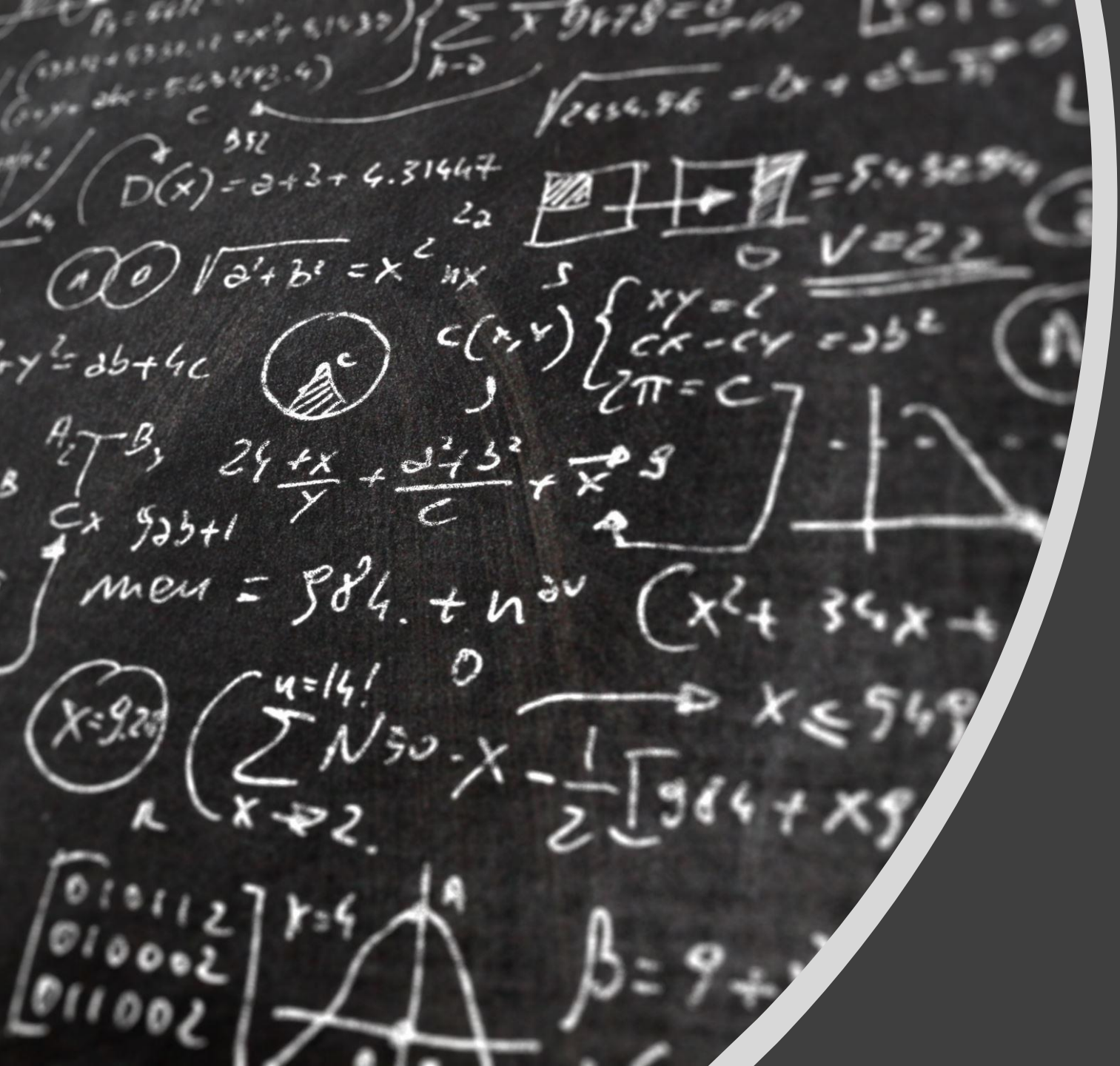


Mathematical Entity Linking (MathEL)

Methods and Applications

Problem:
Formula Concept Retrieval Challenge

$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \left(\frac{m_0 c}{\hbar}\right)^2 \psi = 0$	$u_{tt} + Au + f(u) = 0$
$\partial_{ct}^2 h_n(z,t) - \partial_z^2 h_n(z,t) + \nu_n^2 h_n(z,t) = 0$	$\nabla^a \nabla_a \psi = \mu^2 \psi$
$\frac{\hbar^2}{c^2}\frac{\partial^2\psi}{\partial t^2} - \frac{\hbar^2\partial^2\psi}{\partial x^2} = -2i\hbar\frac{\partial\psi}{\partial\tau}$	$-\hbar^2\frac{\partial^2\psi}{\partial t^2} + c^2\hbar^2\nabla^2\psi = m_0^2 c^4\psi$
$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} - \frac{2\alpha+a}{c^2}\frac{\partial\phi}{\partial t} - \frac{\alpha^2+a\alpha}{c^2}\phi = 0$	$u_{tt} - \Delta u + m^2 u + G'(u) = 0$
$\left(\eta^{\mu\nu}\frac{\partial}{\partial x^\mu}\frac{\partial}{\partial x^\nu} - \left(\frac{mc}{\hbar}\right)^2\right)\phi = 0$	$u_{tt} - \Delta u + mu + P'(u) = 0$
$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \sum_{i=1}^{p}\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^i} - \left(\frac{mc}{\hbar}\right)^2\right)\phi = 0$	$(m > 0, P(u) \geqslant 0)$

Math Entity Linking:

Why and How?

Why: Applications

- Semantic Formula Search
- Math Question Answering
- Math Question Generation

- STEM Document Classification Explainability
- STEM Document Recommender Systems

- Math Citations

How: Methods

- Formula Concept Discovery and Recognition
- Formula Concept Classification and Clustering

- Formula and Identifier Annotation Recommendation

- Formula Concept Benchmarking

What is Mathematical Entity Linking (*MathEL*)?

Definition

Mathematical Entity Linking (*MathEL*) ***links formulae*** *and their constituting entities (identifiers, operators, etc.) to concept names or (**Wikimedia) URLs**.*

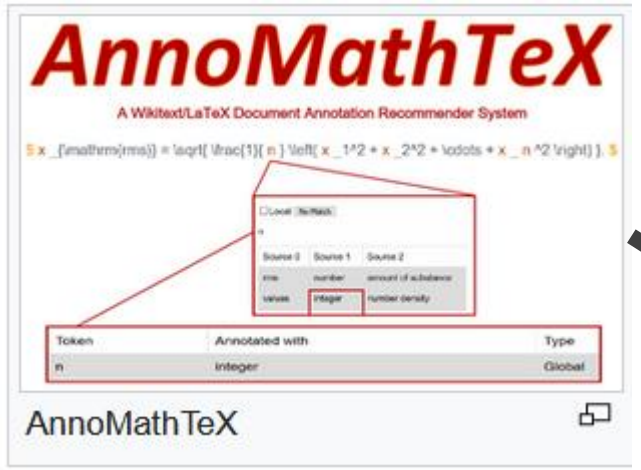
Formula Concept Discovery (*FCD*) *aims at the **definition and exploration** of a **'Formula Concept'** that names bundled equivalent representations of a formula.*

Formula Concept Recognition (*FCR*) *is designed to **match a given formula** to a prior assigned **unique concept identifier**.*

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In *Proceedings of the Web Conference (WWW) 2021*. ACM / IW3C2, April 2021.

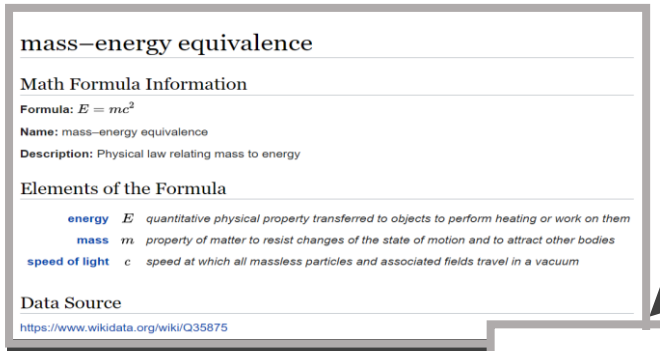
Philipp Scharpf, Moritz Schubotz, Howard S. Cohl, and Bela Gipp. 2019. Towards Formula Concept Discovery and Recognition. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2414)*. CEUR-WS.org, 108–115.

