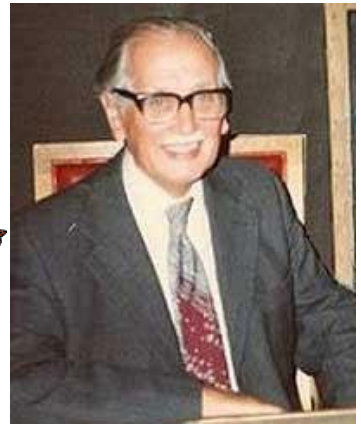
TREC-COVID: Building a Pandemic Information Retrieval Test Collection

Dr. Ellen Voorhees

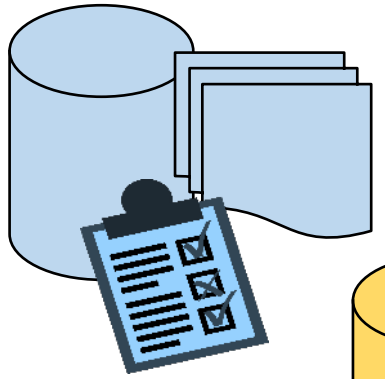
National Institute of Standards and Technology
U.S. Department of Commerce

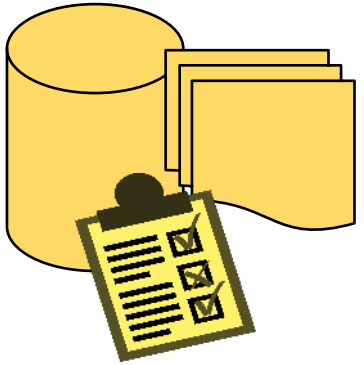
Test Collections Measure Search Engine Effectiveness

