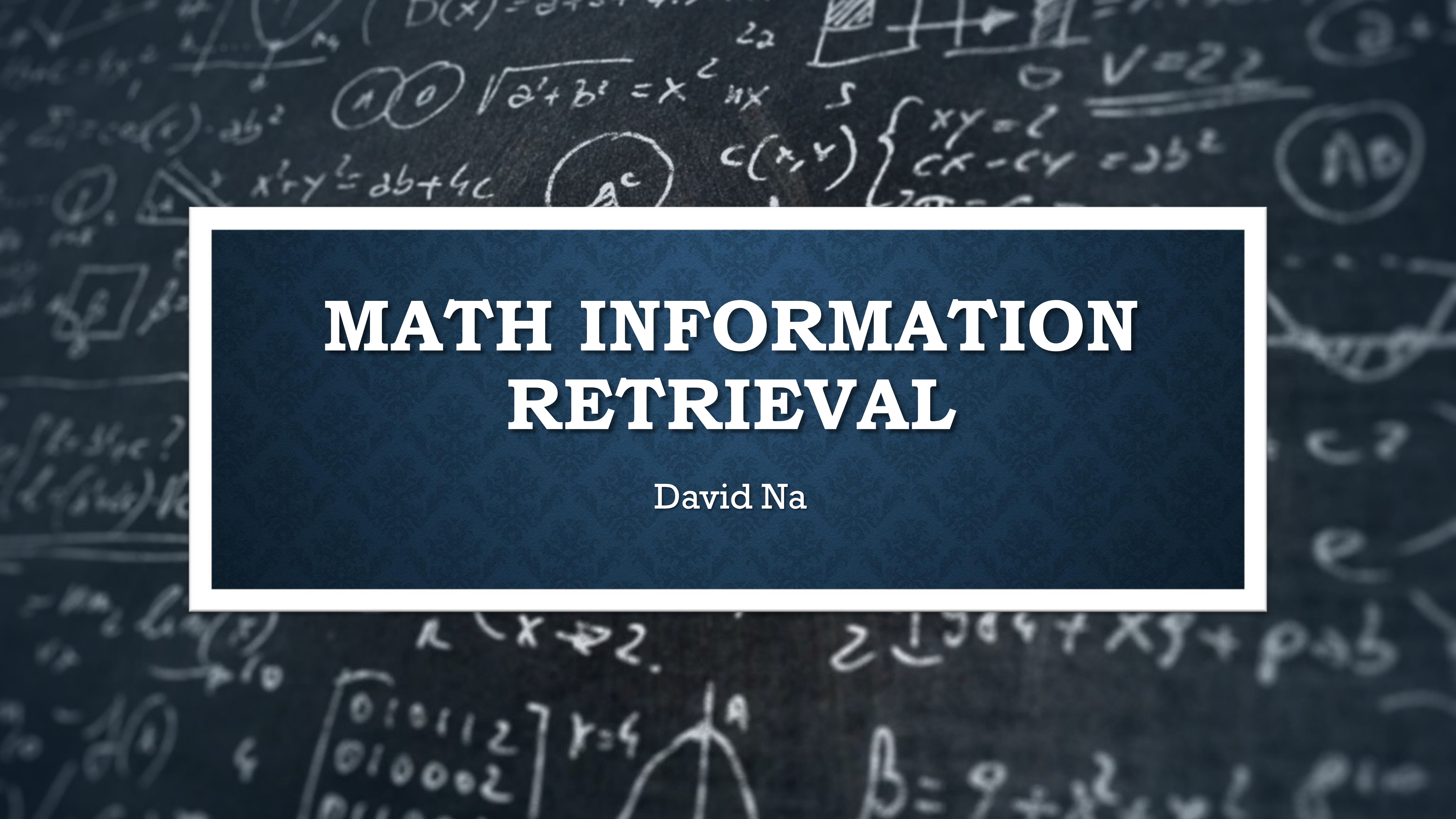


MATH INFORMATION RETRIEVAL

David Na

WHAT IS MATH INFORMATION RETRIEVAL?

Math IR is "concerned with finding information in documents that include mathematics".

Document similarity and retrieval is calculated considering any math keywords and/or formulae found in the collection and query.

It's the same as standard Information Retrieval but instead of querying just strings in documents, you query using some mathematical context.

Math IR is essentially a performing "standard" IR using Math concepts

First Math-aware search engines developed in 2000:

NIST Digital Library of Mathematical Functions (DLMF)

Examples:

Given a mathematical concept in keywords and/or formulae, find the technical papers that use similar mathematical models.

Another example is you can browse for documents containing a given formula.

DIFFICULTIES OF MATH IR

Mathematical notation is dialectic, some variables represent similar concepts but are named differently depending on the subject area

Operations can be defined in different ways depending on context. For instance, a line over a variable can be a division simple or a Boolean negation depending on context.

Math formulae can look very similar but still be very different.

Most of the time Math IR queries are combinations of keywords and formulae.

TYPES OF MATH IR SYSTEMS

Formula Retrieval

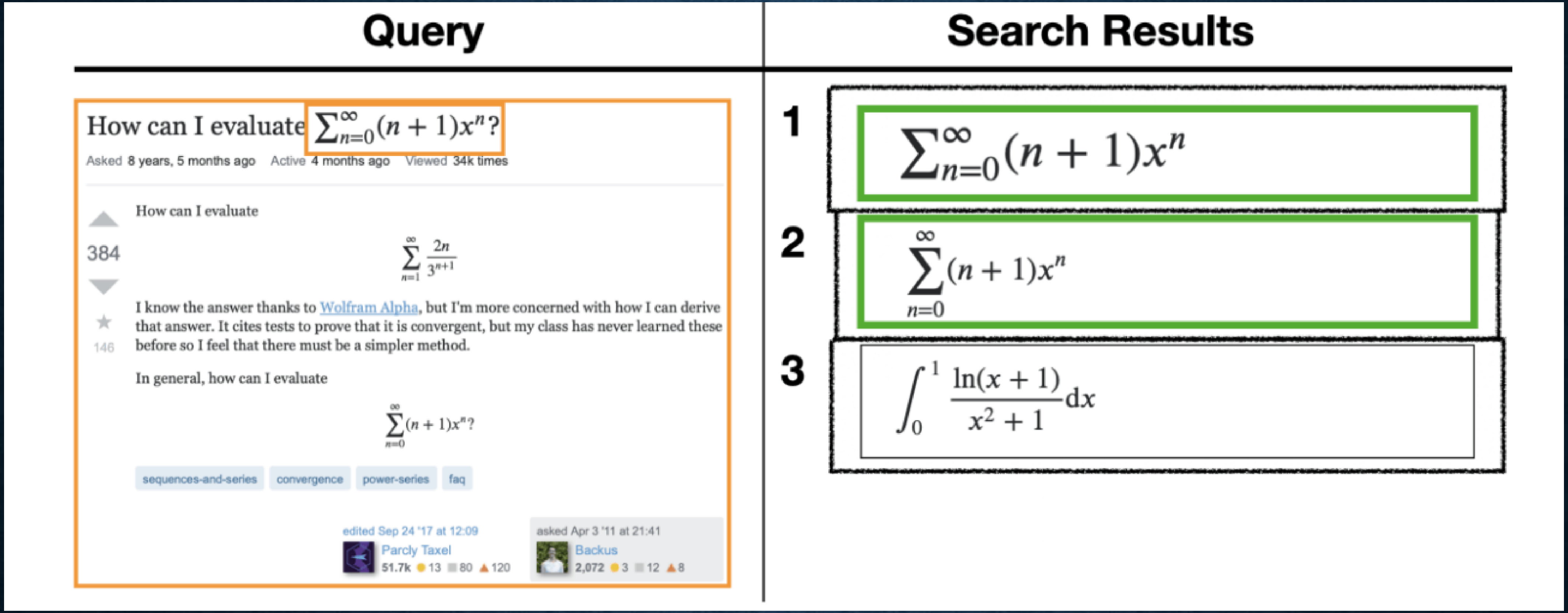
- Related to computation
- Example: How do you compute this formula?

