Topic Similarity in Information Retrieval

Examples and Experience of NLP Centre and LEMMA Projects

Petr Sojka

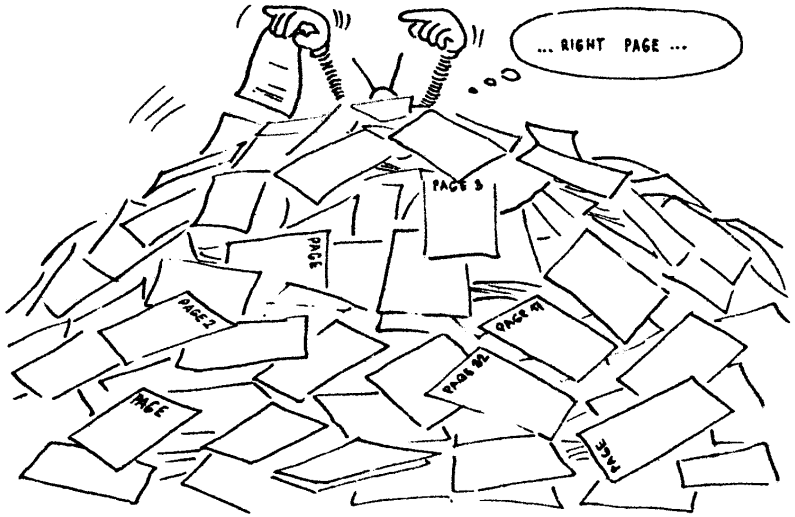
Laboratory of Electronic and Multimedia Applications[1] and
Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

sojka@fi.muni.cz

Seznam day, April 16th, 2014

[1] Donor of today's catering :-)

Coping with Information Overload by Filtering of *Big Data*



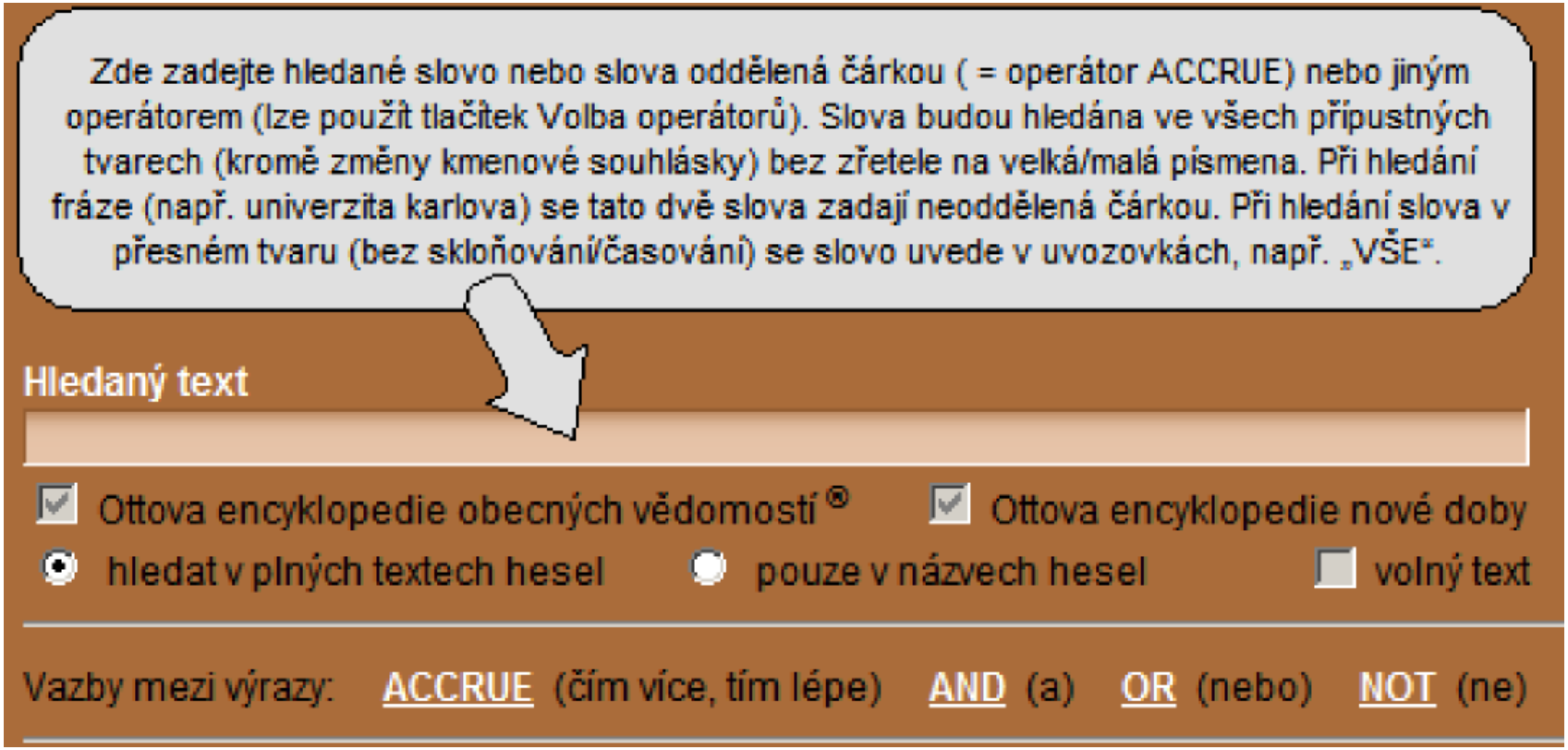
Life is searching: group similar and narrow focus of search in [your] Big Data.

