Network Text Analysis Using Language Models for Mapping Scientific Literature

Petra Vidnerová
Institute of Computer Science, Czech Academy of Sciences

October 17, 2025, Plzeň

Text Analysis

Text analysis is the process of automatically extracting relevant information from unstructured text.

Text analysis has evolved from manual word counting to sophisticated methods that combine statistics, linguistics, semantics, and network theory. Today, it represents an interdisciplinary tool for understanding the meaning and structure of texts in large-scale data.

Text Analysis Tasks

- **Text Classification** – assigning text to predefined categories, e.g., spam-nonspam, sentiment analysis
- **Clustering** – grouping similar texts, e.g., grouping articles by topic, user segmentation based on comments
- **Information Extraction** – identification of keywords or entities from text, e.g., names, dates, relationships between entities
- **Text Summarization** – shortening text while preserving information
- **Topic Modeling** – discovering latent topics in a collection of documents
- **Anomaly and Trend Detection** – identification of unusual patterns, new trends, e.g., fake-news detection

Classical Approaches

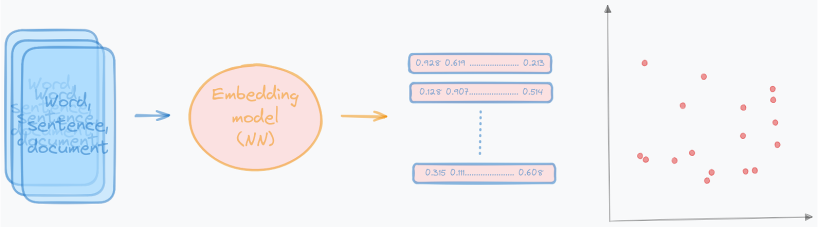
- Based on statistical methods and machine learning algorithms
- Necessity to convert text information into a numerical form (e.g., vectors of numbers)
- Classical approaches: Bag-of-Words, TF-IDF

- **Bag-of-Words** – frequency of word occurrences in a document, simple, fast, loses word order and context
- **TF-IDF** – weighted word count, considers the importance of words in the document and the corpus

- Now - embeddings, e.g., word2vec

Embeddings

- Embeddings are vectors of numbers that capture the meaning of words, sentences, or documents in a high-dimensional space
- The goal is for words with similar meanings to be close to each other in the vector space
- Vectors are trained on large text corpora using neural networks
- Distance between vectors (e.g., cosine similarity) reflects semantic similarity

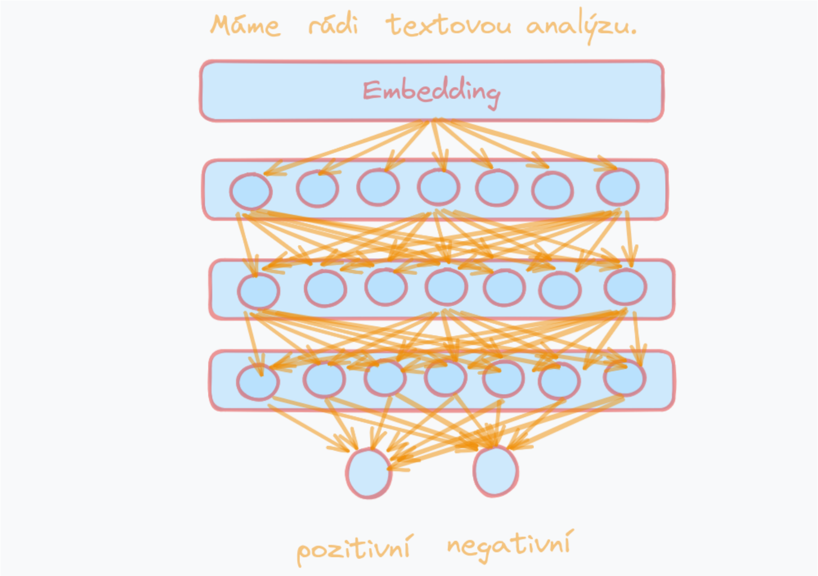


Embeddings

Advantages

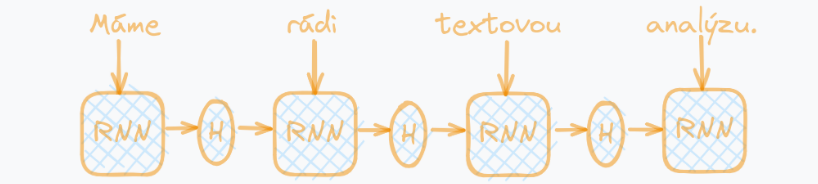
- Capture semantic meaning, not just word count
- More compact representation (lower dimension than BoW while preserving information)
- Enable measuring similarity, clustering, searching, and recommending

