Evaluating Elements of Web-Based Data Enrichment for Pseudo-relevance Feedback Retrieval

Timo Breuer, Melanie Pest, Philipp Schaer

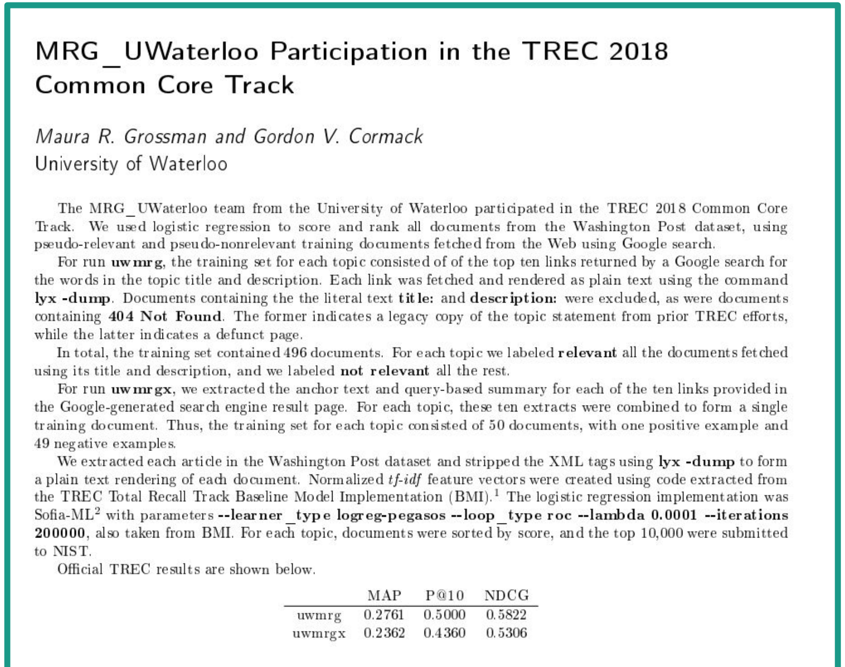
Technology
Arts Sciences
TH Köln

CLEF 2021; September, 21-24, 2021;
online event (from Bucharest - Romania).

Contributions

- **Robustness test** of web-based pseudo-relevance feedback retrieval w.r.t.
  - time
  - web search engine (Google vs. DuckDuckGo)
  - query type (`title` vs. `title+desc`)
  - test collection (Robust 04/05, Core 17/18)
- Open source **reimplementation** of runs by Grossman and Cormack and SERP **dataset**

Approach [Grossman & Cormack, TREC, 2018]



MRG_UWaterloo Participation in the TREC 2018 Common Core Track

Maura R. Grossman and Gordon V. Cormack
University of Waterloo

The MRG_UWaterloo team from the University of Waterloo participated in the TREC 2018 Common Core Track. We used logistic regression to score and rank all documents from the Washington Post dataset, using pseudo-relevant and pseudo-nonrelevant training documents fetched from the Web using Google search.

For run **uwmrg**, the training set for each topic consisted of of the top ten links returned by a Google search for the words in the topic title and description. Each link was fetched and rendered as plain text using the command **lyx -dump**. Documents containing the the literal text **title:** and **description:** were excluded, as were documents containing **404 Not Found**. The former indicates a legacy copy of the topic statement from prior TREC efforts, while the latter indicates a defunct page.

In total, the training set contained 496 documents. For each topic we labeled **relevant** all the documents fetched using its title and description, and we labeled **not relevant** all the rest.

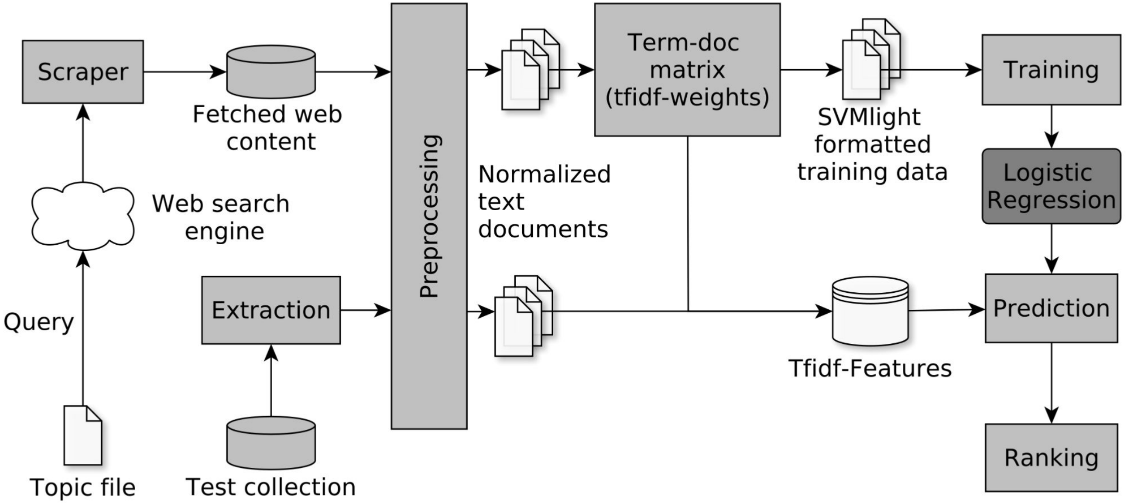
For run **uwmrgx**, we extracted the anchor text and query-based summary for each of the ten links provided in the Google-generated search engine result page. For each topic, these ten extracts were combined to form a single training document. Thus, the training set for each topic consisted of 50 documents, with one positive example and 49 negative examples.

We extracted each article in the Washington Post dataset and stripped the XML tags using **lyx -dump** to form a plain text rendering of each document. Normalized *tf-idf* feature vectors were created using code extracted from the TREC Total Recall Track Baseline Model Implementation (BMI).[1] The logistic regression implementation was Sofia-ML[2] with parameters **--learner_type logreg-pegasos --loop_type roc --lambda 0.0001 --iterations 200000**, also taken from BMI. For each topic, documents were sorted by score, and the top 10,000 were submitted to NIST.

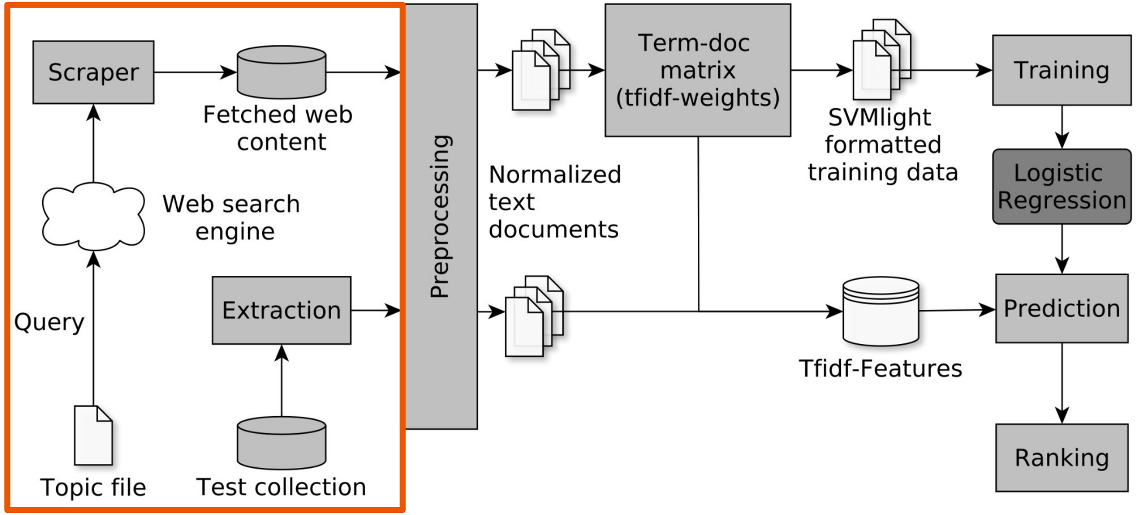
Official TREC results are shown below.

	MAP	P@10	NDCG
uwmrg	0.2761	0.5000	0.5822
uwmrgx	0.2362	0.4360	0.5306

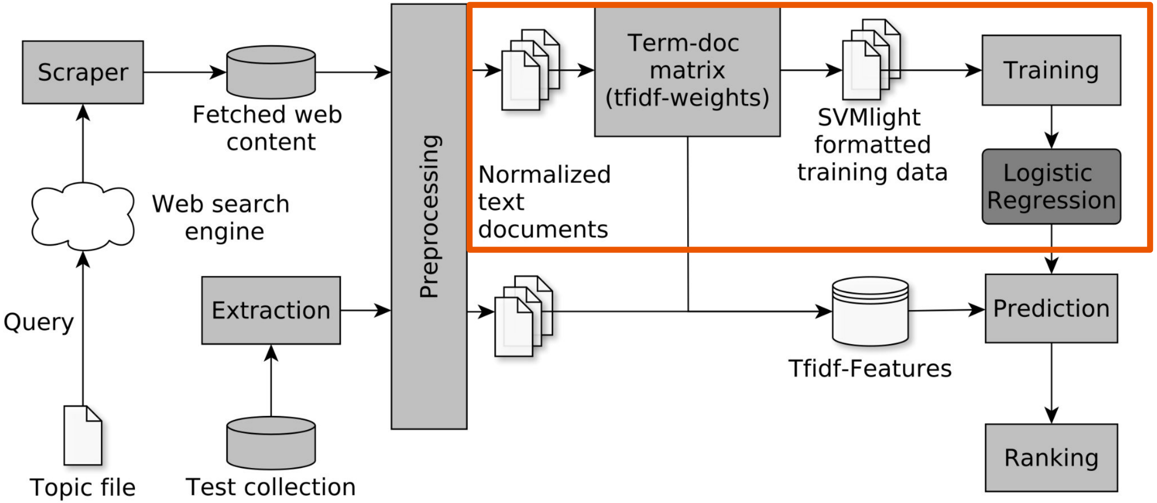
Approach [Grossman & Cormack, TREC, 2018]



