e) In our effort to allow iterative algorithms, e.g. active learning algorithms, to be submitted, we handed the test sets' relevant judgments directly to the participants, which is rather unusual for this type of evaluation exercises. An alternative would be the setup used by the TREC Total Recall, where participants submitted their running algorithms to the organizers. (f) When it comes to evaluation measures there is a large variety of those, all of which take a different often useful view point on the effectiveness of algorithm, but which makes it difficult to decide upon a single golden measure to rank participants' runs.

## References


1. Alharbi, A., Stevenson, M.: Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to clef ehealth 2017 task 2. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
2. van Altena, A.J.: Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In: Working Notes of CLEF 2017

- Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
3. Anagnostou, A., Lagopoulos, A., Tsoumakas, G., Vlahavas, I.: Hybridranksvm: A cost-effective hybrid ltr approach for document ranking. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
4. Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., He, L.: Ecnu at 2017 ehealth task 2: Technologically assisted reviews in empirical medicine. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
5. Cohen, A.M., Hersh, W.R., Peterson, K., Yen, P.Y.: Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association 13(2), 206–219 (2006)
6. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 75–84. SIGIR '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2911451.2911510
7. Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
8. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth evaluation lab overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (September 2017)
9. Hollmann, N., Eickhoff, C.: Relevance-based stopping for recall-centric medical document retrieval. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (Oct 2002), http://doi.acm.org/10.1145/582415.582418
11. Kalphov, V., Georgiadis, G., Azzopardi, L.: Sis at clef 2017 ehealth tar task. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
12. Lee, G.E.: Medical document classification for systematic reviews using convolutional neural networks: Sysreview at clef ehealth 2017. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
13. Norman, C., Leeflang, M., Neveol, A.: Limsi@clef ehealth 2017 task 2: Logistic regression for automatic article ranking. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
14. Nunzio, G.M.D., Beghini, F., Vezzani, F., Henrot, G.: An interactive two-dimensional approach to query aspects rewriting in systematic reviews. ims unipd at clef ehealth task 2. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)

15. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews 4(1), 5 (2015)
16. Scells, H., Zuccon, G., Deacon, A., Koopman, B.: Qut ielab at clef 2017 technology assisted reviews track: Initial experiments with learning to rank. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
17. Scells, H., Zuccon, G., Koopman, B., Deacon, A., Geva, S., Azzopardi, L.: A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In: To appear in Proceedings of the 40th international ACM SIGIR conference on Research and development in Information Retrieval. ACM (2017)
18. Shemilt, I., Khan, N., Park, S., Thomas, J.: Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Systematic Reviews 5(1), 140 (Aug 2016)
19. Singh, G., Marshall, I., Thomas, J., Wallace, B.: Identifying diagnostic test accuracy publications using a deep model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
20. Yu, Z., Menzies, T.: Technologically assisted reviews in empirical medicine: Data balancing or reweighting. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)