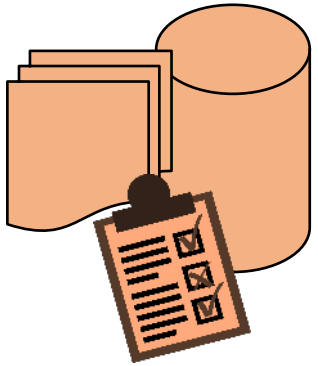
1960's



Cyril Cleverdon

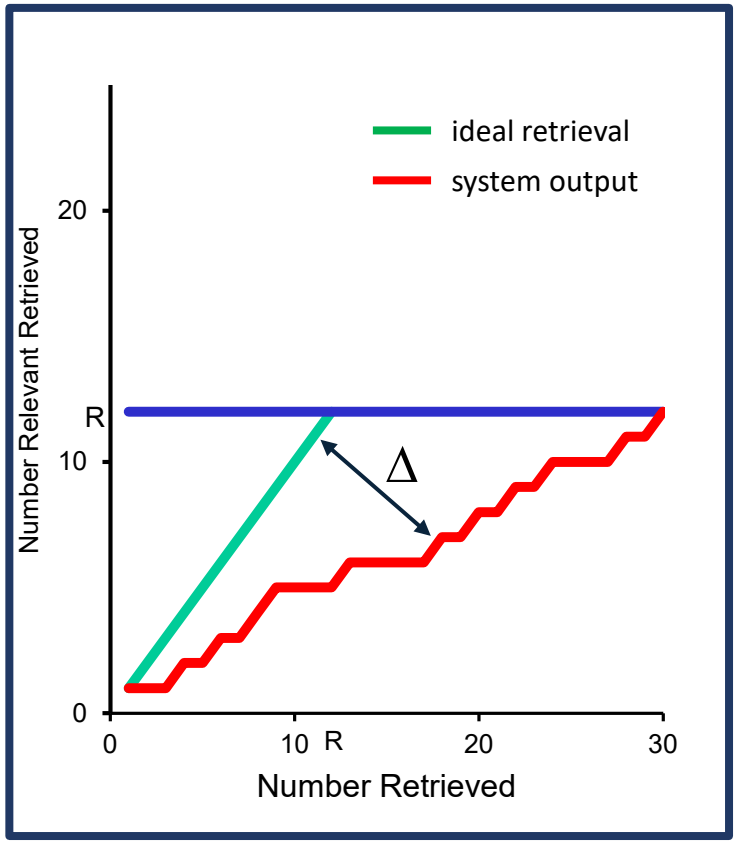






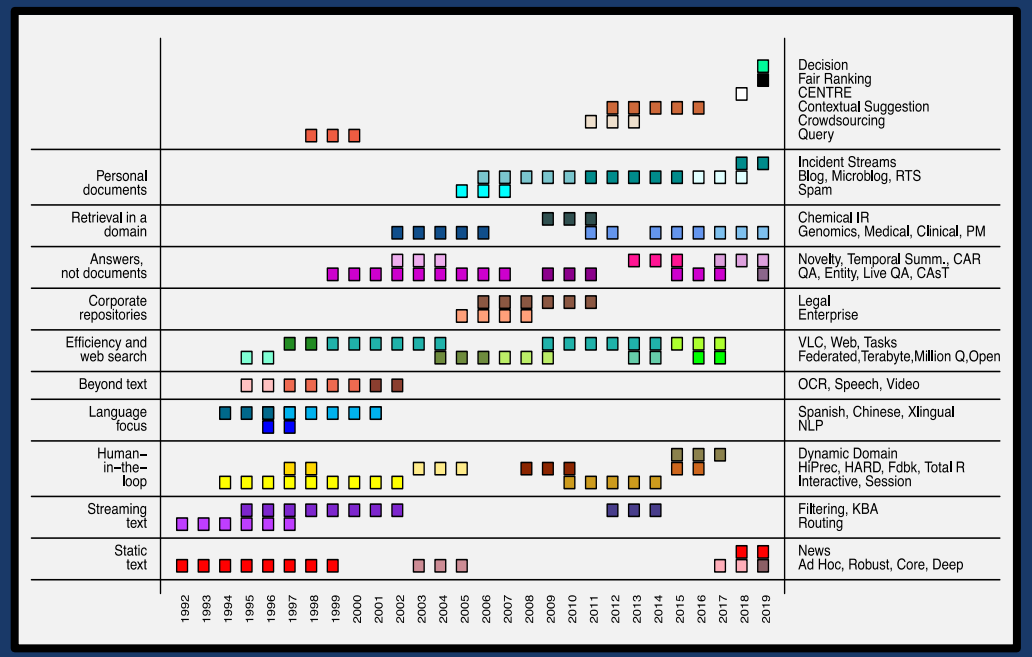
1980's

Test collection defines ideal retrieval so can calculate system's distance from it



Text REtrieval Conference (TREC)

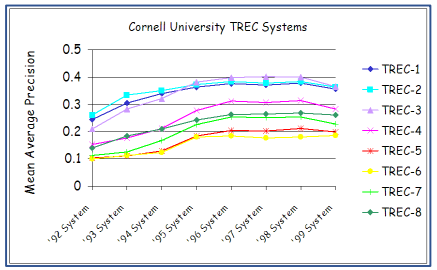
Series of community evaluations that build research infrastructure.



- Pioneered use of “pooling” for building large collections
- Built > 150 test collections for dozens of search tasks
- Hundreds of participant teams world-wide
- Premier venue for determining research methodology
- Model for other efforts in IR and related fields

Community Evaluations

Improve the state of the art



The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in the field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.

Hal Varian
Google Chief Economist
March 4, 2008

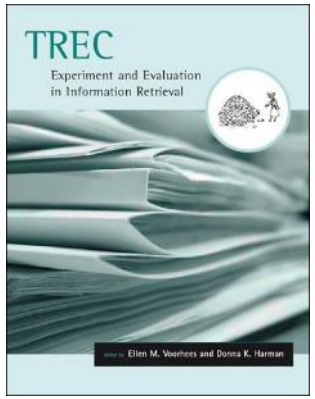
Solidify a research community

Credit: 10th Mountain Films

This project [the TREC Legal track] can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.

