Searching Documents with Text and Mathematical Content Using a Pen-Based Interface

George Labahn and Frank Tompa
University of Waterloo

Michael Feng
Huawei Labs

WHJIL Workshop 2021

Outline

Project: Team and Objectives

Search Engine

Front End

MathBrush-Search System Demo

Future Work

MathBrush-Search Team

University of Waterloo

- **PIs**: George Labahn, Frank Tompa
- **Research Associate**: Mirette Marzouk
- **PhD Students**: Avery Hiebert, Besat Kassaie
- **Masters Students**: Kiki Ng, Kevin Wang
- **Undergraduate Research**: Yining Wang

Huawei

- **Huawei Collaborator**: Michael Feng

Objective

Build **MathBrush-Search**, a math-aware search system

- A math-aware search engine that uses text and mathematical content, combining their semantics, and considering users' provided constraints to get the most relevant search results
- An intuitive front end (recognizer and user interface) that accepts handwritten mathematical formulas and supports use of natural gestures to specify constraints and wildcards

Tangent-L Search Engine

We utilize the Tangent-L search engine

- Based on Lucene framework
- Indexes both text and formulas' syntactic features
- Uses "bag-of-words" semantics (i.e., word order is ignored)
- Performs comparably to state-of-the-art math retrieval systems but is still has room for improvement

Integrated with front-end to show basic search functionality.

New features added in last 12 months...

Search Engine - Wild Cards

- Mathematicians choose variable names (almost) arbitrarily.

  But which symbols in query

  $$\int_{-n}^{n} e^{-x^2} dx$$

  are arbitrary?

Search Engine - Wild Cards

- Mathematicians choose variable names (almost) arbitrarily.

  But which symbols in query

  $$\int_{-n}^{n} e^{-x^2} dx$$

  are arbitrary?

- Wild card can be completely arbitrary or of particular types (variables, numbers, fractions, etc.).

  - e.g., n is a number, x is a variable, e and d are *not* wild.

- New feature added to capture and match repetition patterns

  - e.g., x, n above

Search Engine - Wild Cards

- When indexing expressions,
  - feature with a variable is indexed as a “variable wild card” ($?V$)
  - any number is also indexed as a “number wild card” ($?N$),
  - etc.
- Searches with “variable wild card” match the feature stored with $?V$, etc.
- Searches for “expression (i.e., arbitrary) wild cards” match any type of wild card

Search Engine - Proximity Matching

- Proximity is a strong signal of relevance for a query
  - Keywords contained within a single paragraph
  - Math terms (features) contained within a single formula
  - Keywords and formula appear close together
- Question: which measure of proximity is best?
  - Min, average, or max distance between search terms
  - Minimum span including *at least one of each* term *vs.* smallest span including *all occurrences* of search terms
  - Normalized by document length?
- Rerank documents returned by Tangent-L *vs*. use new ranking within Tangent-L that understands word order?

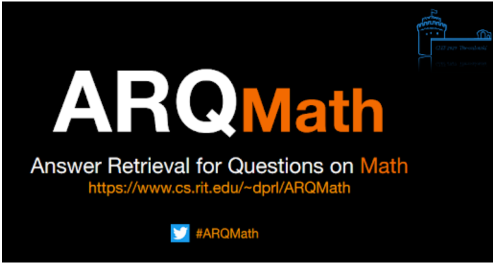
Search Engine - Holistic Formula Matching

- Query formula's features might be matched across multiple formulas in a document (because document parts unordered)

- **Alternative**: try to match whole formulas
  - At index time:
    - Create formula corpus of all visually distinct formulas in database, each with unique formula "key"
    - Index document database using formula keys in place of formulas
  - At query time:
    - Rank all individual formulas based on features
    - Search database using formula keys of top-k ranked formulas
    - Weight matching formula keys by how well query formula matches

2020 ARQMath Lab

- ARQMath - **A**nswer **R**etrieval for **Q**uestions On **Math**
- Held at CLEF (Conference and Labs of the Evaluation Forum)
- We participated with Tangent-L