| | |
|---|---|
| 10.1002/14651858.CD010438.pub2/full | 10.1002/14651858.CD009551.pub3/full |
| 10.1002/14651858.CD010775.pub2/full | 10.1002/14651858.CD012019/full |
| 10.1002/14651858.CD009175.pub2/full | 10.1002/14651858.CD008686.pub2/full |
| 10.1002/14651858.CD011984/full | 10.1002/14651858.CD009020.pub2/full |
| 10.1002/14651858.CD009786.pub2/full | 10.1002/14651858.CD011548/full |
| 10.1002/14651858.CD008643.pub2/full | 10.1002/14651858.CD010896.pub2/full |
| 10.1002/14651858.CD009579.pub2/full | 10.1002/14651858.CD010023.pub2/full |
| 10.1002/14651858.CD009925/full | 10.1002/14651858.CD010772.pub2/full |
| 10.1002/14651858.CD009944.pub2/full | 10.1002/14651858.CD011145.pub2/full |
| 10.1002/14651858.CD007431.pub2/full | 10.1002/14651858.CD010409.pub2/full |
| 10.1002/14651858.CD007427.pub2/full | 10.1002/14651858.CD008054.pub2/full |
| 10.1002/14651858.CD008803.pub2/full | 10.1002/14651858.CD010771.pub2/full |
| 10.1002/14651858.CD008122.pub2/full | 10.1002/14651858.CD009694.pub2/full |
| 10.1002/14651858.CD009593.pub3/full | 10.1002/14651858.CD010705.pub2/full |
| 10.1002/14651858.CD008782.pub4/full | 10.1002/14651858.CD010633.pub2/full |
| 10.1002/14651858.CD009647.pub2/full | 10.1002/14651858.CD010173.pub2/full |
| 10.1002/14651858.CD009135.pub2/full | 10.1002/14651858.CD009591.pub2/full |
| 10.1002/14651858.CD008760.pub2/full | 10.1002/14651858.CD010386.pub2/full |
| 10.1002/14651858.CD011549/full | 10.1002/14651858.CD011021.pub2/full |
| 10.1002/14651858.CD009263.pub2/full | 10.1002/14651858.CD008691.pub2/full |
| 10.1002/14651858.CD009519.pub2/full | 10.1002/14651858.CD010632.pub2/full |
| 10.1002/14651858.CD009372.pub2/full | 10.1002/14651858.CD007394.pub2/full |
| 10.1002/14651858.CD011134.pub2/full | 10.1002/14651858.CD010783.pub2/full |
| 10.1002/14651858.CD010079.pub2/full | 10.1002/14651858.CD010860.pub2/full |
| 10.1002/14651858.CD010276.pub2/full | 10.1002/14651858.CD007424.pub2/full |
| 10.1002/14651858.CD008081.pub3/full | 10.1002/14651858.CD011431/full |
| 10.1002/14651858.CD009185.pub2/full | 10.1002/14651858.CD010339.pub2/full |
| 10.1002/14651858.CD011975/full | 10.1002/14651858.CD010653.pub2/full |
| 10.1002/14651858.CD009323.pub2/full | 10.1002/14651858.CD010542.pub2/full |