Similarity types: from plagiarism (similarity on *n*-grams, narrative similarity, evolved into `http://theses.cz`) to thematic, topical similarity.

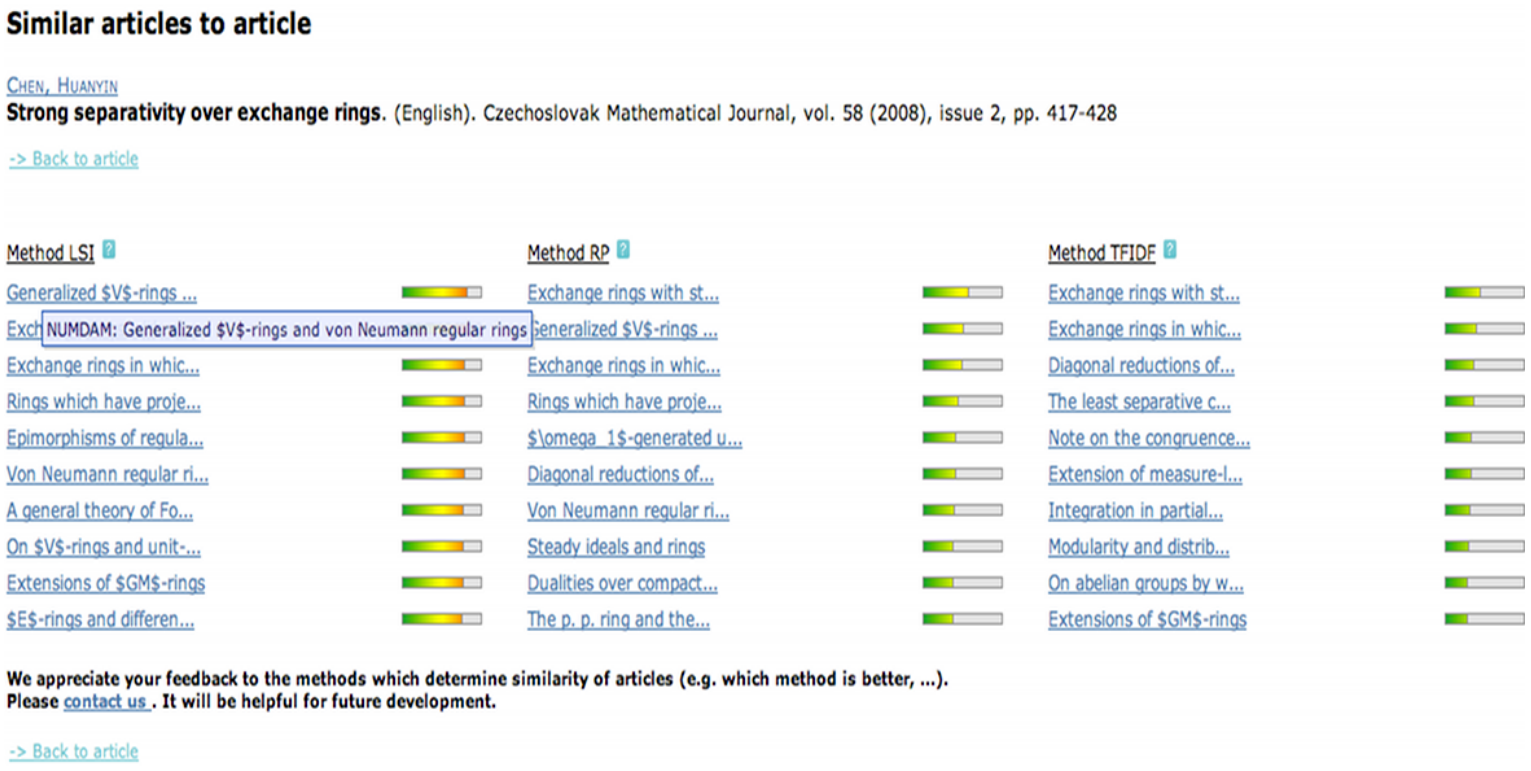
Prehistoric Example: Project Ottův Slovník naučný, 1998

Levels of content processing: strings → words and collocations → semantics (word meaning) → information (knowledge).

Grabing the essence (content) of documents: topical modelling.