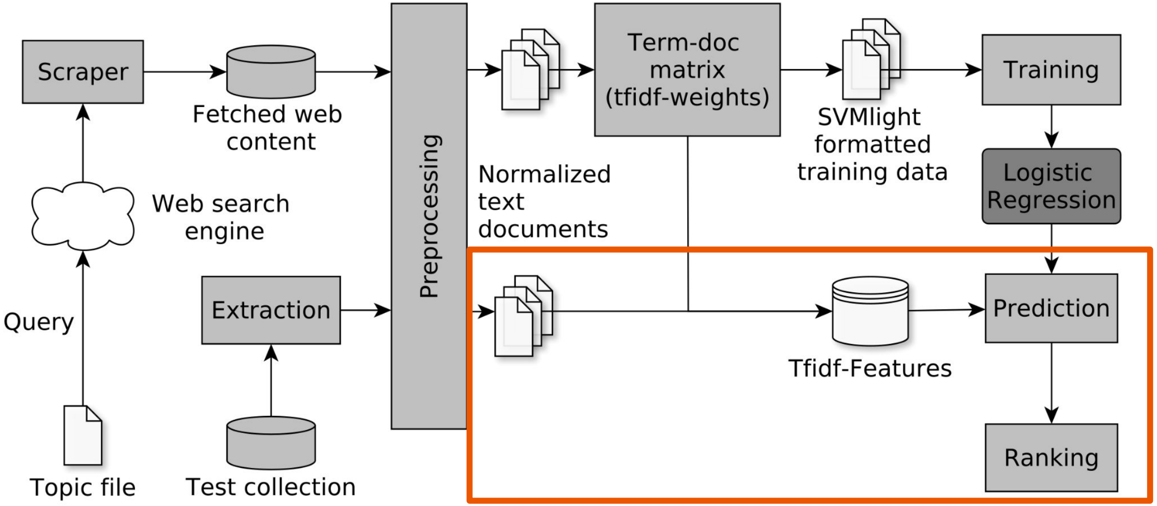
Approach [Grossman & Cormack, TREC, 2018]



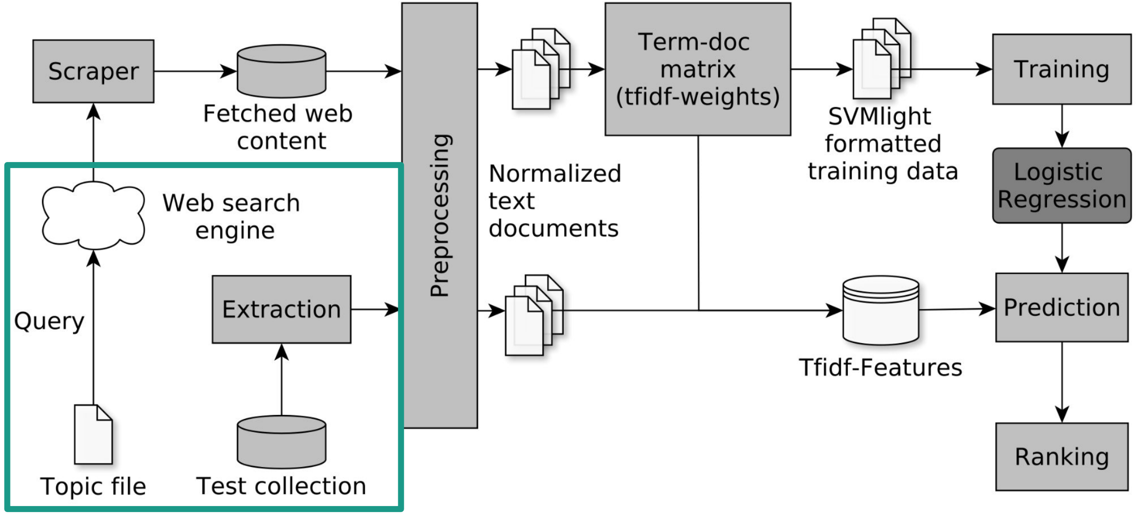
Approach [Grossman & Cormack, TREC, 2018]



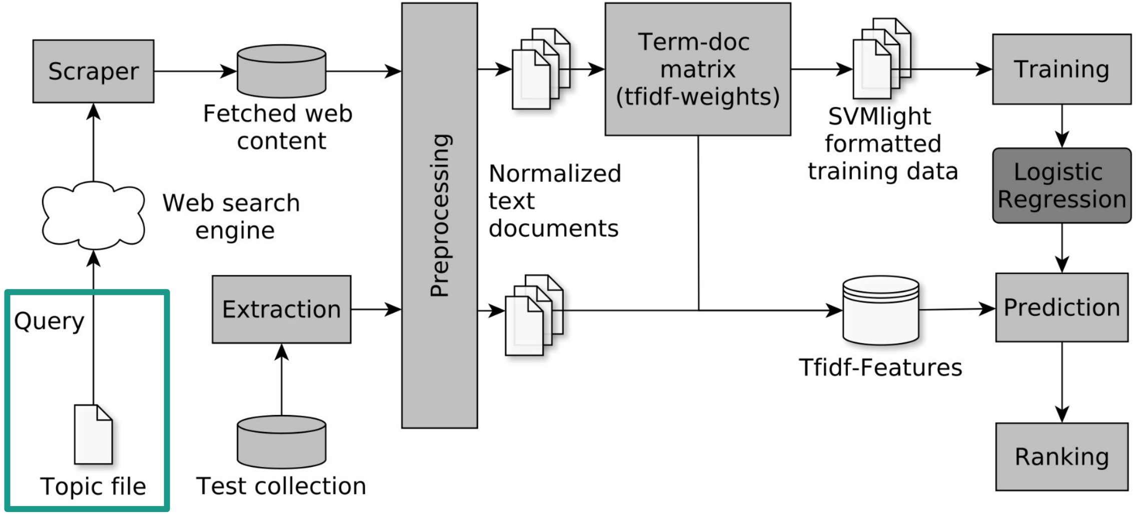
Approach [Grossman & Cormack, TREC, 2018]



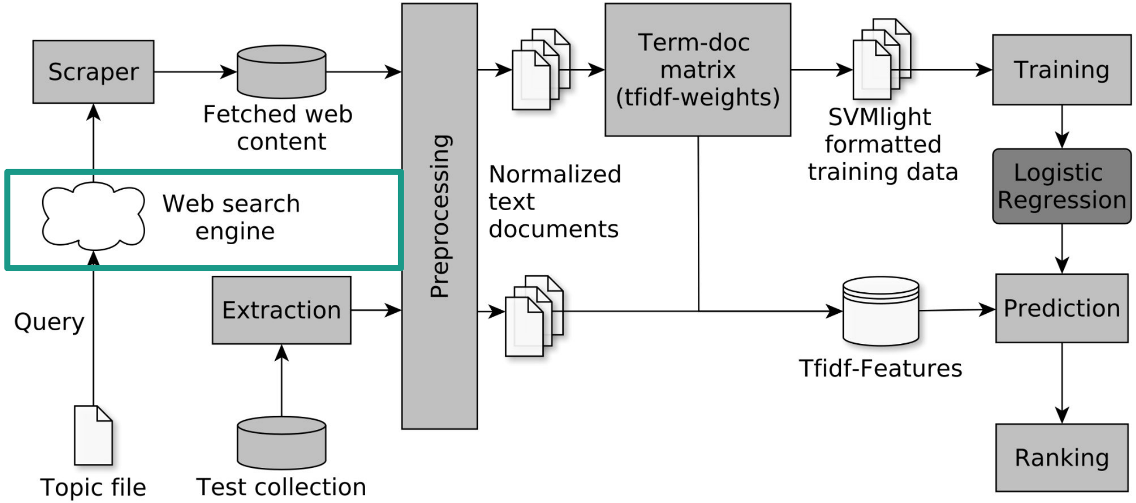
Approach [Grossman & Cormack, TREC, 2018]



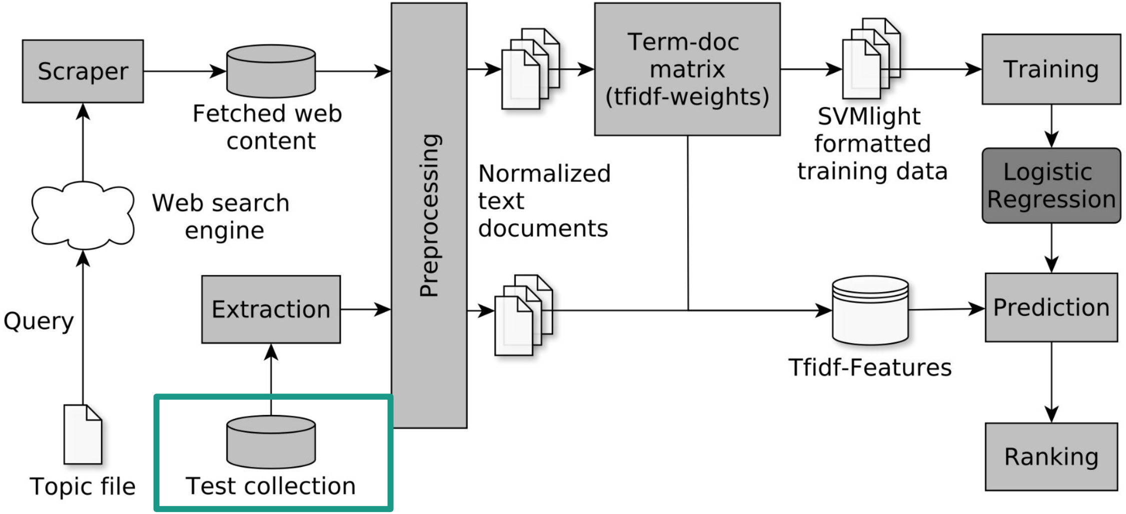
Approach [Grossman & Cormack, TREC, 2018]



Approach [Grossman & Cormack, TREC, 2018]



