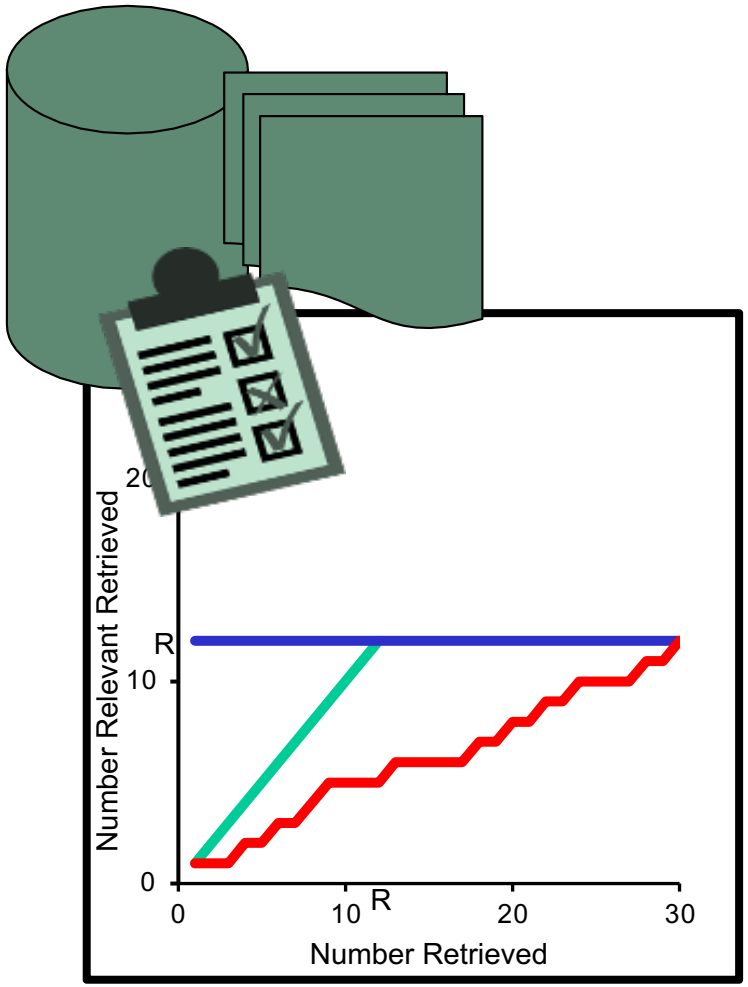
Building Reusable Test Collections

Ellen M. Voorhees

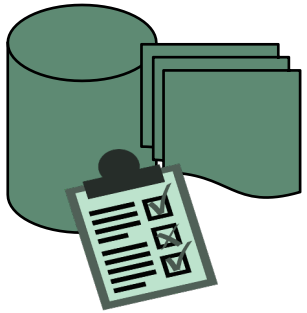
National Institute of Standards and Technology
U.S. Department of Commerce

Test Collections



- Evaluate search effectiveness using test collections
  - set of documents
  - set of questions
  - relevance judgments
- Relevance judgments
  - ideally, complete judgments---all docs for all topics
  - unfeasible for document sets large enough to be interesting
  - so, need to sample, but how?

Problem Statement



Want to build general-purpose, reusable IR test collections at acceptable cost

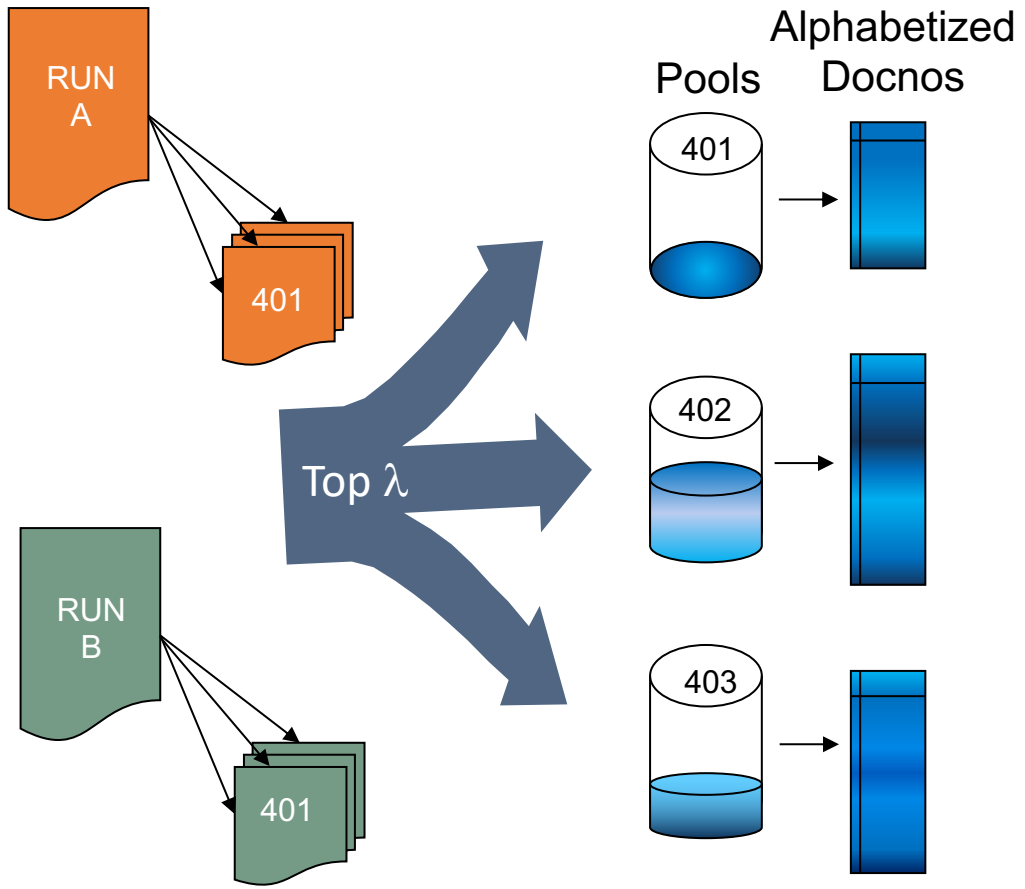
General-purpose: supports a wide range of measures and search scenarios

Reusable: unbiased for systems that were not used to build the collection

Cost: proportional to the number of human judgments required for entire procedure