ARQMath

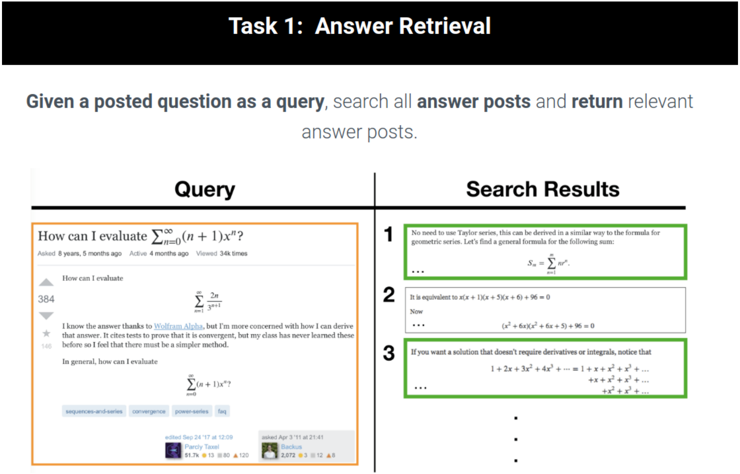
Answer Retrieval for Questions on Math

https://www.cs.rit.edu/~dprl/ARQMath

#ARQMath

2020 ARQMath Lab

- Dataset: Math Stack Exchange posts from 2010 to 2018



Task 1: Answer Retrieval

Given a posted question as a query, search all **answer posts** and **return** relevant answer posts.

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

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

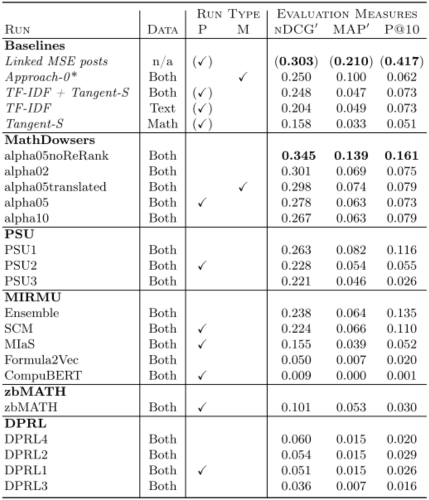
...

.
.
.

Above is an example query (question post) at left, with search results shown as *excerpts* from question answers at right (relevant answers are indicated in green).

2020 ARQMath Lab

- We (MathDowsers) achieved the highest nDCG' and MAP' (these are the primary measure of effectiveness)



		Run Type		Evaluation Measures		
Run	Data	P	M	nDCG′	MAP′	P@10
Baselines						
Linked MSE posts	n/a	(✓)		(**0.303**)	(**0.210**)	(**0.417**)
*Approach-0**	Both		✓	0.250	0.100	0.062
TF-IDF + Tangent-S	Both	(✓)		0.248	0.047	0.073
TF-IDF	Text	(✓)		0.204	0.049	0.073
Tangent-S	Math	(✓)		0.158	0.033	0.051
MathDowsers						
alpha05noReRank	Both			**0.345**	**0.139**	**0.161**
alpha02	Both			0.301	0.069	0.075
alpha05translated	Both		✓	0.298	0.074	0.079
alpha05	Both	✓		0.278	0.063	0.073
alpha10	Both			0.267	0.063	0.079
PSU						
PSU1	Both			0.263	0.082	0.116
PSU2	Both	✓		0.228	0.054	0.055
PSU3	Both			0.221	0.046	0.026
MIRMU						
Ensemble	Both			0.238	0.064	0.135
SCM	Both	✓		0.224	0.066	0.110
MIaS	Both	✓		0.155	0.039	0.052
Formula2Vec	Both			0.050	0.007	0.020
CompuBERT	Both	✓		0.009	0.000	0.001
zbMATH						
zbMATH	Both	✓		0.101	0.053	0.030
DPRL						
DPRL4	Both			0.060	0.015	0.020
DPRL2	Both			0.054	0.015	0.029
DPRL1	Both	✓		0.051	0.015	0.026
DPRL3	Both			0.036	0.007	0.016