Approach [Grossman & Cormack, TREC, 2018]



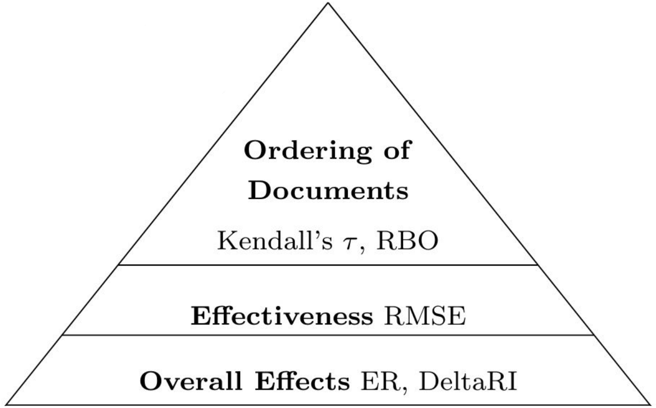
Research questions

RQ1 *How do the components of the workflow, i.e., the query formulation and the web search engine, affect the system performance over time?*

RQ2 *To which extent are the original effects present in different contexts, i.e., with other newswire test collections?*

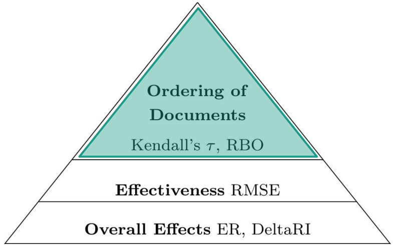
Evaluation Metrics

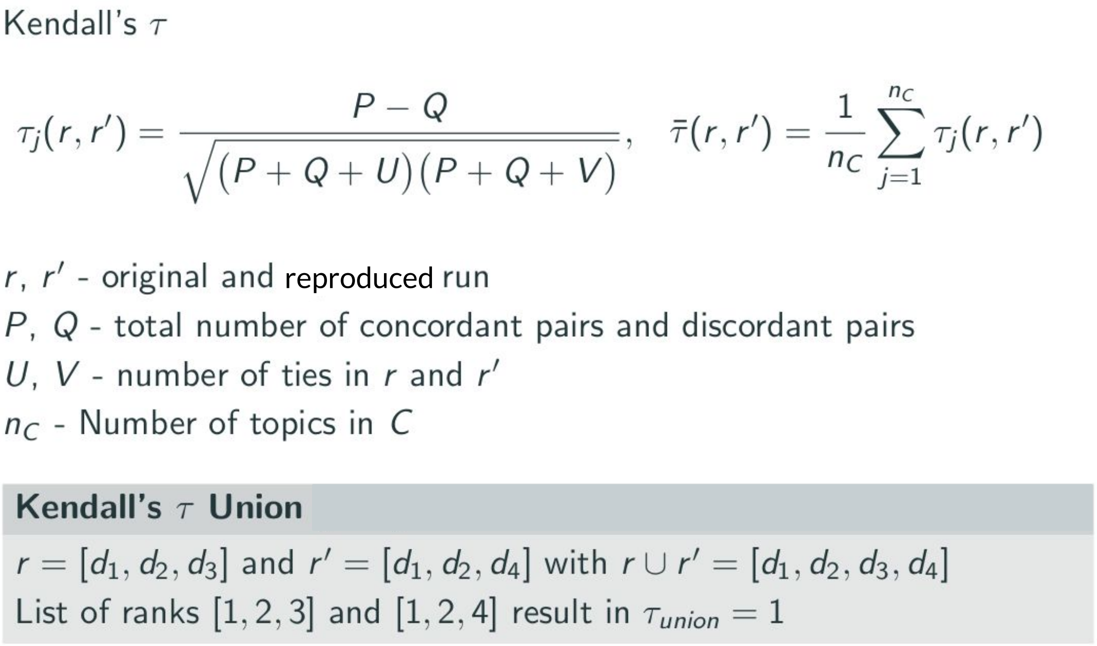
How to Measure the Reproducibility of System-oriented IR Experiments



[Breuer, Ferro, Fuhr, Maistro, Sakai, Schaer, Soboroff, SIGIR, 2020]

Ordering of Documents





Kendall's τ

$$\tau_j(r, r') = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}}, \quad \bar{\tau}(r, r') = \frac{1}{n_C} \sum_{j=1}^{n_C} \tau_j(r, r')$$

r, r' - original and reproduced run

P, Q - total number of concordant pairs and discordant pairs

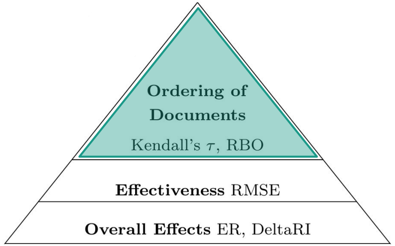
U, V - number of ties in r and r'

n_C - Number of topics in C

Kendall's τ Union

$r = [d_1, d_2, d_3]$ and $r' = [d_1, d_2, d_4]$ with $r \cup r' = [d_1, d_2, d_3, d_4]$

List of ranks $[1, 2, 3]$ and $[1, 2, 4]$ result in $\tau_{union} = 1$



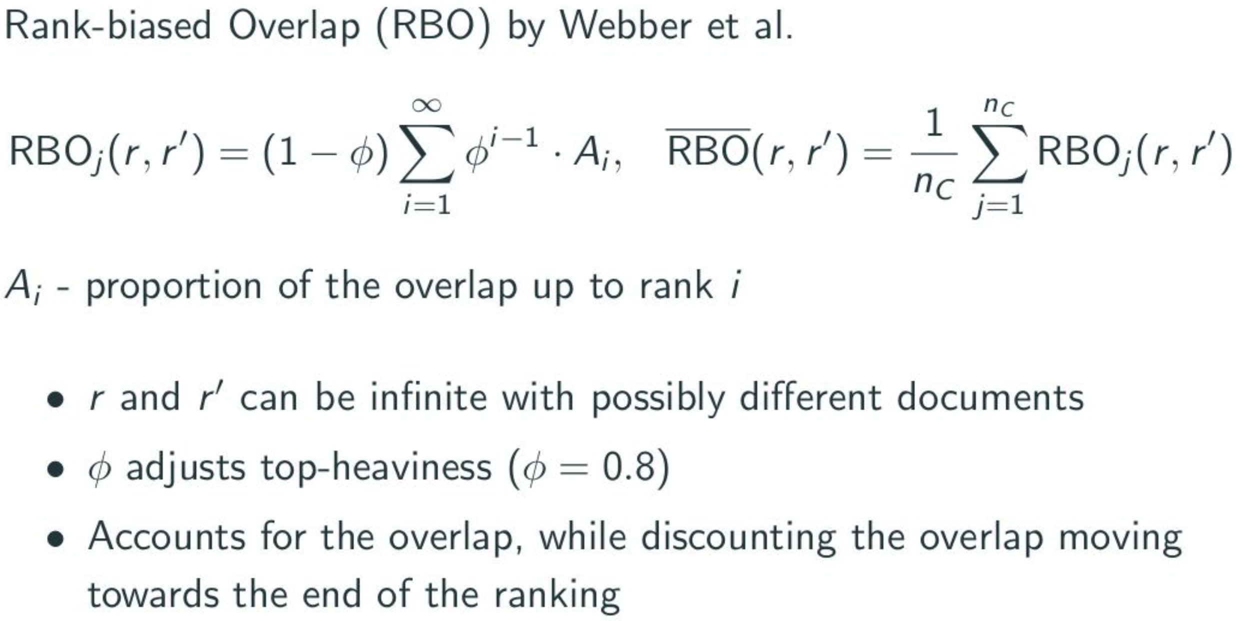
Ordering of Documents

Kendall's τ, RBO

Effectiveness RMSE

Overall Effects ER, DeltaRI

Ordering of Documents



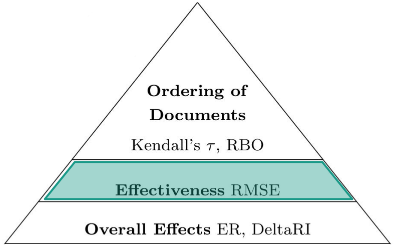
Rank-biased Overlap (RBO) by Webber et al.

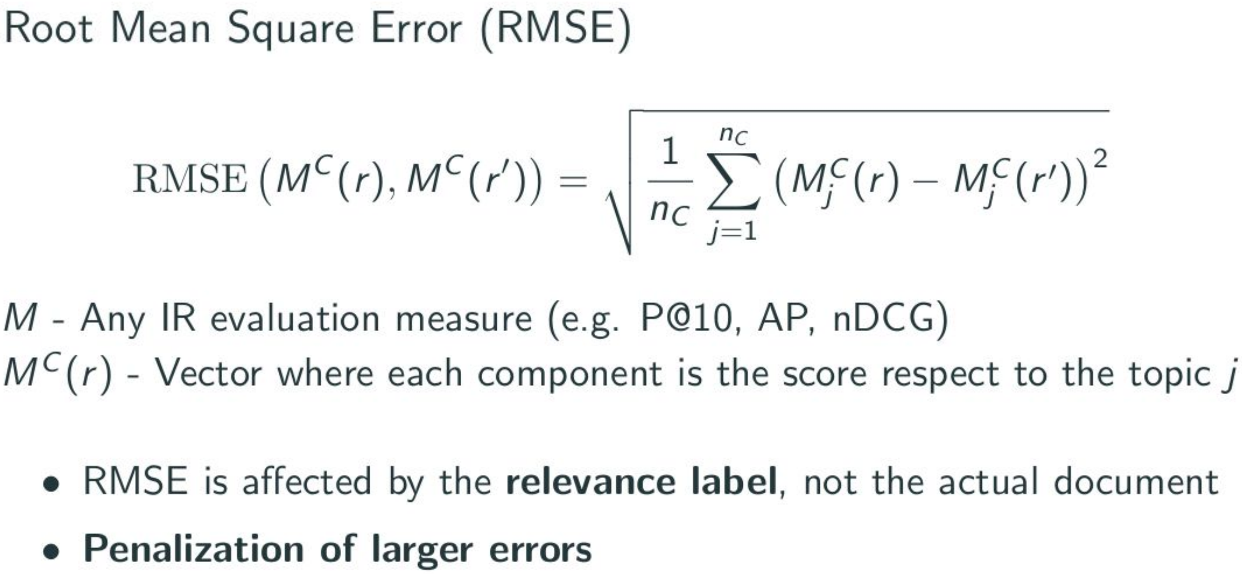
$$\text{RBO}_j(r, r') = (1 - \phi) \sum_{i=1}^{\infty} \phi^{i-1} \cdot A_i, \quad \overline{\text{RBO}}(r, r') = \frac{1}{n_C} \sum_{j=1}^{n_C} \text{RBO}_j(r, r')$$