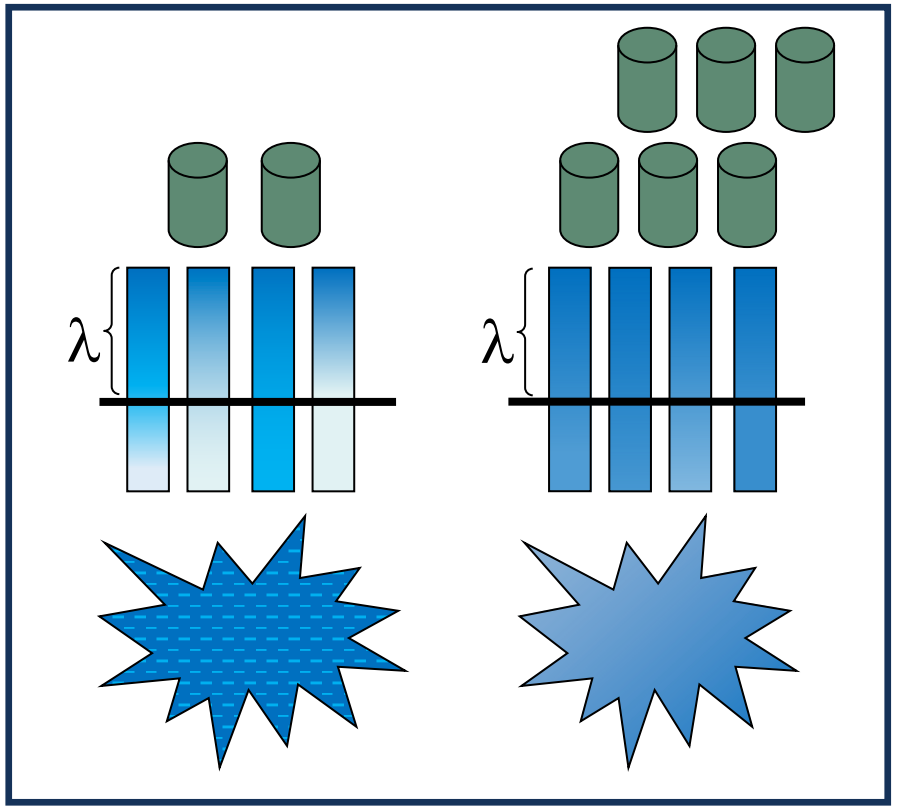
Pooling

- For sufficiently large λ and diverse engines, depth-λ pools produce "essentially complete" judgments
- Unjudged documents are assumed to be not relevant when computing traditional evaluation measures such as average precision (AP)
- Resulting test collections have been found to be both fair and reusable.
  1) fair: no bias against systems used to construct collection
  2) reusable: fair to systems not used in collection construction



Pooling Bias

- Traditional pooling takes top λ documents
  1) intentional bias toward top ranks where relevant are found
  2) λ was originally large enough to reach past swell of topic-word relevant
- As document collection grows, a constant cut-off stays within swell
- Pools cannot be proportional to corpus size due to practical constraints
  1) sample runs differently to build unbiased pools
  2) new evaluation metrics that do not assume complete judgments



C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. *Bias and the limits of pooling for large collections*. **Information Retrieval**, 10(6):491-508, 2007.

LOU Test

"Leave Out Uniques" test of reusability:

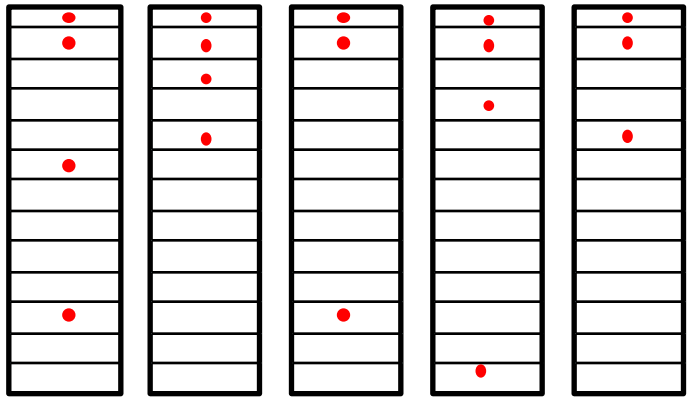
examine effect on test collection if some participating team had not done so

Procedure

- create judgment set that removes all uniquely-retrieved relevant documents for one team
- evaluate all runs using original judgment set and again using newly created set
- compare evaluation results
  - Kendall's τ between system rankings
  - maximum drop in ranking over runs submitted by team

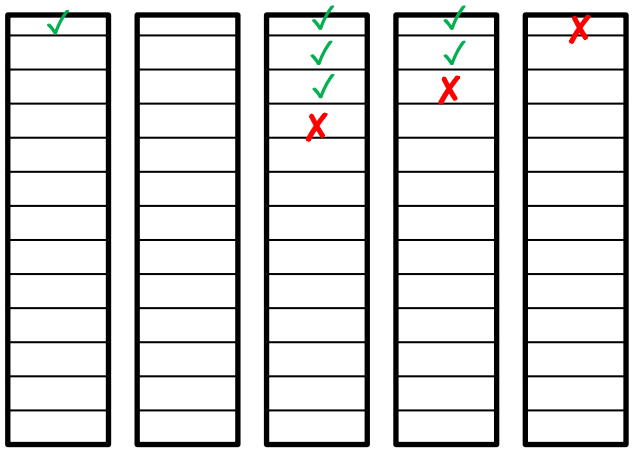
Inferred Measure Sampling

- Stratified sampling where strata are defined by ranks
- Different strata have different probabilities for documents to be selected to be judged
- Given strata and probabilities, estimate AP by inferring which unjudged docs are likely to be relevant
- Quality of estimate varies widely depending on exact sampling strategy
- Fair, but may be less reusable



E. Yilmaz, E. Kanoulas, and J. A. Aslam. *A simple and efficient sampling method for estimating AP and NDCG*. **SIGIR 2008**, pp.603—610.

Multi-armed Bandit Sampling



D. Losada, J. Parapar, A. Barreiro. *Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation*. Proceedings of SAC 2016. pp. 1027-1034.

