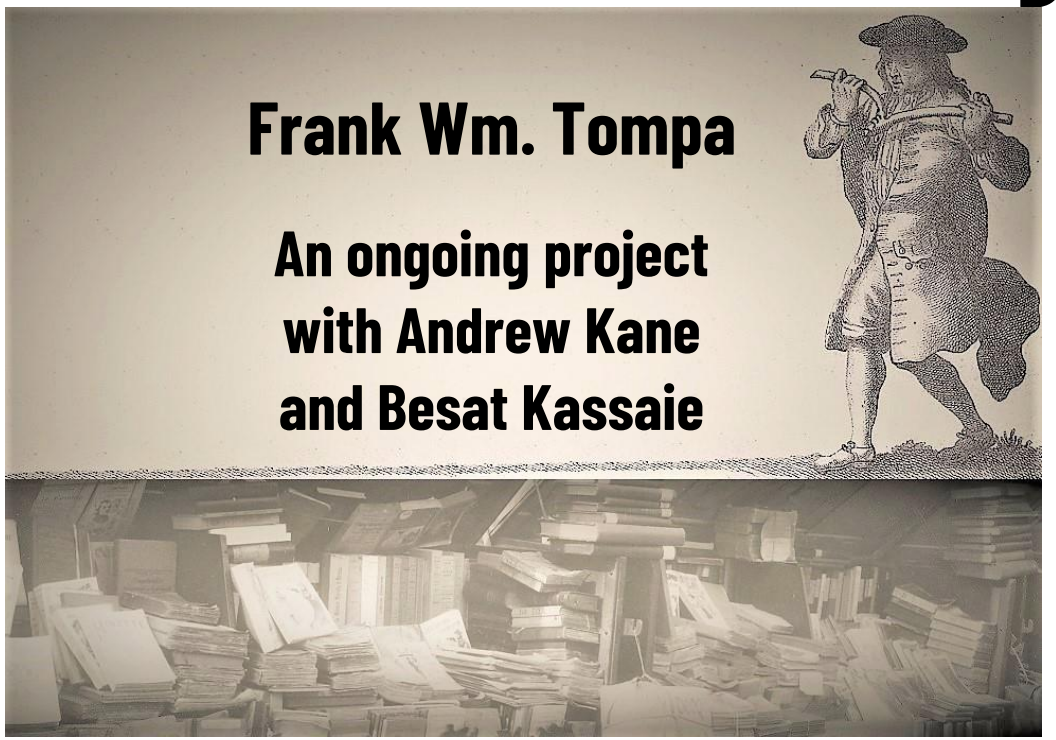
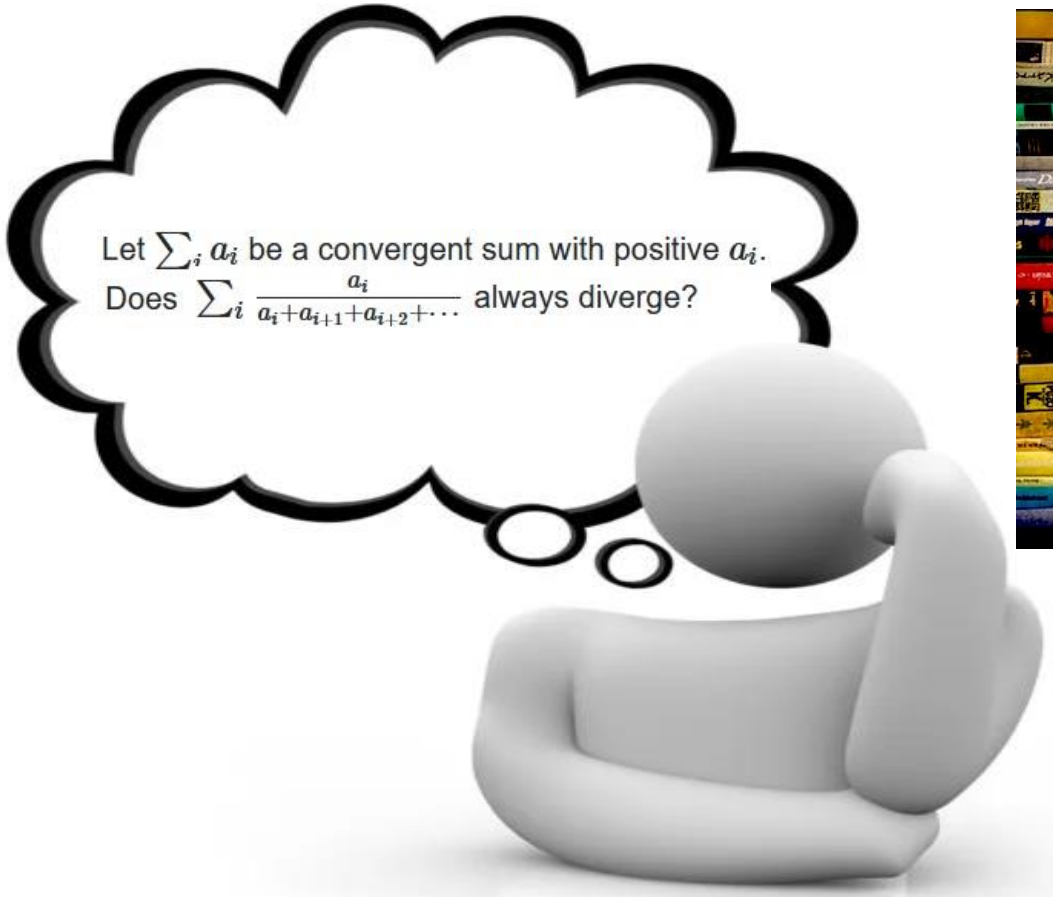
Math Information Retrieval using a Conventional Search Engine