A_i - proportion of the overlap up to rank i

- r and r' can be infinite with possibly different documents
- ϕ adjusts top-heaviness ($\phi = 0.8$)
- Accounts for the overlap, while discounting the overlap moving towards the end of the ranking

Effectiveness





Root Mean Square Error (RMSE)

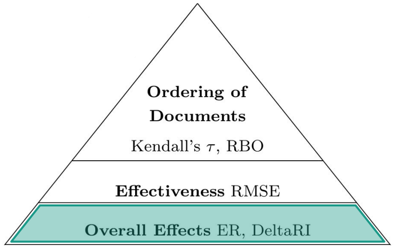
$$\mathrm{RMSE}\left(M^C(r), M^C(r')\right) = \sqrt{\frac{1}{n_C}\sum_{j=1}^{n_C}\left(M_j^C(r) - M_j^C(r')\right)^2}$$

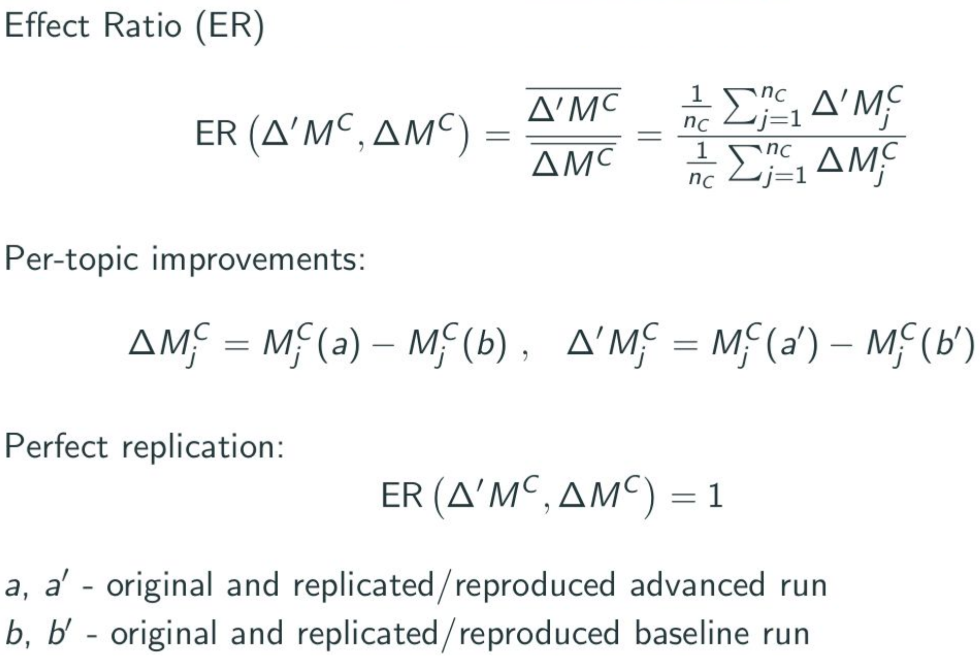
M - Any IR evaluation measure (e.g. P@10, AP, nDCG)

$M^C(r)$ - Vector where each component is the score respect to the topic j

- RMSE is affected by the **relevance label**, not the actual document
- **Penalization of larger errors**

Overall effects





Effect Ratio (ER)

$$\mathrm{ER}\left(\Delta' M^C, \Delta M^C\right) = \frac{\overline{\Delta' M^C}}{\overline{\Delta M^C}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C} \Delta' M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C} \Delta M_j^C}$$

Per-topic improvements:

$$\Delta M_j^C = M_j^C(a) - M_j^C(b)\ , \quad \Delta' M_j^C = M_j^C(a') - M_j^C(b')$$

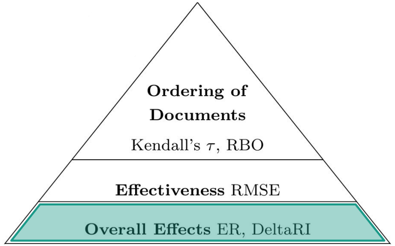
Perfect replication:

$$\mathrm{ER}\left(\Delta' M^C, \Delta M^C\right) = 1$$

a, a' - original and replicated/reproduced advanced run

b, b' - original and replicated/reproduced baseline run

Overall effects



Delta Relative Improvement (DeltaRI)

$$\Delta\text{RI}(\text{RI}, \text{RI}') = \text{RI} - \text{RI}'$$

Relative Improvement:

$$\text{RI} = \frac{\overline{M^C(a)} - \overline{M^C(b)}}{\overline{M^C(b)}}, \qquad \text{RI}' = \frac{\overline{M^C(a')} - \overline{M^C(b')}}{\overline{M^C(b')}}$$

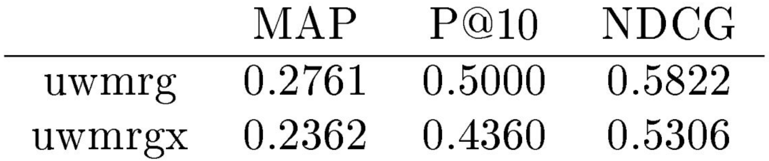
Perfect replication:

$$\Delta\text{RI}(\text{RI}, \text{RI}') = 0$$

Experimental Results

Two run types

Run	Type	Description
uwmrgx	**baseline**	tfidf features based on anchor text and summary
uwmrg	**advanced**	tfidf features based on scraped website texts of the URLs



	MAP	P@10	NDCG
uwmrg	0.2761	0.5000	0.5822
uwmrgx	0.2362	0.4360	0.5306

Average Retrieval Performance (Core18)

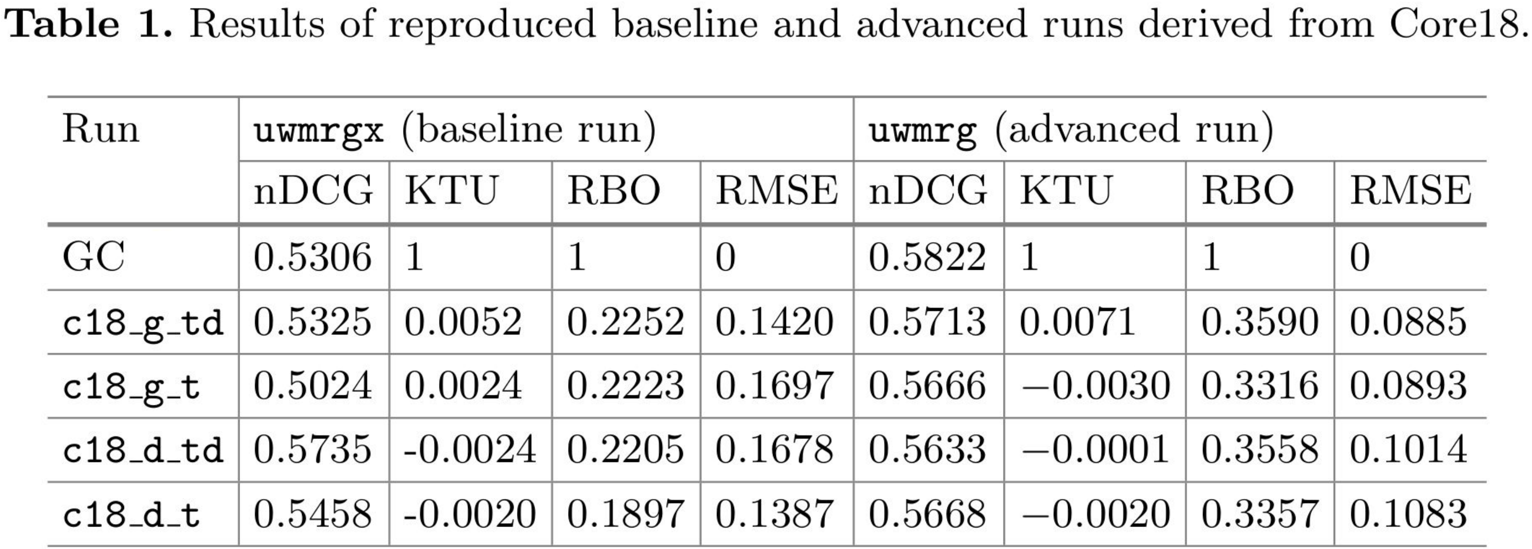


Table 1. Results of reproduced baseline and advanced runs derived from Core18.