Text Answer Retrieval

- Related to a Mathematical Concept
- Example: What is a Hamiltonian Circuit?

Formula and Text Retrieval

- Related to Proofs
- Example: Prove *a* is divisible by *b*



How can I evaluate $\sum_{n=0}^{\infty}(n+1)x^n$?

Asked 8 years, 5 months ago Active 4 months ago Viewed 34k times
384
146

How can I evaluate

$$\sum_{n=1}^{\infty}\frac{2n}{3^{n+1}}$$

I know the answer thanks to Wolfram Alpha, but I'm more concerned with how I can derive that answer. It cites tests to prove that it is convergent, but my class has never learned these before so I feel that there must be a simpler method.

In general, how can I evaluate

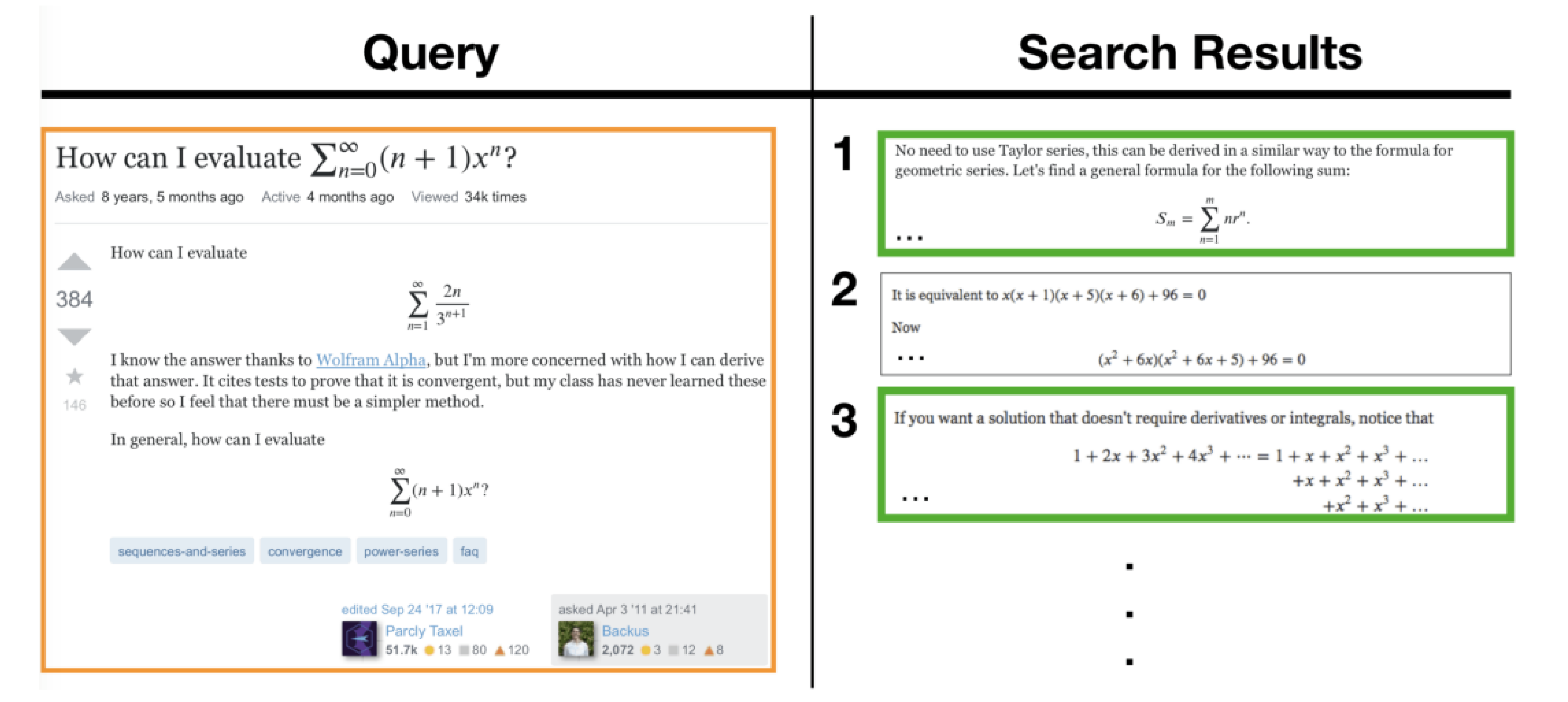
$$\sum_{n=0}^{\infty}(n+1)x^n?$$

1 $\sum_{n=0}^{\infty}(n+1)x^n$

2 $\displaystyle\sum_{n=0}^{\infty}(n+1)x^n$

3 $\displaystyle\int_0^1 \frac{\ln(x+1)}{x^2+1}\mathrm{d}x$

FORMULA RETRIEVAL



Query

How can I evaluate $\sum_{n=0}^{\infty}(n+1)x^n$?

Asked 8 years, 5 months ago Active 4 months ago Viewed 34k times

384

146

How can I evaluate

$$\sum_{n=1}^{\infty}\frac{2n}{3^{n+1}}$$

I know the answer thanks to Wolfram Alpha, but I'm more concerned with how I can derive that answer. It cites tests to prove that it is convergent, but my class has never learned these before so I feel that there must be a simpler method.

In general, how can I evaluate

$$\sum_{n=0}^{\infty}(n+1)x^n?$$

sequences-and-series convergence power-series faq

edited Sep 24 '17 at 12:09
Parcly Taxel
51.7k ● 13 ■ 80 ▲ 120

asked Apr 3 '11 at 21:41
Backus
2,072 ● 3 ■ 12 ▲ 8

Search Results

1 No need to use Taylor series, this can be derived in a similar way to the formula for geometric series. Let's find a general formula for the following sum:

$$S_m = \sum_{n=1}^{m} nr^n.$$

...

2 It is equivalent to $x(x+1)(x+5)(x+6)+96=0$

Now

$$(x^2+6x)(x^2+6x+5)+96=0$$

...

3 If you want a solution that doesn't require derivatives or integrals, notice that

$$\begin{aligned}1+2x+3x^2+4x^3+\cdots = 1+x+x^2+x^3+\ldots\\ +x+x^2+x^3+\ldots\\ +x^2+x^3+\ldots\end{aligned}$$

...

.
.
.

TEXT ANSWER RETRIEVAL

TWO MAIN MATH IR CONFERENCES

NTCIR – NII Testbeds and Community for Information Access Research

- Run by National Institute of Informatics (Tokyo, Japan)
- Math IR was only the focus in workshops 10, 11 and 12
- Held from 2014-2016

ARQMath - Answer Retrieval for Questions on Math

- Hosted by Rochester Institute of Technology and the NSF
- ARQMath has had 3 iterations
- More Recently held in 2019-2021

NTCIR

- NTCIR 10-12 (2014-2016): Math IR
  - 3 Main Tasks:
    - Formula Search: Search given a formula
    - Formula + Text Search: Search documents given keywords and formula
    - Open Information Retrieval: Search the collection given a text query
  - NTCIR-12 focused on the Wikipedia dataset
    - Wikipedia Formula Browsing: Focusing on formula search
    - Formula Similarity Search: Finding formulas related to the query
  - Two Datasets: Wikipedia was introduced in NTCIR 12
    - 100,000 scientific articles from arXiv and 35,000 Wikipedia articles
    - All HTML files
    - Includes 30-60M math formulae
  - Given a set of queries, the Math IR system must return a ranked list of search results from both datasets.
    - Queries will be a set of formulae and keywords and you must search within the document collection