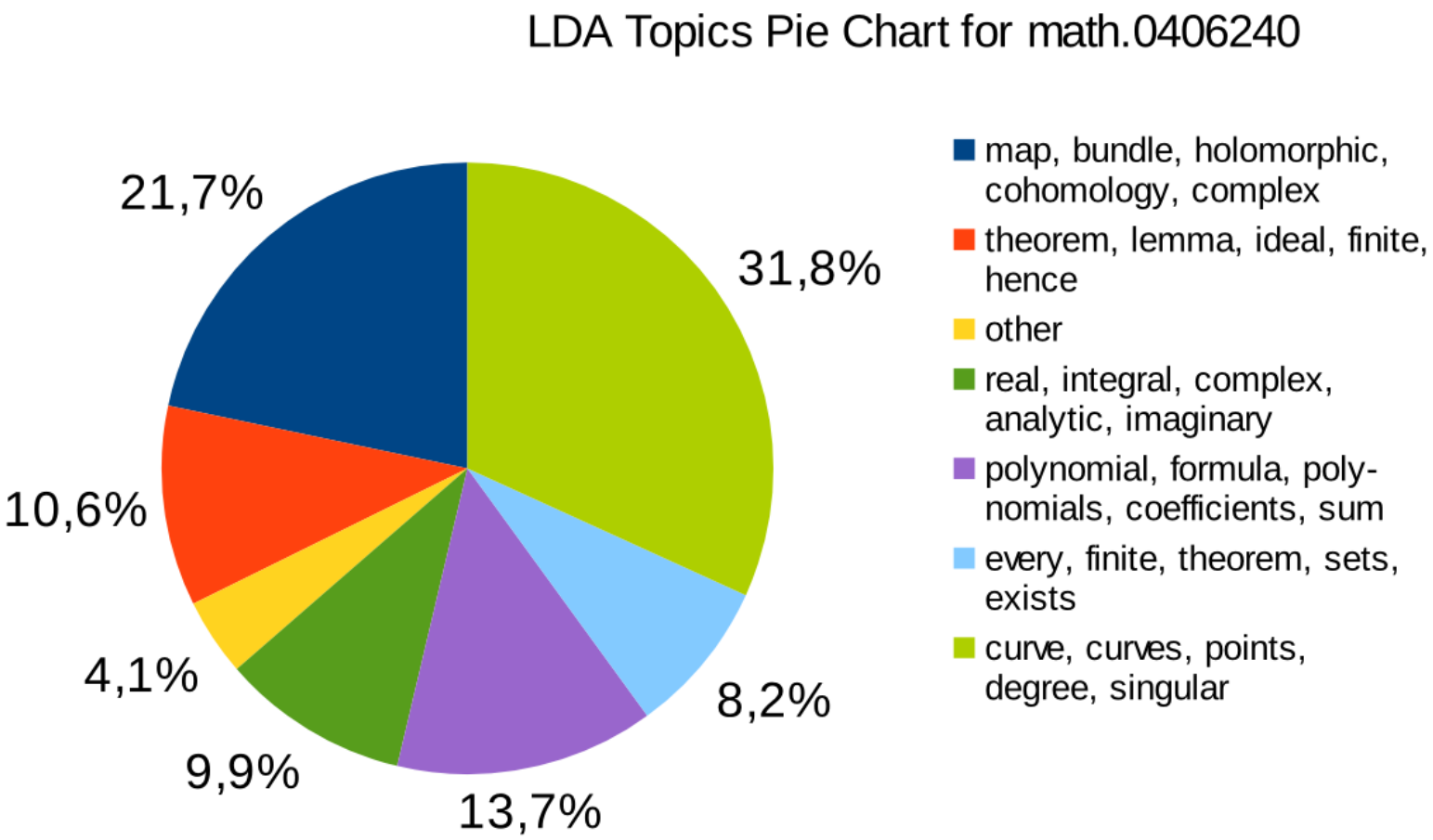


Topical Similarity in Digital Mathematics Library



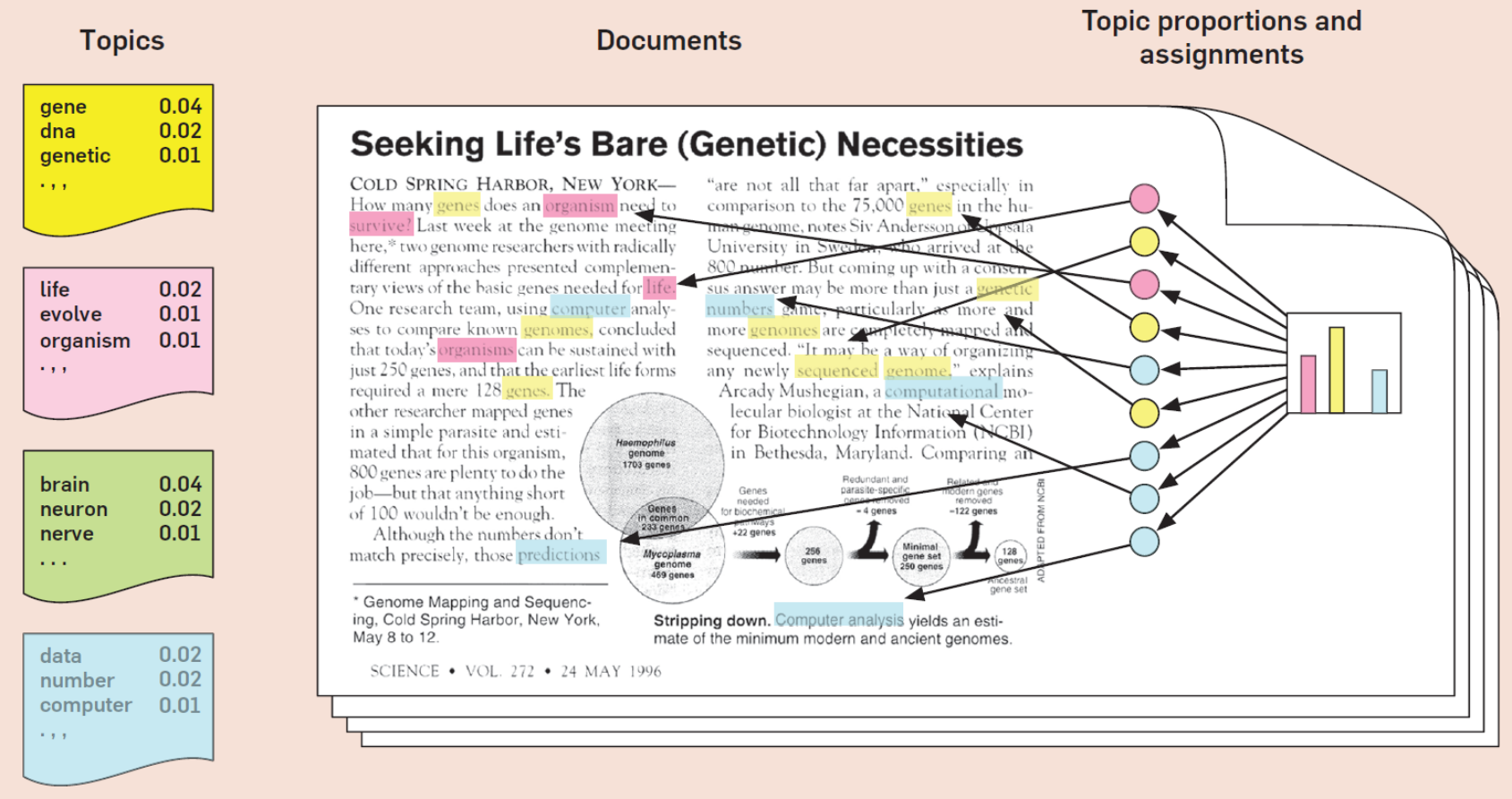
- 2005, GVP, Radim Řehůřek and Jan Pomikálek
- 2006, gensim, different machine learning methods as Random Projections, TFIDF word weighting, Latent Semantic Indexing/Analysis, Latent Dirichlet Allocation
- 50,000+ fulltexts on `http://dml.cz`

Leading Edge Example: Automated Meaning Picking from Texts



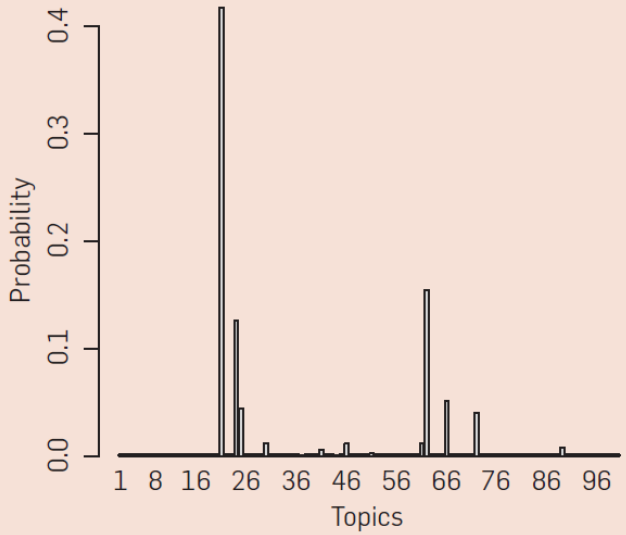
Probabilistic Topical Modelling: Latent Dirichlet Allocation

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



Topical Modelling: 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

Content Similarity Results in EuDML

Within *European Digital Mathematics Library, EuDML*, project EU CIP-ICT-PSP we have developed and delivered technology for similarity (gensim), document conversions (Braille) and accessibility (math OCR), NLP content normalization (Mathml2text).

