

# Dynamic Sampling Applied to TREC Core 2018

TREC 2018

Mustafa Abualsaud, Gordon V. Cormack, Nimesh Ghelani, Amira Ghenai,  
Maura R. Grossman, Shahin Rahbariasl, Mark D. Smucker, and Haotian Zhang



- Large collections make test collection construction difficult:
  - Hard to find all relevant documents, especially for topics with many relevant documents.
  - Human assessors are slow and thus expensive.
- For TREC 2018, our goal was to create a set of relevance judgments of equivalent quality as NIST at far lower cost without the benefit of submitted runs to pool.

# Approach

- **Dynamic Sampling** [Cormack and Grossman, SIGIR 2018]
  - Combines high-recall retrieval with sampling.
  - High-recall retrieval aims to find all or nearly all relevant documents with least effort.
  - Sampling allows a reduced number of judgments while still estimating effectiveness measures.
- **HiCAL** <http://hical.github.io/> [Abualsaud et al., SIGIR 2018]
  - Interactive high recall retrieval system that allows both interactive search and judging (ISJ) and continuous active learning (CAL) [Cormack and Grossman, SIGIR 2014].
  - Modified for dynamic sampling.
  - Paragraph-only judging for CAL. [Zhang et al., CIKM 2018]

- Dynamic sampling is a form of stratified random sampling.
- Stratified random sampling: Rather than uniformly sample a population, divide the population into strata and sample each stratum at different rates.
- Dynamic sampling idea: Stratified random sampling where judgments from one stratum are used to train a classifier to dynamically create the next stratum.