**Table 6.** The DOI's of the studies considered for the construction of the benchmark collection

| Run | Docs Shown | Feedback | Rel Docs Found | Last Rank Rel | wss@100 | wss@95 | Cost w/ Uniform Penalty | Cost w/ Weighted Penalty | Area Under Recall | AP | Recall@ Thresh | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| amc.run | 117548 | 0 | 1857 | 2913 | 0.25 | 0.33 | 3918 | 3918 | 0.76 | 0.13 | 1.00 | 0.54 |
| auth.simple.run1 | 117561 | 41337 | 1857 | 2143 | 0.52 | 0.69 | 6674 | 6674 | 0.93 | 0.30 | 1.00 | 0.54 |
| auth.simple.run2 | 117561 | 41377 | 1857 | 2124 | 0.52 | 0.70 | 6677 | 6677 | 0.92 | 0.29 | 1.00 | 0.54 |
| auth.simple.run3 | 117561 | 23337 | 1857 | 2183 | 0.51 | 0.68 | 5474 | 5474 | 0.92 | 0.28 | 1.00 | 0.54 |
| auth.simple.run4 | 117561 | 41537 | 1857 | 2119 | 0.52 | 0.69 | 6687 | 6687 | 0.92 | 0.29 | 1.00 | 0.54 |
| BASELINE.BM25 | 117550 | 0 | 1857 | 2851 | 0.28 | 0.40 | 3918 | 3918 | 0.81 | 0.17 | 1.00 | 0.54 |
| BASELINE.pubmed.random | 117562 | 0 | 1857 | 3722 | 0.04 | 0.03 | 3918 | 3918 | 0.48 | 0.04 | 1.00 | 0.54 |
| cnrs.abrupt.all | 117557 | 19980 | 1857 | 3414 | 0.17 | 0.24 | 5250 | 5250 | 0.73 | 0.14 | 1.00 | 0.54 |
| cnrs.gradual.all | 117557 | 23683 | 1857 | 3406 | 0.20 | 0.29 | 5497 | 5497 | 0.71 | 0.15 | 1.00 | 0.54 |
| cnrs.noaf.all | 117557 | 0 | 1857 | 2993 | 0.26 | 0.36 | 3918 | 3918 | 0.78 | 0.14 | 1.00 | 0.54 |
| cnrs.noaffull.all | 117557 | 0 | 1857 | 2250 | 0.41 | 0.50 | 3918 | 3918 | 0.84 | 0.18 | 1.00 | 0.54 |
| ecnu.run1 | 117561 | 0 | 1857 | 3633 | 0.10 | 0.12 | 3918 | 3918 | 0.63 | 0.09 | 1.00 | 0.54 |
| ecnu.run2 | 30000 | 0 | 1191 | 699 | 0.07 | 0.16 | 4003 | 6641 | 0.64 | 0.16 | 0.71 | 0.44 |
| ecnu.run3 | 30000 | 0 | 1197 | 725 | 0.08 | 0.17 | 4016 | 6717 | 0.65 | 0.17 | 0.72 | 0.44 |
| eth.m1 | 51640 | 0 | 1686 | 1372 | 0.24 | 0.28 | 2306 | 4740 | 0.81 | 0.22 | 0.93 | 0.20 |
| eth.m2 | 51604 | 5063 | 1702 | 1435 | 0.14 | 0.24 | 2676 | 4720 | 0.80 | 0.21 | 0.90 | 0.14 |
| eth.m4 | 27046 | 0 | 1406 | 785 | 0.12 | 0.16 | 2527 | 5590 | 0.74 | 0.21 | 0.82 | 0.14 |
| iiit.run1 | 15354 | 15354 | 1006 | 548 | 0.11 | 0.14 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run2 | 15354 | 15354 | 1006 | 548 | 0.11 | 0.14 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run3 | 15354 | 15354 | 1006 | 548 | 0.11 | 0.14 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| iiit.run4 | 15354 | 15354 | 1006 | 548 | 0.11 | 0.14 | 3550 | 6685 | 0.68 | 0.16 | 0.74 | 0.15 |
| ncsu.abs | 12942 | 12942 | 1073 | 378 | 0.12 | 0.16 | 4409 | 7695 | 0.61 | 0.11 | 0.71 | 0.33 |
| ncsu.simple | 27950 | 27950 | 1611 | 928 | 0.14 | 0.27 | 4145 | 6964 | 0.68 | 0.11 | 0.83 | 0.18 |
| ntu.run1 | 111170 | 0 | 1795 | 3403 | 0.09 | 0.11 | 3936 | 4130 | 0.61 | 0.08 | 0.98 | 0.55 |
| ntu.run2 | 111170 | 0 | 1795 | 3204 | 0.12 | 0.13 | 3936 | 4130 | 0.59 | 0.06 | 0.98 | 0.55 |
| ntu.run3 | 111196 | 0 | 1795 | 3570 | 0.09 | 0.07 | 3937 | 4130 | 0.54 | 0.05 | 0.98 | 0.55 |
| padua.iafa_m10k150f0m10 | 117557 | 2350 | 1857 | 2269 | 0.41 | 0.51 | 4075 | 4075 | 0.90 | 0.28 | 1.00 | 0.54 |
| padua.iafap_m10p2f0m10 | 117557 | 2367 | 1857 | 2395 | 0.37 | 0.48 | 4076 | 4076 | 0.88 | 0.25 | 1.00 | 0.54 |
| padua.iafap_m10p5f0m10 | 117557 | 5893 | 1857 | 2260 | 0.40 | 0.50 | 4311 | 4311 | 0.89 | 0.27 | 1.00 | 0.54 |
| padua.iafas_m10k50f0m10 | 117557 | 4320 | 1857 | 2304 | 0.41 | 0.52 | 4206 | 4206 | 0.89 | 0.27 | 1.00 | 0.54 |