MathEL – Systems

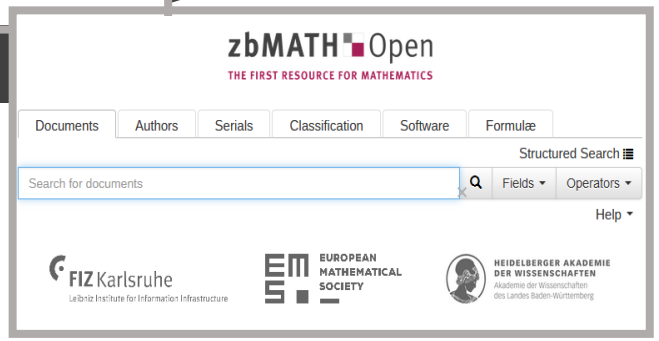


Annotation[1]

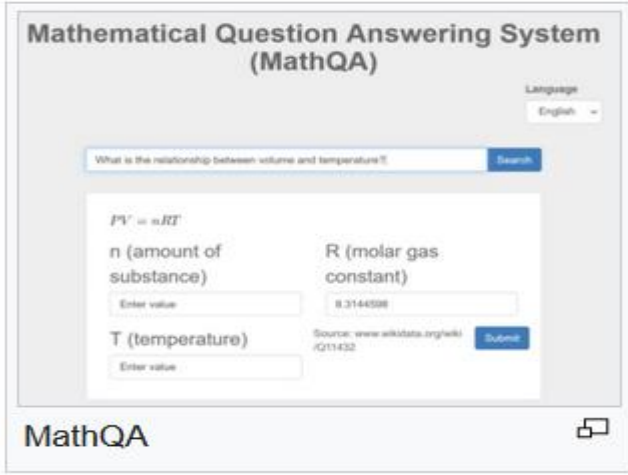
Benchmark[2,3]



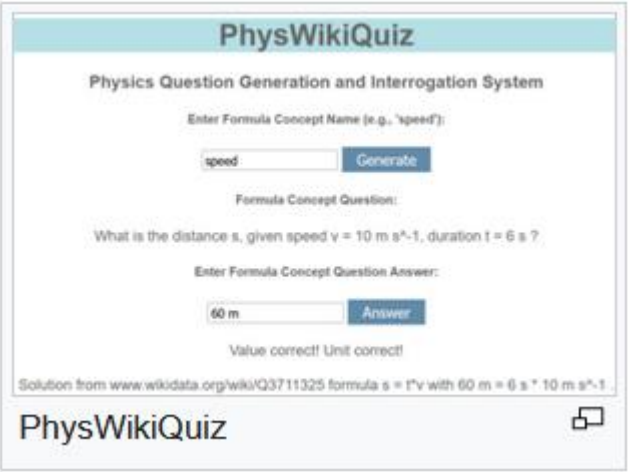
Readability[4]



Classification[5]



Question Answering[6]



Question Generation[7]

[1] https://annomathtex.wmflabs.org
[2] https://mathmlben.wmflabs.org
[3] https://www.wikidata.org
[4] https://www.wikipedia.org
[5] https://zbmath.org
[6] https://mathqa.wmflabs.org
[7] https://physwikiquiz.wmflabs.org

Methods

Problem:
Formula Concept Retrieval Challenge

Issues (1)

1. Different symbols for constants or variables are used. $\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \left(\frac{m_0 c}{\hbar}\right)^2\psi = 0$ $u_{tt} + Au + f(u) = 0$
2. Constants appear in different terms. $-\hbar^2\frac{\partial^2\Psi}{\partial t^2} + c^2\hbar^2\nabla^2\Psi = m_0^2c^4\Psi$
3. Additional terms and functions are introduced. $\frac{\hbar^2}{c^2}\frac{\partial^2\psi}{\partial t^2} - \frac{\hbar^2\partial^2\psi}{\partial^2 x^2} = -2i\hbar\frac{\partial\psi}{\partial\tau}$
4. Signs of the terms differ with the metric signature that is used. $G_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu}$
5. Substitutions, i.e., identifiers are subsumed into others and then appear implicitly.

Problem:
Formula Concept Retrieval Challenge

Issues (2)

6. Additional (index or semantic) sub- or superscripts are introduced. $\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\sum_{i=1}^{p}\frac{\partial}{x^i}\frac{\partial}{x^i}-\left(\frac{mc}{\hbar}\right)^2\right)\varphi=0$

7. Sometimes, a variable dependence is explicitly displayed. $R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\kappa_T(T)T_{\mu\nu}+\Lambda(T)g_{\mu\nu}$

8. Varying derivative notation is used,
e.g., for the time derivative (partial signs, double dot, etc.) of the wave function. $\frac{\partial x}{\partial t}$, $\partial_t x$, x_t

9. Different unit systems are applied.
Constant factors or numbers can be transformed into different unit.

$G_{\mu\nu}+\Lambda g_{\mu\nu}=8\pi T_{\mu\nu}\ (G=c=1)$

$\operatorname{div}\mathbf{E}=4\pi\rho,\ \operatorname{div}\mathbf{B}=0$

10. Differential and integral forms are employed interchangeably.
Maxwell's equations can be written using derivatives.

$\oiint_{\partial\Omega}\mathbf{E}\cdot\mathrm{d}\mathbf{S}=\frac{1}{\varepsilon_0}\iiint_{\Omega}\rho\mathrm{d}V,$

Problem:
Formula Concept Retrieval Challenge

Issues (3)

11. Unification into a single (physics) framework is applied.
Maxwell's equations of electromagnetism combine multiple Formula Concepts:
Gauß' law, Faraday's law, and Ampère's law.

$$\left.\begin{aligned} \operatorname{div}\mathbf{E} &= 4\pi\rho,\ \operatorname{div}\mathbf{B} = 0 \\ \operatorname{rot}\mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t},\ \operatorname{rot}\mathbf{B} = \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \end{aligned}\right\} \partial_\alpha F^{\alpha\beta} = \frac{4\pi}{c} j^\beta$$

12. Tensor notation is used.
Transforming to more compact forms, tensors and indices are introduced.
The electromagnetic field tensor subsumes multiple components of two field vectors.

13. Einstein's summation notation can be used to compactify terms (e.g., derivatives) while omitting summation signs.

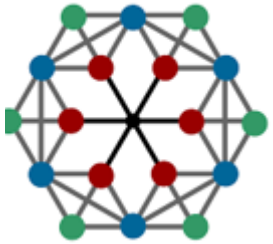
$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \kappa_T(T)T_{\mu\nu} + \Lambda(T)g_{\mu\nu}$$

Formula Concept Clustering

Hubble's law (Q179916)	Equation of state (Q214967)
$p = \omega\rho$	$\dot{a} = aH$
$p = \kappa\rho$	$H_i = \dot{R}/R$
$\omega = p/\rho$	$H = \dot{a}/a$
$p_d = \omega\rho_d$	$H(t) = \dot{a}/a$

Clustering equivalent representations of formulae in the semantic space as named Formula Concept Wikidata items.

Philipp Scharpf, Moritz Schubotz, Howard S. Cohl, and Bela Gipp. 2019. Towards Formula Concept Discovery and Recognition. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2414)*. CEUR-WS.org, 108–115.

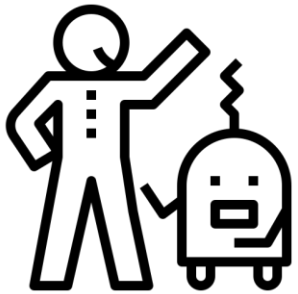


Database:
Wikidata

https://wikidata.org

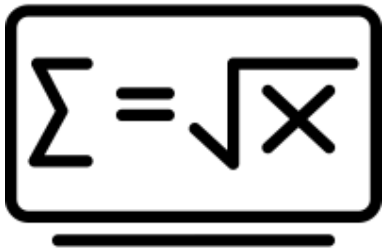
Launched 2014 to connect unique concepts in **language-independent** items

Free, **open** access, collaborative **semantic** knowledge-base

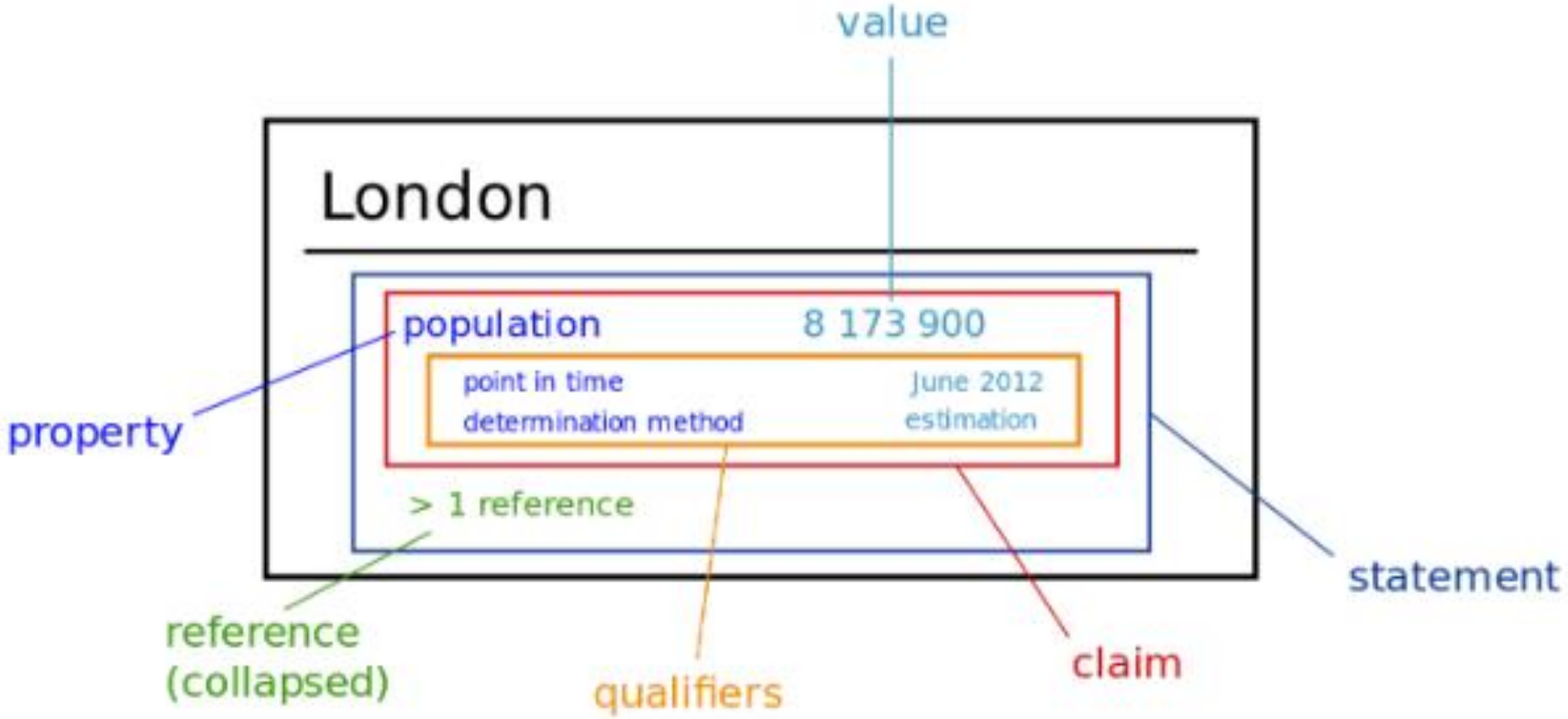


Humans or bots can
create, read, accept, decline or edit item content

~ 5,000 mathematical statements



Database:
Wikidata



https://wikidata.org

Formula Items Wikidata

speed (Q3711325)

magnitude of velocity of motion edit

Statements

defining formula (Property:P2534)