# Dynamic Sampling Applied

1. Interactive search and judging (ISJ) and continuous active learning (CAL).

2. Use judgments to train classifier

3. Classifier ranks collection and top ranked documents form next stratum.

4. Documents are sampled and judged.

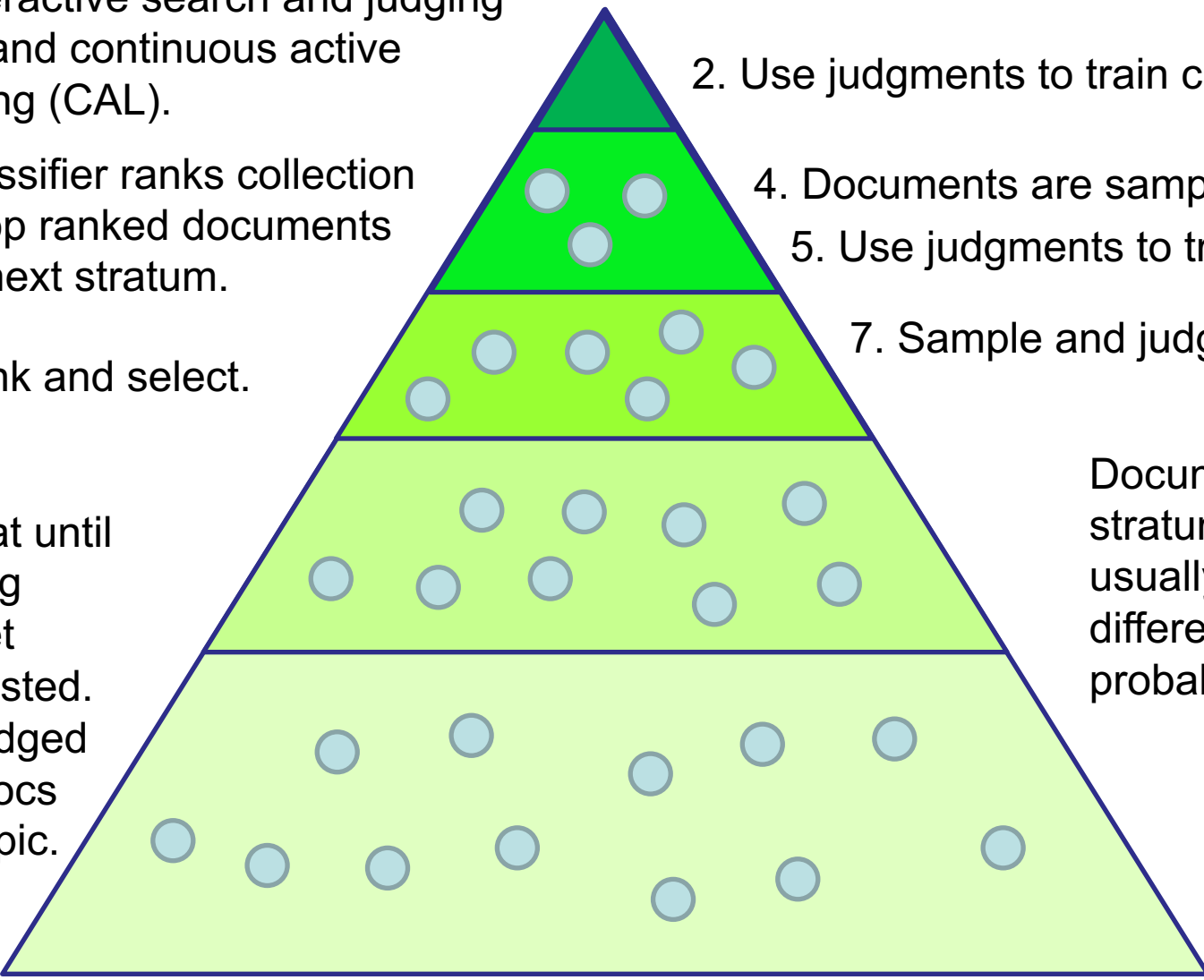
5. Use judgments to train classifier

6. Rank and select.

7. Sample and judge.

Repeat until judging budget exhausted.  
We judged 300 docs per topic.

Documents in each stratum can (and usually do) have a different inclusion probability.



# What Dynamic Sampling Produces

- For each topic, we produce a sampled set of relevance judgments.
- Each sampled document has associated with it an inclusion probability =  $\pi_i$ .
- With these “prels”, one can then estimate the number of relevant documents:

$$R = \sum_{i=1} \frac{r_i}{\pi_i}, \text{ where } r_i = 1 \text{ if relevant, } 0 \text{ o.w.}$$

- In addition, prels can be used for inferred effectiveness measures.

# HiCAL – Search Interface

The screenshot shows a web browser window with the URL `tuna.cs.uwaterloo.ca:9000/search/`. The page title is "HiCAL - Search". A navigation bar at the top contains a home icon, the text "Black Bear Attacks [Click on topic to view full description]", and a "Help" link. On the left side, there is a vertical sidebar with icons for a lightbulb, a magnifying glass, a folder, and a clipboard. The main content area features a search bar with the text "Black bear attack" and a "Search" button. Below the search bar, there are three search results, each with a title, a document ID, a short text snippet, and a relevance indicator. The relevance indicator consists of three colored boxes: "Not relevant" (red), "Relevant" (yellow), and "Highly Relevant" (green). In all three cases, the "Highly Relevant" box is highlighted, and the text "Document has been judged." is displayed below the indicator.

Black Bear Attacks [Click on topic to view full description] Help

Search

Black bear attack Search

Former Marine, 66, survives bloody hand-to-claw combat with **black bear**  
doc\_id: d6ed7028c686e5756ceb0aa0c9b62e0d  
It was just before dawn when 66-year-old Vietnam veteran Larry Yepez locked eyes with a 200-pound **black bear** outside his home near California's Yosemite National Park and knew it meant war. Within seconds, the animal was on him, crushing his hand in

Not relevant Relevant Highly Relevant  
Document has been judged.

Bear mauls woman in Maryland driveway in 'rarest of rare' **attack**  
doc\_id: 6ddf7abe-acb2-11e6-a31b-4b6397e625d0  
She heard a dog barking outside and wondered about the commotion. It appeared to be coming from the yard next door, where relatives lived. Accompanied by a dog on a leash, Karen Osborne ventured into the darkness. But minutes later, a 200-pound **black**

Not relevant Relevant Highly Relevant  
Document has been judged.