**Table 7.** PART I: Evaluation results for submitted runs computed using abstract-level relevance judgments

| Run | Docs Shown | Feedback | Rel Docs Found | Last Rank Rel | wss@100 | wss@95 | Cost w/ Uniform Penalty | Cost w/ Weighted Penalty | Area Under Recall | AP | Recall@ Thresh | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| qut.ca_bool_ltr | 117557 | 0 | 1857 | 3142 | 0.20 | 0.29 | 3918 | 3918 | 0.73 | 0.11 | 1.00 | 0.54 |
| qut.ca_pico_ltr | 117557 | 0 | 1857 | 3344 | 0.21 | 0.29 | 3918 | 3918 | 0.75 | 0.15 | 1.00 | 0.54 |
| qut.rf_bool_ltr | 117557 | 0 | 1857 | 3099 | 0.19 | 0.27 | 3918 | 3918 | 0.70 | 0.11 | 1.00 | 0.54 |
| qut.fr_pico_ltr | 117557 | 0 | 1857 | 3155 | 0.23 | 0.29 | 3918 | 3918 | 0.73 | 0.12 | 1.00 | 0.54 |
| qut.bool_es_test | 69951 | 0 | 1475 | 1972 | 0.10 | 0.11 | 3480 | 4976 | 0.64 | 0.13 | 0.76 | 0.36 |
| qut.pico_es_test | 63018 | 0 | 1414 | 1873 | 0.11 | 0.13 | 3527 | 5168 | 0.62 | 0.12 | 0.74 | 0.34 |
| sheffield.run1 | 117562 | 0 | 1857 | 2678 | 0.31 | 0.42 | 3918 | 3918 | 0.82 | 0.17 | 1.00 | 0.54 |
| sheffield.run2 | 117562 | 0 | 1857 | 2441 | 0.39 | 0.49 | 3918 | 3918 | 0.84 | 0.22 | 1.00 | 0.54 |
| sheffield.run3 | 117562 | 0 | 1857 | 2404 | 0.38 | 0.47 | 3918 | 3918 | 0.84 | 0.20 | 1.00 | 0.54 |
| sheffield.run4 | 117562 | 0 | 1857 | 2382 | 0.40 | 0.49 | 3918 | 3918 | 0.85 | 0.22 | 1.00 | 0.54 |
| ucl.run_abstract | 117562 | 0 | 1857 | 3727 | 0.04 | 0.03 | 3918 | 3918 | 0.48 | 0.04 | 1.00 | 0.54 |
| ucl.run_fulltext | 117562 | 0 | 1857 | 3727 | 0.04 | 0.03 | 3918 | 3918 | 0.48 | 0.04 | 1.00 | 0.54 |
| uos.bm25_threshold1 | 103051 | 0 | 1828 | 2503 | 0.28 | 0.40 | 3454 | 3786 | 0.81 | 0.17 | 0.99 | 0.45 |
| uos.bm25_threshold2.5 | 76104 | 0 | 1758 | 1877 | 0.22 | 0.35 | 2905 | 3902 | 0.79 | 0.17 | 0.94 | 0.27 |
| uos.bm25_threshold2 | 84740 | 0 | 1784 | 2068 | 0.23 | 0.37 | 3117 | 3748 | 0.80 | 0.17 | 0.95 | 0.33 |
| uos.sis.AL30Q_BM25 | 94967 | 0 | 1809 | 2333 | 0.27 | 0.39 | 3280 | 3865 | 0.80 | 0.17 | 0.97 | 0.38 |
| uos.sis.TMAL30Q_BM25 | 117551 | 35432 | 1857 | 2305 | 0.40 | 0.53 | 6280 | 6280 | 0.84 | 0.16 | 1.00 | 0.54 |
| uos.sis.TMBEST_BM25 | 117557 | 0 | 1857 | 3124 | 0.27 | 0.32 | 3918 | 3918 | 0.73 | 0.12 | 1.00 | 0.54 |
| waterloo.A-rank-normal | 117558 | 117558 | 1857 | 1464 | 0.60 | 0.70 | 11755 | 11755 | 0.93 | 0.28 | 1.00 | 0.54 |
| waterloo.A-thresh-normal | 87767 | 87767 | 1842 | 1161 | 0.56 | 0.70 | 8809 | 9543 | 0.93 | 0.28 | 1.00 | 0.50 |
| waterloo.B-rank-normal | 117558 | 117558 | 1857 | 1469 | 0.61 | 0.70 | 11755 | 11755 | 0.93 | 0.32 | 1.00 | 0.54 |
| waterloo.B-thresh-normal | 60936 | 60936 | 1548 | 914 | 0.54 | 0.66 | 6470 | 7150 | 0.91 | 0.31 | 0.97 | 0.43 |