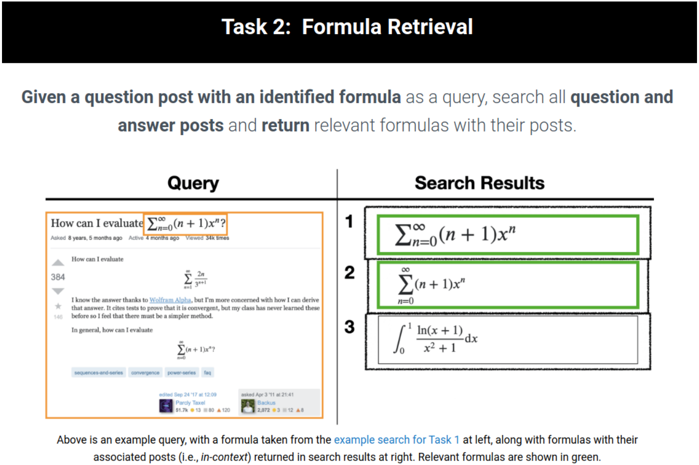
2020 ARQMath Lab

Summary of Findings:

- With proper adaptation (e.g. query extraction to turn math questions into formal queries, informative indexing unit), Tangent-L gives good results when retrieving answers to math questions.
- Compared to other participants' system, our system out-performs for formula-dependent math questions.
- The ARQMath 2020 Evaluation data serves as a benchmark to help us better tune configuration of Tangent-L (such as the relative weight to apply to keyword features vs. math features during query time).

2021 ARQMath Lab

- Same dataset, new set of math questions
- In addition to finding answer to math questions, we also participated in the Formula Retrieval task



Task 2: Formula Retrieval

Given a question post with an identified formula as a query, search all **question and answer posts** and **return** relevant formulas with their posts.

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

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

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

Above is an example query, with a formula taken from the example search for Task 1 at left, along with formulas with their associated posts (i.e., *in-context*) returned in search results at right. Relevant formulas are shown in green.

2021 ARQMath Lab

Objectives:

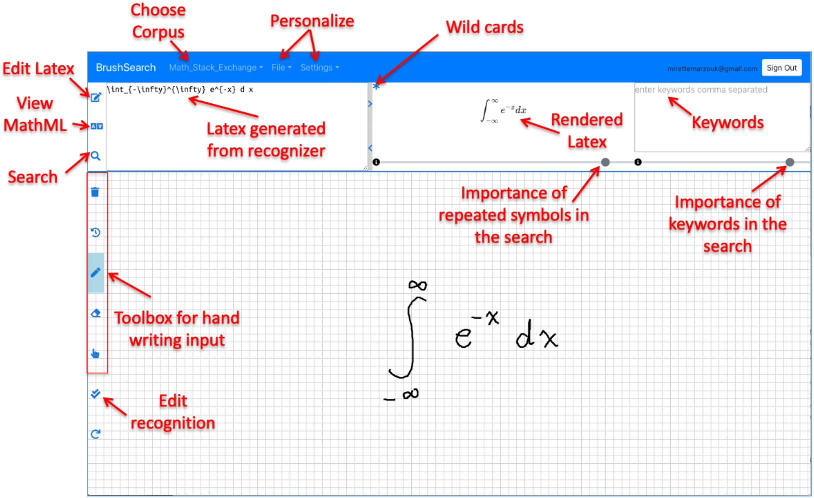
- Participate with the improved Tangent-L
- Investigate proximity matching and holistic formula matching

Preliminary results with last year's queries shows that our new system has:

- More than 10-point gain in the Answer Retrieval task (Task 1)
- Comparable performance to last year's best participant run in the Formula Retrieval task (Task 2)

(Results for 2021 queries not yet available)