How to best survive a **black bear attack**  
doc\_id: f4bb75ea4c8f198c30013cab9eab3be4  
There they were — 22-year-old Darsh Patel and four other men — hiking in the Apshawa Preserve , northwest of New York City, with a 300-pound **black bear** on their trail. When the five friends, all of Edison, N.J., noticed the **bear** in pursuit, they sp

Not relevant Relevant Highly Relevant

Tourists are finding out what happens when bears have a food shortage  
doc\_id: bc66bf6f61cc165f002k720b7c00224

# HiCAL – CAL Interface

The screenshot shows the HiCAL – CAL interface. The main content area displays the title "Tourists trying to take selfies with bears just ruined a Colorado park for everyone" and a paragraph of text. A yellow box with an arrow points to the paragraph, stating: "Only a classifier selected paragraph is shown. This speeds judging with little loss of accuracy." The interface includes a search bar, a "Judge document" section with buttons for "Not relevant", "Relevant", and "Highly relevant", and a "Keyboard shortcuts" section with buttons for "s", "r", "h", and "u".

2015-09-15 docno: 29a711e2ccb306baf5819509f1e305ed

Bears rarely attack humans, and when they do, it often costs the bear its own life. But the first rule of successfully encountering a bear in the wilderness is to not stick around long enough for it to escalate into a bear attack, which is probably why the National Park Service recommends keeping your distance.

**s** Mark as non-relevant  
**r** Mark as relevant  
**h** Mark as highly relevant

**u** Show latest judged documents

Only a classifier selected paragraph is shown. This speeds judging with little loss of accuracy.



# Solutions for Large Test Collection Construction

- Problem: Hard to find relevant documents
- Solution: ISJ + CAL using HiCAL
  
- Problem: Assessors are expensive and thus limit judging budget.
- Solution:
  - Sample: fewer documents to judge
  - Passages: allow high speed judging with fast user interface

# Experiment Details

- Gordan Cormack did an initial round of CAL for each topic.
- Topics with few found relevant documents were manually searched by 5 co-authors.
- These initial judgments formed the zeroth stratum with an inclusion probability of 1.
- 5 co-authors then used dynamic sampling with a budget of 300 judgments.
- Average time to judge a document = 6.6s (33 minutes per topic).

# Results

- 19,161 judgments. 383.2 per topic.
- 39,214 docs in strata. 784.3 per topic.
- Relevant documents not in the strata are missed. Same issue as rel docs outside pool.
- System recall = recall of the docs in strata.
- Average system recall = 0.92

System Recall				
Min	1 <sup>st</sup> Quartile	Median	3 <sup>rd</sup> Quartile	Max
0.53	0.88	0.97	1.0	1.0

# Top 5 Low System Recall Topics

ID	Topic	Sys. Recall	Assessor Agreement		
			TPR	FPR	d-prime
819	U.S. age demographics	0.53	0.17	0.14	0.13
442	heroic acts	0.58	0.19	0.10	0.43
341	Airport Security	0.64	0.14	0.06	0.46
439	inventions, scientific discoveries	0.65	0.47	0.16	0.95
803	declining middle class in U.S.	0.77	0.50	0.37	0.32

- In general, where we had poor system recall is where our assessors failed to agree with NIST.
- Of note, topic 439 had low, but not abysmal agreement with NIST. Assessor noted that CAL seemed to get stuck on a subtopic.

# Submitted Runs

Run	Summary	MAP
UWaterMDS_DS_A	Docs judged rel ordered by classifier	0.37
UWaterMDS_DS_B	Docs judged rel in reverse order by classifier	0.24
UWaterMDS_Rank	Classifier ranking of collection	0.43
UWaterMDS_SEQ	Rel docs in order judged, plus non-rel and unjudged	0.17