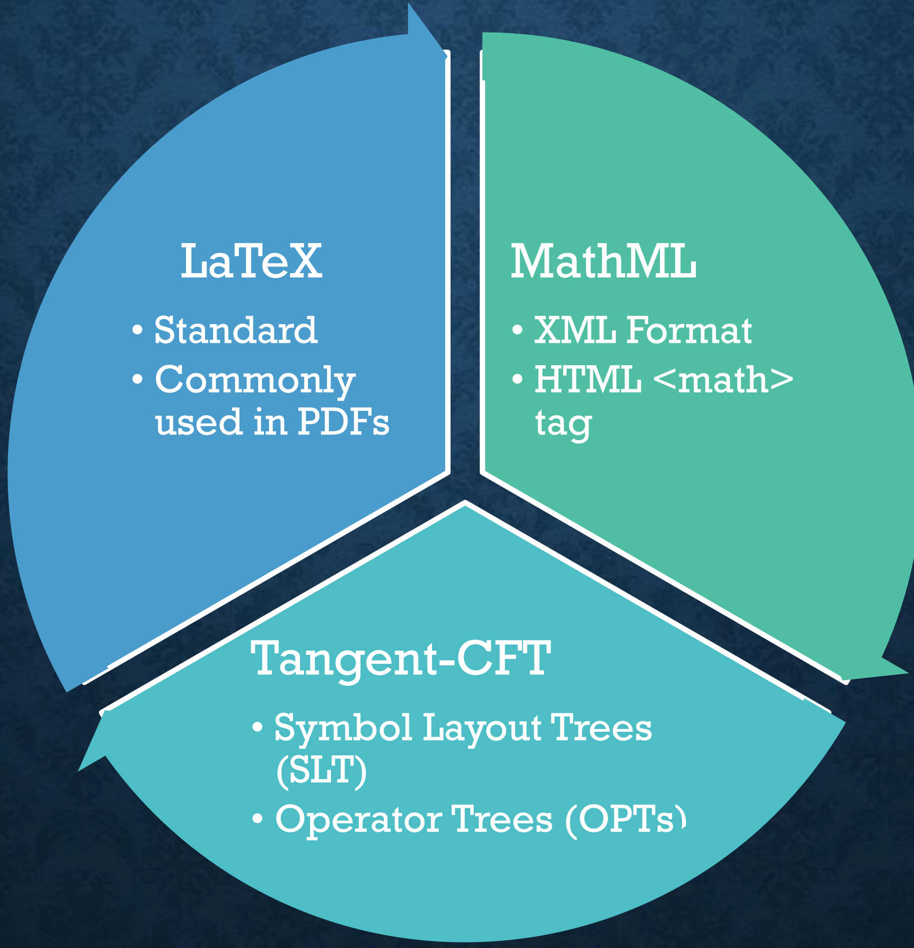
https://www.cs.rit.edu/~rlaz/files/ntcir12-mathir.pdf

ARQMATH

- ARQMath 1, 2 and 3 (2019-2021):
    - Focused directly on Answer Retrieval
    - Dataset: Question + Answer posts from Stack Exchange
        - Data from 2010 to 2020
        - 28 Million mathematical formulas in LaTeX
    - Task 1: Finding relevant answers to a mathematical question
        - Given a math question as a query, search all answers and return relevant answers
    - Task 2: Search based on a formula
        - Given a question with an identified formula as a query, return relevant formulas
    - Task 3: Open Domain Question Answering (ARQMath 3)
        - Like Task 1, but now the question can come from anywhere (even written by hand). Task 1 used questions from the Stack Exchange dataset as input. Task 3 can take as input any query.

http://sigir.org/wp-content/uploads/2020/12/p06.pdf

REPRESENTING MATH FORMULAS



LaTeX

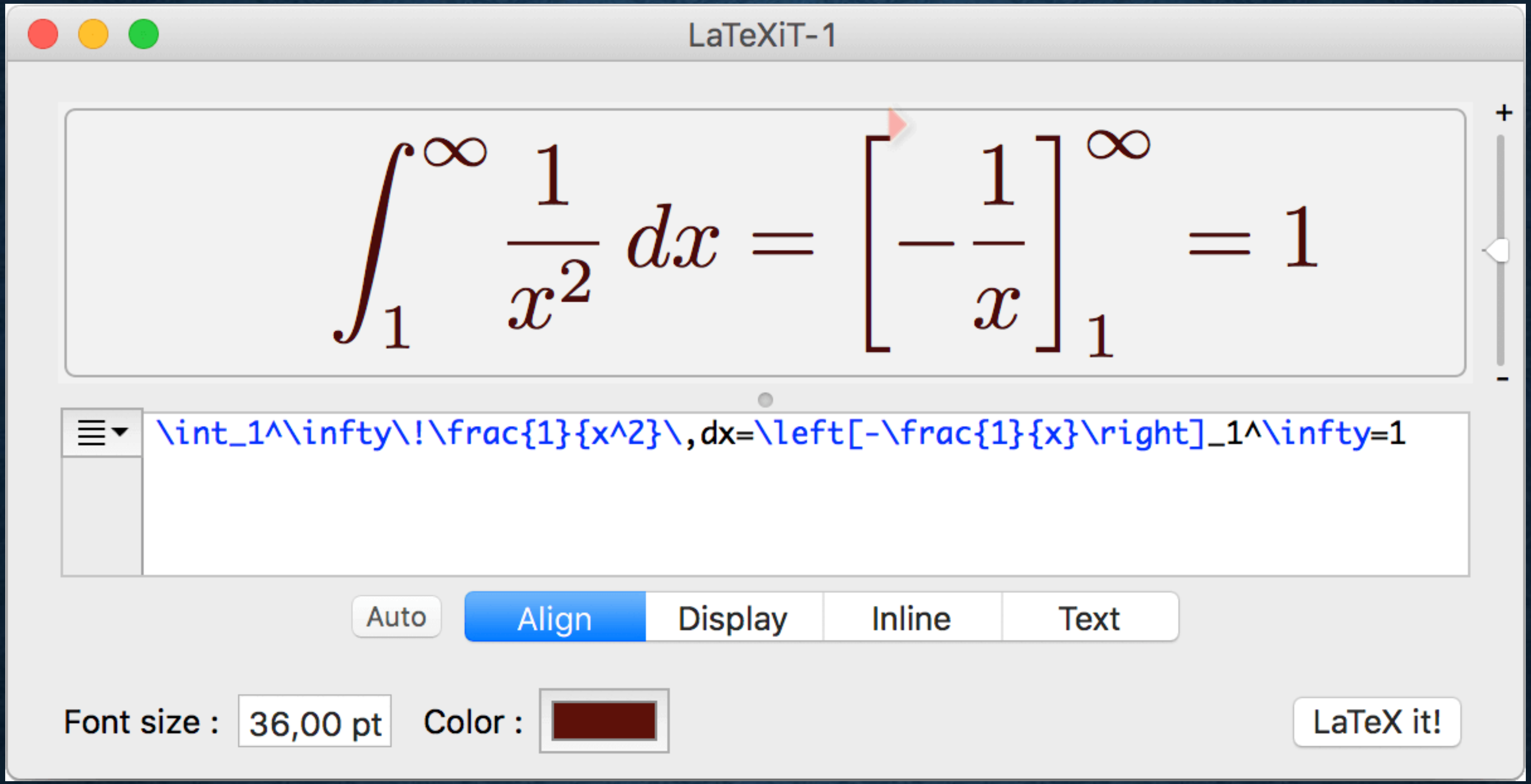
- Standard
- Commonly used in PDFs

MathML

- XML Format
- HTML <math> tag

Tangent-CFT

- Symbol Layout Trees (SLT)
- Operator Trees (OPTs)



