FACULTY
OF INFORMATICS
Masaryk University

Trains and Trees of Thoughts

Towards the Representation of Structural Semantics

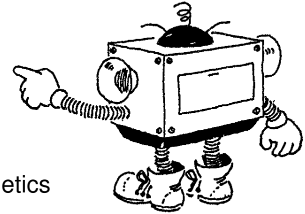
Petr Sojka

sojka@fi.muni.cz

Semiomaths workshop, ETH Zurich, 2018, March 2nd, 2PM

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

Brno, Czech Republic



- town, where Kurt Gödel was born
  and spent his childhood
- town, where Gregor Mendel has founded genetics
- town of ‘Silicon Valley’ of [Central] Europe with high concentration of Computer Science bussinesses (RedHat, IBM, Kiwi, Honeywell), and academia (Faculty of Informatics MU, FIT MU)

Table of Contents

Semio(tics)

Communication axioms [for maths]

Six blind monks and an elephant

The Mental and Semiotics Iceberg

Maths

Math scholarly communication

Novel methodologies in empirical linguistics and corpora

Higher-level representations derived from [math] corpora

Examples: LDA, formulae indexing, citation parsing

Embeddings in vector spaces

References

FACULTY
OF INFORMATICS
Masaryk University

Semiotics

Study of signs as means of language or *communication*.

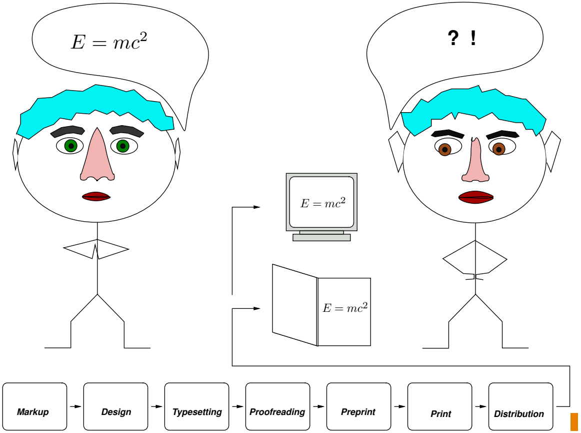
Is math communication specific?

Are the signs used in math specific?

Are computer programming languages specific?

How mathematics and computer science differ?

Scholarly Communication via Digital Mathematics Libraries (DMLs): DML-CZ, EuDML project participation



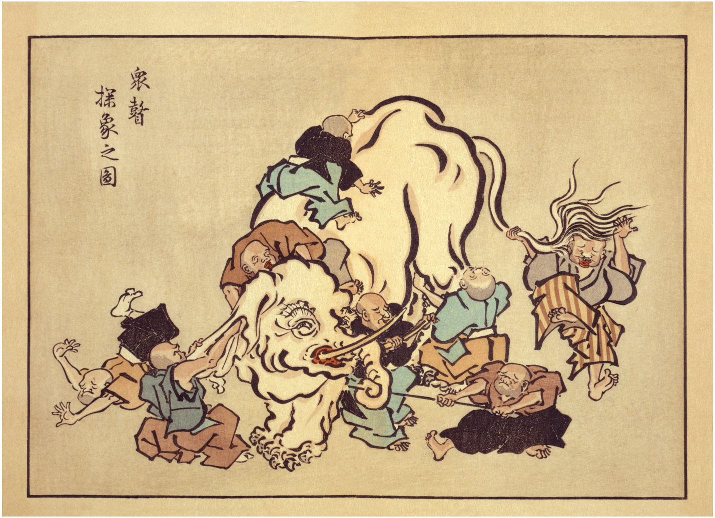
Markup → Design → Typesetting → Proofreading → Preprint → Print → Distribution