Math Search Interface EuDML

Demo of math search in EuDML

Help About

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

How to write query

```
<math><mrow><msup><mi>x</mi> <mn>2</mn> </msup><mo>+</mo><msup><mi>y</mi> <mn>2</mn> </msup></mrow></math>
```

Canonicalized MathML query:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup> <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup> <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

Search in: MREC 2011.4.439 Search

Total hits: 36817, showing 1- 30. Searching time: 116 ms

Finite Precision Measurement Nullifies Euclid's Postulates

... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...

score = 3.2980976

arxiv.org/abs/quant-ph/0310035 - cached XHTML

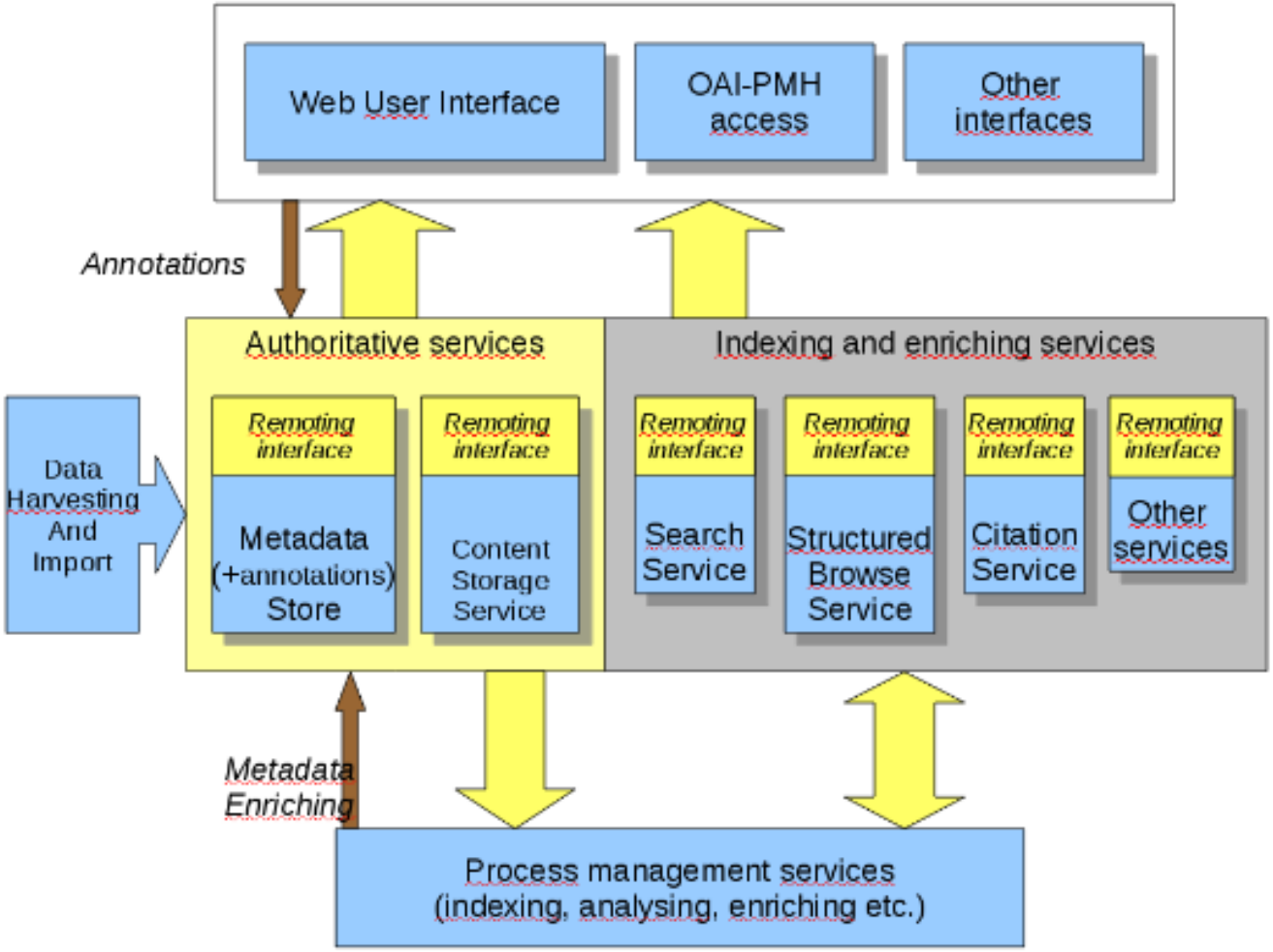
COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88

... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.