MathBrush-Search Front End



```
\int_{-\infty}^{\infty} e^{-x} d x
```

$\int_{-\infty}^{\infty} e^{-x} dx$

Choose Corpus

Personalize

Wild cards

Edit Latex

View MathML

Search

Latex generated from recognizer

Rendered Latex

Keywords

Importance of repeated symbols in the search

Importance of keywords in the search

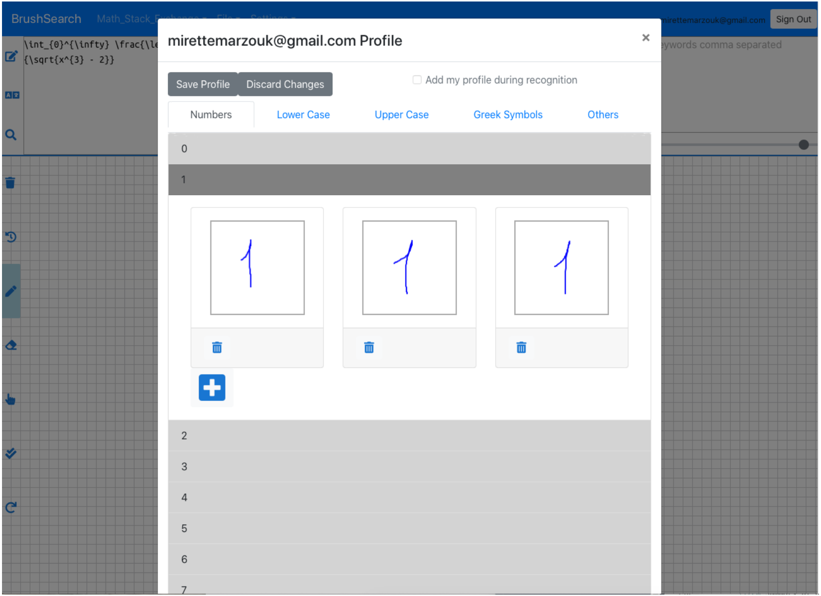
Toolbox for hand writing input

Edit recognition

Front End - Math Recognizer

- Want data driven approach for math recognition
  - Previous recognizer uses grammar based approach
- Past work:
  - Attempted Transformer architecture for HMER
  - Synthetic handwritten expression data
- Ongoing work
  - Handling per-user recognizer customization
  - Transfer learning taking advantage of non-handwritten expression data
  - Interpretability (incl. interpretable vector representations for formulas)

Front End - Training Recognizer

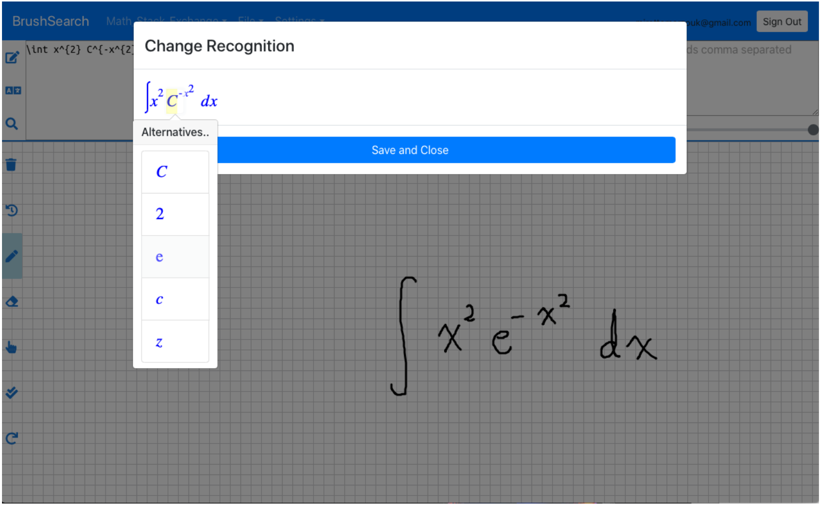


```
\int_{0}^{\infty} \frac{\le
{\sqrt{x^{3} - 2}}
```

mirettemarzouk@gmail.com Profile

□ Add my profile during recognition

Front End - Recognition Correction



```
\int x^{2} C^{-x^{2}
```

$\int x^2 C^{-x^2}\ dx$

Generating Synthetic Data

Motivation:

- New trends in mathematical recognition systems use deep learning and neural networks
- A very large number of diverse handwritten expressions are needed for training and testing

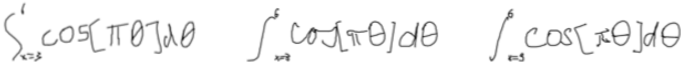
Generating Synthetic Data

Approach:

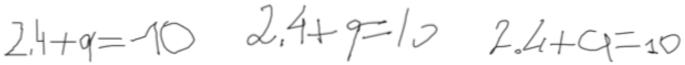
- Convert typeset expression into a Symbol Layout Tree (SLT), capturing how formula pieces are laid out when printed
- Traverse SLT and construct layout based on edge types and symbols spatial information
- Query a Unicode font for spatial symbol information
- Sample normalized handwritten symbols from a data set and insert into the layout
- Apply local and global distortion models to guarantee the variability of output expressions

Generating Synthetic Data

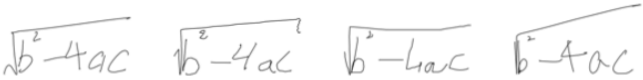
```
\int_{x = 3}^{6} \cos{\left[\pi \theta \right]} d \theta
```



```
2.4 + q = 10
```



```
\sqrt{b^2 - 4ac}
```

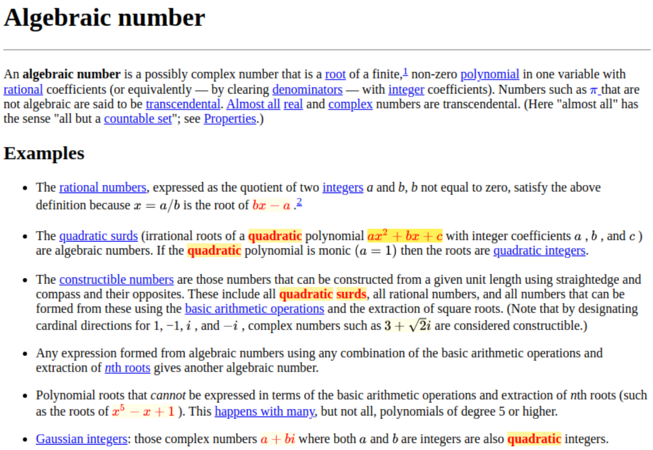


Front End - Math Highlighting

- Highlighting search results would be helpful for users to locate their desired information
- Support both keyword highlighting and formula highlighting

Front End - Math Highlighting

- Highlighting search results would be helpful for users to locate their desired information
- Support both keyword highlighting and formula highlighting



Algebraic number

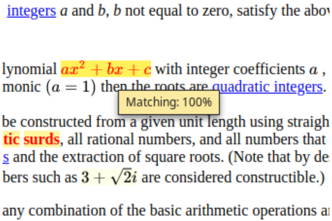
An **algebraic number** is a possibly complex number that is a root of a finite,[1] non-zero polynomial in one variable with rational coefficients (or equivalently — by clearing denominators — with integer coefficients). Numbers such as π that are not algebraic are said to be transcendental. Almost all real and complex numbers are transcendental. (Here "almost all" has the sense "all but a countable set"; see Properties.)

Examples

- The rational numbers, expressed as the quotient of two integers *a* and *b*, *b* not equal to zero, satisfy the above definition because $x = a/b$ is the root of $bx - a$.[2]
- The quadratic surds (irrational roots of a **quadratic** polynomial $ax^2 + bx + c$ with integer coefficients a , b , and c) are algebraic numbers. If the **quadratic** polynomial is monic $(a = 1)$ then the roots are quadratic integers.
- The constructible numbers are those numbers that can be constructed from a given unit length using straightedge and compass and their opposites. These include all **quadratic surds**, all rational numbers, and all numbers that can be formed from these using the basic arithmetic operations and the extraction of square roots. (Note that by designating cardinal directions for 1, −1, i , and $-i$, complex numbers such as $3 + \sqrt{2}i$ are considered constructible.)
- Any expression formed from algebraic numbers using any combination of the basic arithmetic operations and extraction of *n*th roots gives another algebraic number.
- Polynomial roots that *cannot* be expressed in terms of the basic arithmetic operations and extraction of *n*th roots (such as the roots of $x^5 - x + 1$). This happens with many, but not all, polynomials of degree 5 or higher.
- Gaussian integers: those complex numbers $a + bi$ where both a and b are integers are also **quadratic** integers.