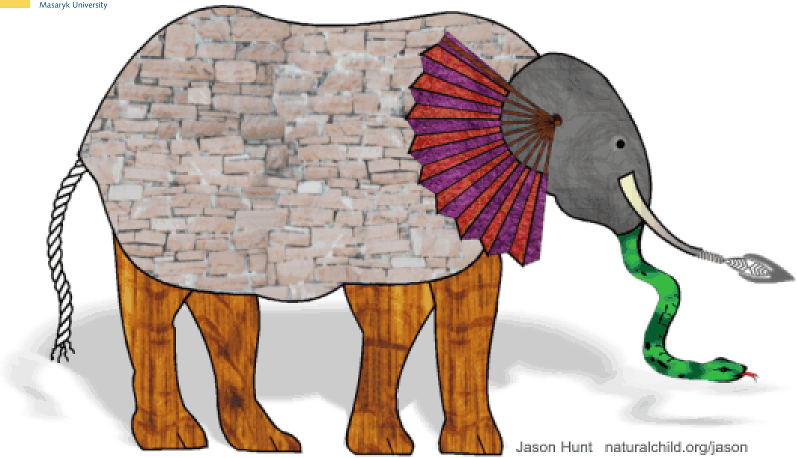
Paul Watzlawick’s Five Axioms of Communication

1. (cannot not) One cannot not communicate.
2. (content & relationship) Every communication has a content and relationship aspect such that the latter classifies the former and is therefore a meta-communication.
3. (punctuation) The nature of a relationship is dependent on the punctuation of the partners’ communication procedures.
4. (digital & analogic, discrete & continuous) Human communication involves both digital and analogic modalities.
5. (symmetric or complementary) Inter-human communication procedures are either symmetric or complementary, depending on whether the relationship of the partners is based on differences or parity.

Do they hold for math-specific discourse?

Six blind monks examining an elephant on a trip





Q: Is elephant a wall (belly), hand fan (ear), solid pipe (tusk), pillar (leg), rope (tail) or tree branch (trunk)?

Let the animal on the road is meaning-conveying [math] communication in the form of scientific papers stored in the digital libraries like the EuDML or arXiv, digital library with math content.

Six blind “monks” quarreling after the trip

1. (Shannon) touched information-theoretic properties of communication
2. (Chomsky) stresses formal grammars in communication languages
3. (Gödel) states limitations in formal expressiveness communication, being incomplete or inconsistent
4. (Rogers) person-centered communication
5. (Locke) views that knowledge comes primarily from sensory experience: empiricism
6. (Lakoff) argues that conceptual metaphors are basis for embodied minds’ communication

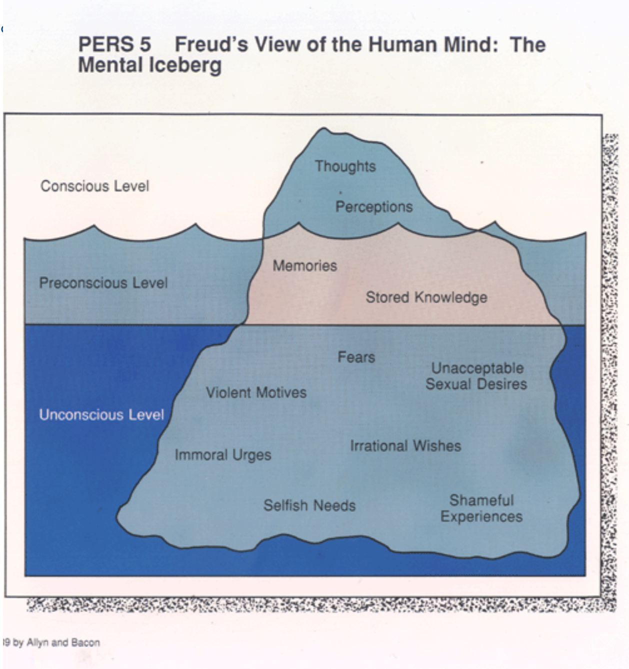
Who is right? Does the question makes sense at all? Whom do you identify with on your ride of thoughts?

Metaphor of an Iceberg

- Icebergs have small visible and big invisible part.
- Icebergs with the same visible parts may have quite different invisible parts.
- Iceberg does have internal structure and qualities that differ.
- The glaciers swim, struggle against each other, melt down, diminish, grow, …

***

- Signs are created in human minds: they reflect human's (objective) conscious thoughts, but also human's (subjective) unconscious personality.
- Freud's view of the human mind is *mental iceberg*.