$$\int_1^\infty\!\frac{1}{x^2}\,dx=\left[-\frac{1}{x}\right]_1^\infty=1$$

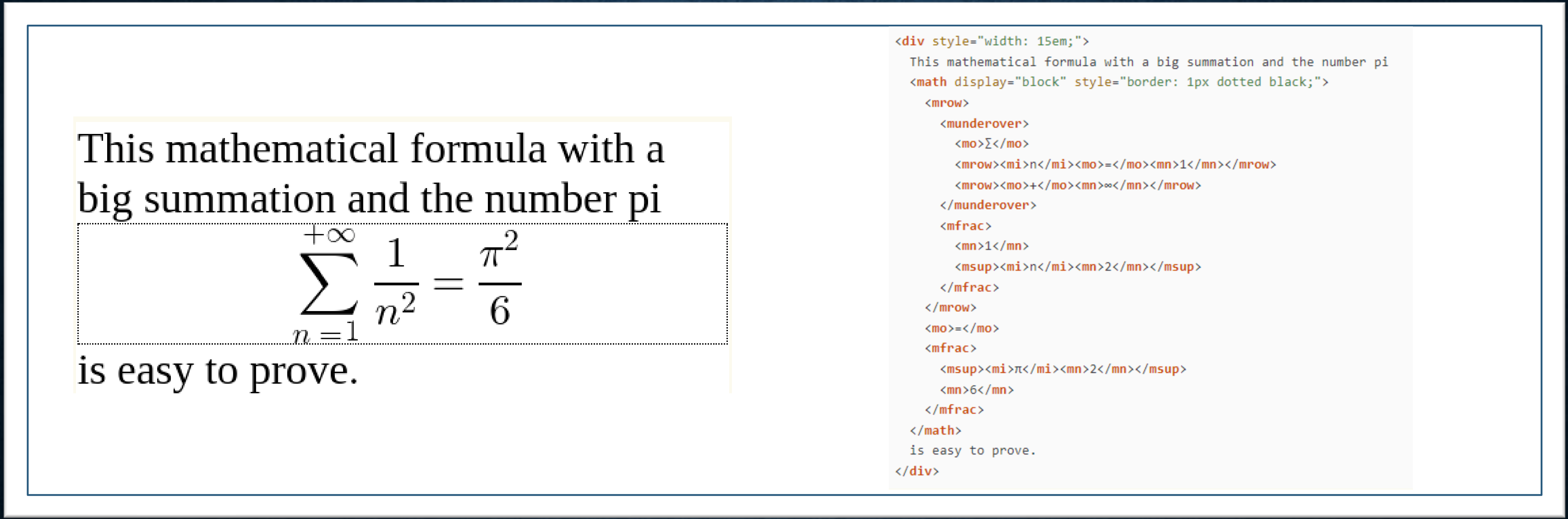
```
\int_1^\infty\!\frac{1}{x^2}\,dx=\left[-\frac{1}{x}\right]_1^\infty=1
```

Font size : 36,00 pt

Color : 🟥

LATEX

Reference: http://www.malinc.se/math/latex/basiccodeen.php



This mathematical formula with a big summation and the number pi

$$\sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

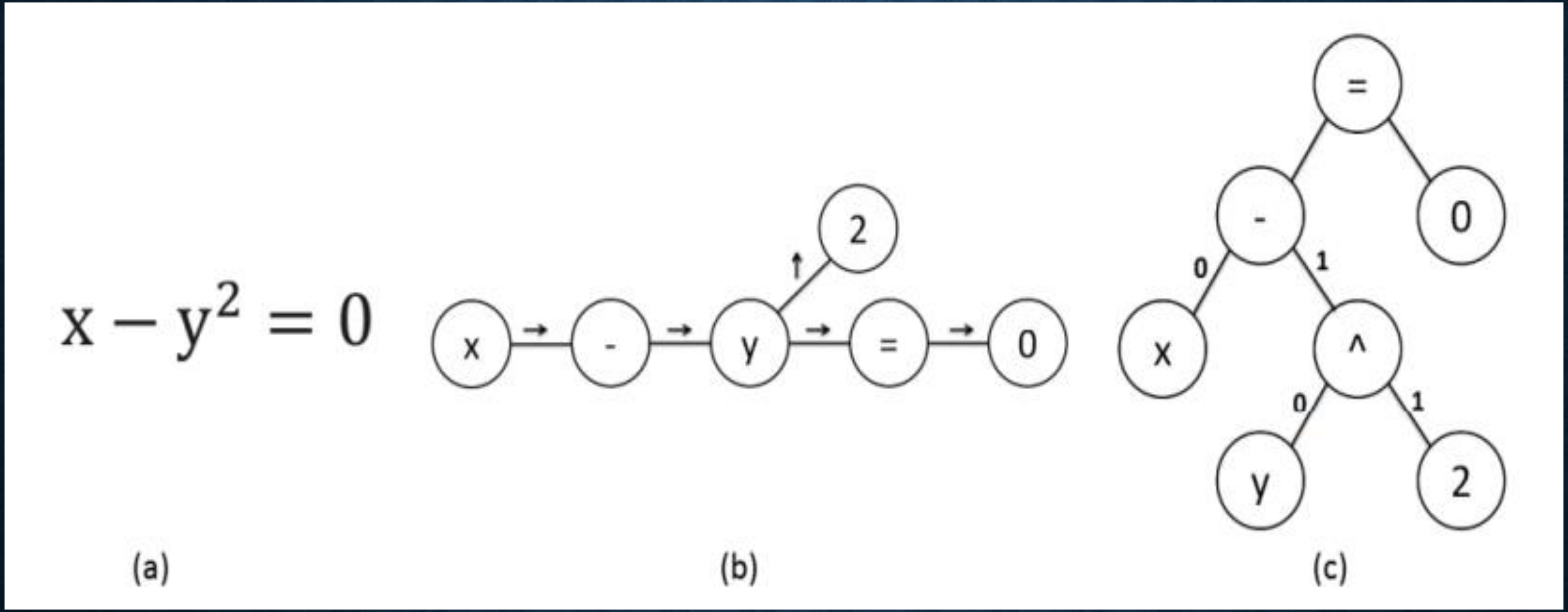
is easy to prove.

```
<div style="width: 15em;">
  This mathematical formula with a big summation and the number pi
  <math display="block" style="border: 1px dotted black;">
    <mrow>
      <munderover>
        <mo>∑</mo>
        <mrow><mi>n</mi><mo>=</mo><mn>1</mn></mrow>
        <mrow><mo>+</mo><mn>∞</mn></mrow>
      </munderover>
      <mfrac>
        <mn>1</mn>
        <msup><mi>n</mi><mn>2</mn></msup>
      </mfrac>
    </mrow>
    <mo>=</mo>
    <mfrac>
      <msup><mi>π</mi><mn>2</mn></msup>
      <mn>6</mn>
    </mfrac>
  </math>
  is easy to prove.
</div>
```

