Doc. RNDr. Petr Sojka, Ph.D.

1. Obsah výzkumu

Our main research topic is representation learning of structures: structural representation of documents, mathematical formulae, human motion (gait), word/sentence/paragraph meanings. Learned representations are optimized for information retrieval, classification and clustering tasks. We use unsupervised methods (our modifications of word2vec, fasttext) for learning natural language semantics and document similarity, with supervised methods for representation normalization (math-aware information retrieval). For learning robust, universal gait features we use a modification of the Fisher's Linear Discriminant Analysis with Maximum Margin Criterion.

2. Klíčoví výzkumníci

doc. RNDr. Petr Sojka, Ph.D.
RNDr. Michal Balážia, Ph.D.
RNDr. Michal Růžička, Ph.D.

3. Klíčové metody a technologie

natural language processing, representation learning, deep learning, information retrieval, query answering, math-aware information retrieval, gait recognition from motion capture data, knowledge representation, digital libraries, vector space representation by [word] embeddings

4. Top 3 výsledky

Články:

1. Petr Sojka, Michal Růžička, and Vít Novotný. 2018. MIaS: Math-Aware Retrieval in Digital Mathematical Libraries. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18). ACM, New York, NY, USA, 1923-1926. DOI: https://doi.org/10.1145/3269206.3269233
2. Balazia M., Sojka P.: Gait Recognition from Motion Capture Data. In: ACM Transactions on Multimedia Computing (TOMM), special issue on Representation, Analysis and Recognition of 3D Humans, ACM, volume 14(1s), pp 22:1-22:18, New York, USA, February 2018. DOI: https://doi.org/10.1145/3152124
3. ŘEHŮŘEK, Radim a Petr SOJKA. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. s. 46--50, 5 s. ISBN 2-9517408-6-7.

Aplikované výsledky:

1. Topic modelling for humans: Semantic Similarity of Text Documents; based on Gensim (https://mir.fi.muni.cz/gensim/); 1000+ citations and library usage in both commerce and academia
2. MIaS (Math Indexer and Searcher) is a maths-aware full-text based search engine. It is based on the state-of-the-art system Apache Lucene, however, its maths processing capabilities are standalone and can be easily integrated into any Lucene/Solr based system, as in EuDML search service. https://mir.fi.muni.cz/mias/

3. MathML Canonicalizer is a tool for unification of different forms of MathML codding of equal formulae. It is being primary developed to meet the needs of our mathematical search engine MIaS. However, it might be useful as a general purpose tool for MathML encoding normalization. https://mir.fi.muni.cz/mathml-normalization/

5. Top 5 projektů

1. TAČR Omega project TD03000295 Inteligentní software pro sémantické hledání dokumentů https://www.muni.cz/vyzkum/projekty/33886; Cooperation with RaRe Technologies on the ScaleText system; https://scaletext.com for semantic indexing and searching
2. FP7 EU project of The European Digital Mathematical Library EuDML (https://www.eudml.org): up and running digital library with core technologies from our group
3. DocSim: Semantic Similarity of Text Documents based on Gensim (https://mir.fi.muni.cz/gensim/); 1000+ citations and library usage in both commerce and academia
4. Czech Digital Mathematics Library (https://mir.fi.muni.cz/dmlcz/); up and running with Czech Digital Mathematics Library