**Table 8.** PART II: Evaluation results for submitted runs computed using abstract-level relevance judgments.

| Run | Docs Shown | Feedback | Rel Docs Found | Last Rank Rel | wss@100 | wss@95 | Cost w/ Uniform Penalty | Cost w/ Weighted Penalty | Area Under Recall | AP | Recall@ Thresh | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| amc.run | 109547 | 0 | 607 | 1742 | 0.51 | 0.51 | 3777 | 3777 | 0.84 | 0.10 | 1.00 | 0.74 |
| auth.simple.run1 | 109559 | 39337 | 607 | 853 | 0.80 | 0.82 | 6490 | 6490 | 0.95 | 0.23 | 1.00 | 0.74 |
| auth.simple.run2 | 109559 | 39377 | 607 | 857 | 0.79 | 0.81 | 6493 | 6493 | 0.94 | 0.21 | 1.00 | 0.74 |
| auth.simple.run3 | 109559 | 22337 | 607 | 839 | 0.80 | 0.82 | 5318 | 5318 | 0.95 | 0.22 | 1.00 | 0.74 |
| auth.simple.run4 | 109559 | 39537 | 607 | 858 | 0.79 | 0.81 | 6504 | 6504 | 0.94 | 0.21 | 1.00 | 0.74 |
| BASELINE.BM25 | 109549 | 0 | 607 | 1664 | 0.54 | 0.57 | 3777 | 3777 | 0.85 | 0.14 | 1.00 | 0.74 |
| BASELINE.pubmed.random | 109560 | 0 | 607 | 3316 | 0.09 | 0.07 | 3777 | 3777 | 0.48 | 0.02 | 1.00 | 0.74 |
| cnrs.abrupt.all | 109555 | 19980 | 607 | 2619 | 0.35 | 0.39 | 5155 | 5155 | 0.80 | 0.11 | 1.00 | 0.74 |
| cnrs.gradual.all | 109555 | 22684 | 607 | 2384 | 0.41 | 0.46 | 5342 | 5342 | 0.77 | 0.11 | 1.00 | 0.74 |
| cnrs.noaf.all | 109555 | 0 | 607 | 2263 | 0.42 | 0.50 | 3777 | 3777 | 0.82 | 0.10 | 1.00 | 0.74 |
| cnrs.noaffull.all | 109555 | 0 | 607 | 1678 | 0.59 | 0.64 | 3777 | 3777 | 0.89 | 0.13 | 1.00 | 0.74 |
| ecnu.run1 | 109559 | 0 | 607 | 2905 | 0.26 | 0.27 | 3777 | 3777 | 0.66 | 0.06 | 1.00 | 0.74 |
| ecnu.run2 | 29000 | 0 | 426 | 515 | 0.29 | 0.30 | 3069 | 5286 | 0.72 | 0.12 | 0.79 | 0.49 |