Run	uwmrgx (baseline run)				uwmrg (advanced run)			
	nDCG	KTU	RBO	RMSE	nDCG	KTU	RBO	RMSE
GC	0.5306	1	1	0	0.5822	1	1	0
c18_g_td	0.5325	0.0052	0.2252	0.1420	0.5713	0.0071	0.3590	0.0885
c18_g_t	0.5024	0.0024	0.2223	0.1697	0.5666	−0.0030	0.3316	0.0893
c18_d_td	0.5735	-0.0024	0.2205	0.1678	0.5633	−0.0001	0.3558	0.1014
c18_d_t	0.5458	-0.0020	0.1897	0.1387	0.5668	−0.0020	0.3357	0.1083

Average Retrieval Performance (Core18)

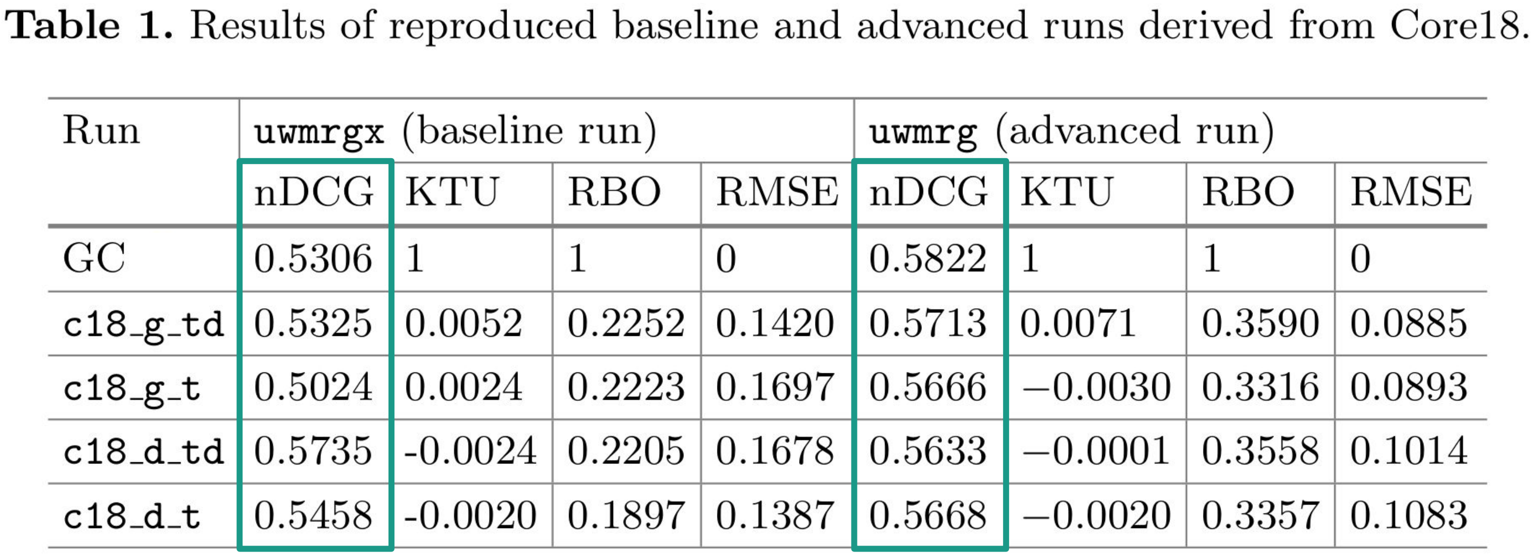


Table 1. Results of reproduced baseline and advanced runs derived from Core18.

Run	uwmrgx (baseline run)				uwmrg (advanced run)			
	nDCG	KTU	RBO	RMSE	nDCG	KTU	RBO	RMSE
GC	0.5306	1	1	0	0.5822	1	1	0
c18_g_td	0.5325	0.0052	0.2252	0.1420	0.5713	0.0071	0.3590	0.0885
c18_g_t	0.5024	0.0024	0.2223	0.1697	0.5666	−0.0030	0.3316	0.0893
c18_d_td	0.5735	-0.0024	0.2205	0.1678	0.5633	−0.0001	0.3558	0.1014
c18_d_t	0.5458	-0.0020	0.1897	0.1387	0.5668	−0.0020	0.3357	0.1083

Average Retrieval Performance (Core18)

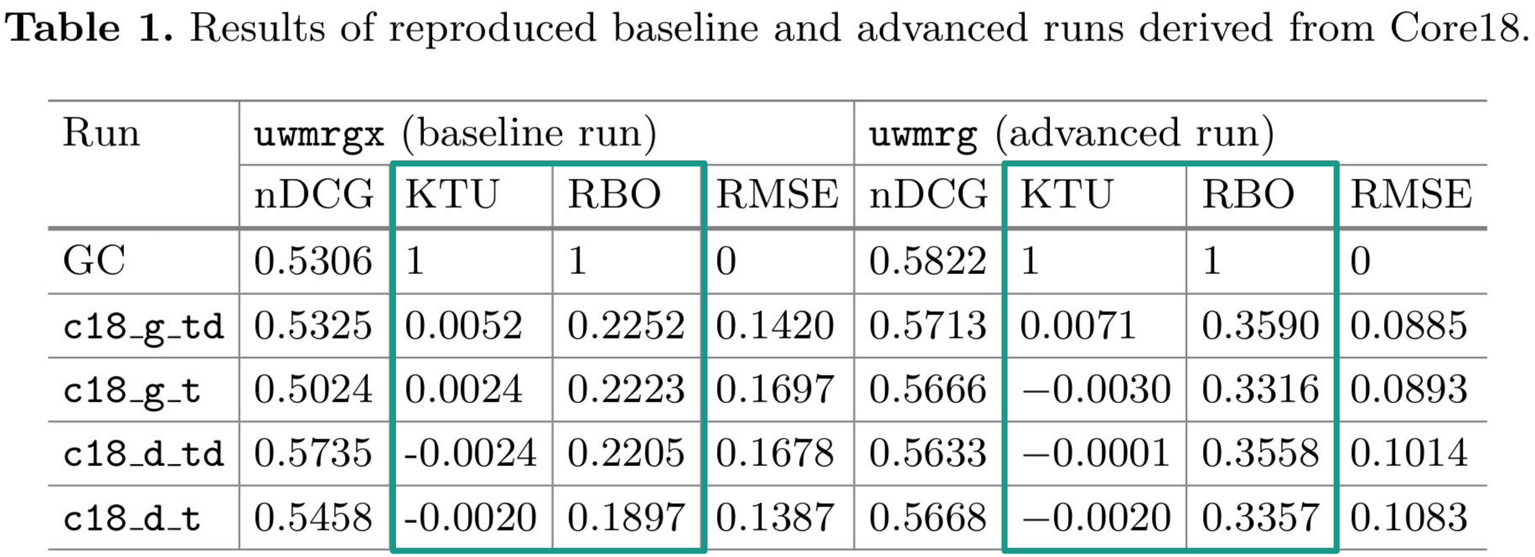


Table 1. Results of reproduced baseline and advanced runs derived from Core18.

Run	uwmrgx (baseline run)				uwmrg (advanced run)			
	nDCG	KTU	RBO	RMSE	nDCG	KTU	RBO	RMSE
GC	0.5306	1	1	0	0.5822	1	1	0
c18_g_td	0.5325	0.0052	0.2252	0.1420	0.5713	0.0071	0.3590	0.0885
c18_g_t	0.5024	0.0024	0.2223	0.1697	0.5666	−0.0030	0.3316	0.0893
c18_d_td	0.5735	-0.0024	0.2205	0.1678	0.5633	−0.0001	0.3558	0.1014
c18_d_t	0.5458	-0.0020	0.1897	0.1387	0.5668	−0.0020	0.3357	0.1083

Average Retrieval Performance (Core18)

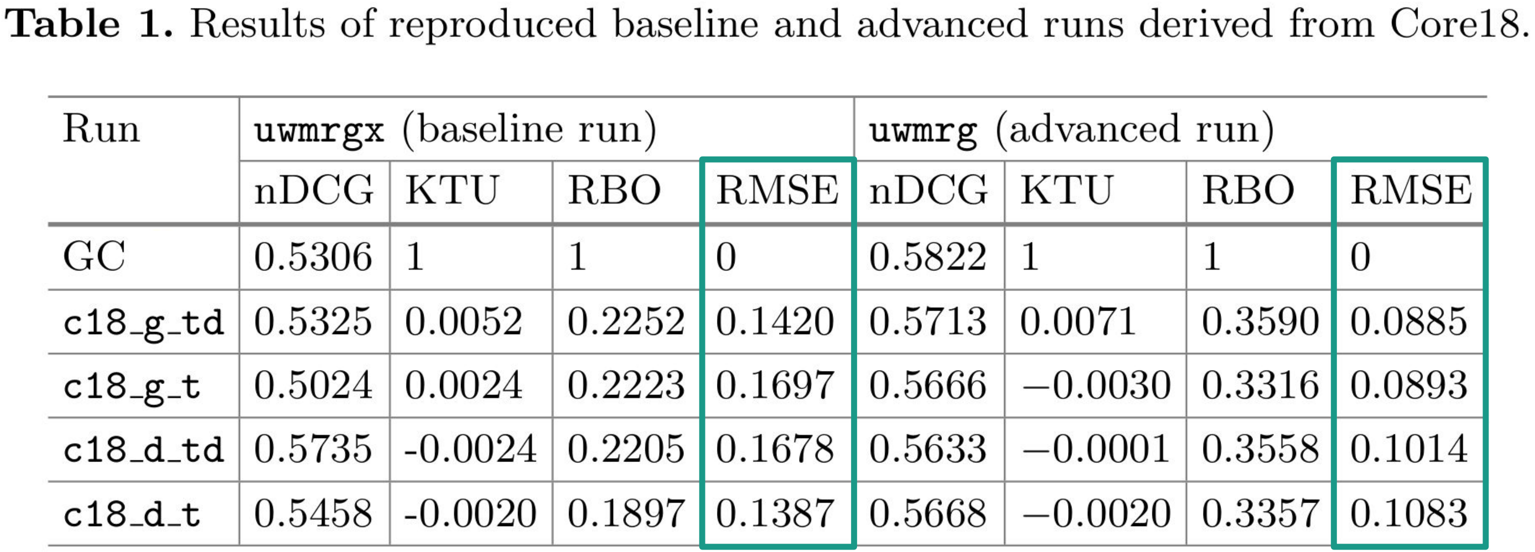
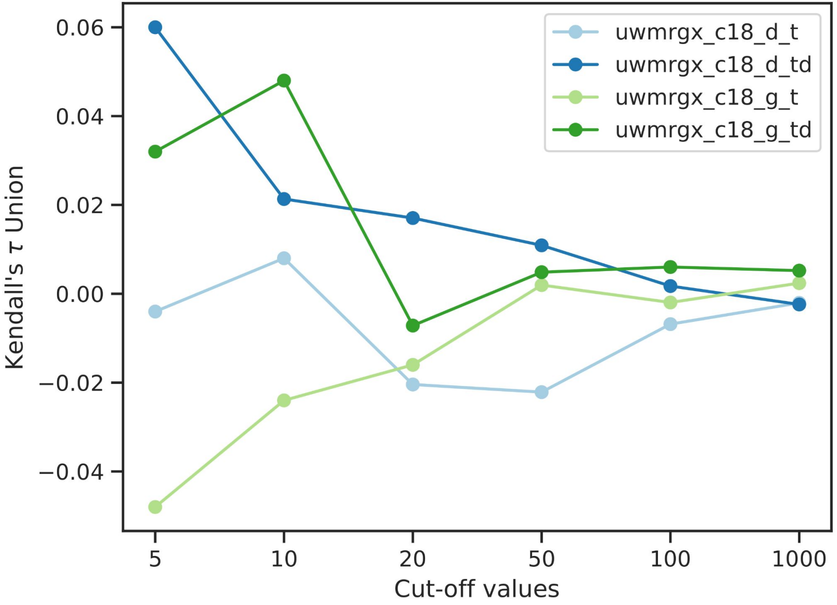
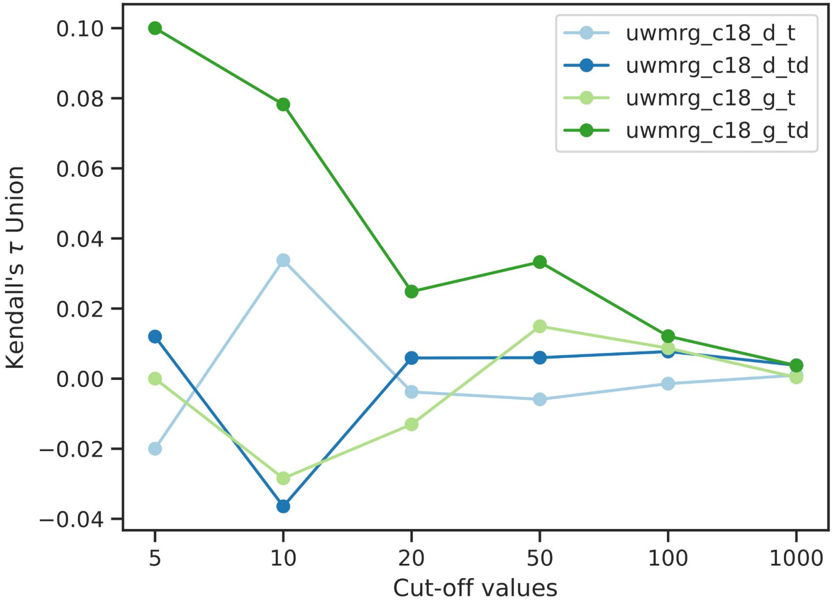