MATHML

Reference: https://www.w3.org/TR/mathml-core/

TOKENIZING MATHEMATICAL FORMULAE



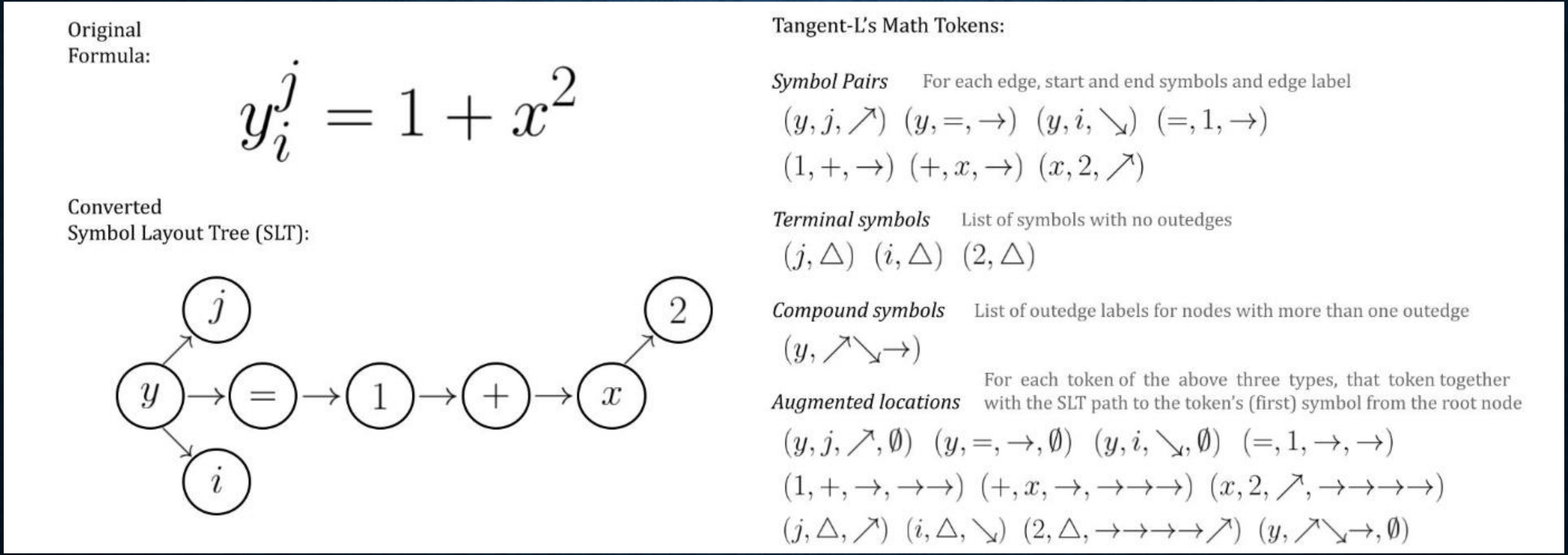
Formula

Symbol Layout Tree

Operator Tree

Reference: https://clgiles.ist.psu.edu/pubs/IDTIR2019.pdf

TOKENIZING MATHEMATICAL FORMULAE – TANGENT-L AND SLT



Original Formula:

$$y_i^j = 1 + x^2$$

Converted Symbol Layout Tree (SLT):

Tangent-L's Math Tokens:

Symbol Pairs For each edge, start and end symbols and edge label

$(y, j, \nearrow)$ $(y, =, \rightarrow)$ $(y, i, \searrow)$ $(=, 1, \rightarrow)$

$(1, +, \rightarrow)$ $(+, x, \rightarrow)$ $(x, 2, \nearrow)$

Terminal symbols List of symbols with no outedges

$(j, \triangle)$ $(i, \triangle)$ $(2, \triangle)$

Compound symbols List of outedge labels for nodes with more than one outedge

$(y, \nearrow\searrow\rightarrow)$

Augmented locations For each token of the above three types, that token together with the SLT path to the token's (first) symbol from the root node

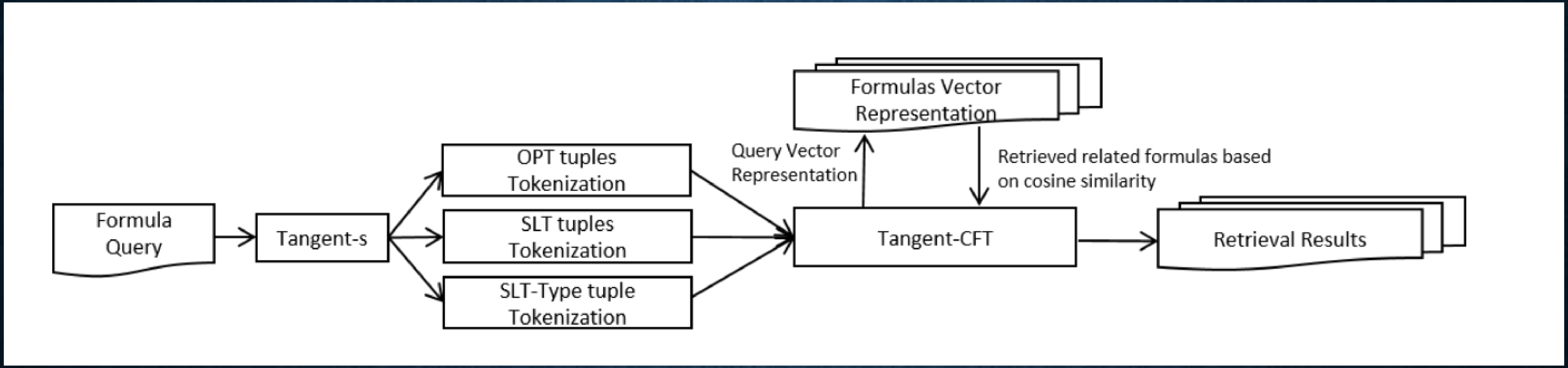
$(y, j, \nearrow, \emptyset)$ $(y, =, \rightarrow, \emptyset)$ $(y, i, \searrow, \emptyset)$ $(=, 1, \rightarrow, \rightarrow)$

$(1, +, \rightarrow, \rightarrow\rightarrow)$ $(+, x, \rightarrow, \rightarrow\rightarrow\rightarrow)$ $(x, 2, \nearrow, \rightarrow\rightarrow\rightarrow\rightarrow)$

$(j, \triangle, \nearrow)$ $(i, \triangle, \searrow)$ $(2, \triangle, \rightarrow\rightarrow\rightarrow\rightarrow\nearrow)$ $(y, \nearrow\searrow\rightarrow, \emptyset)$

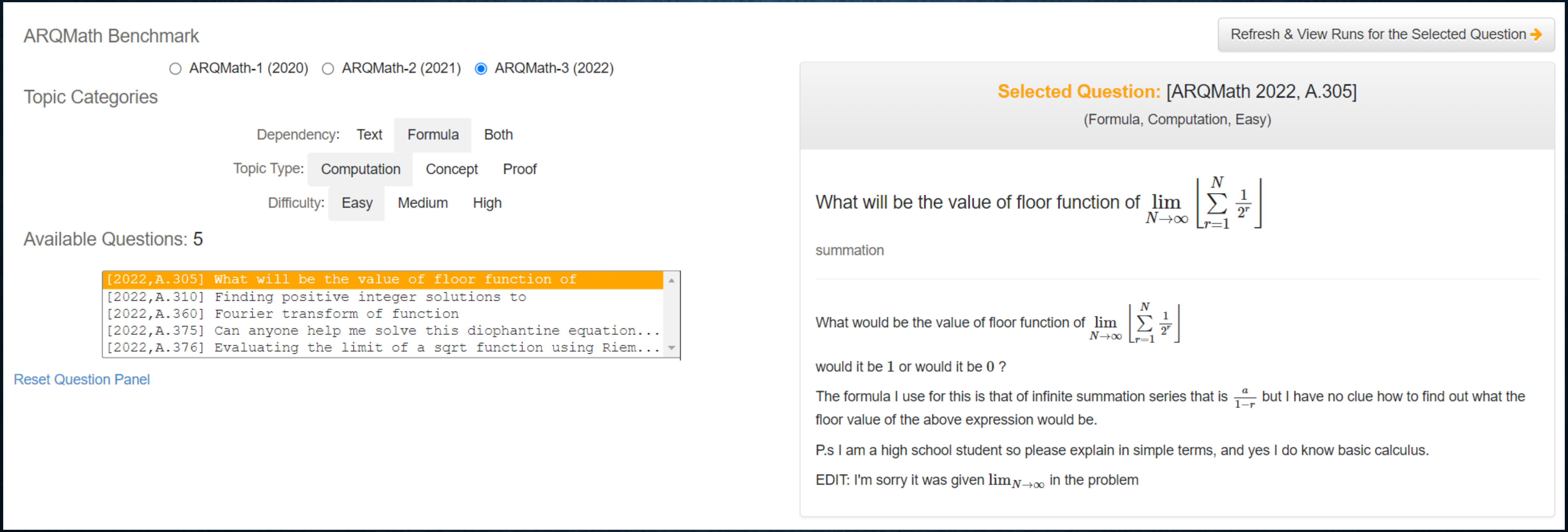
A detailed animation provided here: https://www.scg.uwaterloo.ca/brushsearch/tangent-l/