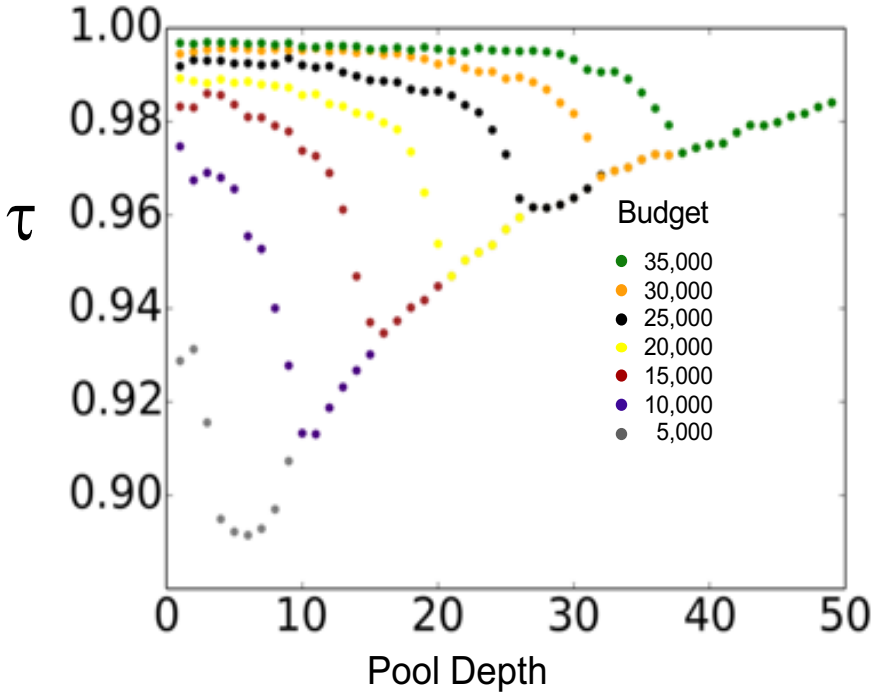
- Bandit techniques trade-off between exploiting known good “arms” and exploring to find better arms. For collection building, each run is an arm, and reward is finding a relevant doc
- Simulations suggest can get similar-quality collections as pooling but with many fewer judgments
- TREC 2017 Common Core track first attempt to build new collection using bandit technique

bandit selection method:
2017: MaxMean 2018: MTF

Implementing a practical bandit approach

How does assessor learn topic?

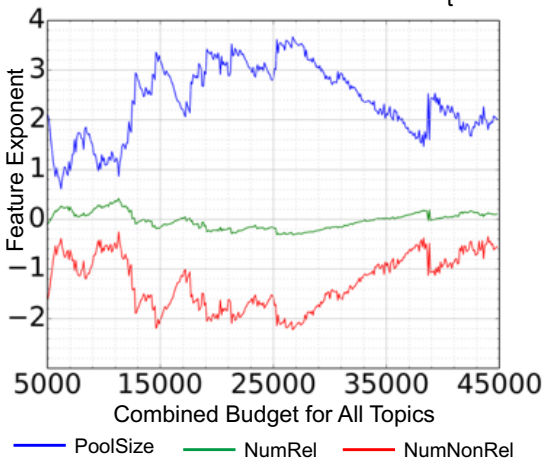
- allocating some budget to shallow pools causes minimal degradation over "pure" bandit method



How should overall budget be divided among topics?

- use features of top-10 pools, to predict per-topic minimum judgments needed

$$\text{Estimate}_t = \frac{\text{PoolSize}_t}{\sqrt{\text{NumNonRel}_t}}$$

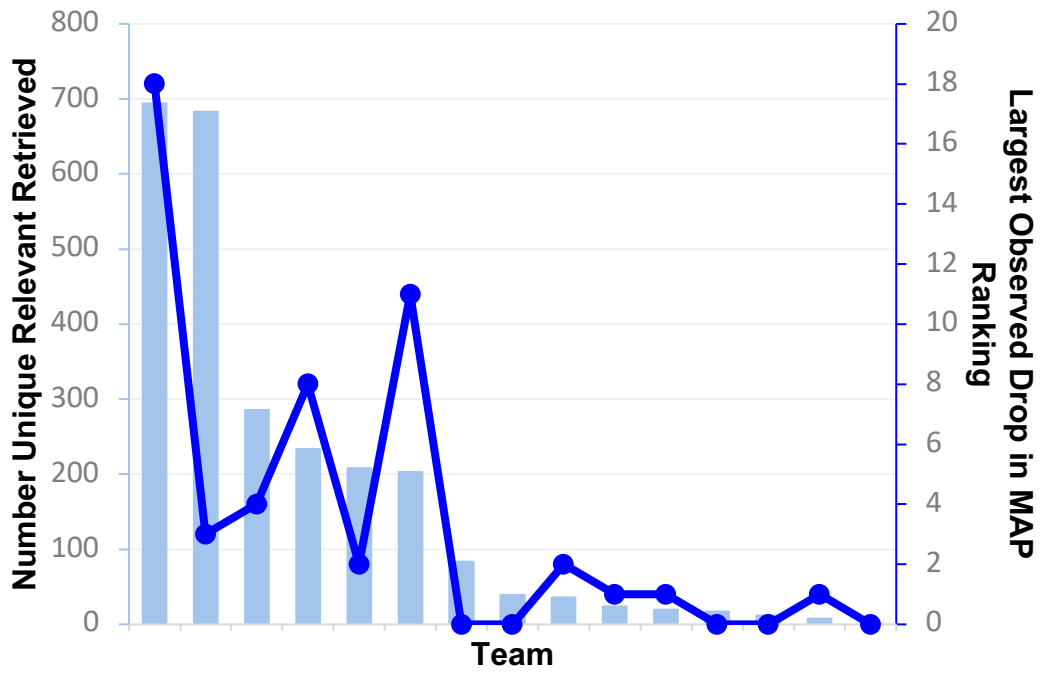


- results in a conservative, but reasonable, allocation of budget across topics for historical collections

Collection Quality

2017 Common Core collection less *reusable* than hoped (just too few judgments)

Additional experiments demonstrate greedy bandit methods can be *UNFAIR*



LOU-results for TREC 2017 Common Core collection

	MAP		Precision(10)	
	τ	Drop	τ	Drop
MaxMean	.980	2	.937	11
Inferred	.961	7	.999	1

Fairness test: build collection from judgments on small inferred-sample or on equal number of documents selected by MaxMean bandit approach (average of 300 judgments per topic). Evaluate runs using respective judgment sets and compare run rankings to full collection rankings. Judgment budget is *small enough that R exceeds budget for some topics*.

Example: topic 389 with R=324, 45% of which are uniques; one run has 98 relevant in top 100 ranks, so 1/3 relevant in bandit set came from this single run to the exclusion of other runs.

An Aside

- Note that this is a concrete example of why the goal in building a collection is **NOT** to maximize the number of relevant found!
- The goal is actually to find an unbiased set of relevant.
- We don't know how to build a guaranteed unbiased judgment set, nor prove that an existing set is unbiased, but sometimes less is more.