Table 1. Results of reproduced baseline and advanced runs derived from Core18.

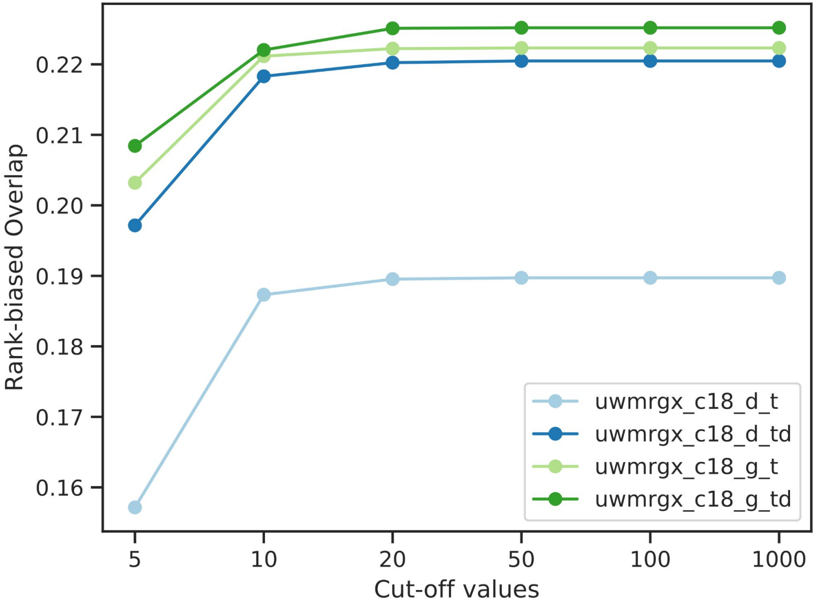
Run	uwmrgx (baseline run)				uwmrg (advanced run)			
	nDCG	KTU	RBO	RMSE	nDCG	KTU	RBO	RMSE
GC	0.5306	1	1	0	0.5822	1	1	0
c18_g_td	0.5325	0.0052	0.2252	0.1420	0.5713	0.0071	0.3590	0.0885
c18_g_t	0.5024	0.0024	0.2223	0.1697	0.5666	−0.0030	0.3316	0.0893
c18_d_td	0.5735	-0.0024	0.2205	0.1678	0.5633	−0.0001	0.3558	0.1014
c18_d_t	0.5458	-0.0020	0.1897	0.1387	0.5668	−0.0020	0.3357	0.1083

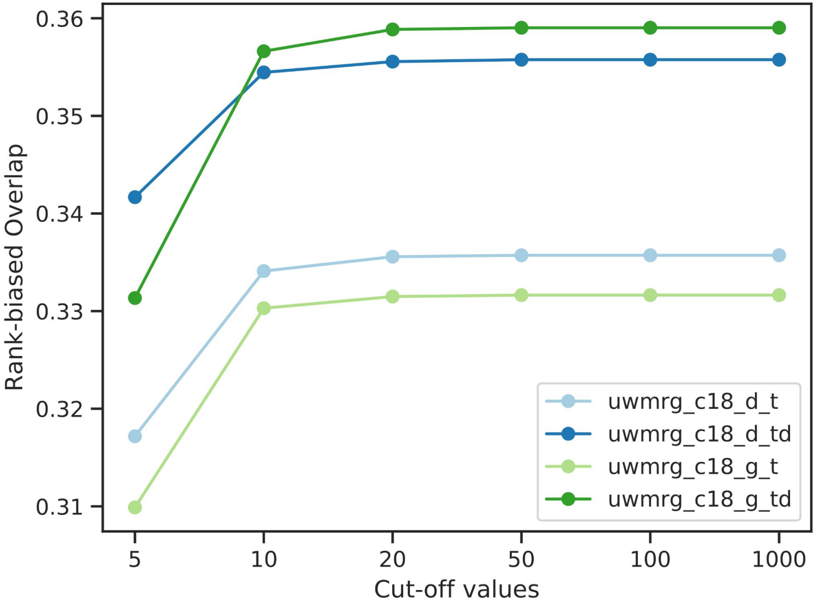
Document orderings - Kendall’s tau Union



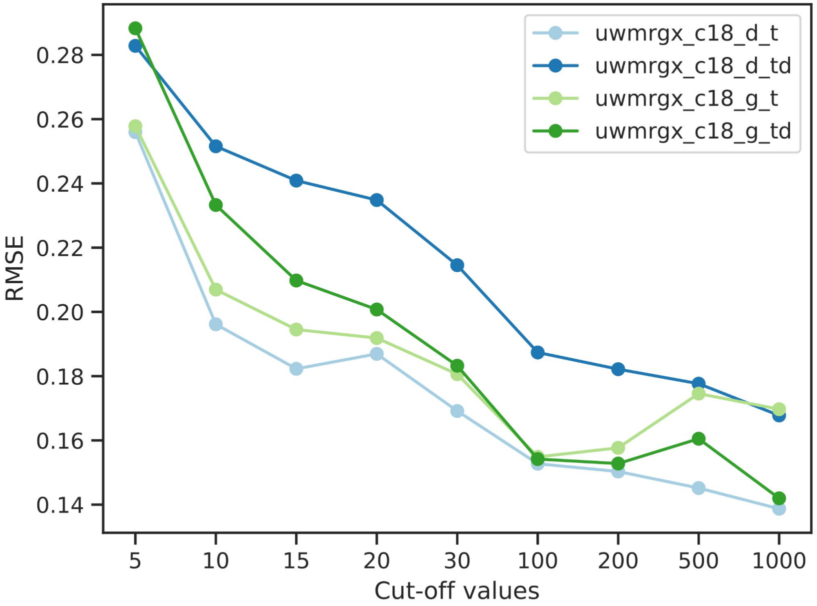


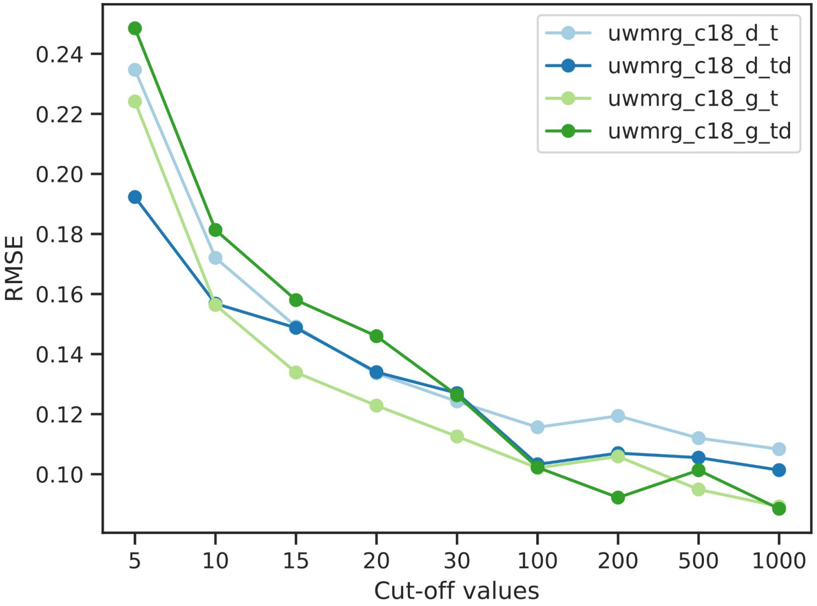
Document orderings - Rank-biased Overlap