BASELINE MATH IR SYSTEM – TANGENT-CFT


Formula Query
Tangent-s
OPT tuples Tokenization
SLT tuples Tokenization
SLT-Type tuple Tokenization
Query Vector Representation
Formulas Vector Representation
Retrieved related formulas based on cosine similarity
Tangent-CFT
Retrieval Results

Reference: https://clgiles.ist.psu.edu/pubs/IDTIR2019.pdf

INSPIRATION: MATHDOWSERS



ARQMath Benchmark

○ ARQMath-1 (2020) ○ ARQMath-2 (2021) ● ARQMath-3 (2022)

Topic Categories

Dependency: Text **Formula** Both

Topic Type: **Computation** Concept Proof

Difficulty: **Easy** Medium High

Available Questions: 5

```
[2022,A.305] What will be the value of floor function of
[2022,A.310] Finding positive integer solutions to
[2022,A.360] Fourier transform of function
[2022,A.375] Can anyone help me solve this diophantine equation...
[2022,A.376] Evaluating the limit of a sqrt function using Riem...
```

Reset Question Panel

Selected Question: [ARQMath 2022, A.305]

(Formula, Computation, Easy)

What will be the value of floor function of $\lim_{N\to\infty}\left\lfloor\sum_{r=1}^{N}\frac{1}{2^r}\right\rfloor$

summation

What would be the value of floor function of $\lim_{N\to\infty}\left\lfloor\sum_{r=1}^{N}\frac{1}{2^r}\right\rfloor$

would it be 1 or would it be 0 ?

The formula I use for this is that of infinite summation series that is $\frac{a}{1-r}$ but I have no clue how to find out what the floor value of the above expression would be.

P.s I am a high school student so please explain in simple terms, and yes I do know basic calculus.

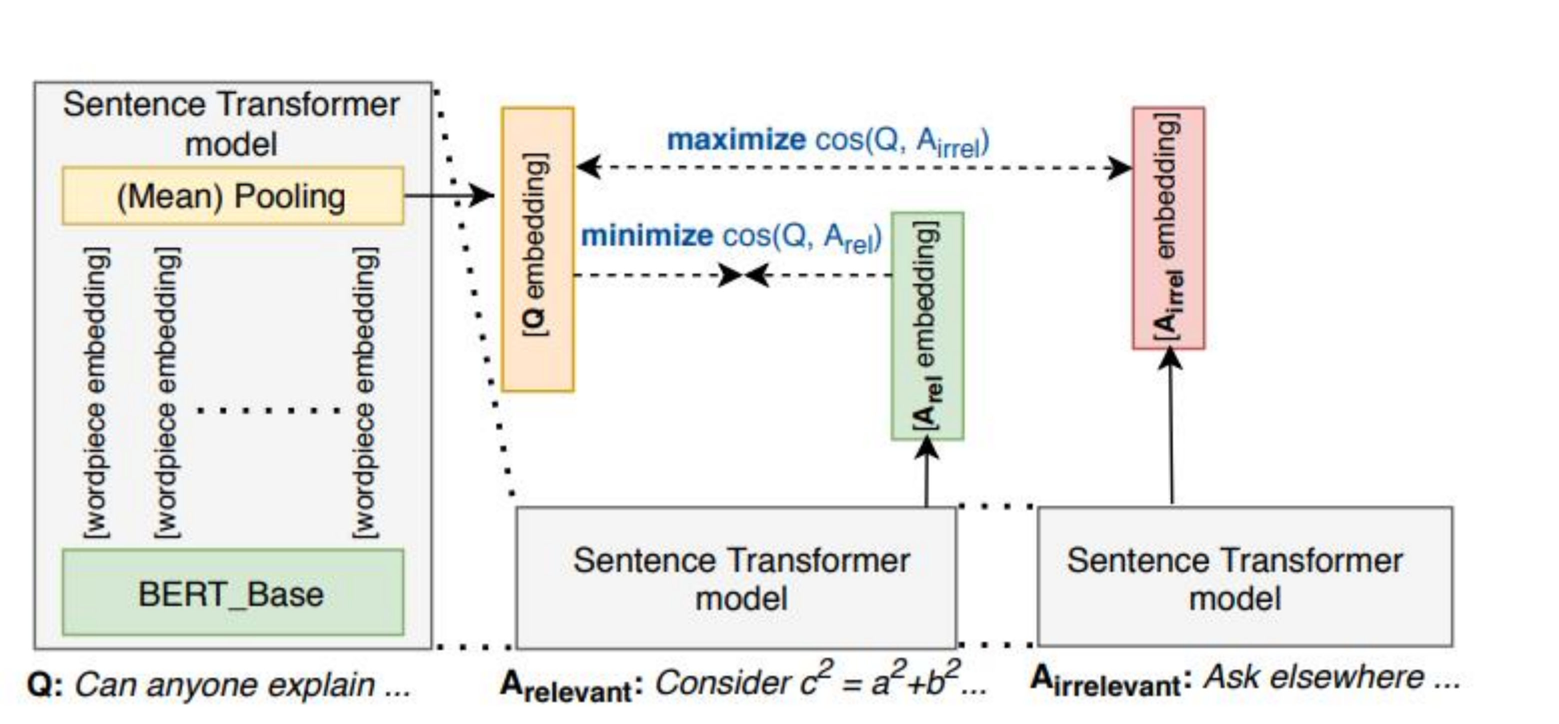
EDIT: I'm sorry it was given $\lim_{N\to\infty}$ in the problem

Reference: https://cs.uwaterloo.ca/~yk2ng/MathDowsers-ARQMath/ & https://ceur-ws.org/Vol-3180/paper-03.pdf

MY MATH IR SYSTEM

- GitHub: https://github.com/davidna22/math-IR-ARQMath-CompuBERT
- Overview:
  - My math IR system was created following the ARQMath workshop. My primary goal was to create an IR system capable of vectorizing q+a pairs from the ARQMath dataset, in order to be able to take in a math query and return a ranked list of documents matching that query. I chose to develop my system following Task 1 of the ARQMath workshop.
- Dataset:
  - Math Stack Exchange posts from 2010-2021
  - Data preprocessing was provided by ARQMath and found on their github
  - I implemented their preprocessing, added some others that I found online in papers
- Architecture:
  - I use a variation of a model called CompuBERT (Novotny et al.) that I found very interesting. It utilizes pre-trained Transformer models such as BERT to create feature embeddings for question answer pairs. Using those feature embeddings, I can compute cosine similarity scores and return a ranked list of documents. Most of my code was taken from their open-source implementation (https://github.com/MIR-MU/CompuBERT) as it was one of the first transformer-encoder models and easiest to implement.

COMPUBERT



$$\text{minimize} \quad \sum_{i=1}^{|Qs|} \sum_{j=1}^{|A_i|} \left| \left(1 - \cos(q_i, a_{ij})\right) - \text{dist}_{\text{exp}}(q_i, a_{ij}) \right|$$

Reference: https://ceur-ws.org/Vol-2696/paper_235.pdf

The Model relies on the feature vector embeddings generated by the BERT model and uses cosine similarity to find the best answers per query

The overall results of the this first transformer model was poor relative to other submissions, so part of my experiment was to test how it will perform with a newer pre-trained model

Since CompuBERT, other Transformer models have been implemented to success. This (given my lack of experience) was the easiest to implement and thus, a good starting point for me.