$v = s/t$ edit

0 references

+ add reference

+ add value

in defining formula (Property:P7235)

v edit

0 references

+ add reference

s edit

symbol represents distance

0 references

+ add reference

t edit

symbol represents duration

0 references

+ add reference

+ add value

ISQ dimension (Property:P4020)

LT^{-1} edit

0 references

+ add reference

+ add value

Formula Pages Wikipedia

Mass–energy equivalence

ARTICLE

From Wikipedia, the free encyclopedia

In physics, **mass–energy equivalence** is the principle that mass is a form of energy and that in the rest frame, mass and energy are equivalent and differ only by a constant.[1][2] The principle is fundamental to many fields of physics, including nuclear and particle physics and is described by Albert Einstein's famous formula:[3]

Mass–energy relation

$$E = m c^2$$

FORMULA DETAIL PAGE

Math Formula Information

Formula: $E = mc^2$

Name: mass–energy equivalence

Description: Physical law relating mass to energy

Elements of the Formula

energy	E	*quantitative physical property transferred to objects to perform heating or work on them*
mass	m	*property of matter to resist changes of the state of motion and to attract other bodies*
speed of light	c	*speed at which all massless particles and associated fields travel in a vacuum*

Data Source

https://www.wikidata.org/wiki/Q35875

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In *Proceedings of the Web Conference (WWW) 2021*. ACM / IW3C2, April 2021.

LaTeX Wikidata Macros and MathML Wikidata Content Dictionaries

Formula Annotation Formats

- LaTeX

```
\w{Q11379}{E} = \w{Q11423}{m} \w{Q2111}{c}^2
```

- MathML

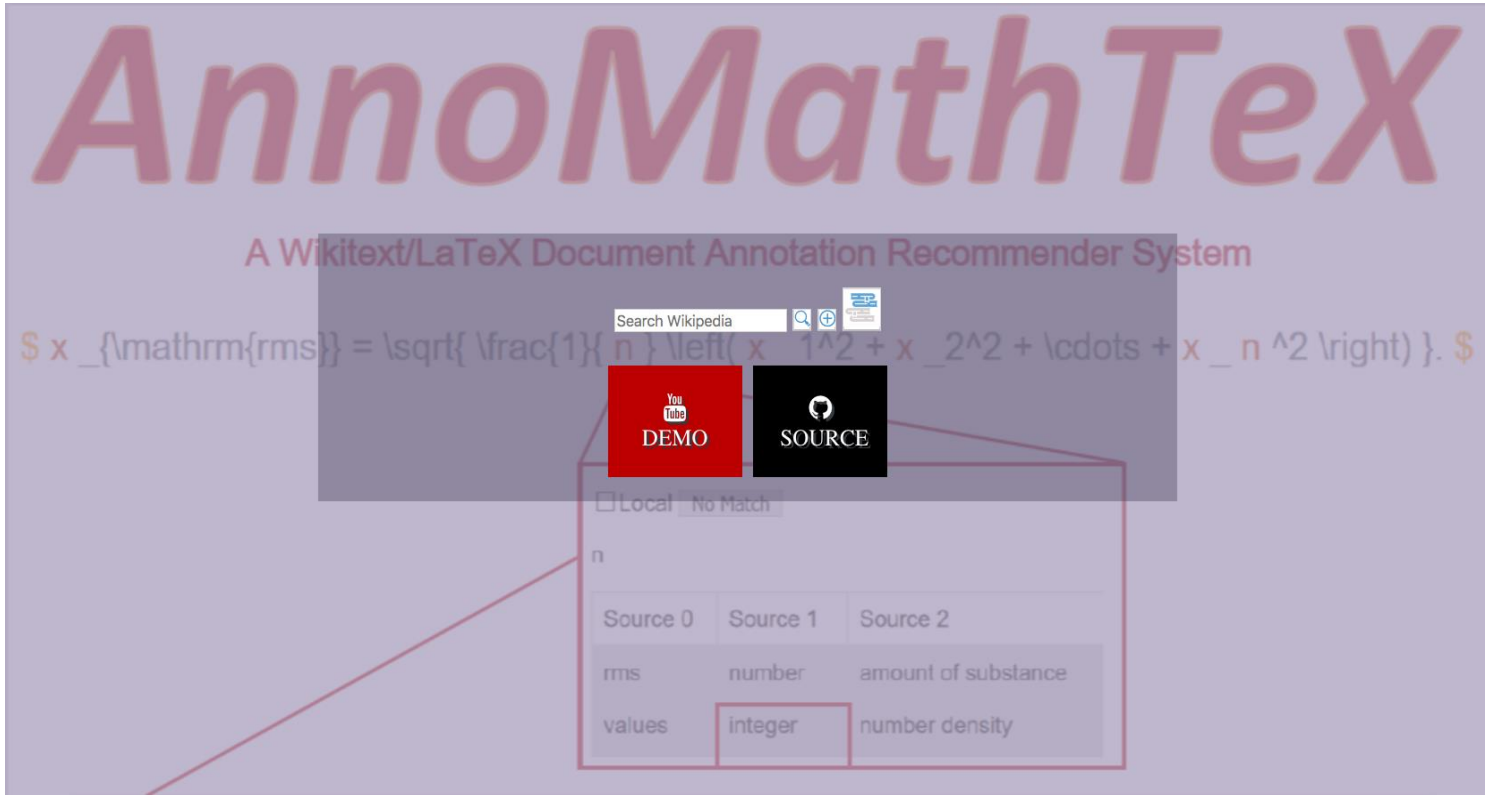
```
<mrow id="p1.1.m1.1.6" xref="p1.1.m1.1.6.cmml">
   <mi id="p1.1.m1.1.1" xref="p1.1.m1.1.1.cmml">E</mi>
   <mo id="p1.1.m1.1.2" xref="p1.1.m1.1.2.cmml">=</mo>
   <mrow id="p1.1.m1.1.6.1" xref="p1.1.m1.1.6.1.cmml">
      <mi id="p1.1.m1.1.3" xref="p1.1.m1.1.3.cmml">m</mi>
      <mo id="p1.1.m1.1.6.1.1" xref="p1.1.m1.1.6.1.1.cmml"></mo>
      <msup id="p1.1.m1.1.6.1.2" xref="p1.1.m1.1.6.1.2.cmml">
         <mi id="p1.1.m1.1.4" xref="p1.1.m1.1.4.cmml">c</mi>
         <mn id="p1.1.m1.1.5.1" xref="p1.1.m1.1.5.1.cmml">2</mn>
      </msup>
   </mrow>
</mrow>
```

```
<annotation-xml encoding="MathML-Content" id="p1.1.m1.1b">
   <apply id="p1.1.m1.1.6.cmml" xref="p1.1.m1.1.6">
      <eq id="p1.1.m1.1.2.cmml" xref="p1.1.m1.1.2"/>
      <csymbol cd="wikidata" id="p1.1.m1.1.1.cmml" xref="p1.1.m1.1.1">Q11379</csymbol>
      <apply id="p1.1.m1.1.6.1.cmml" xref="p1.1.m1.1.6.1">
         <times id="p1.1.m1.1.6.1.1.cmml" xref="p1.1.m1.1.6.1.1"/>
         <csymbol cd="wikidata" id="p1.1.m1.1.3.cmml" xref="p1.1.m1.1.3">Q11423</csymbol>
         <apply id="p1.1.m1.1.6.1.2.cmml" xref="p1.1.m1.1.6.1.2">
            <power id="p1.1.m1.1.6.1.2.1.cmml" xref="p1.1.m1.1.6.1.2"/>
            <csymbol cd="wikidata" id="p1.1.m1.1.4.cmml" xref="p1.1.m1.1.4">Q2111</csymbol>
            <cn id="p1.1.m1.1.5.1.cmml" type="integer" xref="p1.1.m1.1.5.1">2</cn>
         </apply>
      </apply>
   </apply>
</annotation-xml>
```

```
<annotation encoding="application/x-tex" id="p1.1.m1.1c">E=mc^{2}</annotation>
```

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. 2018. Representing Mathematical Formulae in Content MathML using Wikidata. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2132)*. CEUR-WS.org, 46–59.

Annotation Recommendation



$ E = m c ^2 $

Identifiers	Time (seconds)
Recommendation	2.6
Manual	6.3
Formulae	**Time (seconds)**
Recommendation	2.8
Manual	4.0

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In *Proceedings of the Web Conference (WWW) 2021*. ACM / IW3C2, April 2021.

https://annomathtex.wmflabs.org

Annotation Recommendation

FORMULA ANNOTATION

Local
No match: Submit

Formula: E=mc^2

$ E = m c ^2 $

Not a formula

Wikidata1	Wikidata2	WordWindow	FormulaConceptDB	Manual
mass–energy equivalence (Q35875)	mass–energy equivalence (Q35875)	formula (N/A)	mass–energy equivalence (Q35875)	
	time-independent Schrödinger equation (Q25829357)	einstein (N/A)		

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In *Proceedings of the Web Conference (WWW) 2021*. ACM / IW3C2, April 2021.