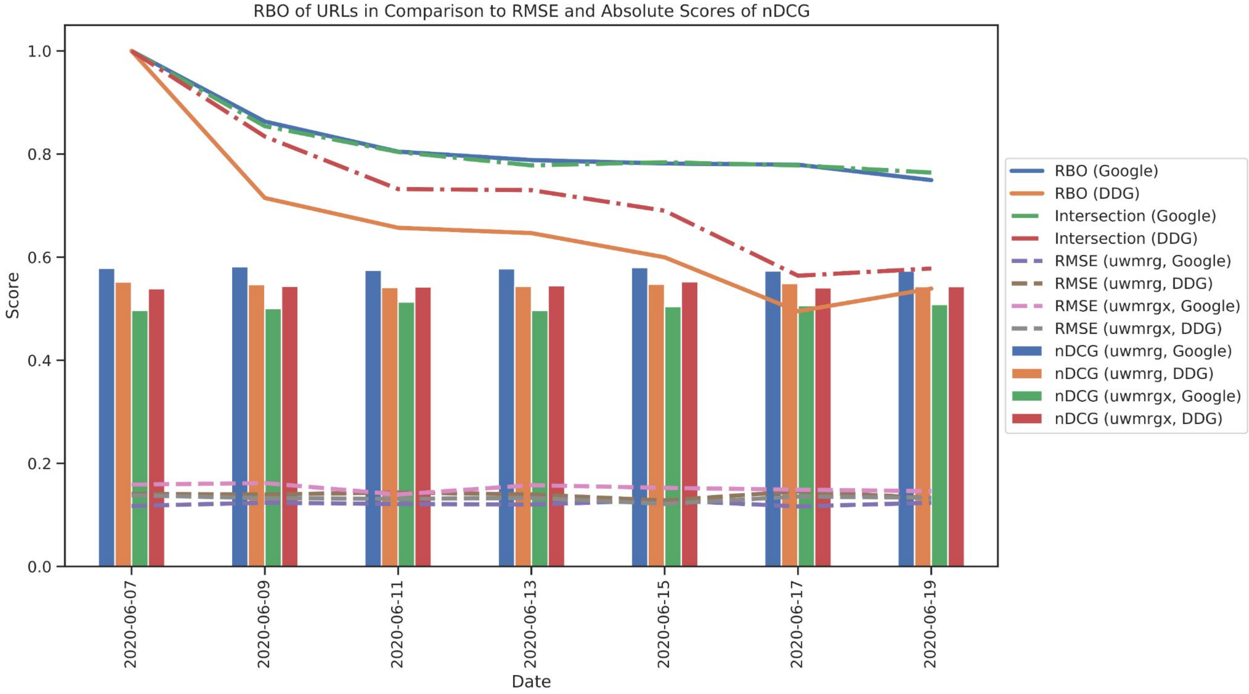


Effectiveness - RMSE

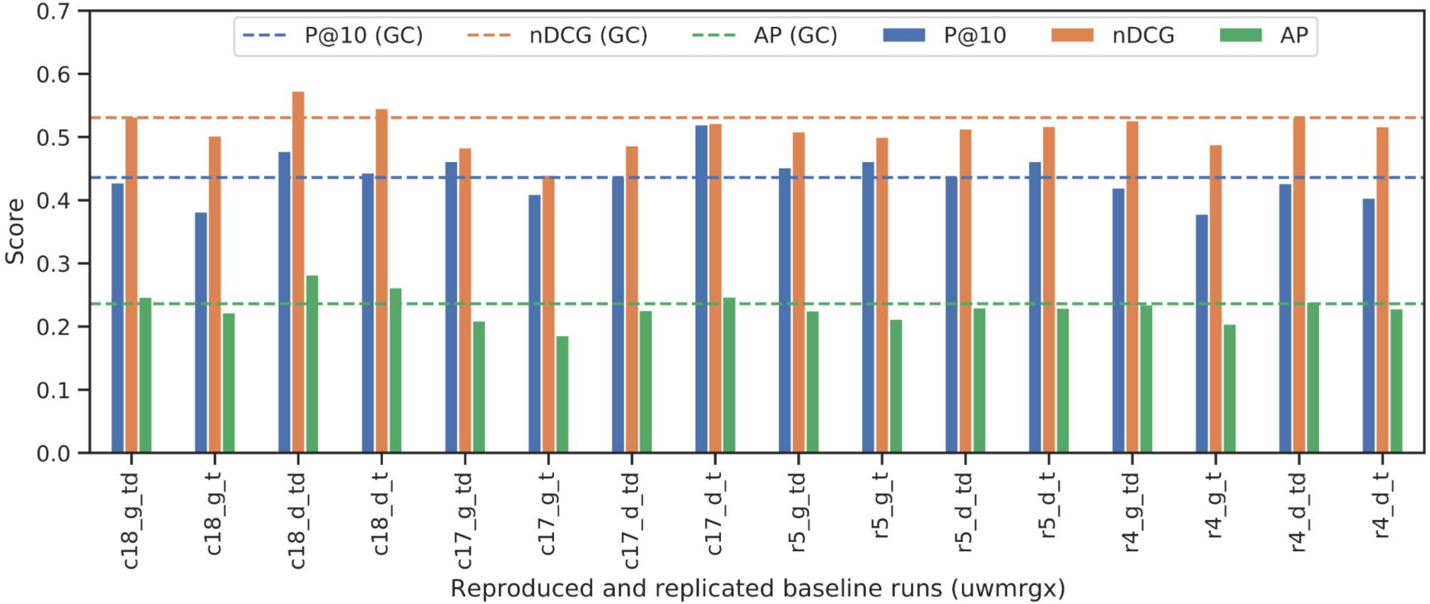




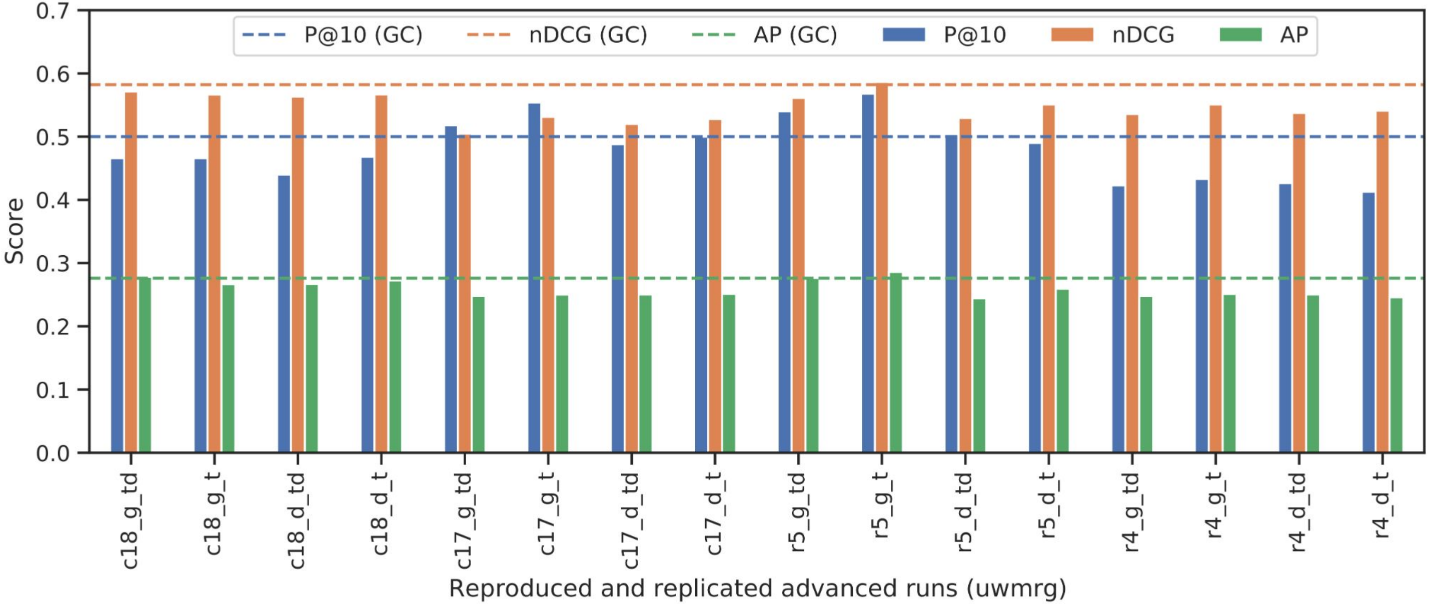
Time analysis



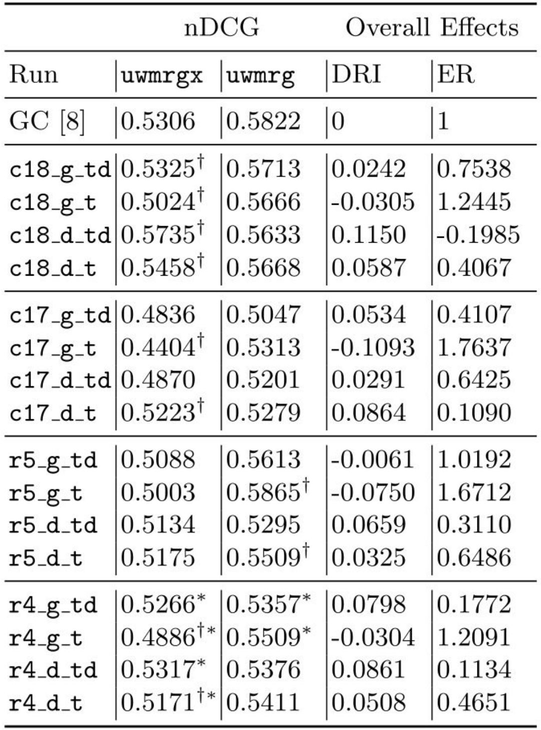
Average Retrieval Performance - uwmrgx



Average Retrieval Performance - uwmrg

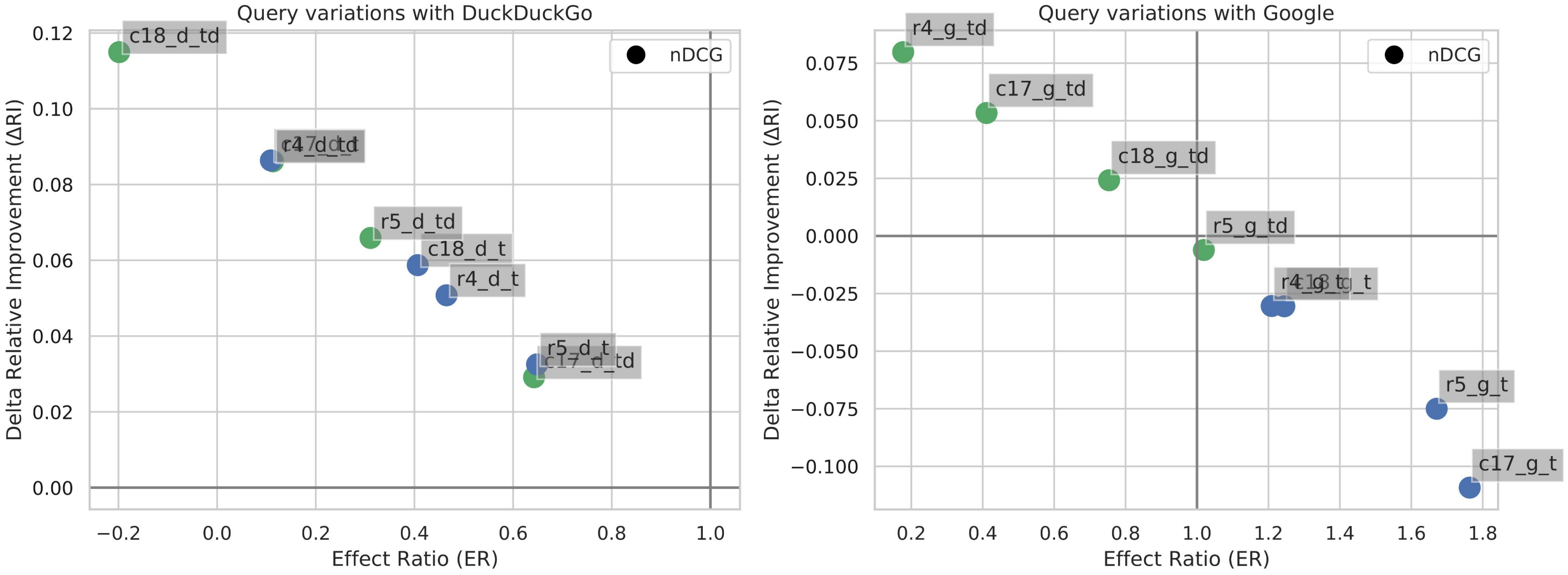


Overall effects

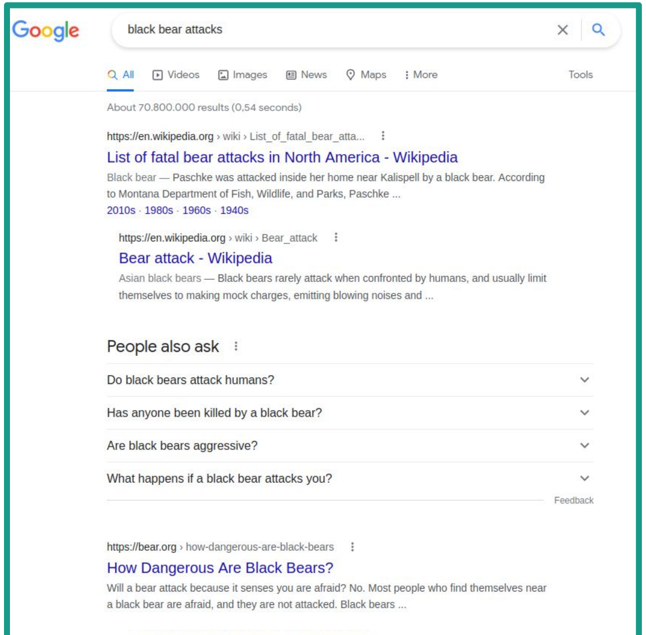


	nDCG		Overall Effects	
Run	uwmrgx	uwmrg	DRI	ER
GC [8]	0.5306	0.5822	0	1
c18_g_td	0.5325†	0.5713	0.0242	0.7538
c18_g_t	0.5024†	0.5666	-0.0305	1.2445
c18_d_td	0.5735†	0.5633	0.1150	-0.1985
c18_d_t	0.5458†	0.5668	0.0587	0.4067
c17_g_td	0.4836	0.5047	0.0534	0.4107
c17_g_t	0.4404†	0.5313	-0.1093	1.7637
c17_d_td	0.4870	0.5201	0.0291	0.6425
c17_d_t	0.5223†	0.5279	0.0864	0.1090
r5_g_td	0.5088	0.5613	-0.0061	1.0192
r5_g_t	0.5003	0.5865†	-0.0750	1.6712
r5_d_td	0.5134	0.5295	0.0659	0.3110
r5_d_t	0.5175	0.5509†	0.0325	0.6486
r4_g_td	0.5266*	0.5357*	0.0798	0.1772
r4_g_t	0.4886$^{\dagger *}$	0.5509*	-0.0304	1.2091
r4_d_td	0.5317*	0.5376	0.0861	0.1134
r4_d_t	0.5171$^{\dagger *}$	0.5411	0.0508	0.4651

Overall effects



Overall effects



black bear attacks

About 70.800.000 results (0,54 seconds)

https://en.wikipedia.org › wiki › List_of_fatal_bear_atta...

List of fatal bear attacks in North America - Wikipedia

Black bear — Paschke was attacked inside her home near Kalispell by a black bear. According to Montana Department of Fish, Wildlife, and Parks, Paschke ...

2010s · 1980s · 1960s · 1940s

https://en.wikipedia.org › wiki › Bear_attack

Bear attack - Wikipedia

Asian black bears — Black bears rarely attack when confronted by humans, and usually limit themselves to making mock charges, emitting blowing noises and ...

People also ask

Do black bears attack humans?

Has anyone been killed by a black bear?

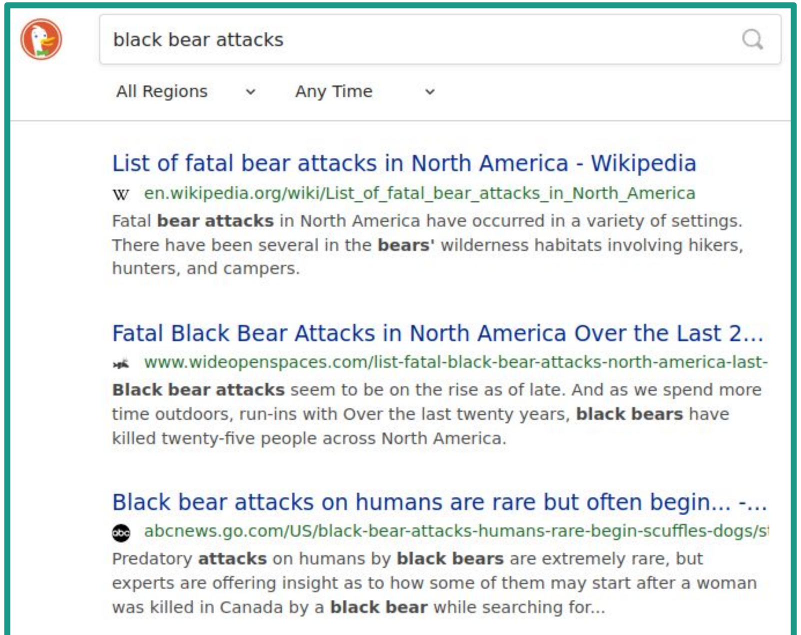
Are black bears aggressive?

What happens if a black bear attacks you?

https://bear.org › how-dangerous-are-black-bears