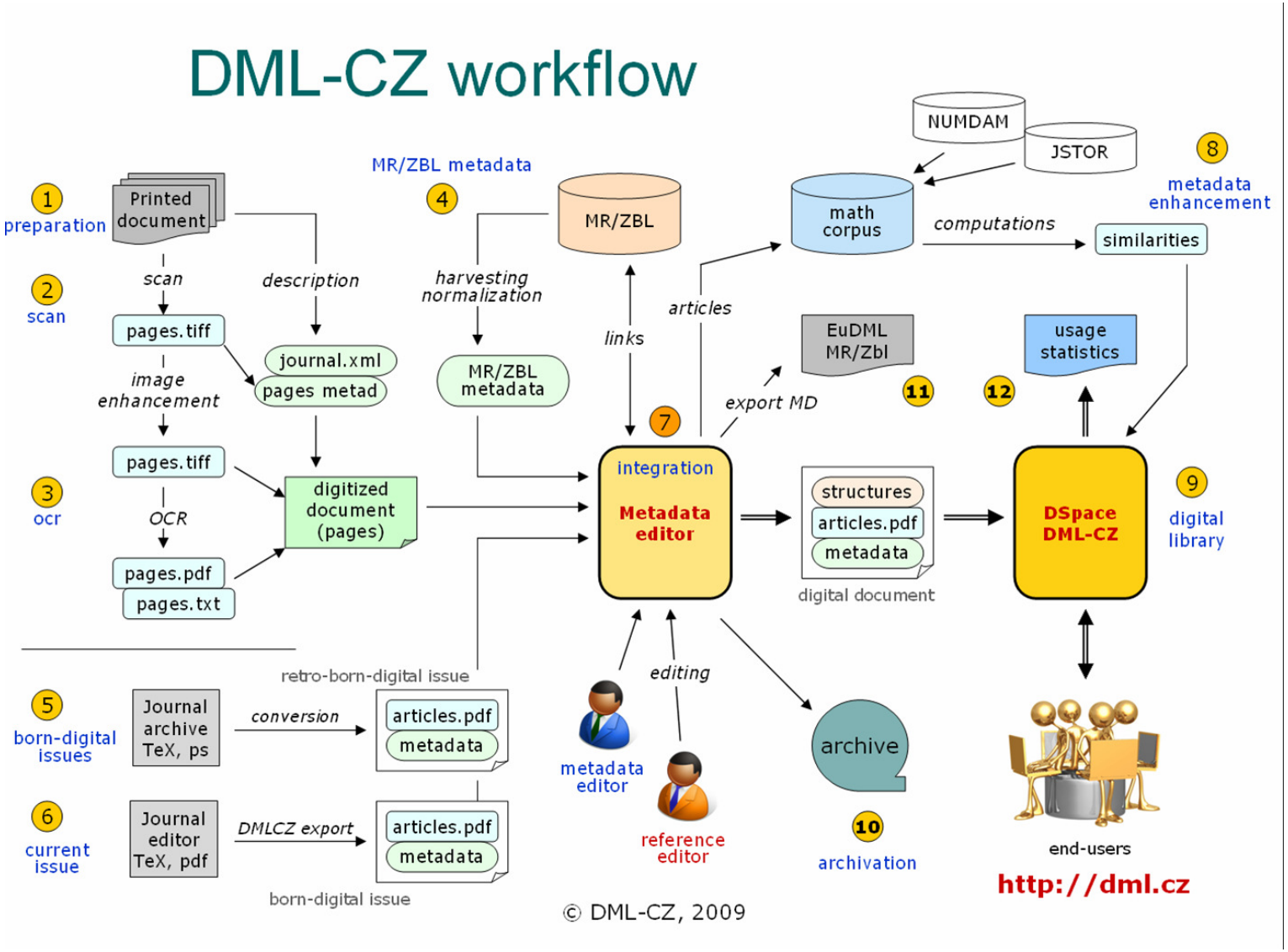
score = 1.6912912

Digital Library Service Architecture and Workflow (EuDML)

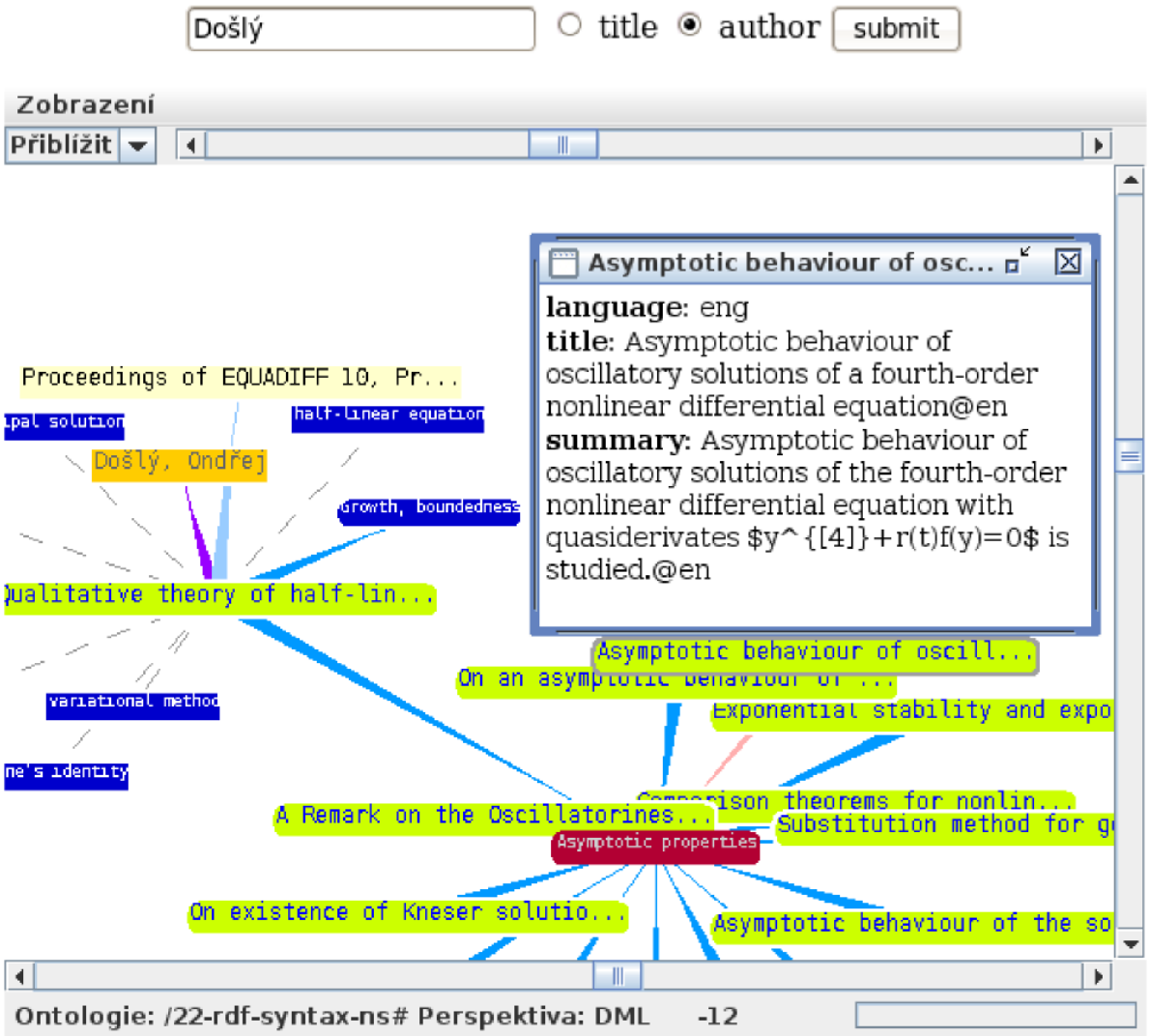
Document engineering and workflows including [Math] OCR.