PERS 5 Freud’s View of the Human Mind: The Mental Iceberg

9 by Allyn and Bacon

The Mental Iceberg metaphor: conscious, preconscious and unconsious levels

Only 10% of an iceberg is visible (conscious) whereas the other 90% is beneath the water: the preconscious is allotted approximately 10%–15% whereas the unconscious is allotted an overwhelming 75%–80%.

Conscious $\equiv$ visible (surface) words/texts, same for all.

Preconscious $\equiv$ structured information deducible from (surface) text based on language and common knowledge, personal (e.g. different and subjective based on previous occurences). Mostly present *latently*.

Unconscious $\equiv$ yet unnamed unknown relations and knowledge indirectly related to the meaning of the conveyed, visible part of mind: expressed messages.

The Semiotics Iceberg Metaphor: layers of *visible* signs and hidden *shared* and *personal* structures linked to them

Only 10% of communication is on a) **surface** (written text with visual marks and punctuation) whereas the other 90% is beneath the water: ‘preconscious’ b) **shared common** sense layer, and ‘subconscious’ level of c) **personal** association and connotations. agreed notions (arithmetics) and notation (formulas) ground preconscious and unconscious).

a) $\equiv$ visible (surface) words/texts, formulas.

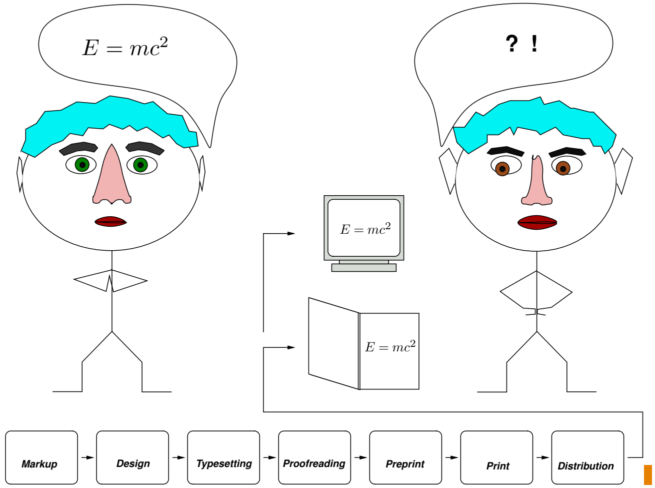
b) $\equiv$ structured information deducible from (surface) text based on language and common knowledge, latently present in the communication. Mostly present *latently*.

c) $\equiv$ subjective, possibly emotional connotation the conveyed message raises.

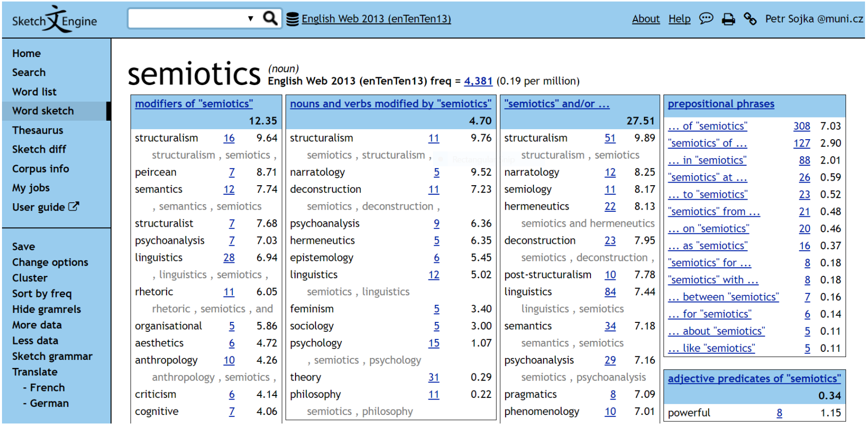
The Semiotics Iceberg Metaphor: $E = mc^2$

- *visible* part: $E = mc^2$
- signs and hidden *shared* part: notions of mass, energy, speed of light and their notations
- *personal* structures and connotations linked to them