Annotation Recommendation

IDENTIFIER ANNOTATION

☐ Local
No match: Submit

$ E = m c ^2 $

Identifier: m

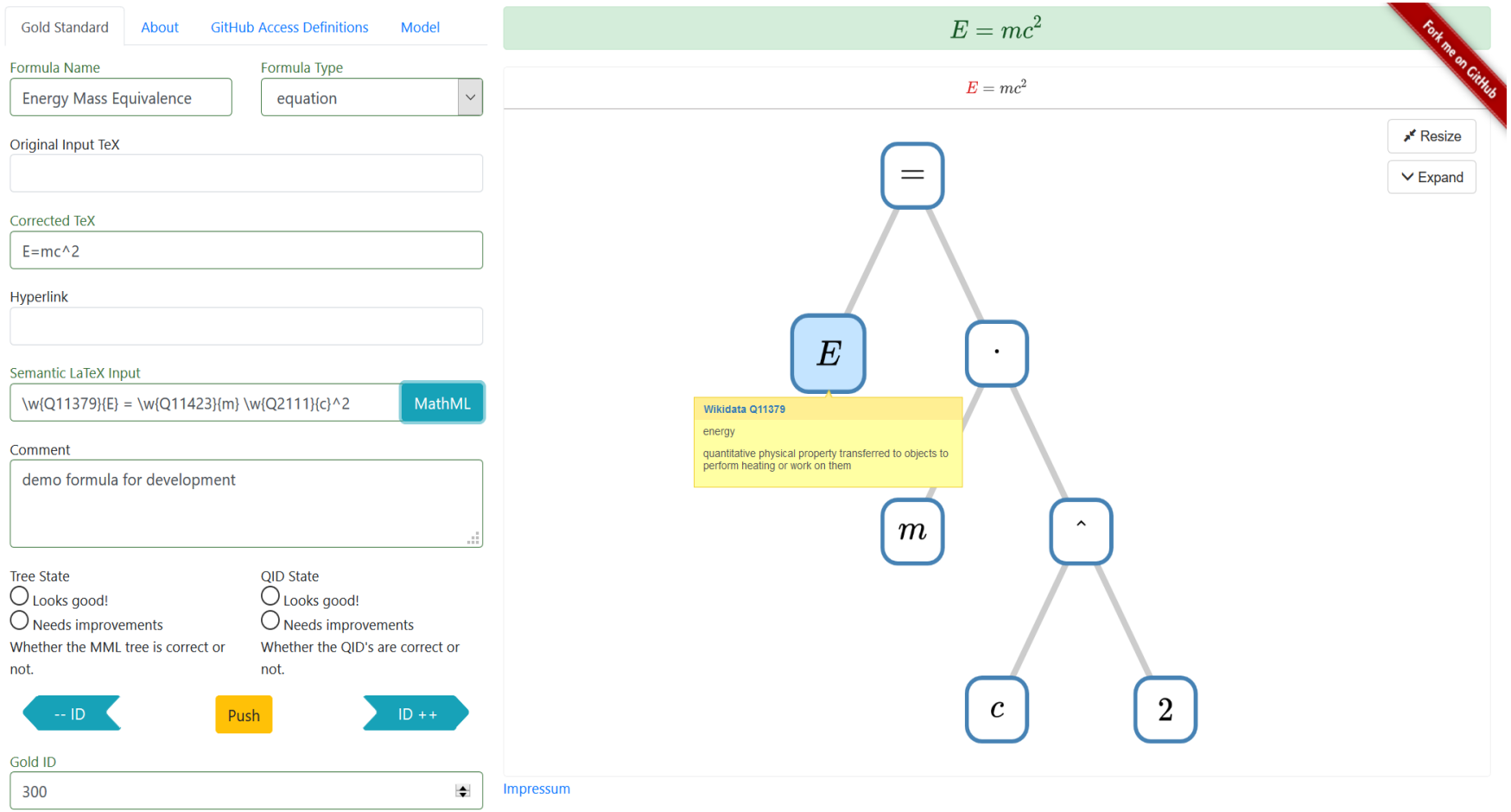
Not an identifier

Annotated Identifiers: 61/96

Source 1	Source 2	Source 3	Source 4	Source 5
mass (Q11423)	hypothesis tests (N/A)	N/A (N/A)	motion (N/A)	exhaust gas mass flow (Q320176)
field (Q185674)	edges (N/A)		law (N/A)	mass (Q11423)
integer (Q12503)	mass (Q11423)		newton (N/A)	
number (N/A)	message (N/A)		acceleration (Q11376)	
particle (N/A)	rest mass (N/A)		mass (Q11423)	

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Fast linking of mathematical wikidata entities in wikipedia articles using annotation recommendation. In *Proceedings of the Web Conference (WWW) 2021*. ACM / IW3C2, April 2021.

Benchmark: *MathMLben*



https://mathmlben.wmflabs.org

Schubotz, M., Greiner-Petter, A., Scharpf, P., Meuschke, N., Cohl, H.S., Gipp, B.: Improving the representation and conversion of mathematical formulae by considering their textual context. In: JCDL. pp. 233–242. ACM (2018)

Mining the arXiv NTCIR Dataset

http://ntcir-math.nii.ac.jp/data/

National Institute of Informatics Testbeds and Community for Information Access Research Project (NTCIR) (Aizawa et al., 2014, Zanibbi et al., 2016)

104,062 TEI document section files

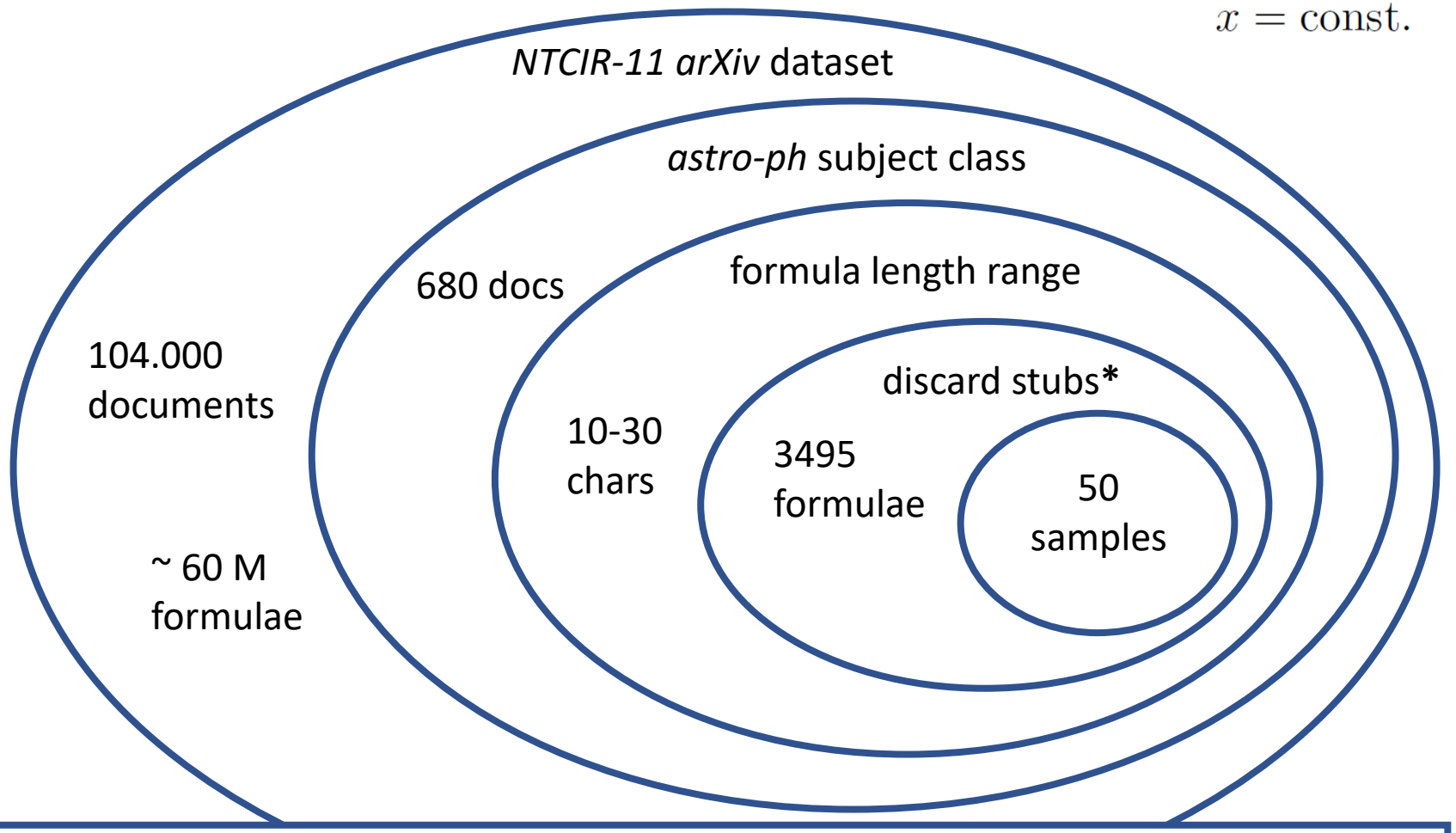
'astro-ph', 'cond-mat', 'cs', 'gr-qc', 'hep-lat', 'hep-ph', 'hep-th', 'math-ph', 'math', 'nlin', 'quant-ph', 'physics', 'alg-geom', 'q-alg'

Data Selection

* $x = x(t)$

$x = y$

$x = \text{const.}$



Aizawa, A., Kohlhase, M., Ounis, I., Schubotz, M.: NTCIR-11 math-2 task overview. In: NTCIR. National Institute of Informatics (NII) (2014)

Equivalent Representations

Using k-Nearest Neighbor Retrieval

#	Formula	Name (QID)	d / $\hat{d}$	$s_{e_m}, s_{\hat{e}_m}, s_{e_s}, s_{\hat{e}_s}$	Encoding: sample formula
1	$H = \dot{a}/a$	hubble parameter (Q179916)	32 / 32	0.0, 0.1, 0.0, 0.9	$\hat{e}_s$: $H_i = \dot{R}/R$
2	$p = \omega\rho$	equation of state (Q214967)	6 / 5	0.3, 0.0, 0.1, 0.6	e_s: $p_d = w\rho_d$
3	$\omega = p/\rho$	accelerating universe (Q1049613)	4 / 3	0.7, 0.0, 0.0, 0.3	e_m: $p = \omega\rho$
4	$p = -A/\rho^{\alpha}$	dark fluid (Q5223514)	4 / 4	0.7, 0.0, 0.3, 0.0	e_m: $p = -\frac{A}{\rho^{\alpha}}$
5	$p_d = w\rho_d$	dark energy (Q18343)	4 / 3	0.3, 0.0, 0.3, 0.3	e_s: $p_X = \omega_X \rho_X$

...

math2vec (e_m)**: 70%**
semantics tf-idf $(\hat{e}_s)$**: 15%**
semantics2vec (e_s)**: 11%**
math tf-idf $(\hat{e}_m)$**: 4%**

Overall, for 34/50 = 68% of the sample formulae, we could retrieve equivalent representations

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. 2018. Representing Mathematical Formulae in Content MathML using Wikidata. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2132)*. CEUR-WS.org, 46–59.