| ecnu.run3 | 29000 | 0 | 426 | 486 | 0.30 | 0.31 | 3069 | 5286 | 0.73 | 0.12 | 0.79 | 0.49 |
| eth.m1 | 47500 | 0 | 561 | 1000 | 0.44 | 0.58 | 1876 | 2170 | 0.86 | 0.16 | 0.97 | 0.24 |
| eth.m2 | 46538 | 4662 | 553 | 1050 | 0.42 | 0.55 | 2196 | 2489 | 0.84 | 0.15 | 0.93 | 0.18 |
| eth.m4 | 25381 | 0 | 476 | 596 | 0.31 | 0.36 | 1850 | 3739 | 0.80 | 0.15 | 0.87 | 0.15 |
| iiit.run1 | 15234 | 15234 | 406 | 501 | 0.15 | 0.19 | 3583 | 5094 | 0.70 | 0.12 | 0.77 | 0.18 |
| iiit.run2 | 15234 | 15234 | 406 | 501 | 0.15 | 0.19 | 3583 | 5094 | 0.70 | 0.12 | 0.77 | 0.18 |
| iiit.run3 | 15234 | 15234 | 406 | 501 | 0.15 | 0.19 | 3583 | 5094 | 0.70 | 0.12 | 0.77 | 0.18 |
| iiit.run4 | 15234 | 15234 | 406 | 501 | 0.15 | 0.19 | 3583 | 5094 | 0.70 | 0.12 | 0.77 | 0.18 |
| ncsu.abs | 12682 | 12682 | 480 | 356 | 0.35 | 0.38 | 3354 | 5978 | 0.69 | 0.07 | 0.81 | 0.31 |
| ncsu.simple | 27950 | 27950 | 607 | 960 | 0.66 | 0.66 | 2891 | 2891 | 0.80 | 0.06 | 1.00 | 0.13 |
| ntu.run1 | 103170 | 0 | 606 | 2954 | 0.20 | 0.22 | 3606 | 3557 | 0.65 | 0.05 | 1.00 | 0.72 |
| ntu.run2 | 103170 | 0 | 606 | 2779 | 0.20 | 0.20 | 3606 | 3557 | 0.62 | 0.04 | 1.00 | 0.72 |
| ntu.run3 | 103194 | 0 | 606 | 3050 | 0.15 | 0.14 | 3607 | 3558 | 0.55 | 0.02 | 1.00 | 0.72 |
| padua.iafa_m10k150f0m10 | 109555 | 2286 | 607 | 1055 | 0.71 | 0.71 | 3935 | 3935 | 0.93 | 0.22 | 1.00 | 0.74 |
| padua.iafap_m10p2f0m10 | 109555 | 2206 | 607 | 1007 | 0.66 | 0.69 | 3929 | 3929 | 0.92 | 0.20 | 1.00 | 0.74 |
| padua.iafap_m10p5f0m10 | 109555 | 5492 | 607 | 838 | 0.71 | 0.70 | 4156 | 4156 | 0.93 | 0.21 | 1.00 | 0.74 |
| padua.iafas_m10k50f0m10 | 109555 | 4170 | 607 | 990 | 0.71 | 0.72 | 4065 | 4065 | 0.93 | 0.19 | 1.00 | 0.74 |