Scholarly Communication *via DMLs using rich KB*



Sketch Engine: from statistics to insight



semiotics *(noun)*
English Web 2013 (enTenTen13) freq = 4,381 (0.19 per million)

modifiers of "semiotics"		12.35
structuralism	16	9.64
structuralism , semiotics ,		
peircean	7	8.71
semantics	12	7.74
, semantics , semiotics		
structuralist	7	7.68
psychoanalysis	7	7.03
linguistics	28	6.94
, linguistics , semiotics ,		
rhetoric	11	6.05
rhetoric , semiotics , and		
organisational	5	5.86
aesthetics	6	4.72
anthropology	10	4.26
anthropology , semiotics ,		
criticism	6	4.14
cognitive	7	4.06

nouns and verbs modified by "semiotics"		4.70
structuralism	11	9.76
semiotics , structuralism ,		
narratology	5	9.52
deconstruction	11	7.23
semiotics , deconstruction ,		
psychoanalysis	9	6.36
hermeneutics	5	6.35
epistemology	6	5.45
linguistics	12	5.02
semiotics , linguistics		
feminism	5	3.40
sociology	5	3.00
psychology	15	1.07
, semiotics , psychology		
theory	31	0.29
philosophy	11	0.22
semiotics , philosophy		

"semiotics" and/or ...		27.51
structuralism	51	9.89
structuralism , semiotics		
narratology	12	8.25
semiology	11	8.17
hermeneutics	22	8.13
semiotics and hermeneutics		
deconstruction	23	7.95
semiotics , deconstruction ,		
post-structuralism	10	7.78
linguistics	84	7.44
linguistics , semiotics		
semantics	34	7.18
semantics , semiotics		
psychoanalysis	29	7.16
semiotics , psychoanalysis		
pragmatics	8	7.09
phenomenology	10	7.01

prepositional phrases		
... of "semiotics"	308	7.03
"semiotics" of ...	127	2.90
... in "semiotics"	88	2.01
"semiotics" at ...	26	0.59
... to "semiotics"	23	0.52
"semiotics" from ...	21	0.48
... on "semiotics"	20	0.46
... as "semiotics"	16	0.37
"semiotics" for ...	8	0.18
"semiotics" with ...	8	0.18
... between "semiotics"	7	0.16
... for "semiotics"	6	0.14
... about "semiotics"	5	0.11
... like "semiotics"	5	0.11

adjective predicates of "semiotics"		0.34
powerful	8	1.15

Which insight math corpora could give us? Pros and cons

Firthian linguistics: You shall know a word by the company it keeps (Firth, J.R., 1957).

You shall know a sign/ notion/ math formulae by the company it keeps.

- Sketches of math signs similarly generated as word sketches?
- Math corpora reveals globality or locality of sign/ notion usage.
- Putting knowledge on one place allows for new killer application: search, similar sign search, similar phrase search, similar formulae search, similar plagiarism search based on word n-grams similar theorem search similar thoughts and structures search similar XXX search.

Towards higher level content representations – knowledge bases

NLP processing from strings via words to meaning, including *math-awareness*
math specifics: structures and abstractions

- to allow searching (semantically) similar papers, precise [semantic] indexing: search as a gate to knowledge
- to allow exploration of a DML by intelligent browsing of (semantically) man is known by the company he keepssimilar papers: distributional semantics topic modeling as Latent Semantic Indexing, Latent Dirichlet Allocation
- to allow personalization and domain specifics, e.g. semantic faceted search (formulae,…)
- to track ‘train of thought’ – narrative qualities of papers, proofs (Mizar type of paper)