Concept Name Candidates

Using Surrounding Text Retrieval

#	Formula	Name (QID)
1	$H = \dot{a}/a$	hubble parameter (Q179916)
2	$p = \omega\rho$	equation of state (Q214967)
3	$\omega = p/\rho$	accelerating universe (Q1049613)
4	$p = -A/\rho^{\alpha}$	dark fluid (Q5223514)
5	$p_d = w\rho_d$	dark energy (Q18343)

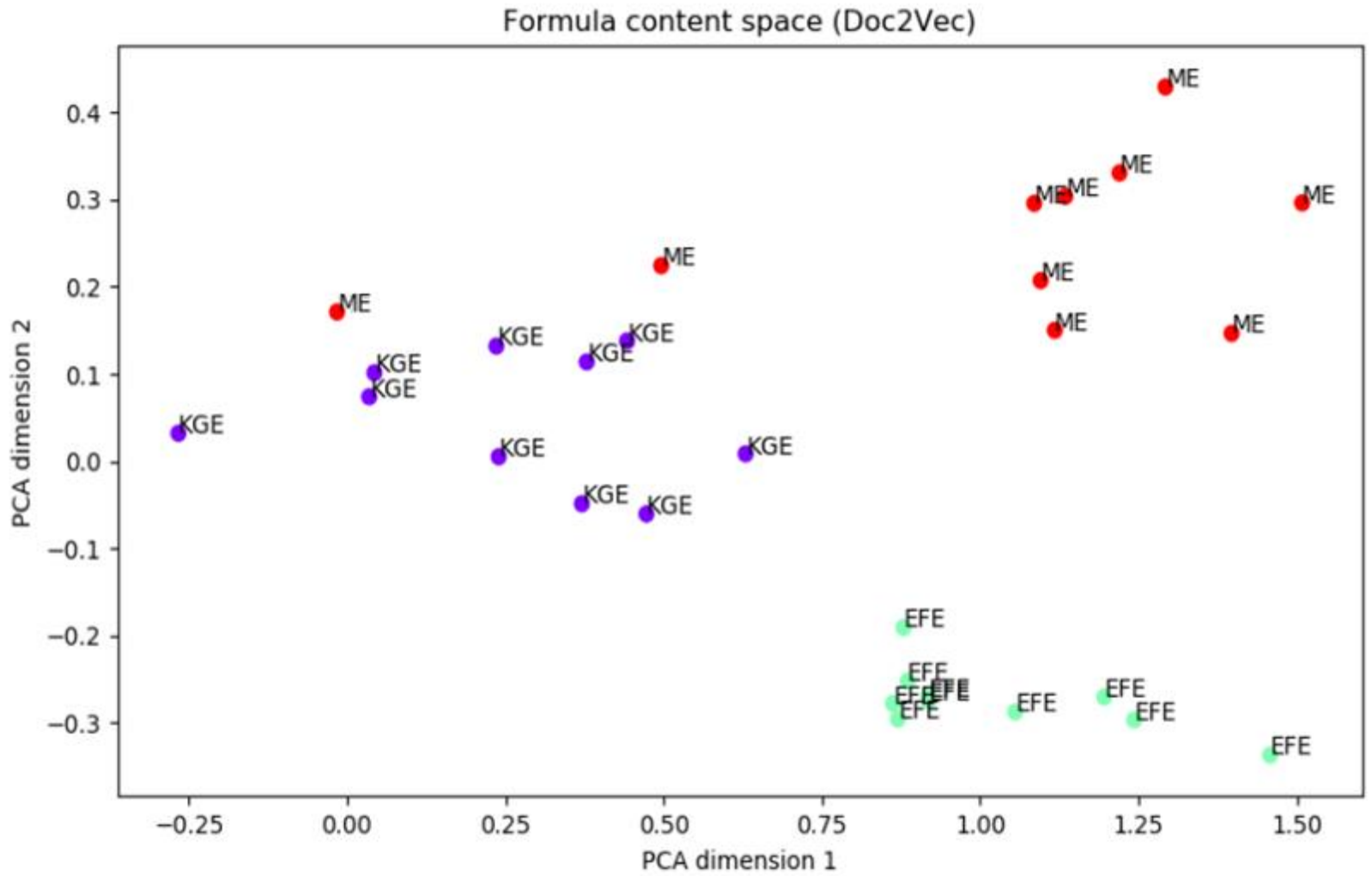
...

We achieve a *recall* of 36/50 = 72% for the formula name recommendations

For 41/50 = 82% of the retrieved name candidates, there was a *Wikidata QID* available

Philipp Scharpf, Moritz Schubotz, and Bela Gipp. 2018. Representing Mathematical Formulae in Content MathML using Wikidata. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2132)*. CEUR-WS.org, 46–59.

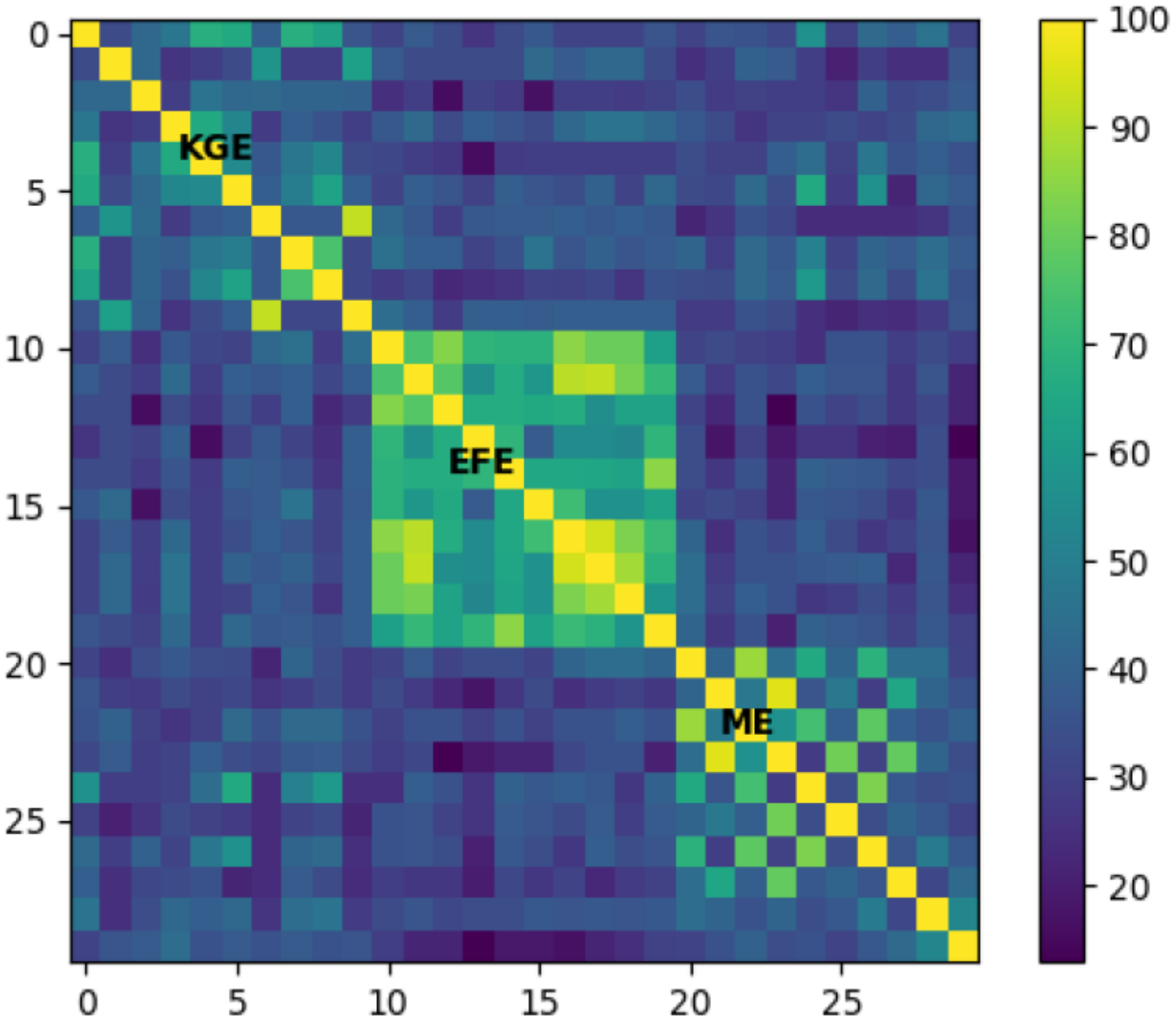
Formula (Concept) Clustering



Philipp Scharpf, Moritz Schubotz, Howard S. Cohl, and Bela Gipp. 2019. T… Formula Concept Discovery and Recognition. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2414)*. CEUR-WS.org, 108–115.