TRANSFORMER-ENCODER AND DECODER MDOELS

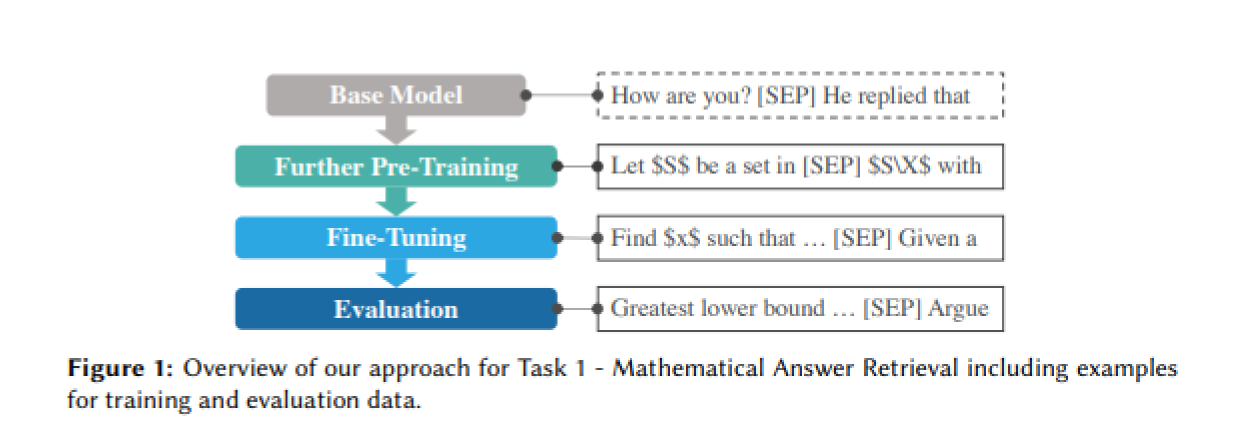
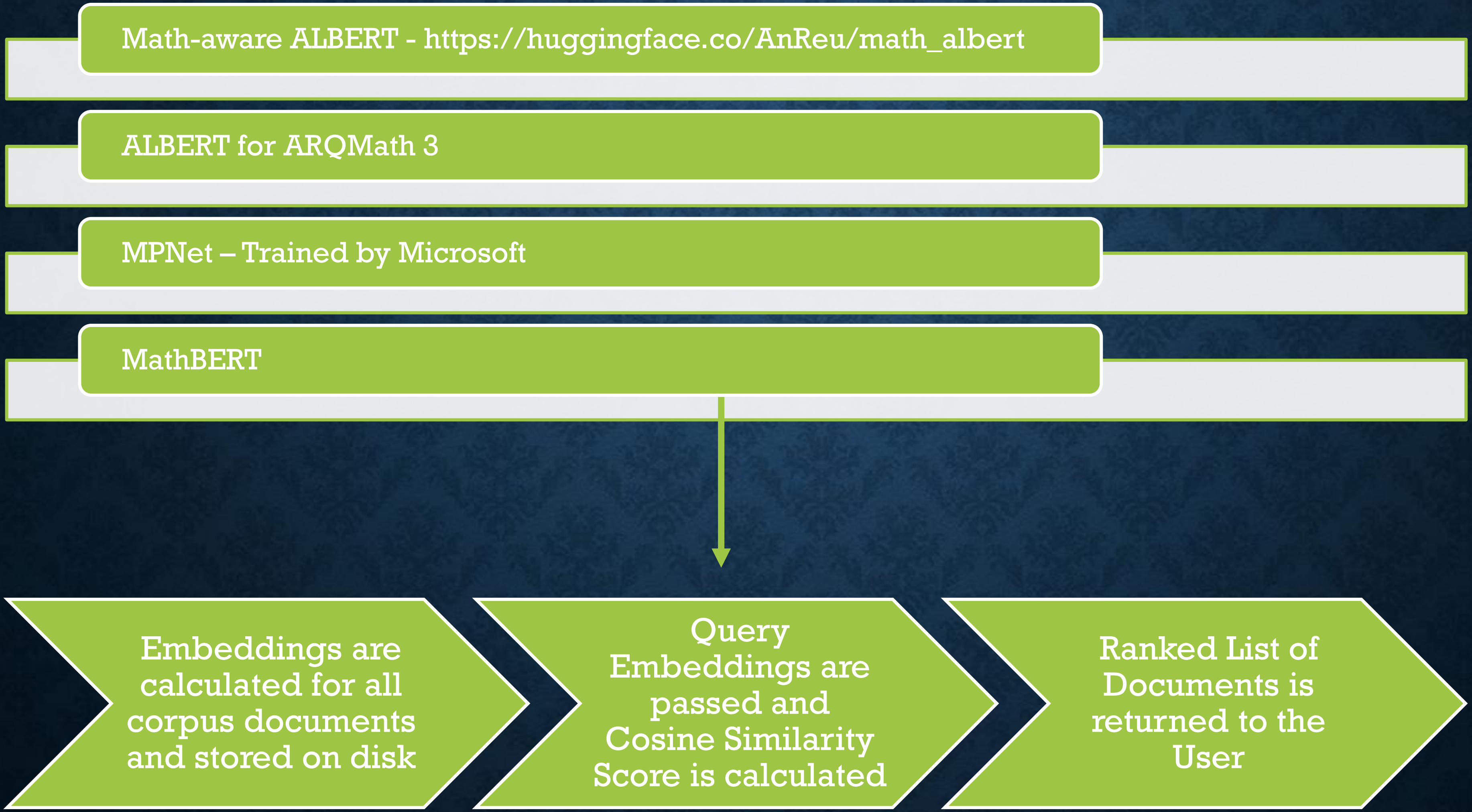

Base Model
How are you? [SEP] He replied that
Further Pre-Training
Let S be a set in [SEP] $S\X$ with
Fine-Tuning
Find x such that … [SEP] Given a
Evaluation
Greatest lower bound … [SEP] Argue

Figure 1: Overview of our approach for Task 1 - Mathematical Answer Retrieval including examples for training and evaluation data.