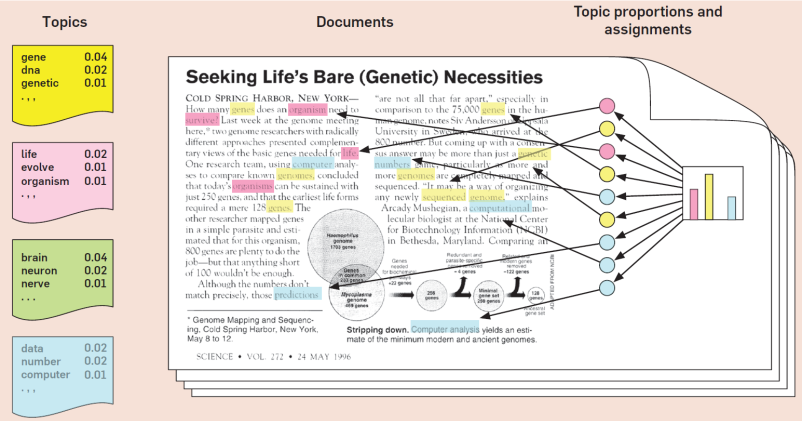
Motivation for example I

```
From: Shayan A Tabrizi <shayantabrizi@gmail.com>
Subject: [Corpora-List] Dataset for Different Research Areas

I want to find the relevance of each of the research papers of
my dataset to each of the research areas such as Physics, CS,
Math, Social Sciences, etc.
Thus, I need a dataset consisting of all research areas and
some sample texts (preferably papers) in that area, to estimate
the similarity of each of my papers to each of the areas.
               *Is there any such dataset?*
Some points:
   1. It is much much better if the dataset has areas in different
   granularities. e.g. in one level: Mathematics, Physics, CS, etc.
   and in a more fine-grained level divides CS to Networks,
   Artificial Intelligence, etc.
   2. Even if the dataset only consists of a specific domain
   (especially CS) and its sub-domains it is still usable.
```

Probabilistic Topical Modeling: Latent Dirichlet Allocation

- topic: weighted list of words
- document: weighted list of topics


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
, , ,
life 0.02
evolve 0.01
organism 0.01
, , ,
brain 0.04
neuron 0.02
nerve 0.01
. . .
data 0.02
number 0.02
computer 0.01
, , ,
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

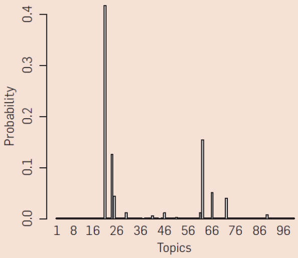
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

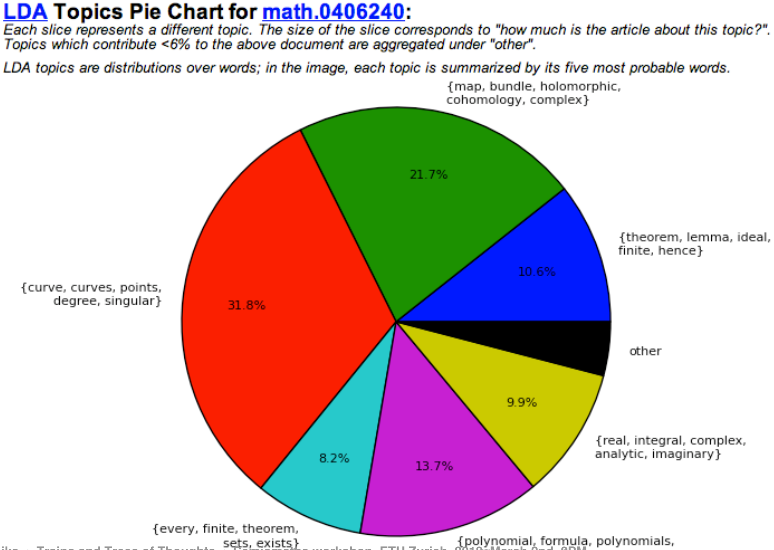
Topical Modeling: Latent Dirichlet Allocation II

- all topics computed automatically from document corpora



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Example I: Automated Meaning Picking from Texts



LDA Topics Pie Chart for math.0406240:

Each slice represents a different topic. The size of the slice corresponds to "how much is the article about this topic?". Topics which contribute <6% to the above document are aggregated under "other".