Doc2Vec Encodings, PCA reduced, Kmeans clustered (purity 0.94)

Formula (Concept) Clustering



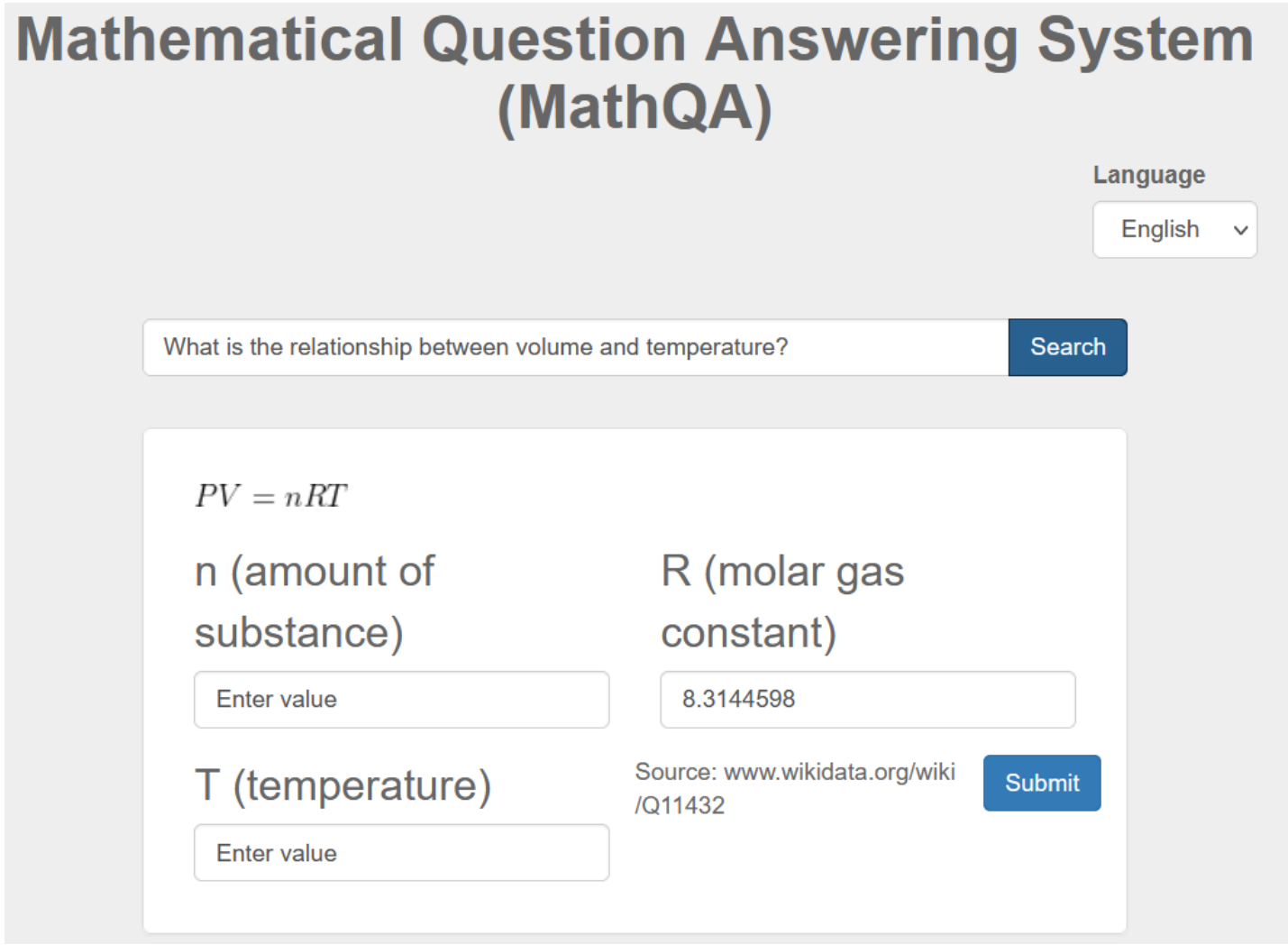
Fuzzy string similarity percentages

Philipp Scharpf, Moritz Schubotz, Howard S. Cohl, and Bela Gipp. 2019. Towards Formula Concept Discovery and Recognition. In *BIRNDL@SIGIR (CEUR Workshop Proceedings, Vol. 2414)*. CEUR-WS.org, 108–115.

Applications

Application: Mathematical Question Answering (MathQA)