How Dangerous Are Black Bears?

Will a bear attack because it senses you are afraid? No. Most people who find themselves near a black bear are afraid, and they are not attacked. Black bears ...



black bear attacks

List of fatal bear attacks in North America - Wikipedia

en.wikipedia.org/wiki/List_of_fatal_bear_attacks_in_North_America

Fatal **bear attacks** in North America have occurred in a variety of settings. There have been several in the **bears'** wilderness habitats involving hikers, hunters, and campers.

Fatal Black Bear Attacks in North America Over the Last 2...

www.wideopenspaces.com/list-fatal-black-bear-attacks-north-america-last-

Black bear attacks seem to be on the rise as of late. And as we spend more time outdoors, run-ins with Over the last twenty years, **black bears** have killed twenty-five people across North America.

Black bear attacks on humans are rare but often begin... -...

abcnews.go.com/US/black-bear-attacks-humans-rare-begin-scuffles-dogs/s

Predatory **attacks** on humans by **black bears** are extremely rare, but experts are offering insight as to how some of them may start after a woman was killed in Canada by a **black bear** while searching for...

In sum

RQ1 *How do the components of the workflow, i.e., the query formulation and the web search engine, affect the system performance over time?*

- no substantial differences in **average retrieval performance**
- **performance is robust** over time and different ranking lists

RQ2 *To which extent are the original effects present in different contexts, i.e., with other newswire test collections?*

- **short queries** with **Google** lead to **stronger overall effects**
- **low overall effects** with **DuckDuckGo** due to high baseline scores
- overall effects of longer queries stay below those of the original experiments

Data & Code

https://zenodo.org/record/4105885

https://github.com/irgroup/clef2021-web-prf/

Thank you for your attention!