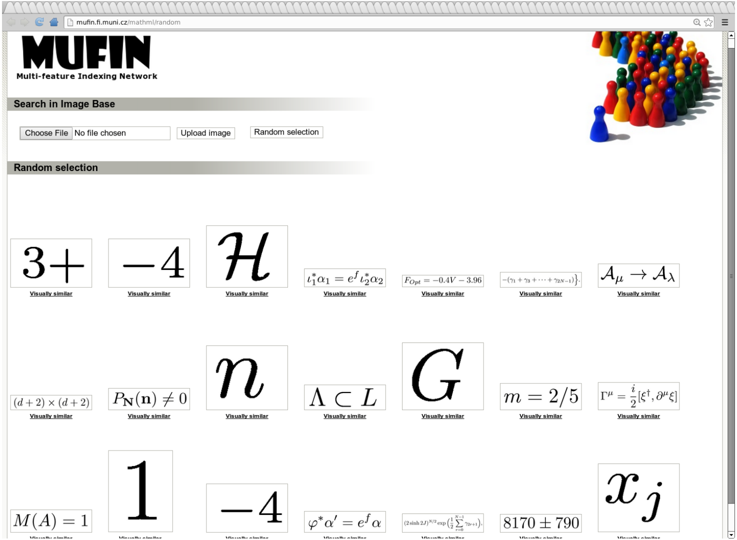
LDA topics are distributions over words; in the image, each topic is summarized by its five most probable words.

{map, bundle, holomorphic, cohomology, complex}
21.7%
{theorem, lemma, ideal, finite, hence}
10.6%
{curve, curves, points, degree, singular}
31.8%
other
9.9%
{real, integral, complex, analytic, imaginary}
8.2%
13.7%
{every, finite, theorem, sets, exists}
{polynomial, formula, polynomials, coefficients, sum}

Picking characteristic words from ‘top of iceberg’: LDA topics for given arXiv article

<https://mir.fi.muni.cz/eudmldemo/gensim-arxiv/>

Example II: metric-based indexing of math (+text)



mufin.fi.muni.cz/mathml/random

MUFIN
Multi-feature Indexing Network

Search in Image Base

Choose File | No file chosen | Upload image | Random selection

Random selection

$3+$ Visually similar
-4 Visually similar
$\mathcal{H}$ Visually similar
$\iota_1^*\alpha_1 = e^f \iota_2^*\alpha_2$ Visually similar
$F_{Opt} = -0.4V - 3.96$ Visually similar
$-(\gamma_1 + \gamma_3 + \cdots + \gamma_{2N-1})\}.$ Visually similar
$\mathcal{A}_\mu \to \mathcal{A}_\lambda$ Visually similar

$(d+2) \times (d+2)$ Visually similar
$P_{\mathbf{N}}(\mathbf{n}) \neq 0$ Visually similar
n Visually similar
$\Lambda \subset L$ Visually similar
G Visually similar
$m = 2/5$ Visually similar
$\Gamma^\mu = \frac{i}{2}[\xi^\dagger, \partial^\mu \xi]$ Visually similar

$M(A) = 1$ Visually similar
1 Visually similar
-4 Visually similar
$\varphi^* \alpha' = e^f \alpha$ Visually similar
$(2\sinh 2J)^{N/2} \exp\left(\frac{1}{2}\sum_{r=0}^{N-1} \gamma_{2r+1}\right),$ Visually similar
8170 ± 790 Visually similar
x_j Visually similar

Example III: text parsing with ParsCit

From OCR we get:

```
[5] Lambe, L., Stasheff, J.: Applications
of perturbation theory to iterated fibrations.
Manuscripta Math. 58 (1987), 363–376.
```

Parsing citations with ParsCit

```
<algorithms version=”110505”>
  <algorithm name=”ParsCit” version=”110505”>
    <citationList>
      <citation valid=”true”>
        <authors>
          <author>L Lambe</author>
          <author>J Stasheff</author>
        </authors>
        <title>Applications of perturbation theory to iterated
               fibrations.</title>
        <date>1987</date>
        <journal>Manuscripta Math.</journal>
        <volume>58</volume>
        <pages>363--376</pages>
        <marker>[5]</marker>
        <rawString>Lambe, L., Stasheff, J.: Applications of
                   perturbation theory to iterated fibrations.
                   Manuscripta Math. 58 (1987), 363-376.</rawString>
      </citation>
    </citationList>
  </algorithm>
</algorithms>
```

Word representations: Word2vec

- Tomas Mikolov from Brno came with the idea of machine learnt representation of words in high-dimensional spaces.
- The representation *capture* both syntactic and semantic properties of word usage in context.
- Only global properties are represented (not outliers).
- This continuous representation proved superior to previous discrete representation of words (of Wordnet type).