Digital Library Service Architecture and Workflow (DML-CZ)



Data Visualization and Representation

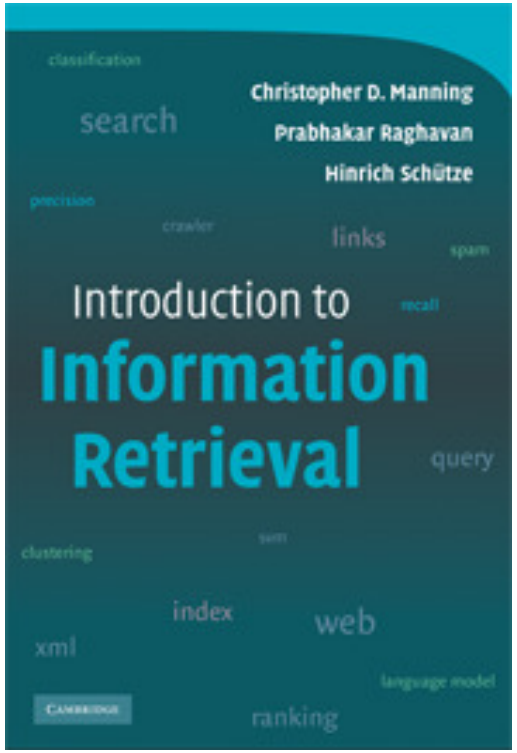


Award Winning Topic Similarity Framework gensim

- Semantic similarity indexing and search of big (continuous stream of) data. Client (search) and server (indexing) architecture.
- Developed by NLPlab PG student Radim Řehůřek (awarded in Česká hlava competition in 2011).
- Leading edge machine learning methods implemented.
- Used in 50+ local, EU or worldwide projects, 88+ citations.
- Typical deployment and fine-tuning scenario: expressing data as words (features) $\rightarrow$ configuration of topic modelling of features $\rightarrow$ setting of gensim methods and tuning parametres $\rightarrow$ usage in an application with proper vizualization interface.

Teaching Laboratory build with Constructivism Principles

- new course PV211, *Úvod do získávání informací*, *Introduction to Information Retrieval*



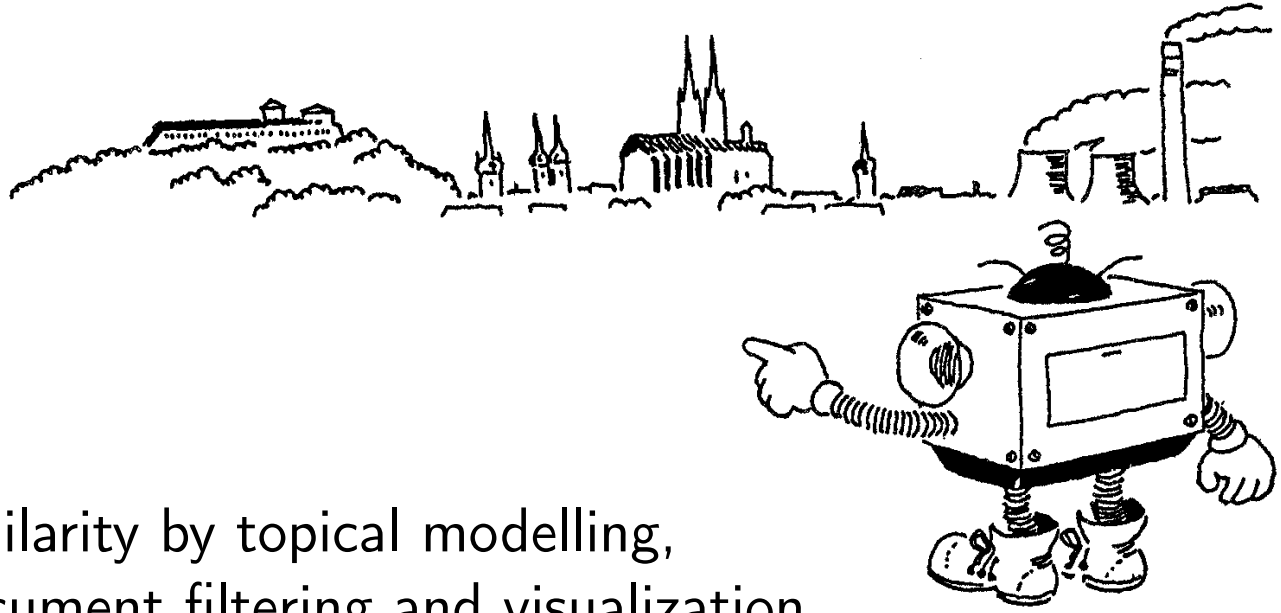
- most work done by students themselves with agile techniques, XP