**Table 9.** PART I: Evaluation results for submitted runs computed using document-level relevance judgments.

| Run | Docs Shown | Feedback | Rel Docs Found | Last Rank Rel | wss@100 | wss@95 | Cost w/ Uniform Penalty | Cost w/ Weighted Penalty | Area Under Recall | AP | Recall@ Thresh | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| qut.ca_bool_ltr | 109555 | 0 | 607 | 2582 | 0.33 | 0.36 | 3777 | 3777 | 0.75 | 0.08 | 1.00 | 0.74 |
| qut.ca_pico_ltr | 109555 | 0 | 607 | 2638 | 0.36 | 0.40 | 3777 | 3777 | 0.78 | 0.11 | 1.00 | 0.74 |
| qut.rf_bool_ltr | 109555 | 0 | 607 | 2477 | 0.33 | 0.35 | 3777 | 3777 | 0.72 | 0.06 | 1.00 | 0.74 |
| qut.rf_pico_ltr | 109555 | 0 | 607 | 2610 | 0.35 | 0.38 | 3777 | 3777 | 0.76 | 0.09 | 1.00 | 0.74 |
| qut.bool_es | 65389 | 0 | 465 | 1595 | 0.22 | 0.25 | 3217 | 4041 | 0.68 | 0.10 | 0.81 | 0.43 |
| qut.pico_es | 58456 | 0 | 451 | 1424 | 0.21 | 0.23 | 3251 | 4519 | 0.67 | 0.09 | 0.78 | 0.40 |
| sheffield.run1 | 109560 | 0 | 607 | 1801 | 0.52 | 0.54 | 3777 | 3777 | 0.84 | 0.12 | 1.00 | 0.74 |
| sheffield.run2 | 109560 | 0 | 607 | 1928 | 0.53 | 0.58 | 3777 | 3777 | 0.87 | 0.18 | 1.00 | 0.74 |
| sheffield.run3 | 109560 | 0 | 607 | 1902 | 0.52 | 0.59 | 3777 | 3777 | 0.87 | 0.15 | 1.00 | 0.74 |
| sheffield.run4 | 109560 | 0 | 607 | 1846 | 0.54 | 0.59 | 3777 | 3777 | 0.87 | 0.18 | 1.00 | 0.74 |
| ucl.run_abstract | 109560 | 0 | 607 | 3472 | 0.13 | 0.12 | 3777 | 3777 | 0.51 | 0.04 | 1.00 | 0.74 |
| ucl.run_fulltext | 109560 | 0 | 607 | 3505 | 0.14 | 0.12 | 3777 | 3777 | 0.51 | 0.04 | 1.00 | 0.74 |
| uos.bm25_threshold1 | 95721 | 0 | 601 | 1548 | 0.53 | 0.57 | 3311 | 3406 | 0.85 | 0.14 | 0.99 | 0.59 |
| uos.bm25_threshold2.5 | 73548 | 0 | 591 | 1483 | 0.52 | 0.56 | 2580 | 2926 | 0.85 | 0.13 | 0.98 | 0.36 |
| uos.bm25_threshold2 | 80976 | 0 | 597 | 1506 | 0.53 | 0.57 | 2820 | 3037 | 0.85 | 0.14 | 0.99 | 0.45 |
| uos.sis.AL30Q_BM25 | 109549 | 33300 | 607 | 906 | 0.68 | 0.69 | 6074 | 6074 | 0.90 | 0.16 | 1.00 | 0.74 |
| uos.sis.TMAL30Q_BM25 | 109550 | 33002 | 607 | 1228 | 0.65 | 0.67 | 6053 | 6053 | 0.87 | 0.11 | 1.00 | 0.74 |
| uos.sis.TMBEST_BM25 | 109555 | 0 | 607 | 1980 | 0.50 | 0.49 | 3777 | 3777 | 0.76 | 0.09 | 1.00 | 0.74 |
| waterloo.A-rank-normal | 109556 | 109556 | 607 | 461 | 0.82 | 0.81 | 11333 | 11333 | 0.95 | 0.19 | 1.00 | 0.74 |
| waterloo.A-thresh-normal | 79765 | 79765 | 607 | 461 | 0.82 | 0.81 | 8251 | 8251 | 0.95 | 0.19 | 1.00 | 0.66 |
| waterloo.B-rank-normal | 109556 | 109556 | 607 | 469 | 0.83 | 0.82 | 11333 | 11333 | 0.95 | 0.23 | 1.00 | 0.74 |
| waterloo.B-thresh-normal | 52934 | 52934 | 575 | 375 | 0.78 | 0.77 | 5765 | 6559 | 0.94 | 0.23 | 0.98 | 0.53 |

**Table 10.** PART II: Evaluation results for submitted runs computed using document-level relevance judgments