Representational learning

“Le silence eternel de ces espaces infinis m’effraie.”

“Those eternal silence of these infinite spaces terrifies me.”

Blaise Pascal, 1670

- representations in different spaces (vector ones, hyperbolic ones …)
- curse of dimensionality, projections, serialization
- ambiguity raises in low dimensions (is it practical or not?)

Learning hierarchical representations: adding structural qualitites of communicated texts

- Poincare embeddings (Nickel, Kiela, 2017) could learn latent hierarchical structural qualities (collocations, phrases, formulae trees, …).
- Trajectories of these representations (points on an *n*-dimensional Poincaré ball) could represent complex hierarchical entities ‘structural signs’: thoughts.

Table of Contents

Semio(tics)

Communication axioms [for maths]

Six blind monks and an elephant

The Mental and Semiotics Iceberg

Maths

Math scholarly communication

Novel methodologies in empirical linguistics and corpora

Higher-level representations derived from [math] corpora

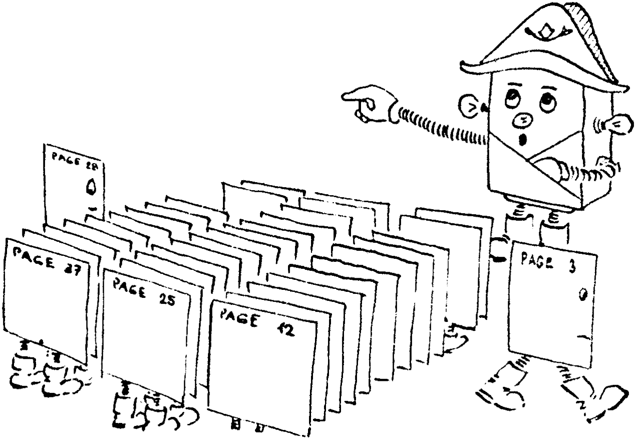
Examples: LDA, formulae indexing, citation parsing

Embeddings in vector spaces

References

Six blind monks examining an elephant on a trip

Questions?



Acknowledgements

Organizers of Semiomaths workshop for invitation.

FACULTY
OF INFORMATICS
Masaryk University

References

[1] Archambault, D., Moço, V.: Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) Computers Helping People with Special Needs, Lecture Notes in Computer Science, vol. 4061, pp. 1191–1198. Springer Berlin / Heidelberg (2006), <https://doi.org/10.1007/11788713_172>

[2] Grimm, J.: Producing MathML with Tralics. In: Sojka [4], pp. 105–117, <https://dml.cz/dmlcz/702579>

[3] MREC – Mathematical REtrieval Collection, <https://mir.fi.muni.cz/MREC/index.html>

[4] Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <https://www.fi.muni.cz/ sojka/dml-2010-program.html>

[5] Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2011), <https://www.fi.muni.cz/ sojka/dml-2011-program.html>

References (cont.)

[6] Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Rabe, F., Urban, J. (eds.) Proceedings of CICM Conference 2011 (Calculemus/MKM). Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 228–243. Springer-Verlag, Berlin, Germany (Jul 2011)

[7] Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: MathML-aware Article Conversion from LaTeX. In: Sojka, P. (ed.) Proceedings of DML 2009. pp. 109–120. Masaryk University, Grand Bend, Ontario, CA (Jul 2009), <https://dml.cz/dmlcz/702561>

[8] Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science 3, 299–307 (2010), <https://doi.org/10.1007/s11786-010-0024-7>

[9] Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [4], pp. 11–24, <https://dml.cz/dmlcz/702569>

References (cont.)

[10] Líška, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math T ask. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task. pp. 686-691. NII, Tokyo, 2013. PDF

[11] D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pp. 91–103, Aachen, 2012.

[12] Sojka, Petr and Martin Líška. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>

References (cont.)

[13] Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec.
Web Interface and Collection for Mathematical Retrieval.
In Petr Sojka and Thierry Bouche, editors, *Proceedings of DML 2011*, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University.
<https://dml.cz/dmlcz/702604>.

[14] Credits for LDA pictures goes to David M. Blei.

[15] Credits for illustrations goes to Jiří Franek.