Magistrate Judge Paul Grimm
Victor Stanley v. Creative Pipe

Establish research methodology



Credit: MIT Press

TREC is an annual benchmarking exercise that has become a de facto standard in Information Retrieval evaluation.

Stephen Robertson
Microsoft
SIGIR 2007

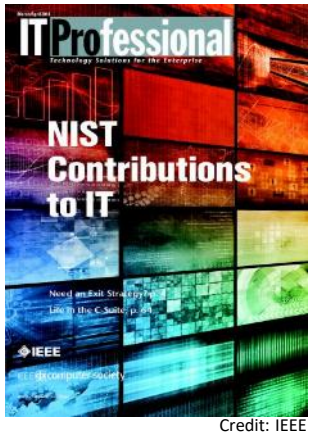
Facilitate technology transfer

Credit: Atomic Taco, CC BY-SA 2.0

TREC has proven to be a valuable forum in which IBM Research has contributed to an improved understanding of search, while at the same time the insights obtained by participating in TREC have helped to improve IBM's products and services.

Alan Marwick, et al.
IBM chapter of the TREC book
2005

Amortize the costs of infrastructure



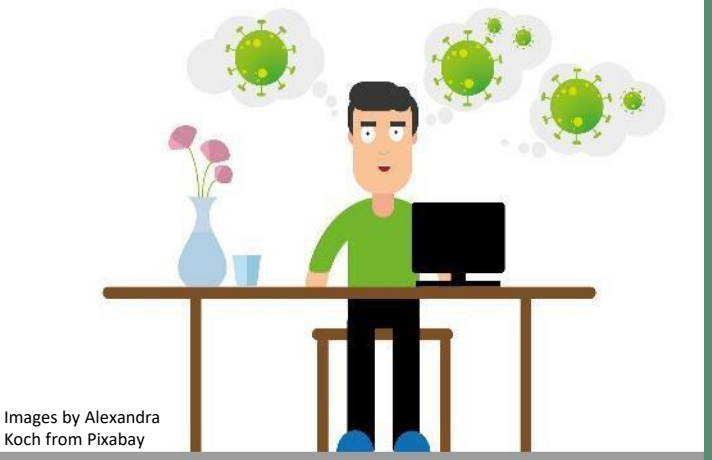
Credit: IEEE

In other words, for every $1 NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers…These responses suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study's economic benefits.

RTI International
Economic Impact Assessment of NIST's TREC Program
December 2010

Effective search helps clinical personnel get the evidence-based answers they need

In response to an OSTP call for action, TREC-COVID is building the infrastructure to improve search systems for future biomedical crises…



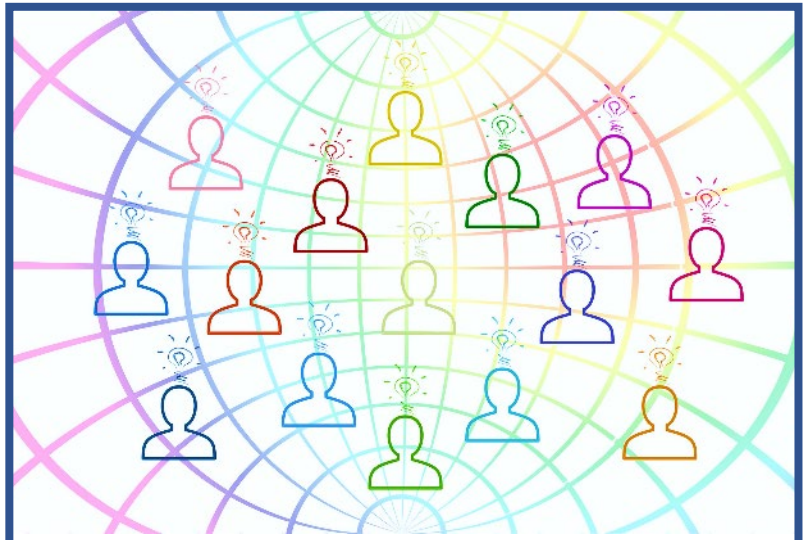
…while providing human-curated answer document sets for today's questions.