Front End - Math Highlighting

- Shades of the highlight color reflect how well a document formula matches the query formulas
- The matching percentage shows how much the query formula has been matched for this document formula in terms of the number of matching symbols



integers a and b, b not equal to zero, satisfy the abov

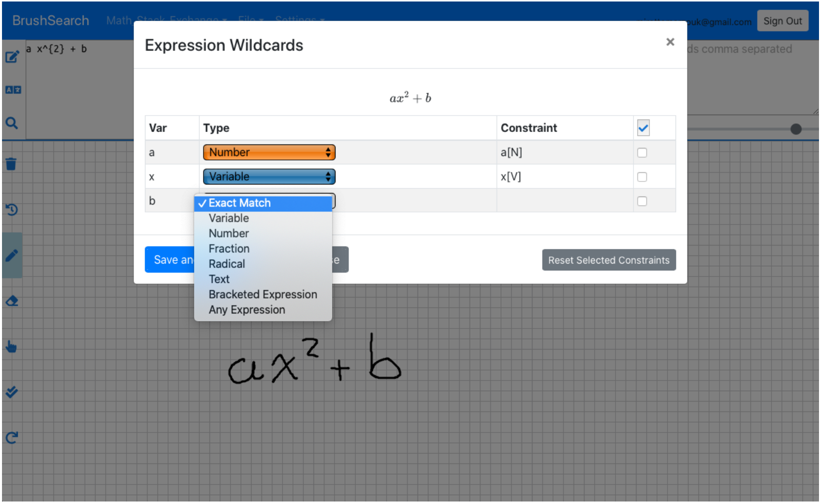
lynomial $ax^2 + bx + c$ with integer coefficients a ,
monic $(a = 1)$ then the roots are quadratic integers.

Matching: 100%

be constructed from a given unit length using straigh
tic surds, all rational numbers, and all numbers that
s and the extraction of square roots. (Note that by de
bers such as $3 + \sqrt{2}i$ are considered constructible.)