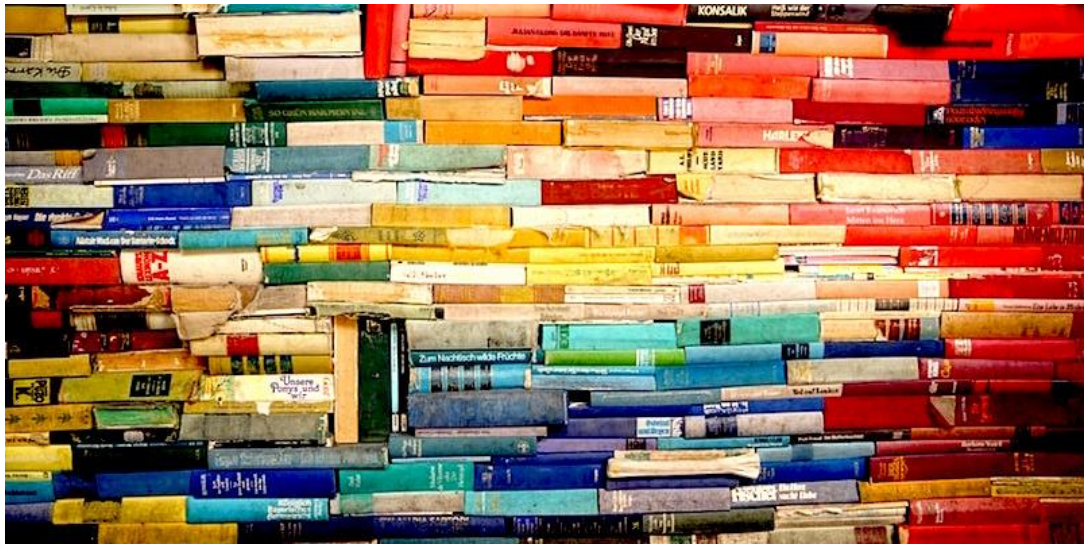


UNIVERSITY OF WATERLOO | DSg Data Systems Group

Motivation

- Searching a STEM corpus requires awareness of math formulas and math terminology





Searching Math Stack Exchange

- Given a query from Math Stack Exchange (MSE), find suitable answers in the corpus of other MSE questions and answers and return a ranked list.

Evaluate the definite integral: $\int_0^\infty e^{-hx^2}\,\mathrm{d}x$

calculus , integration , definite-integrals

where $h > 0$. Could someone explain to me how to solve it? I searched the internet and I found the result is $\frac{\sqrt{\pi}}{2\sqrt{h}}$ but I couldn't undersand Gauss error function - that is involved in solving.

Ranked list of answers

Human judgements

1 Change variables. Let $z = x^2$. We find $\int_0^\infty e^{-x^2}dx = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$.

Addendum: Setting $z = 1/2$ in Euler's reflection formula, $\Gamma(1-z)\Gamma(z) = \pi/\sin \pi z$, we find $\Gamma(1/2) = \sqrt{\pi}$.

Medium Relevance

2 First solve by simple substitution the integral $\int_0^\infty e^{-ax^2}dx$ and the differentiate that result twice with respect to a. The answer is a special case of $\int_{-\infty}^{\infty} x^{2m}e^{-\beta x^{2m}}dx = \Gamma(v)/(n\beta^v)$ with $v \equiv (2m+1)/(2n)$.