Images by Alexandra Koch from Pixabay

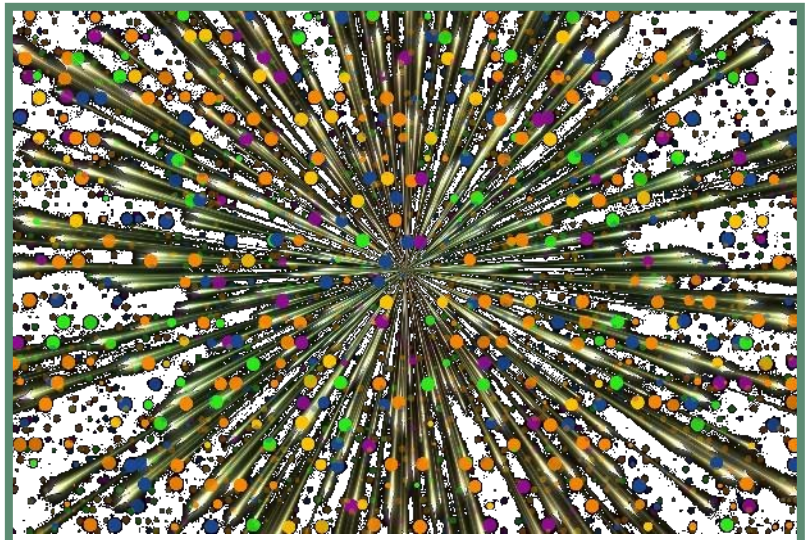
TREC-COVID

Why is a new collection needed?

Because the information landscape for clinical researchers during a crisis event is different from other scientific literature search.



Rapid evolution of topics of interest



Literature in area explodes

Large percentage of literature not peer reviewed and evolving

Images by Gerd Altmann from Pixabay

TREC-COVID

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

Allen Institute for AI

NIH National Library of Medicine

UTHealth
The University of Texas Health Science Center at Houston
School of Biomedical Informatics

OREGON HEALTH & SCIENCE UNIVERSITY
OHSU

Documents

CORD-19, an open-access collection of the scientific literature about COVID-19 and other corona viruses. New snap-shots released weekly.

Questions

Inspired by search logs at medical libraries including NLM. New questions added in each round.

Relevance Judgments

Subset of retrieved documents annotated each round by individuals with biomedical expertise.

TREC-COVID

Two rounds completed; third in progress.