any combination of the basic arithmetic operations a

Front End - Wildcards



a x^{2} + b

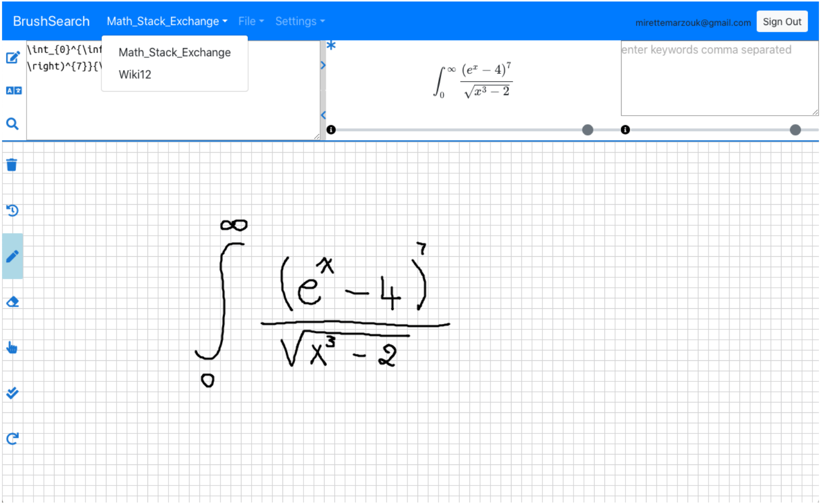
Expression Wildcards

$$ax^2 + b$$

Var	Type	Constraint	☑
a	Number	a[N]	□
x	Variable	x[V]	□
b	✓ Exact Match		□

$ax^2 + b$

Front End - Multiple Corpora

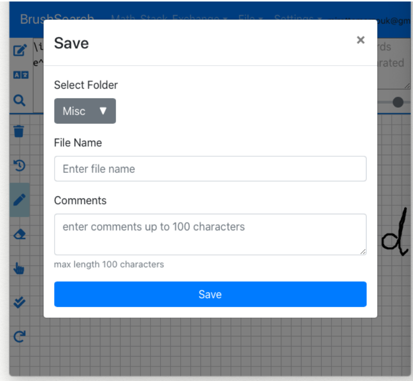


```
\int_{0}^{\inf
\right)^{7}}{\
```

$$\int_0^\infty \frac{(e^x-4)^7}{\sqrt{x^3-2}}$$

$$\int_0^\infty \frac{(e^x-4)^7}{\sqrt{x^3-2}}$$

Front End - Save/Load



Save

Select Folder

Misc ▼

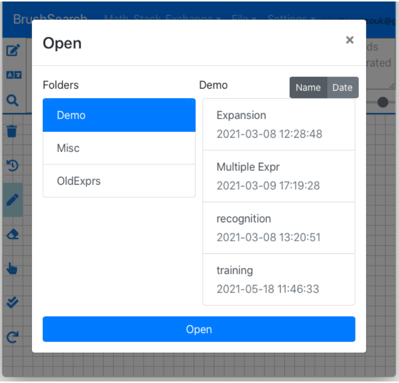
File Name

Enter file name

Comments

enter comments up to 100 characters

max length 100 characters



Open

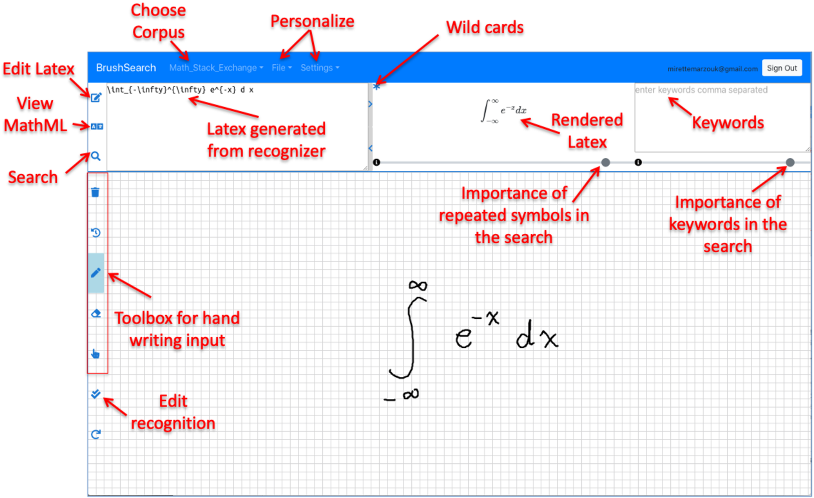
Folders

- Demo
- Misc
- OldExprs

Demo

Expansion	2021-03-08 12:28:48
Multiple Expr	2021-03-09 17:19:28
recognition	2021-03-08 13:20:51
training	2021-05-18 11:46:33

MathBrush-Search System Demo



```
\int_{-\infty}^{\infty} e^{-x} d x
```

$\int_{-\infty}^{\infty} e^{-x}dx$

$\int_{-\infty}^{\infty} e^{-x} dx$

Choose Corpus

Personalize

Wild cards

Edit Latex

View MathML

Search

Latex generated from recognizer

Rendered Latex

Keywords

Importance of repeated symbols in the search

Importance of keywords in the search

Toolbox for hand writing input

Edit recognition

Future Work

- Continue to evaluate the effectiveness of search query using pen-based input.
- Build and test a recognizer using machine learning techniques.
- Improve techniques to highlight matches in searched text, including partial matches within formulas.
- Incorporate proximity matching into the search engine and provide pen-based mechanisms in the front end to help guide users in specifying semantic aspects.
- Implement a web scraper to build multiple corpora for searching to extend system usability and testing.
- Utilize users' feedback to improve both the recognition and the ranking of matches to queries.