Irrelevant

3 In order to solve the integral by polar coordinates first consider $I_s = \int_{-\infty}^{\infty} \mathrm{e}^{-sx^2}\,\mathrm{d}x$. The integral you seek will be obtained by differentiation as $-\frac{\mathrm{d}}{\mathrm{d}s} I_s\big|_{s=1}$.

Now, to evaluate I_s:

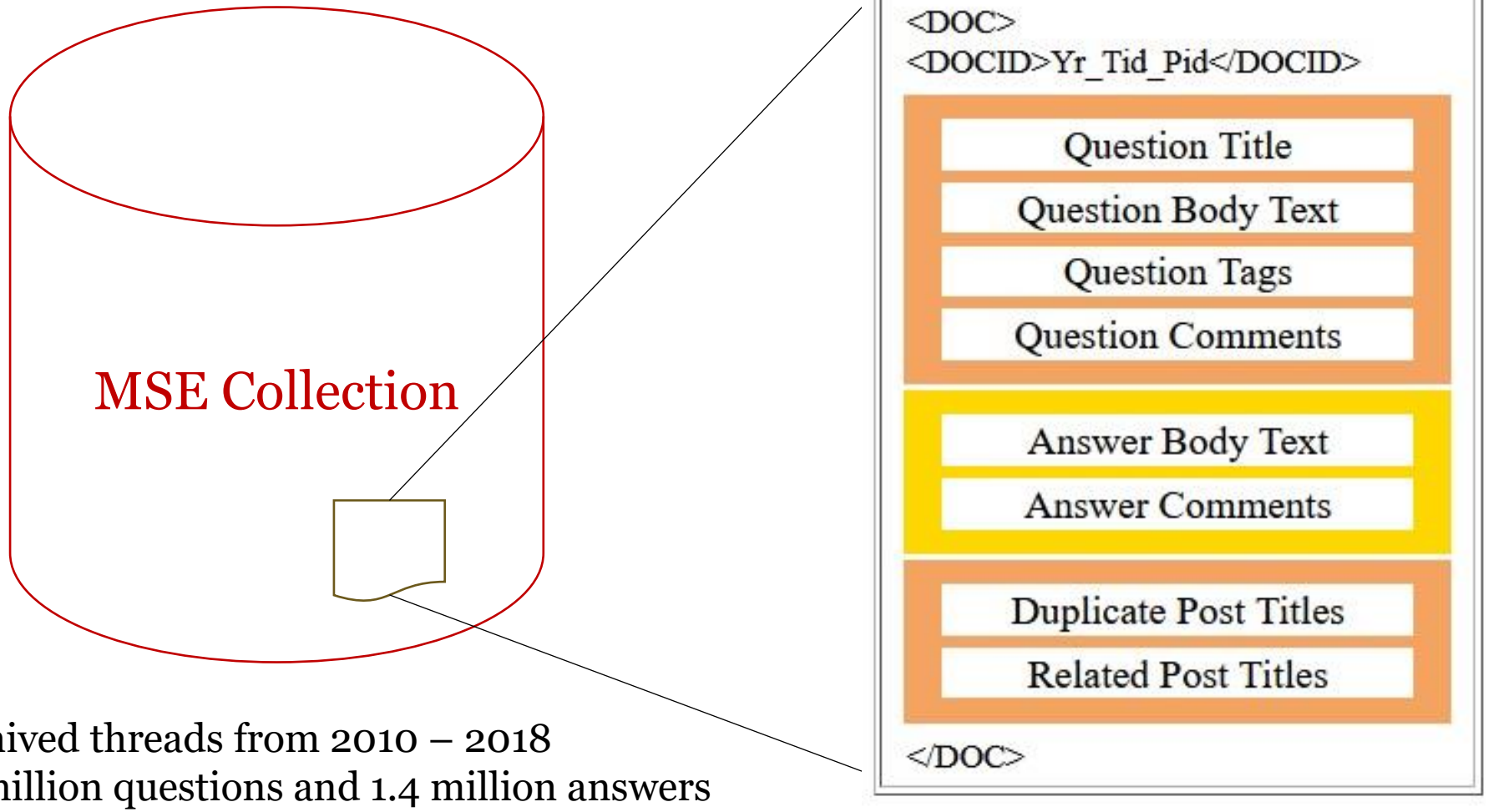
$I_s^2 = \int_{-\infty}^{\infty} \mathrm{e}^{-sx^2}\,\mathrm{d}x \cdot \int_{-\infty}^{\infty} \mathrm{e}^{-sy^2}\,\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-s(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y$

Now change variables into polar coordinates $x = r\sin\theta$ and $y = r\cos\theta$. ...