Image: Eunice De Faria/Pixabay

Bandit Conclusions

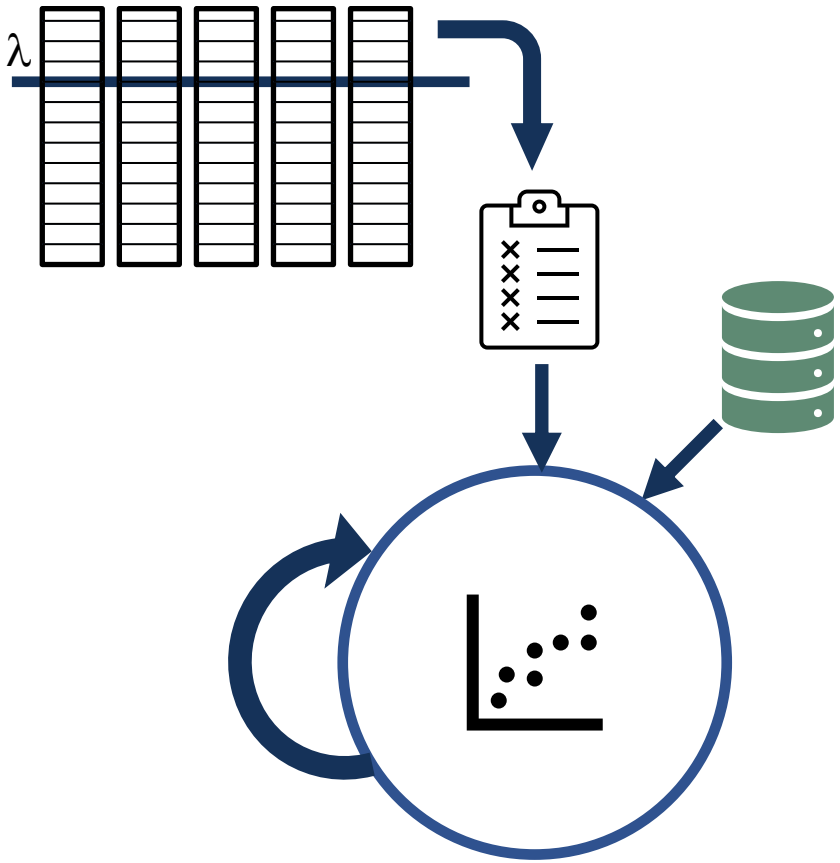
Can be unfair when budget is small relative to (unknown) number of relevant

- must reserve some of budget for quality control, so operative number of judgments is less than B

- Does not provide practical means for coordination among assessors
  - multiple human judges working at different rates and at different times
  - subject to a common overall budget
  - stopping criteria depends on outcome of process

Image: Pascal/flickr

HiCal



Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark Smucker, Gordon Cormack and Maura Grossman. *A System for Efficient High-Recall Retrieval*, **SIGIR 2018**. (https://hical.github.io/)

- TREC 2019 and 2020 Deep Learning track used modification of U. of Waterloo's HiCAL system
- HiCAL dynamic method that builds model of relevance based on available judgments. Suggests first most-likely-to-be-relevant unjudged document as next to judge.
- Modified version used in tracks: start with depth-10 pools
- Judge initial pools and
  - 2019: 300 document sample selected by StatMap; estimate R
  - 2020: 100 additional docs selected by HiCAL

Iterate until $2estR+100<|J|$ or $estR\sim|J|$

HiCAL Collection Quality?

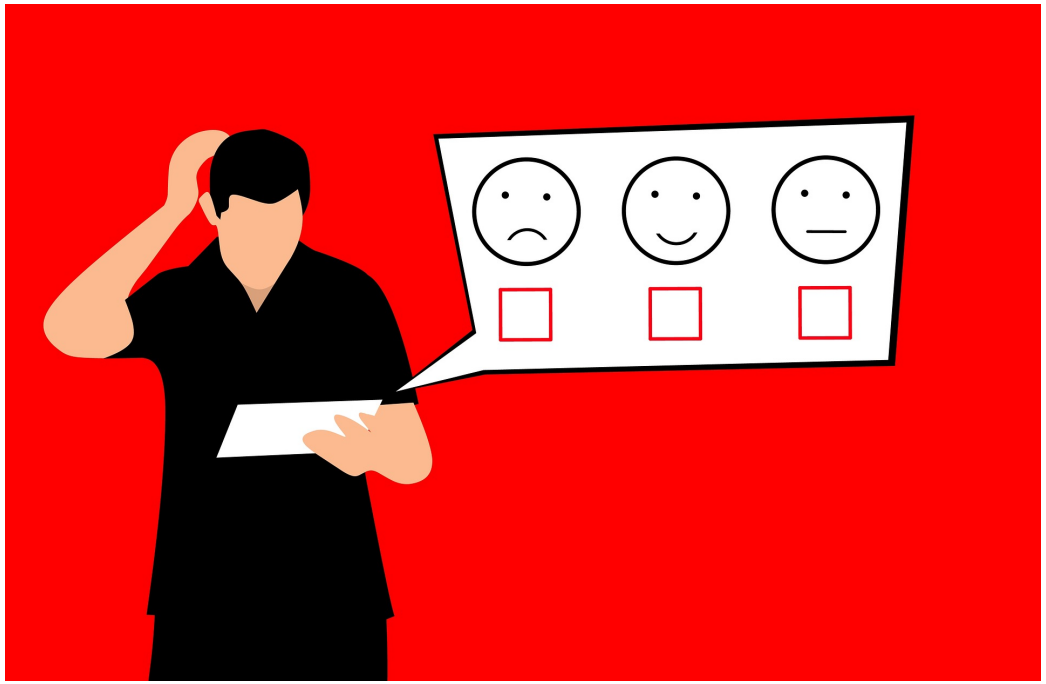
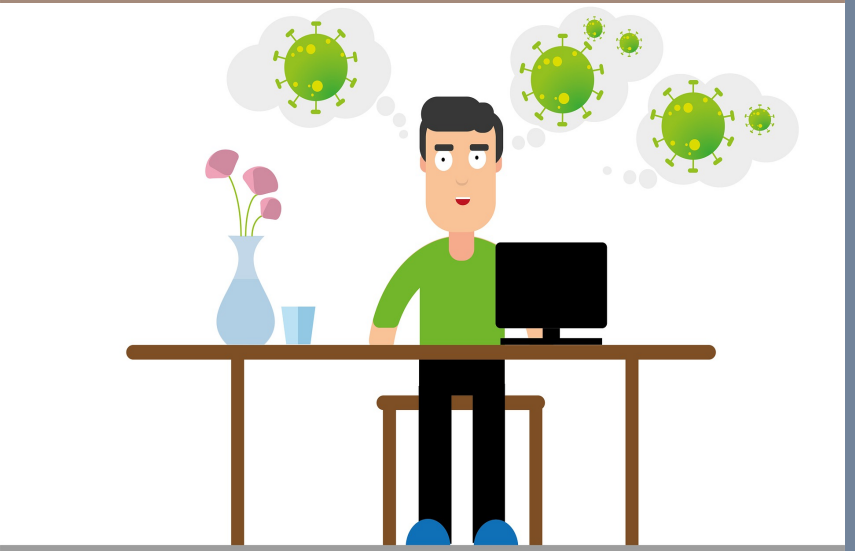