Round 1

April 15—April 23

April 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—June 3

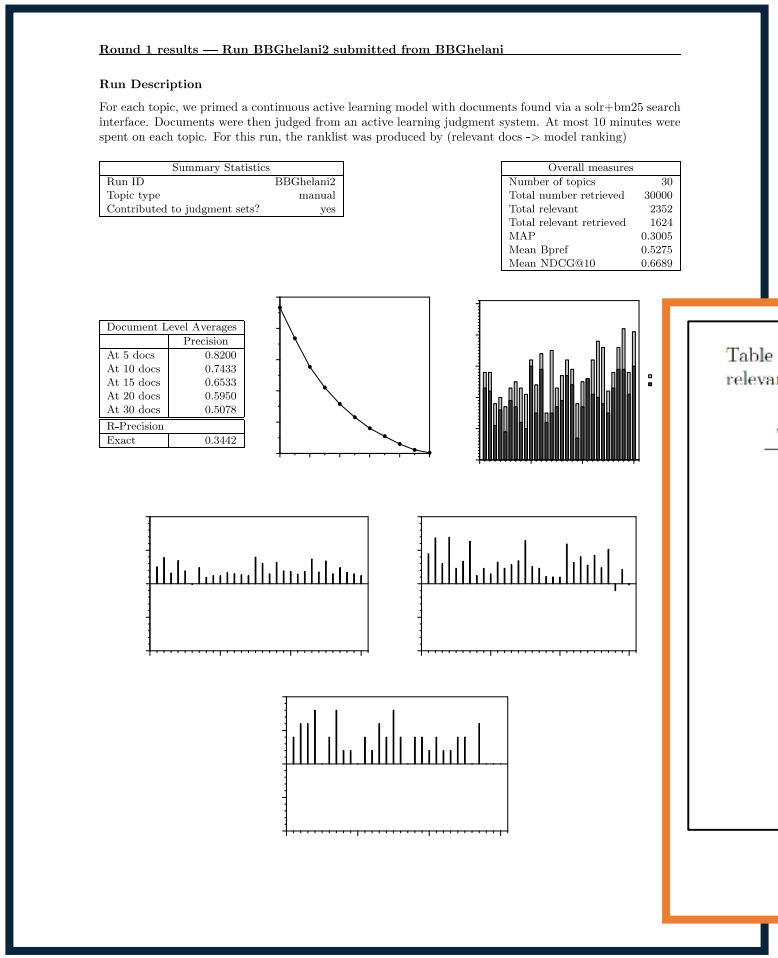
May 19 release of CORD-19, ~128k articles

40 topics

addition of a Kaggle task running in parallel

TREC-COVID (Interim) Results

Participant's score report for one submission



Round 1 results — Run BBGhelani2 submitted from BBGhelani

Run Description

For each topic, we primed a continuous active learning model with documents found via a solr+bm25 search interface. Documents were then judged from an active learning judgment system. At most 10 minutes were spent on each topic. For this run, the ranklist was produced by (relevant docs -> model ranking)

Summary Statistics	
Run ID	BBGhelani2
Topic type	manual
Contributed to judgment sets?	yes

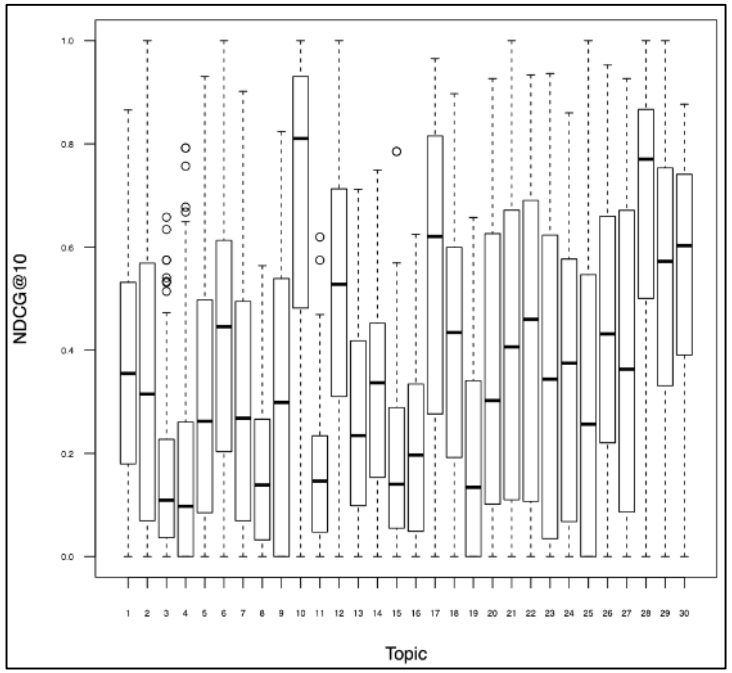
Overall measures	
Number of topics	30
Total number retrieved	30000
Total relevant	2352
Total relevant retrieved	1624
MAP	0.3005
Mean Bpref	0.5275
Mean NDCG@10	0.6689

Document Level Averages	
	Precision
At 5 docs	0.8200
At 10 docs	0.7433
At 15 docs	0.6533
At 20 docs	0.5950
At 30 docs	0.5078
R-Precision	
Exact	0.3442

Table 1: Counts of total numbers of judged documents and number of relevant documents per relevant is the fraction of judged documents that are some form of relevant.