High Relevance

...

Corpus Document Format



- Archived threads from 2010 – 2018
- 1.1 million questions and 1.4 million answers

https://github.com/kiking0501/MathDowsers-ARQMath

Conventional Search Engine

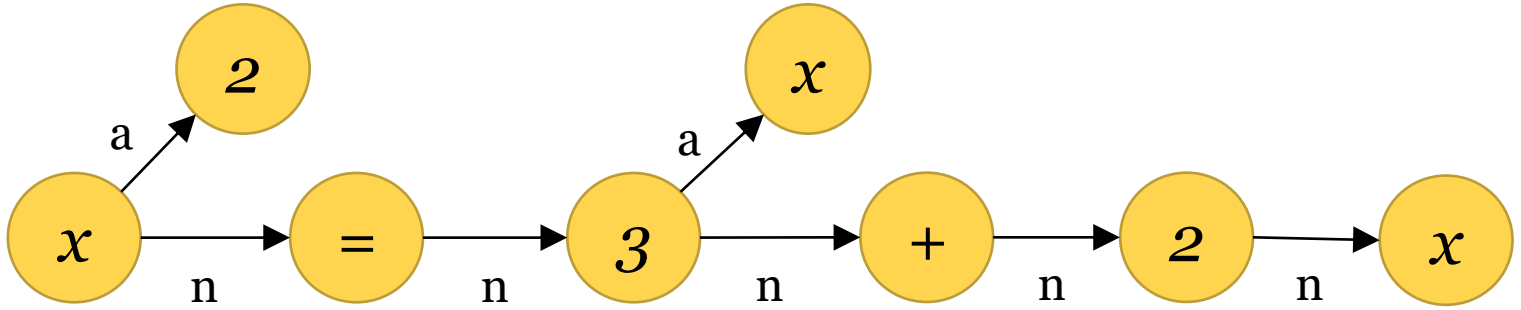
- Convert natural language text to tokens:
  - Treat hyphens as spaces (e.g., “Knuth-Morris-Pratt” → “Knuth Morris Pratt”
  - Deal with other punctuation (e.g., “Beth’s tableaux.” → “Beth tableux”)
  - Use spaces as token boundaries (e.g., “Abelian group” → “Abelian”, “group”)
  - Fold to lower case (e.g., “Graph Theory” → “graph theory”)
  - Apply stemming (e.g., “solving” | “solve” | “solves” → “solv”)
  - Remove stop words (e.g., “the”, “in”)

} extended to Unicode

- Treat document as *bag of tokens*
  - Postings lists for each token comprised of docid/frequency pairs
- Use BM25 (variant of tf*idf) for searching and scoring matches

Handle Formulas by Converting to Bags of Math Tuples

- $x^2 = 3^x + 2x$



- Tuples from the Symbol Layout Tree:
  - 7 symbol pairs (s1,s2,R)
  - 3 terminal symbols (s,!0)
  - 2 compound symbols (s,R1R2…Rk)
  - 3 duplicate symbols (s,P) or (s,P1,P2) + 3 wilds (w,P) or (w,P1,P2)
  - 18 augmented with locations

https://github.com/fwtompa/mathtuples.git

Converting a question to a formal query

- Q: Evaluate the definite integral: $\int_0^\infty e^{-hx^2} dx$ where h>0. Could someone explain to me how to solve it? I searched the internet and I found the result is $\frac{\sqrt{\pi}}{2\sqrt{h}}$ but I couldn't undersand Gauss error function - that is involved in solving.
- Tags: calculus, integration, definite-integrals
- Search terms: all formulas and “mathy” words