Image: mohamed Hassan/Pixabay

- Hard to say in the absence of Truth
- Concept of uniquely retrieved relevant docs not defined, so no LOU testing
  - can leave out team from entire process, but HiCAL able to recover those docs
  - of 5760 tests for cross product of {team}X{map,P_10}X{trec8,robust,deep}X{stopping criterion} exactly one τ was less than 0.92
- Very few topics enter a second iteration
- So: Deep Learning track collections are fair, probably (?) reusable, but unknown effect of topic sample

TREC-COVID

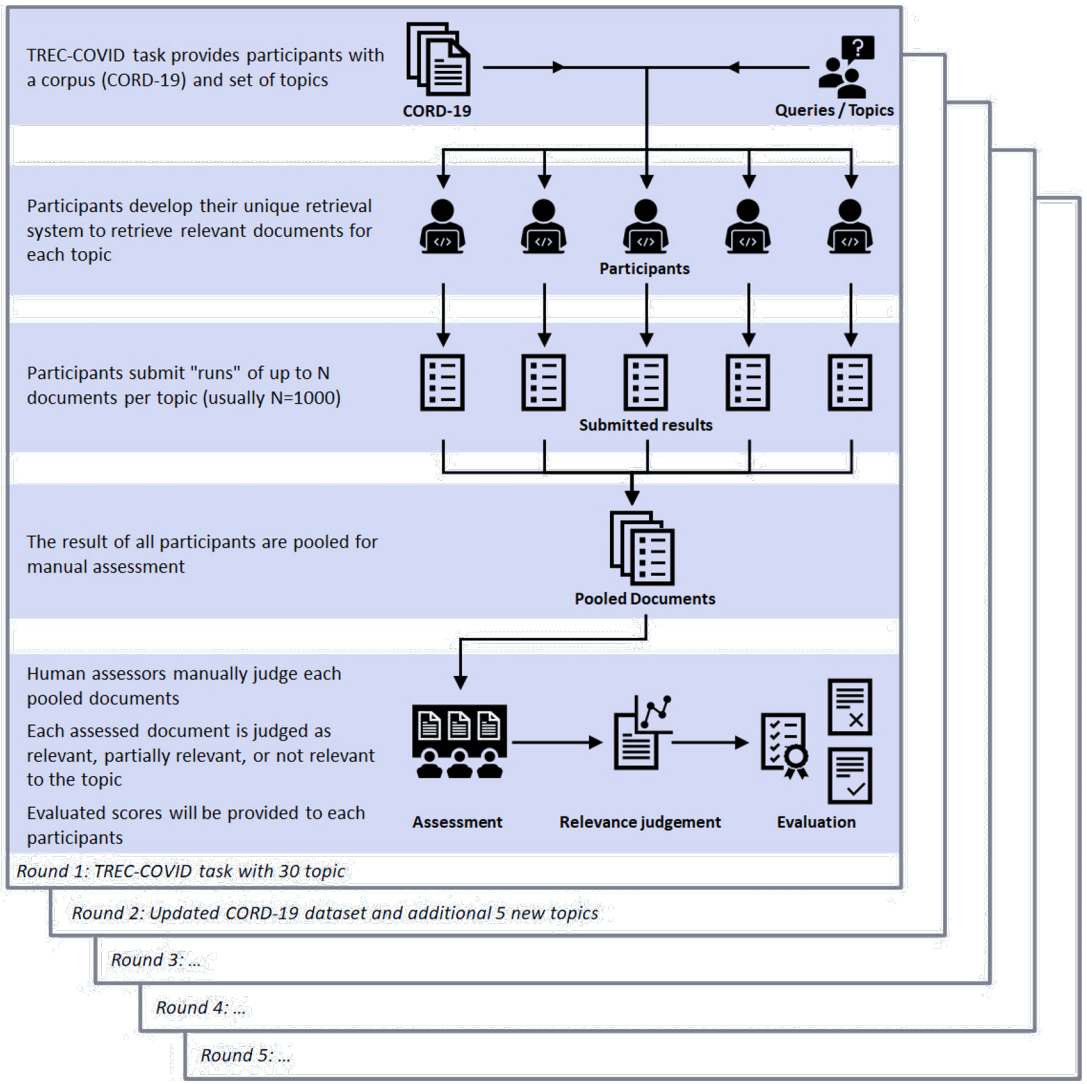


Images: Alexandra Koch/Pixabay

TREC-COVID: build a pandemic test collection for current and future biomedical crises…

…in a very short time frame using open-source literature on COVID-19

TREC-COVID



- Structured as a series of rounds, where each round uses a superset of previous rounds' document and question sets.
- The document set is CORD-19 maintained by AI2. Questions came from search logs of medical libraries. Judgments from people with biomedical expertise.

TREC-COVID Rounds

Round 1

- Apr 15–Apr 23
- Apr 10 release of CORD-19; ~47k articles
- 30 topics
- 56 teams, 143 submissions
- ~8.5k judgments

Round 2

- May 4—May 13
- May 1 release of CORD-19; ~60k articles
- 35 topics
- 51 teams, 136 submissions
- ~20k cumulative judgments

Round 3

- May 26—Jun 3
- May 19 release of CORD-19; ~128k articles
- 40 topics
- 31 teams, 79 submissions
- ~33k cumulative judgments