Reference: https://ceur-ws.org/Vol-3180/paper-07.pdf

MY MODEL



Math-aware ALBERT - https://huggingface.co/AnReu/math_albert

ALBERT for ARQMath 3

MPNet – Trained by Microsoft

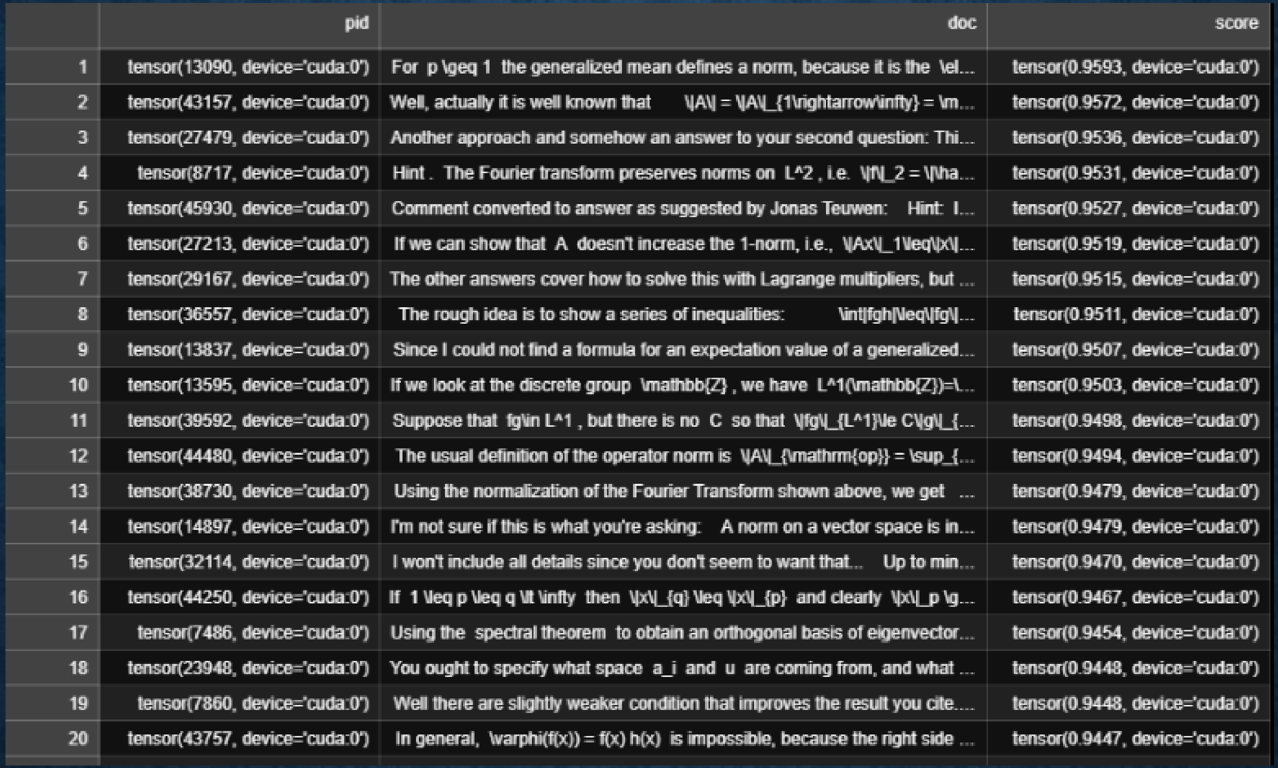
MathBERT

Embeddings are calculated for all corpus documents and stored on disk

Query Embeddings are passed and Cosine Similarity Score is calculated

Ranked List of Documents is returned to the User

RESULTS



<table>
<tr><th></th><th>pid</th><th>doc</th><th>score</th></tr>
<tr><td>1</td><td>tensor(13090, device='cuda:0')</td><td>For p \geq 1 the generalized mean defines a norm, because it is the \el...</td><td>tensor(0.9593, device='cuda:0')</td></tr>
<tr><td>2</td><td>tensor(43157, device='cuda:0')</td><td>Well, actually it is well known that   \|A\| = \|A\|_{1\rightarrow\infty} = \m...</td><td>tensor(0.9572, device='cuda:0')</td></tr>
<tr><td>3</td><td>tensor(27479, device='cuda:0')</td><td>Another approach and somehow an answer to your second question: Thi...</td><td>tensor(0.9536, device='cuda:0')</td></tr>
<tr><td>4</td><td>tensor(8717, device='cuda:0')</td><td>Hint . The Fourier transform preserves norms on L^2 , i.e. \|f\|_2 = \|\ha...</td><td>tensor(0.9531, device='cuda:0')</td></tr>
<tr><td>5</td><td>tensor(45930, device='cuda:0')</td><td>Comment converted to answer as suggested by Jonas Teuwen: Hint: I...</td><td>tensor(0.9527, device='cuda:0')</td></tr>
<tr><td>6</td><td>tensor(27213, device='cuda:0')</td><td>If we can show that A doesn't increase the 1-norm, i.e., \|Ax\|_1\leq\|x\|...</td><td>tensor(0.9519, device='cuda:0')</td></tr>
<tr><td>7</td><td>tensor(29167, device='cuda:0')</td><td>The other answers cover how to solve this with Lagrange multipliers, but ...</td><td>tensor(0.9515, device='cuda:0')</td></tr>
<tr><td>8</td><td>tensor(36557, device='cuda:0')</td><td>The rough idea is to show a series of inequalities: \int|fgh|\leq\|fg\|...</td><td>tensor(0.9511, device='cuda:0')</td></tr>
<tr><td>9</td><td>tensor(13837, device='cuda:0')</td><td>Since I could not find a formula for an expectation value of a generalized...</td><td>tensor(0.9507, device='cuda:0')</td></tr>
<tr><td>10</td><td>tensor(13595, device='cuda:0')</td><td>If we look at the discrete group \mathbb{Z} , we have L^1(\mathbb{Z})=\...</td><td>tensor(0.9503, device='cuda:0')</td></tr>
<tr><td>11</td><td>tensor(39592, device='cuda:0')</td><td>Suppose that fg\in L^1 , but there is no C so that \|fg\|_{L^1}\le C\|g\|_{...</td><td>tensor(0.9498, device='cuda:0')</td></tr>
<tr><td>12</td><td>tensor(44480, device='cuda:0')</td><td>The usual definition of the operator norm is \|A\|_{\mathrm{op}} = \sup_{...</td><td>tensor(0.9494, device='cuda:0')</td></tr>
<tr><td>13</td><td>tensor(38730, device='cuda:0')</td><td>Using the normalization of the Fourier Transform shown above, we get   ...</td><td>tensor(0.9479, device='cuda:0')</td></tr>
<tr><td>14</td><td>tensor(14897, device='cuda:0')</td><td>I'm not sure if this is what you're asking: A norm on a vector space is in...</td><td>tensor(0.9479, device='cuda:0')</td></tr>
<tr><td>15</td><td>tensor(32114, device='cuda:0')</td><td>I won't include all details since you don't seem to want that...   Up to min...</td><td>tensor(0.9470, device='cuda:0')</td></tr>
<tr><td>16</td><td>tensor(44250, device='cuda:0')</td><td>If 1 \leq p \leq q \lt \infty then \|x\|_{q} \leq \|x\|_{p} and clearly \|x\|_p \g...</td><td>tensor(0.9467, device='cuda:0')</td></tr>
<tr><td>17</td><td>tensor(7486, device='cuda:0')</td><td>Using the spectral theorem to obtain an orthogonal basis of eigenvector...</td><td>tensor(0.9454, device='cuda:0')</td></tr>
<tr><td>18</td><td>tensor(23948, device='cuda:0')</td><td>You ought to specify what space a_i and u are coming from, and what ...</td><td>tensor(0.9448, device='cuda:0')</td></tr>
<tr><td>19</td><td>tensor(7860, device='cuda:0')</td><td>Well there are slightly weaker condition that improves the result you cite....</td><td>tensor(0.9448, device='cuda:0')</td></tr>
<tr><td>20</td><td>tensor(43757, device='cuda:0')</td><td>In general, \varphi(f(x)) = f(x) h(x) is impossible, because the right side ...</td><td>tensor(0.9447, device='cuda:0')</td></tr>
</table>

	MPNet	Math-Aware Albert	Albert for Math 3	MathBERT
f1 Scores	0.699	0.626	0.691	0.644
Precision	0.544	0.532	0.54	0.527
Accuracy	0.546	0.51	0.538	0.506

TO-DO LIST

- Trying different Transformer Models
  - There have been new iterations of CompuBERT that have had high performance. An example is found in this paper, Transformer-Encoder and Decoder Models for Questions on Math (https://ceur-ws.org/Vol-3180/paper-07.pdf).
- Training on a larger subsample of the dataset
  - Training on the entire dataset would have taken me about a full month of constant training given the size of the corpus and the size of the models I chose.
  - Bunch of limitations such as GPU RAM and compute power that I would need to pay for via cloud to train effectively
- Applying different transformations in text preprocessing
  - Math text preprocessing is very different from standard IR, so I would love to do more research and apply different preprocessing techniques
- Attempting Task 2 and 3 of the ARQMath competition
  - Task 2 is very different as I would need to fully implement a math-only textual embedding. Most likely requiring the use of symbol trees (Tangent-CFT) and something called Formula2Vec.

REFERENCES

- NTCIR-12 MathIR Task Overview
  - https://www.cs.rit.edu/~rlaz/files/ntcir12-mathir.pdf
- NTCIR Website
  - https://ntcir-math.nii.ac.jp/introduction/
- ARQMath: A New Benchmark for Math-Aware CQA and Math Formula Retrieval
  - http://sigir.org/wp-content/uploads/2020/12/p06.pdf
- ARQMath Website
  - https://www.cs.rit.edu/~dprl/ARQMath/2021/
- Tangent-CFT: An Embedding Model for Mathematical Formulas
  - https://clgiles.ist.psu.edu/pubs/IDTIR2019.pdf
- Math ML Docs
  - https://www.w3.org/TR/mathml-core/
- Latex Docs
  - http://www.malinc.se/math/latex/basiccodeen.php
- CompuBERT
  - https://github.com/MIR-MU/CompuBERT
- Three is Better than One: Ensembling Math Information Retrieval Systems
  - https://ceur-ws.org/Vol-2696/paper_235.pdf
- Transformer-Encoder and Decoder Models for Questions on Math
  - https://ceur-ws.org/Vol-3180/paper-07.pdf