Converting a question to a formal query

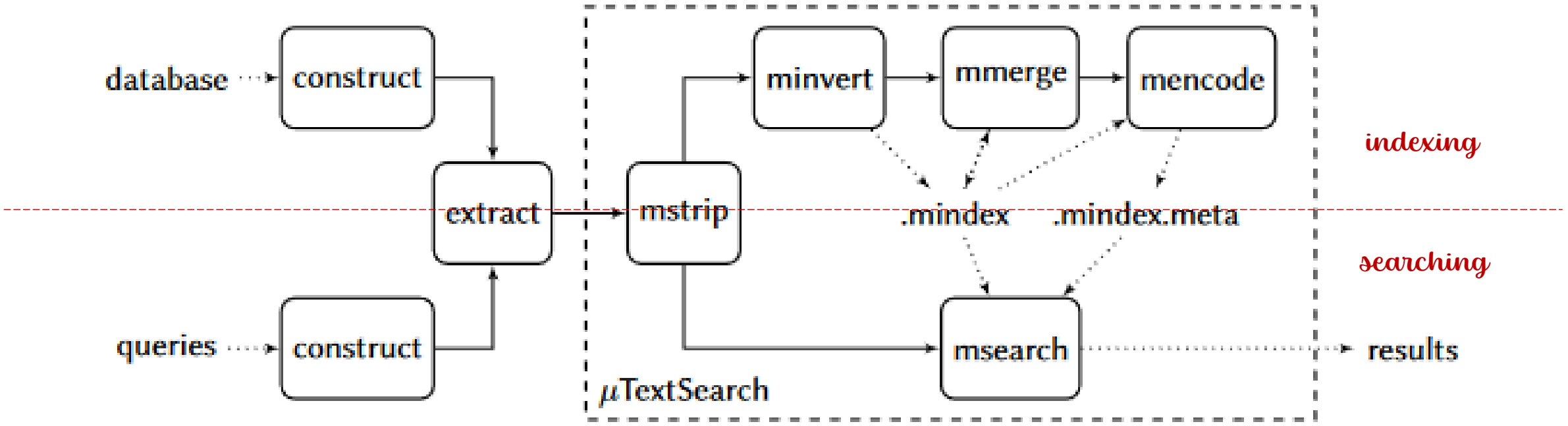
Evaluate definite integral $\int_0^\infty e^{-hx^2}\,dx$ h>0
explain solve searched internet
found result $\frac{\sqrt{\pi}}{2\sqrt{h}}$ Gauss
error function involved solving
calculus integration definite integrals

- Search terms: all formulas and "mathy" words

where $\int_0^\infty e^{-hx^2}\,dx$ → #(∫,v!e,n)# #(∫,∞,a)# #(∫,n!0,b)# #(∫,[nab],-)# #(∞,!0)# …

N.B. Unlike traditional search, query has *many* search terms!

Core engine



Adaptations: (1) recognize math tokens in minvert, and (2) use α to balance math vs. text tokens in msearch

https://github.com/andrewrkane/mtextsearch

Efficiency

- Indexing time

Step	Processing	Indexing Speed † (sec)	
1.	construct - generate corpus	46581	Python
2.	extract math tuples	15272	
3 a.	mstrip	2376	C++
3 b.	minvert	4176	
3 c.	mencode	154	

- Corpus size

  Data: 15.9 GB compressed to 1.6 GB (gzip)

  Index: 1.9 GB from minvert + 174 MB from mencode

- Approx. 11 seconds per query (non-optimized, exhaustive-OR)

---

† Mac OSX 10.11.6 laptop; Intel Core i7-4770HQ Processor (4 Cores 8 Threads, 2.2-3.4Ghz); 16GB RAM; 256GB flash

Answer Retrieval for Questions on Math (ARQMath) Benchmarks

	ARQMath-1 (77 topics)			ARQMath-2 (71 topics)			ARQMath-3 (78 topics)		
	nDCG'	MAP'	P'@10	nDCG'	MAP'	P'@10	nDCG'	MAP'	P'@10
MathDowsers 2023:	**0.515**	**0.265**	0.309	**0.523**	**0.231**	0.269	0.498	0.181	0.263
MathDowsers 2022:	0.511	0.261	0.307	0.510	0.223	0.265	0.474	0.164	0.247
MathDowsers 2021:	0.457	0.207	0.267	0.462	0.187	0.241	0.447	0.159	0.236
MathDowsers 2020:	0.345	0.139	0.162						
Approach0 2022 (M):	0.462	0.244	**0.321**	0.460	0.226	**0.296**	**0.514**	**0.219**	**0.349**
MSM 2022:	0.422	0.172	0.197	0.381	0.119	0.152	0.504	0.157	0.241

- Year-over-year improvement in effectiveness
- Some unexpected deterioration in effectiveness on third benchmark

Ongoing investigation

- Natural language text
  - Selection: Improved keyword and keyphrase extraction
  - Augmentation: ChatGBT to augment query terms?
- Formulas
  - Representation: Extract features from Content MathML (operator tree)
  - Features: Select most effective features from symbol layout and operator trees
- Execution
  - Efficiency: Implement MaxScore dynamic pruning and split-lists
  - Effectiveness: Support weighted fields in documents and queries

UNIVERSITY OF WATERLOO | DSg Data Systems Group

Thank You