Round 4

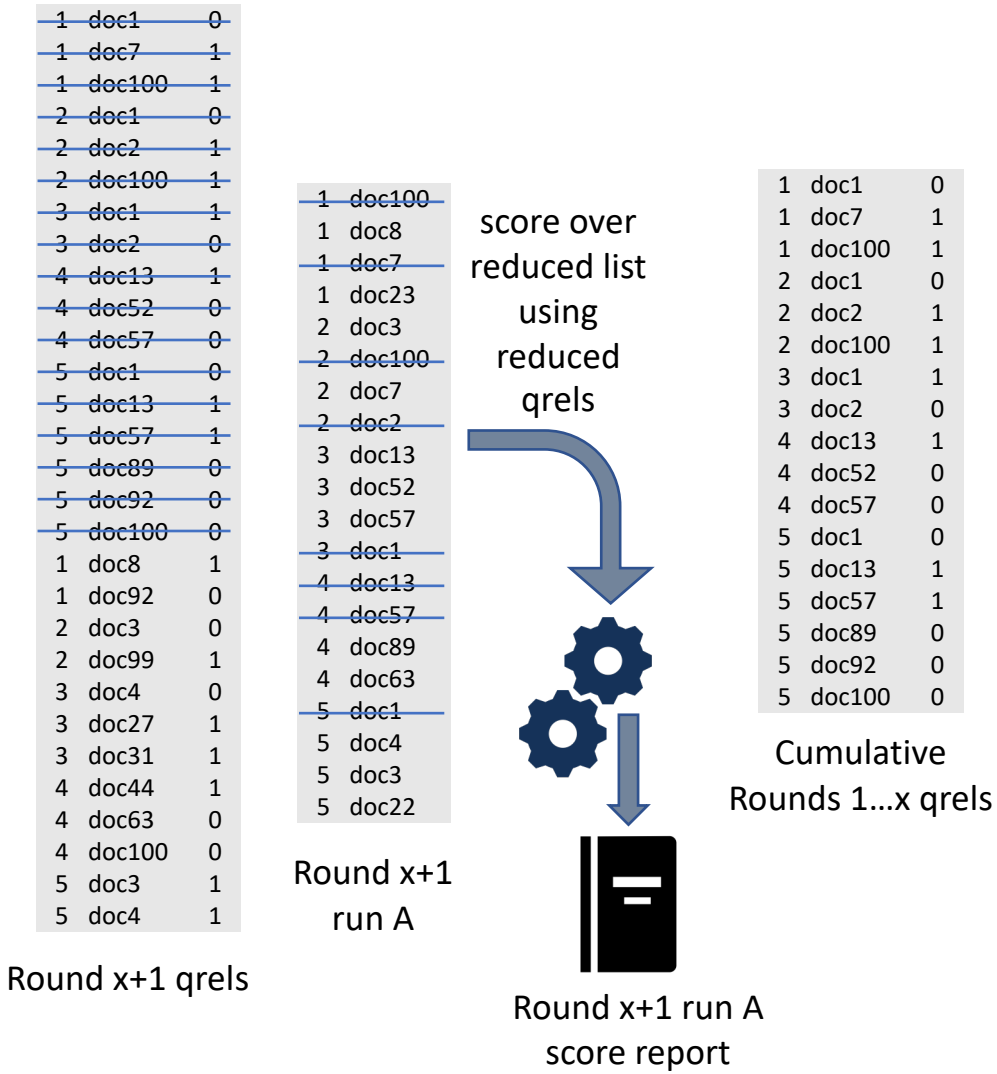
- Jun 26—Jul 6
- Jun 19 release of CORD-19; ~158k articles
- 45 topics
- 27 teams, 72 submissions
- ~46k cumulative judgments

Round 5

- Jul 22—Aug 3
- Jul 16 release of CORD-19; ~191k articles
- 50 topics
- 28 teams, 126 submissions
- ~69k cumulative judgments

Residual Collection Eval

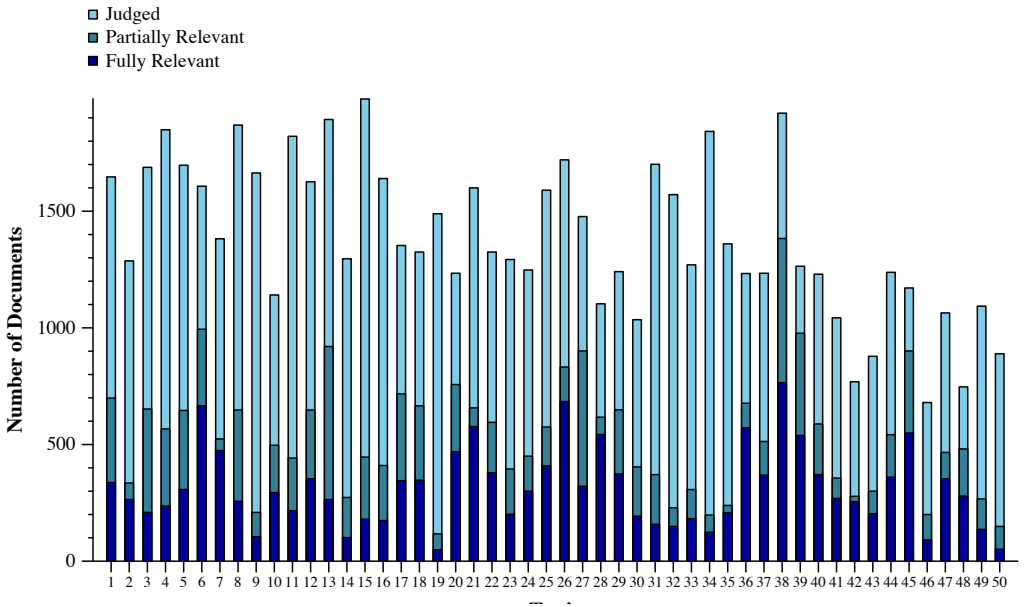
- Using residual collection evaluation: all previously judged documents for a given topic are removed from the collection for scoring
  - prevents gaming of scoring metrics (including inadvertent correlations)
  - can depress absolute values of scores, but relative scores okay
  - bookkeeping of previously judged documents important
  - each rounds' submissions scored only on that round's judgments, not cumulative qrels, so less stable



Judgment Sets

- Per-round judgment period initially short
  - two “half-rounds” of judging per round of TREC-COVID
  - less than 10 days per half-round
  - estimated could judge about 100 documents per topic per half-round
  - SHALLOW pools over large, diverse run sets
- Extended judgment period in later round
  - fewer runs, but still diverse and largely effective (feedback)
  - individual rounds’ pools and thus qrels still just shallow pools because of residual collection evaluation
- TREC-COVID Complete
  - approximately 69k judgments built from multiple rounds of feedback runs
  - 50 topics: topics added in later rounds received relatively more judgments in subsequent rounds to even out assessment effort

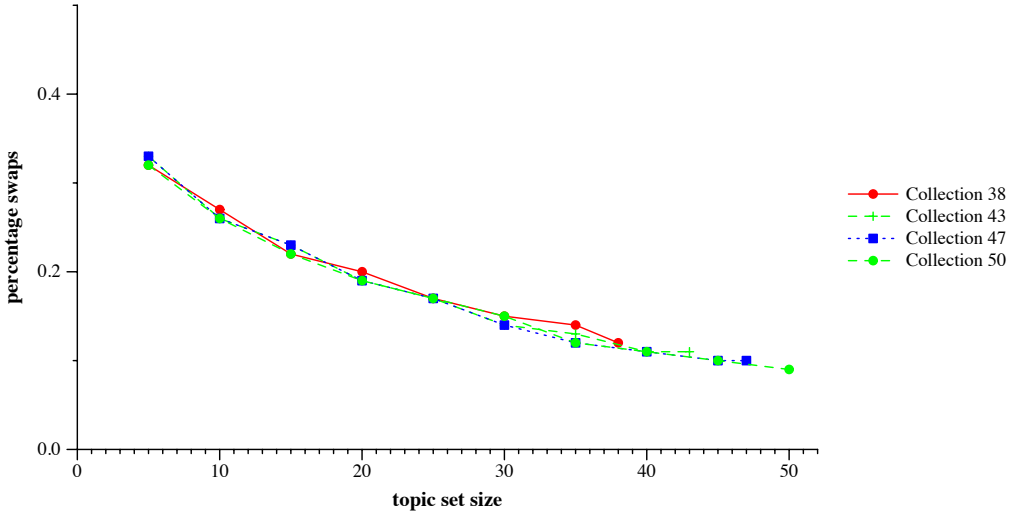
Percentage Relevant as a Quality Indicator