Topic	Total Judged	Partially Relevant	Relevant	Percent Relevant	Topic	Total Judged	Partially Relevant	Relevant	
1	323	45	56	0.313	16	340	42	11	
2	284	21	26	0.165	17	243	32	45	
3	337	66	24	0.267	18	267	79	32	
4	357	32	27	0.165	19	301	27	16	
5	336	35	96	0.390	20	247	41	25	
6	321	80	83	0.508	21	319	15	70	
7	275	2	47	0.178	22	259	17	30	
8	360	46	30	0.211	23	256	4	22	
9	298	25	16	0.138	24	249	14	19	
10	191	35	50	0.445	25	308	9	62	
11	344	67	5	0.209	26	312	19	106	
12	324	76	126	0.623	27	300	30	44	
13	373	97	49	0.391	28	180	9	29	
14	222	24	5	0.131	29	218	42	58	0.459
15	348	45	12	0.164	30	199	39	16	0.276

Analysis of collection quality



Summary of effectiveness over all submissions

Topics, judgments, submissions, score reports all archived on public TREC-COVID website,

http://ir.nist.gov/covidSubmit/

TREC-COVID Future

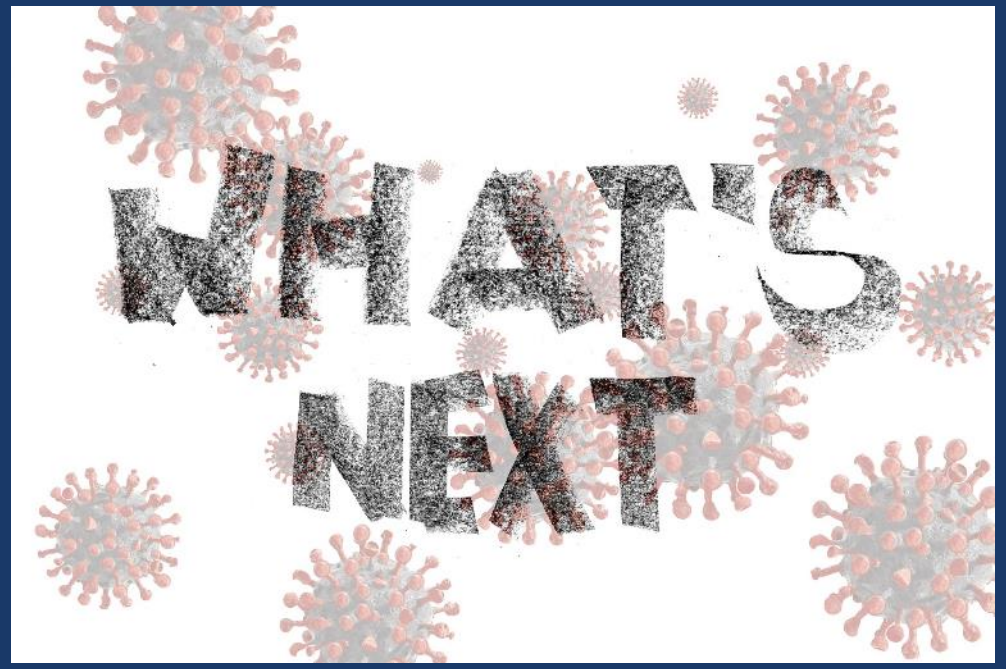


Image by Gerd Altmann from Pixabay

Plan is for 5 rounds total

Conclude mid-August on current schedule.

Final collection

Cumulative set of 50 topics and and ~50k judgments. Metadata will allow series of individual rounds to be recovered as needed.

Rich trove of data to study collection-building under adverse conditions

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