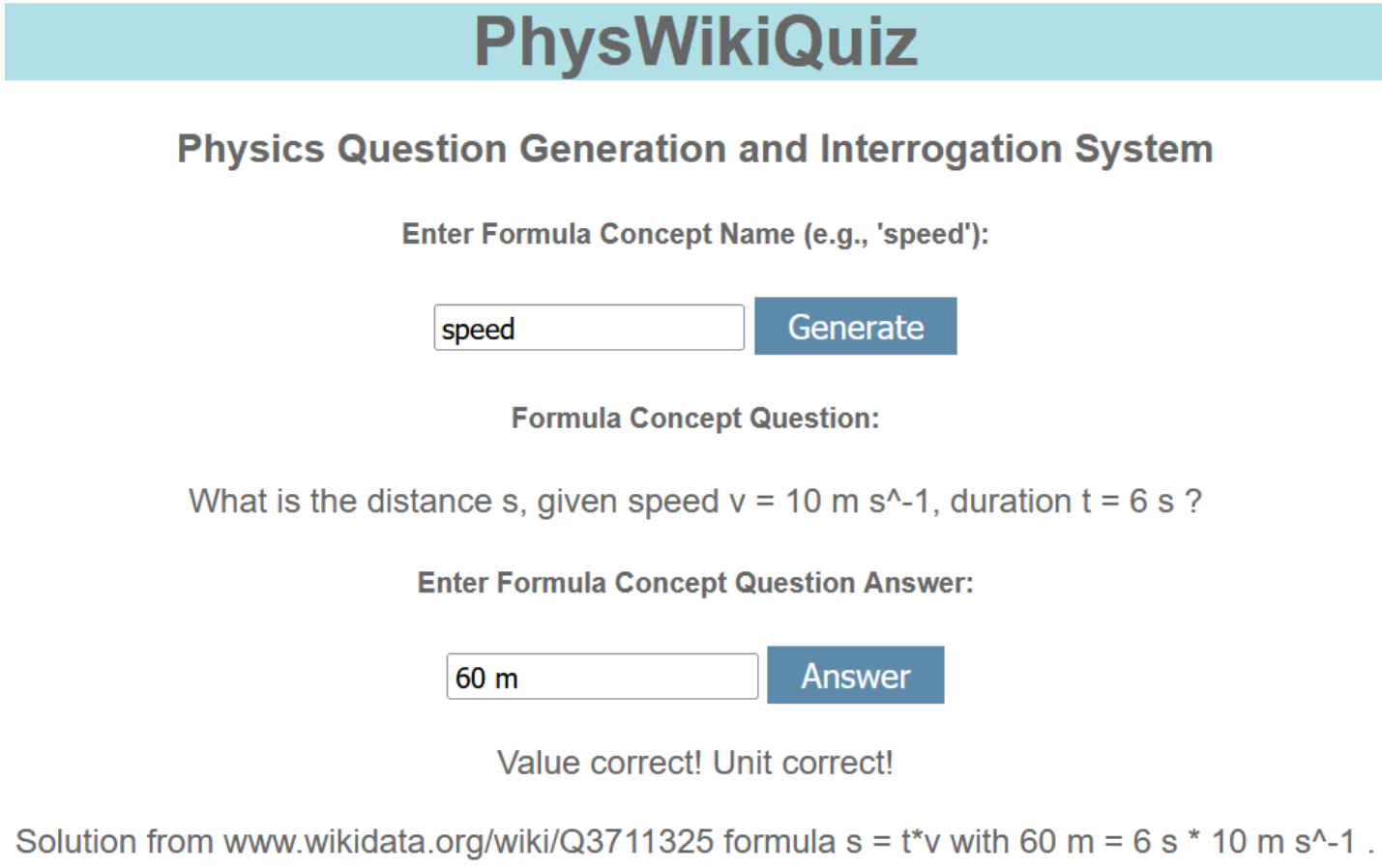
https://mathqa.wmflabs.org



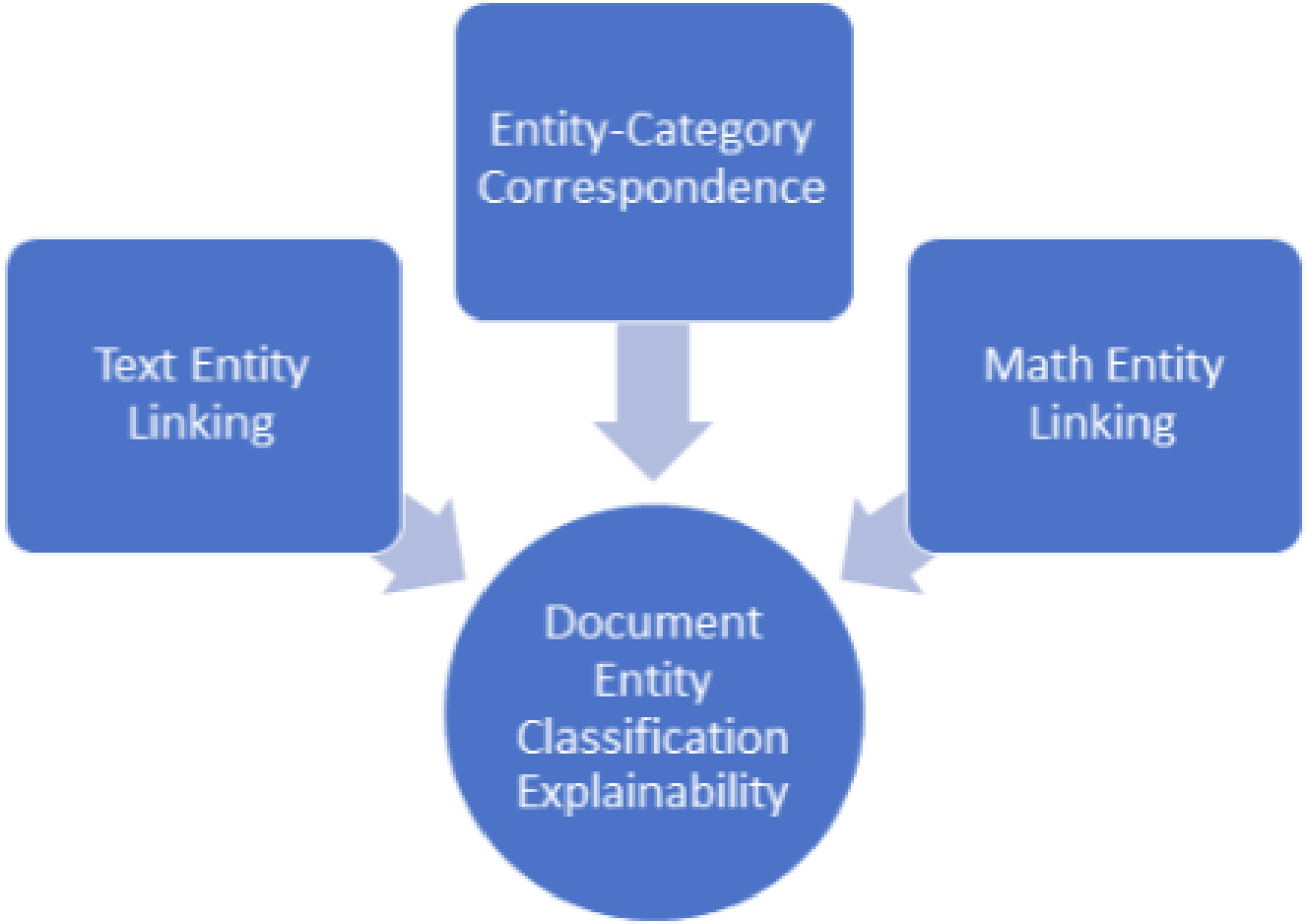
Moritz Schubotz, Philipp Scharpf, Kaushal Dudhat, Yash Nagar, Felix Hamborg, and Bela Gipp. Introducing mathqa - A math-aware question answering system. *Information Discovery and Delivery*, 42, No. 4:214–224, 2019.

Application: Physics Question Generation (PhysWikiQuiz)

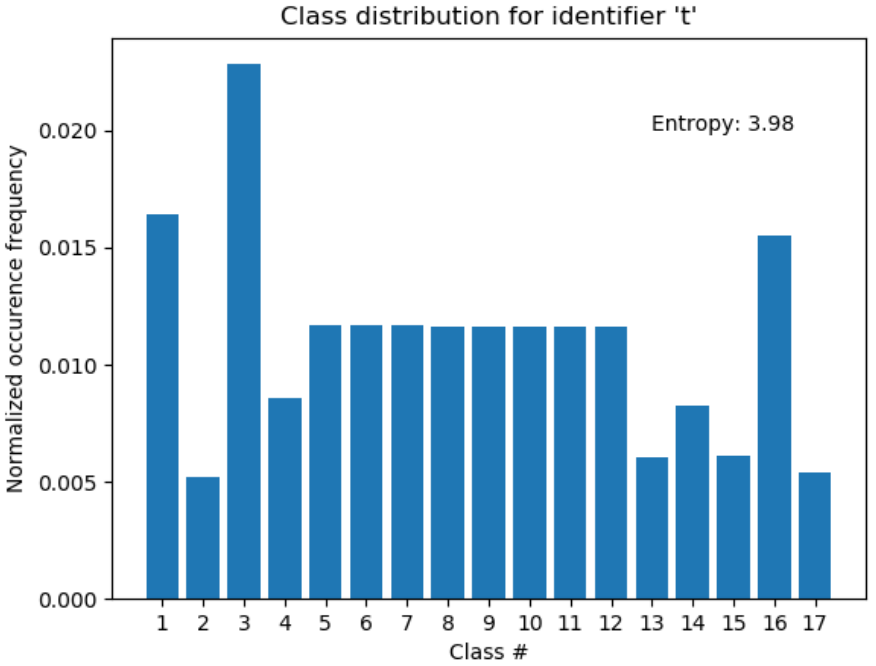
https://physwikiquiz.wmflabs.org



Application: STEM Document Classification Explainability

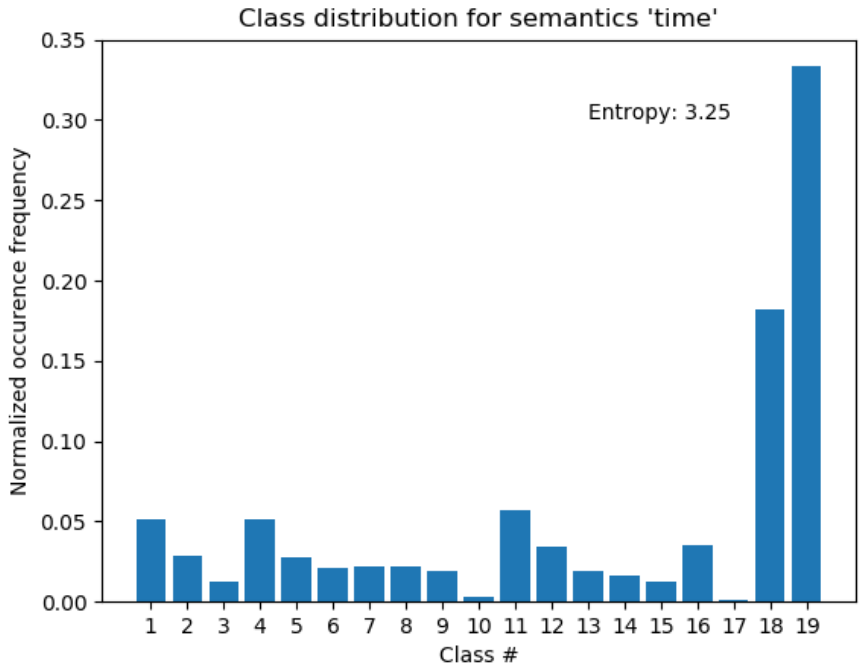


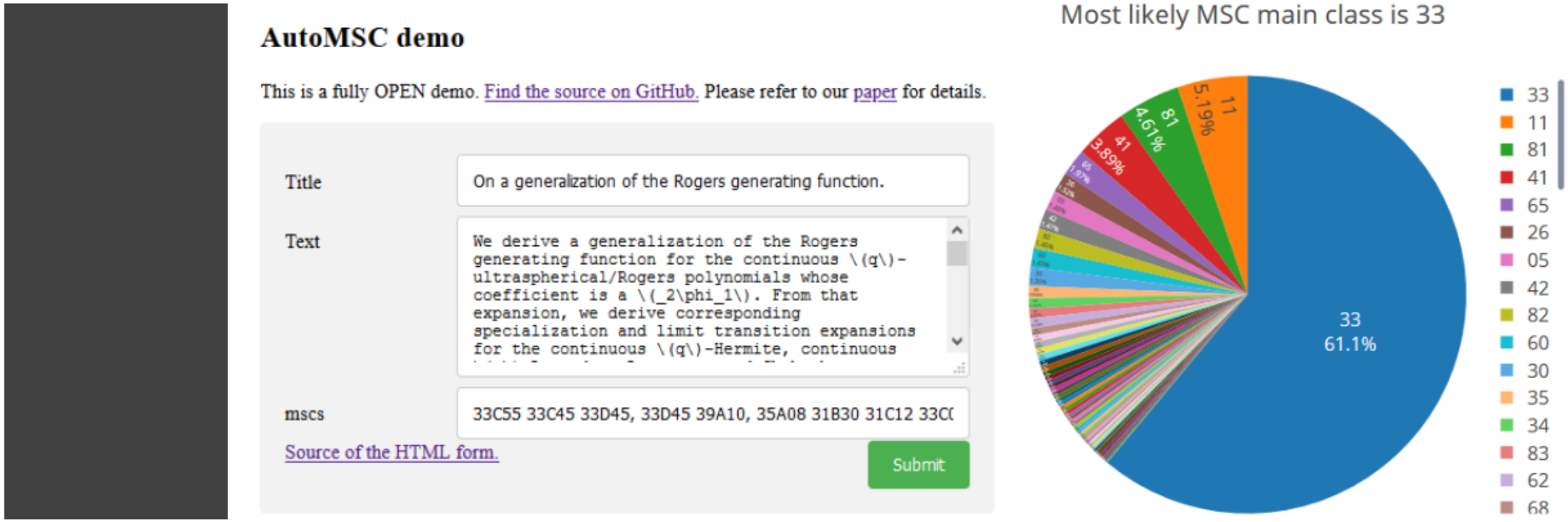
Application: STEM Document Classification Explainability