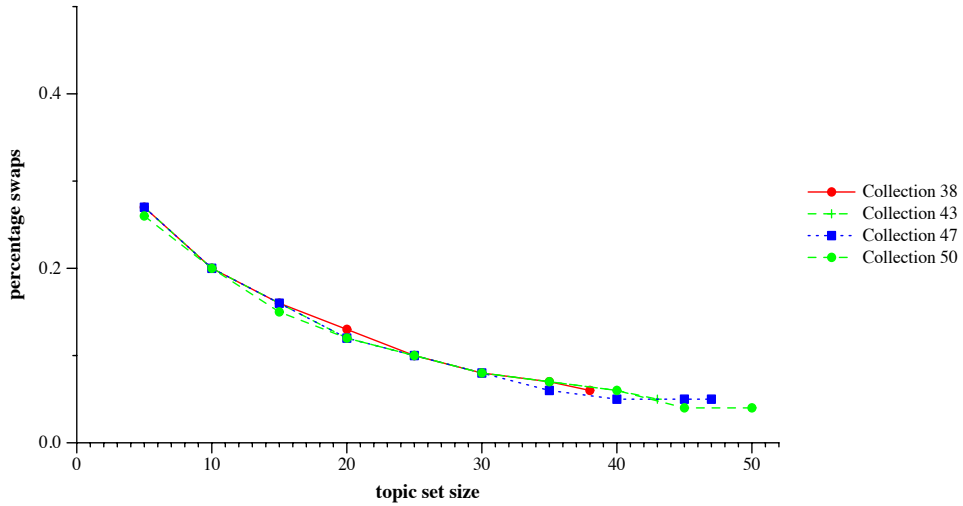
Total number of documents judged and number of documents judged partially or fully relevant in TREC-COVID Complete collection.

- For some topics, almost ¾ of judged documents are relevant!
- Yet stability tests suggest that the collection including those topics is fine.
- Why?
  - about 1% of document collection judged for some topics (enormous percentage)
  - most runs quite effective: pools not filled with chuff & lots of the relevance space explored
  - duplicates in the document collection
  - “relevant” included partially relevant

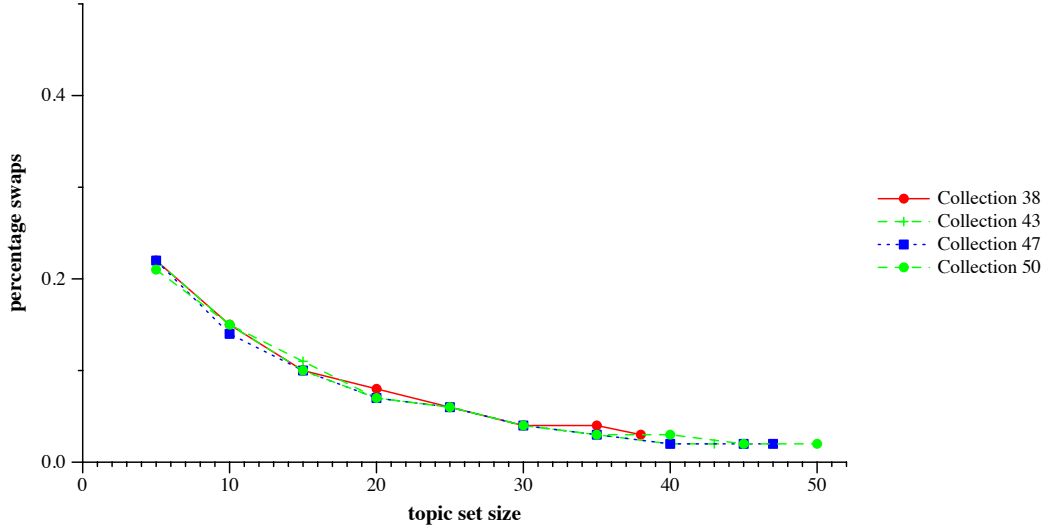
Collection Stability



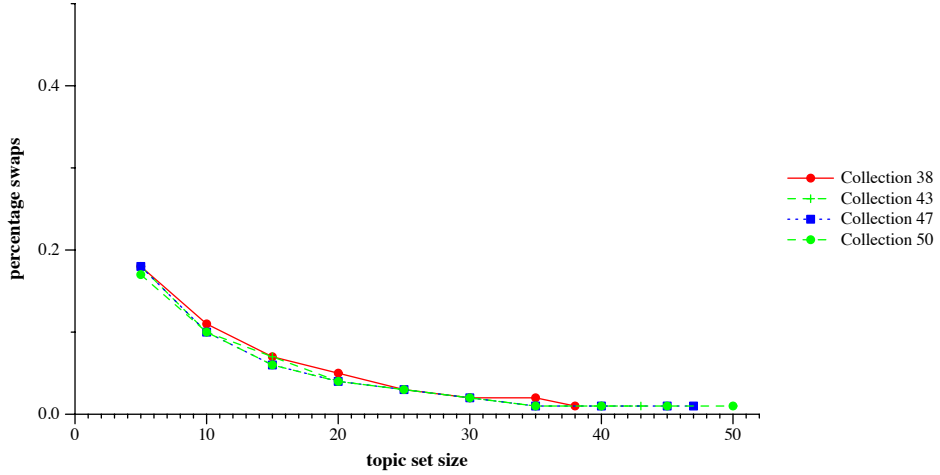
Percentage swaps for map bin 2 (run scores difference between 0.02 and 0.03)



Percentage swaps for map bin 3 (run scores difference between 0.03 and 0.04)