New course PV211: Introduction to Information Retrieval

- In Spring 2014 for the first time: 100 students registered, 60 enrolled
- Invited lecture by Seznam (Roman Rožnik)[2]
- students motivated by Khan Academy movies, premium tasks,...
- further cooperation, continuation course?

[2] cca 10 years ago lectured in Brno Štěpán Škrob and Ivo Lukaševič flied from Prague to brainhunt our students

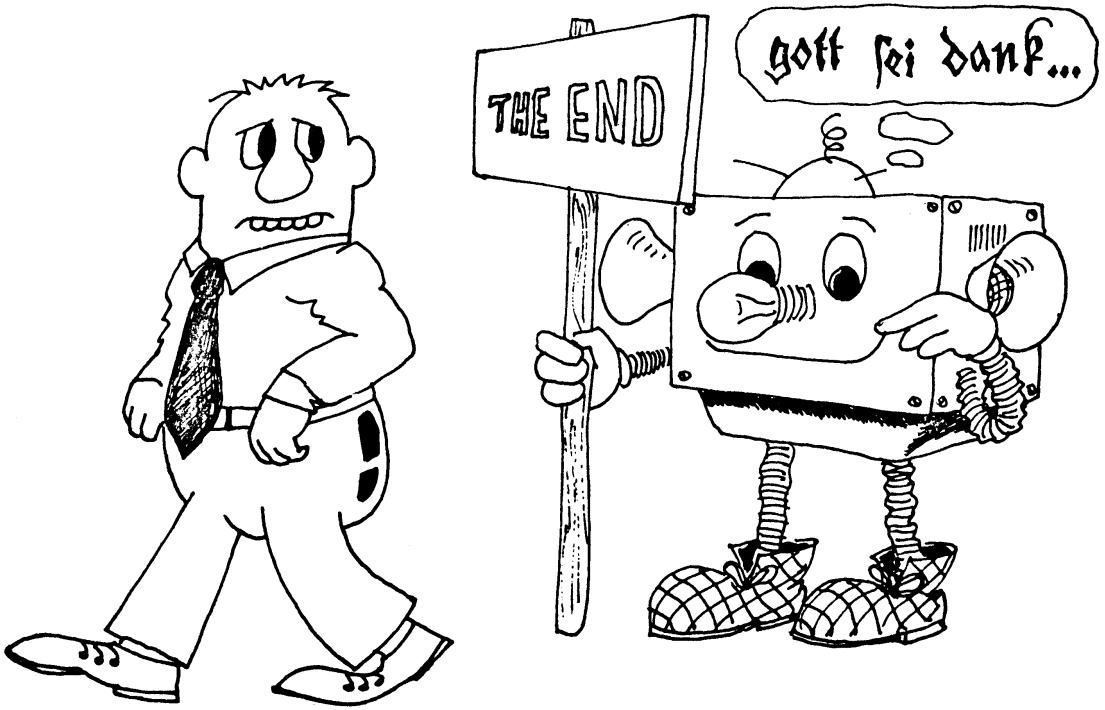
Conclusions and Mutual Research Interests



- similarity by topical modelling, document filtering and visualization
- semantic, meaning computations and modelling of natural language texts (natural NLP)
- new information retrieval course
- personal research interests: random walking for desambiguation, math (tree) indexing and similarity

That's it!

Yes, we can!



Credits: Jiří Franek (illustrations)

Links

- NLP Centre: `http://nlp.fi.muni.cz/`
- Topical modelling: `https://mir.fi.muni.cz/gensim/`
- Math Information Retrieval: `https://mir.fi.muni.cz`
- DML-CZ project: `http://dml.cz`, `http://project.dml.cz`
- EuDML project: `http://eudml.cz`, `http://project.eudml.cz`
- LEMMA: `http://www.fi.muni.cz/lemma/`