Identifier symbol ambiguity

Identifier name expressivity



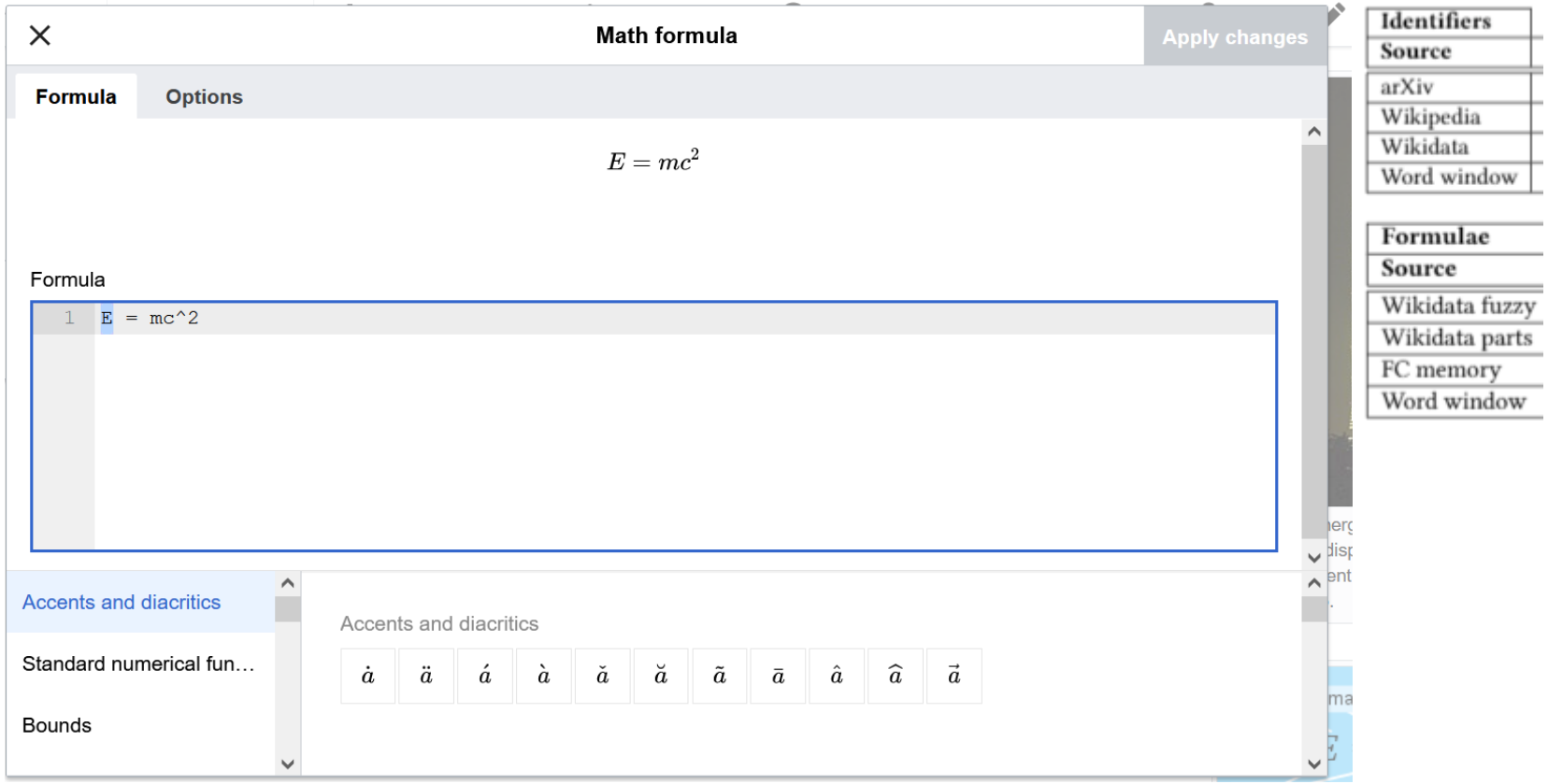


https://automscbackend.formulasearchengine.com

Schubotz et al., *AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels*, CICM 2020

Outlook

Wikipedia Annotation Integration



Identifiers
Source
arXiv
Wikipedia
Wikidata
Word window

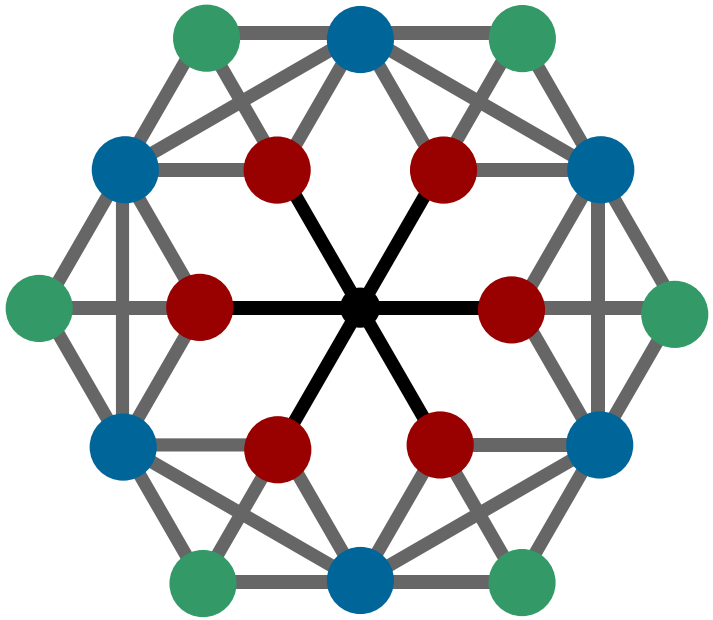
Formulae
Source
Wikidata fuzzy
Wikidata parts
FC memory
Word window

Math π Link Wiki

https://github.com/philsMINT/MathWikiLink

Wikidata knowledge-base population

Connect Wikidata to zbMATH Knowledge Graph



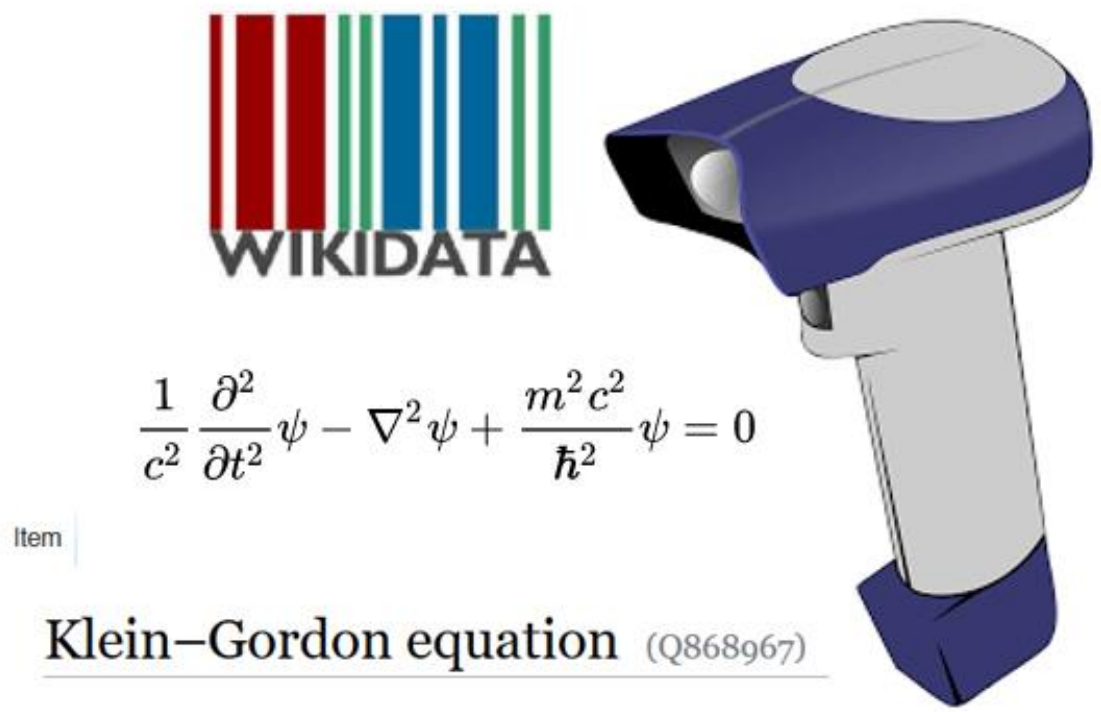
WIKIDATA

zbMATH Open

THE FIRST RESOURCE FOR MATHEMATICS

Formula Concept Image Recognition

$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \left(\frac{m_0 c}{\hbar}\right)^2\psi = 0$	$u_{tt} + Au + f(u) = 0$
$\partial_{ct}^2 h_n(z,t) - \partial_z^2 h_n(z,t) + \nu_n^2 h_n(z,t) = 0$	$\nabla^a\nabla_a\psi = \mu^2\psi$
$\frac{\hbar^2}{c^2}\frac{\partial^2\psi}{\partial t^2} - \frac{\hbar^2\partial^2\psi}{\partial x^2} = -2i\hbar\frac{\partial\psi}{\partial\tau}$	$-\hbar^2\frac{\partial^2\psi}{\partial t^2} + c^2\hbar^2\nabla^2\psi = m_0^2c^4\psi$
$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} - \frac{2\alpha+a}{c^2}\frac{\partial\phi}{\partial t} - \frac{\alpha^2+a\alpha}{c^2}\phi = 0$	$u_{tt} - \Delta u + m^2u + G'(u) = 0$
$\left(\eta^{\mu\nu}\frac{\partial}{\partial x^\mu}\frac{\partial}{\partial x^\nu} - \left(\frac{mc}{\hbar}\right)^2\right)\phi = 0$	$u_{tt} - \Delta u + mu + P'(u) = 0$
$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \sum_{i=1}^{p}\frac{\partial}{\partial x^i}\frac{\partial}{\partial x^i} - \left(\frac{mc}{\hbar}\right)^2\right)\phi = 0$	$(m > 0, P(u) \geqslant 0)$



PhysWikiQuiz Applied Questions

- *A car drives with a speed of 3 m/s.*

→ *How much distance can it cover in 1 s?*

- *The mass of an electron is 9.1 times 10^-31 kg.*

→ *How much is the rest energy of an electron in electron volts?*

- *Two masses of 5 kg and 1 kg are located at a distance of 6 m away from each other.*

→ *Find the center of mass of the system. Assume that the heavier mass is located at the origin and the line joining the two masses is the x-axis.*