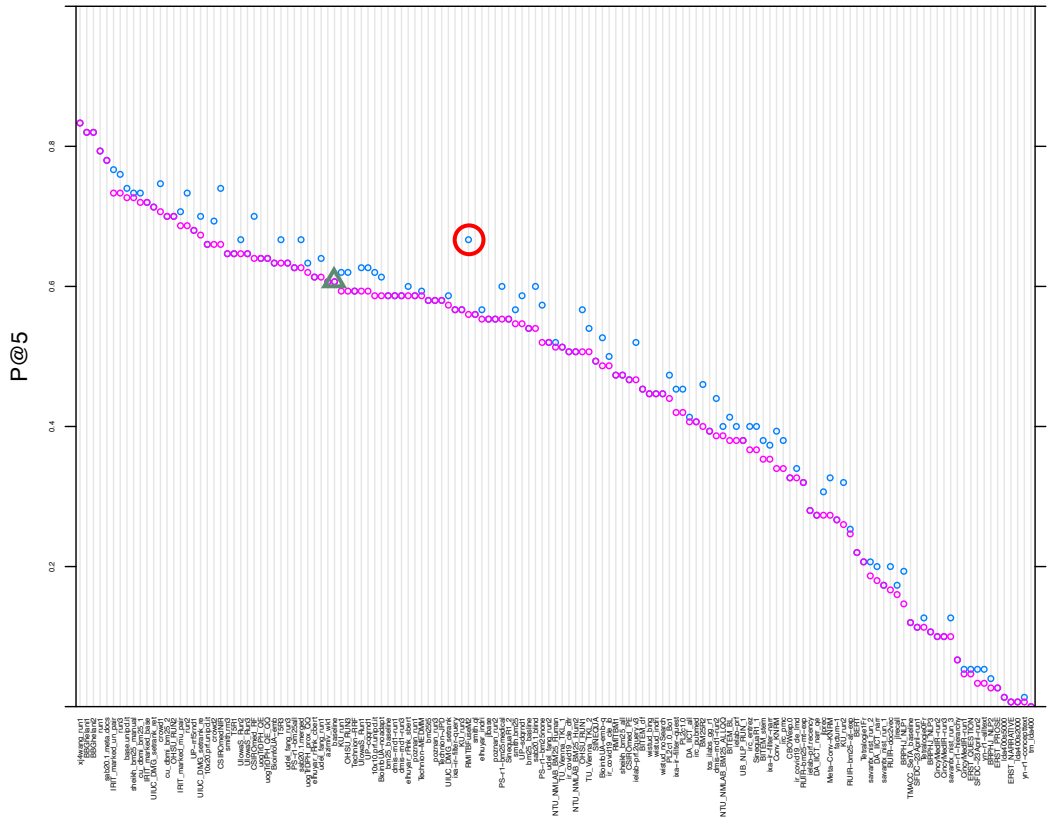
Percentage swaps for map bin 4 (run scores difference between 0.04 and 0.05)



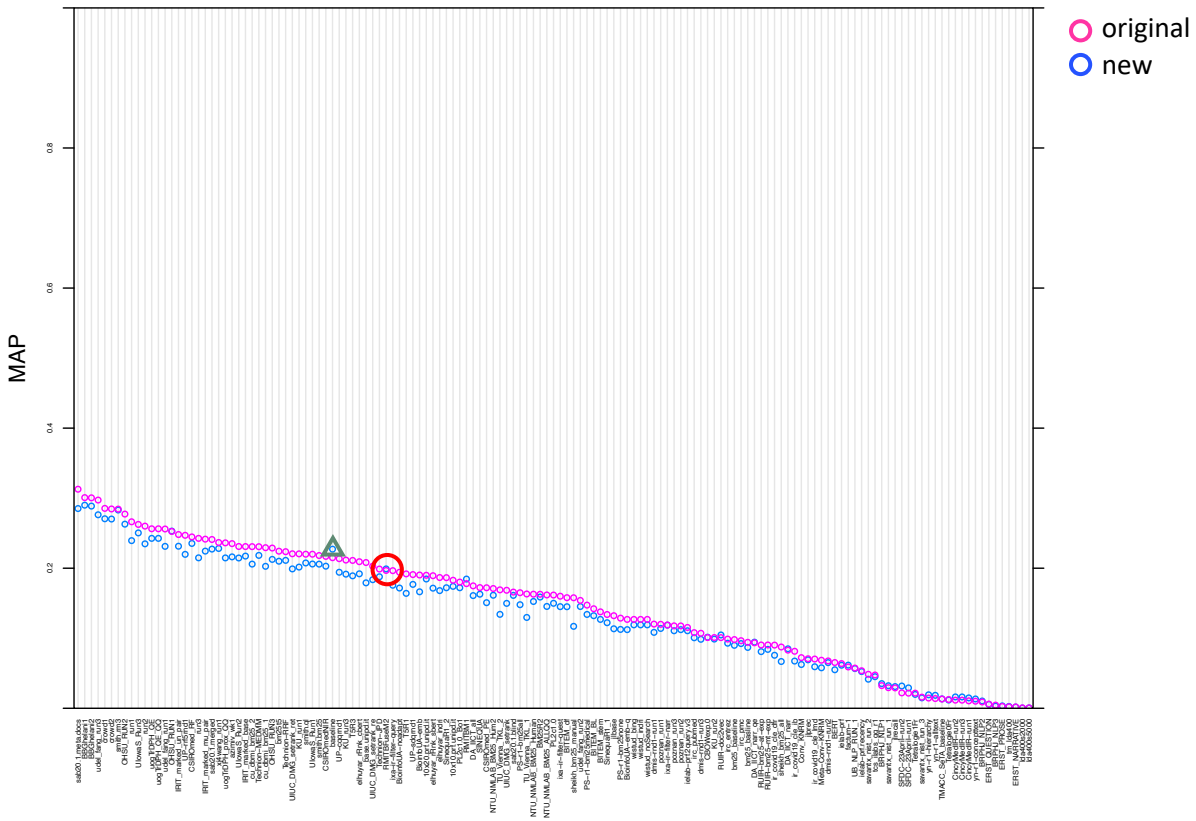
Percentage swaps for map bin 5 (run scores difference between 0.05 and 0.06)

Reusability

Round 1 submissions evaluated using just Round 1 judgments (“original”) and again using cumulative judgments through Round 2 (“new”). Systems are ranked by decreasing score using original judgments (so order is different in different graphs).



P@5: Kendall τ = 0.9452; max Δ = 33



MAP: Kendall τ = 0.9505; max Δ = 19

Building Reusable Collections

- Shared test collections continue to be vital infrastructure for IR research
- We lack effective ways of assessing the quality of a test collection
  - some tests that can detect some problems
  - incremental, diagnostic tests would help in collection creation
  - simulations need to address pragmatic constraints
    - perceived fairness for participants
    - initial learning period of human assessor per topic
    - budget allocation across topics (i.e., stopping conditions)
- Current state-of-the-practice
  - diverse sets of high-quality runs are really helpful in building good collections
  - quality heuristics more context dependent than previously realized
  - we know how to build collections for topics with small R ('small' relative to B)
    - …but can't know size of R until appreciable assessing has been done