Neural Networks in Text Analysis



Neural Networks in Text Analysis

- **Recurrent Neural Networks (RNN, LSTM, GRU)** – work with sequences, model dependencies in text

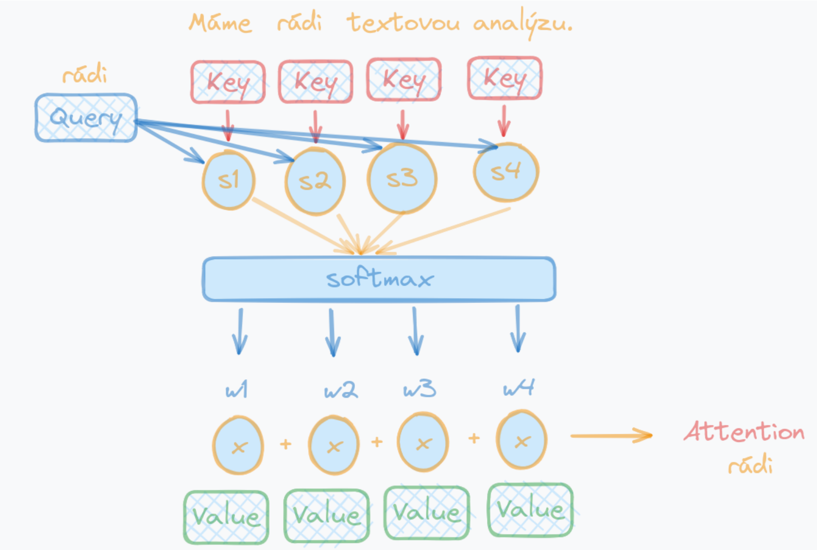


- **Transformers (BERT, GPT)** – process the entire context at once, can understand the meaning of words based on their surroundings

Attention Mechanism

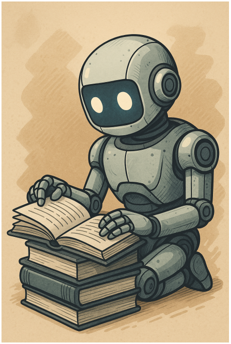
- Each word (token) looks at all other words in the sentence
- The model determines which words are important to it - assigns them higher weights
- It understands dependencies and context better than older models

Attention



Máme rádi textovou analýzu.
rádi
Query
Key
Key
Key
Key
s1
s2
s3
s4
softmax
w1
w2
w3
w4
×
+
×
+
×
+
×
Attention
rádi
Value
Value
Value
Value

(Large) Language Models (LLMs)



- Models based on neural networks
- Trained on enormous volumes of data
- Conversation with the user, text generation, translation, code generation
- Fundamental shift in computational text processing
- Applicable for a whole range of text analysis tasks

- Practical aspect: access via API or a local smaller model

Using LLMs in Text Analysis

Advantages

- Easy to use
- No data preprocessing is needed (except for trimming due to the context window)
- Mechanisms for generating structured output (for later machine processing)

Using LLMs in Text Analysis

Disadvantages

- Sensitive to query formulation (prompt), need to specify the task precisely
- Prejudices and biases contained in the training data are also reflected in the model's answers (gender and racial stereotypes)
- Typically slower response (several seconds) than classical methods
- Limited context window (especially for smaller models)
- Ethical questions, leakage of sensitive data when using via API

LLM Sensitivity to Prompt Formulation - Experiment

- LLMs to some extent model human behavior, including stereotypes, biases, etc.
- Different query framing can lead to a different result
- Experiment: The LLM receives information about a new drug and patient characteristics. We ask if the patient will use the drug

- Necessity to formulate queries precisely and neutrally

Network Text Analysis

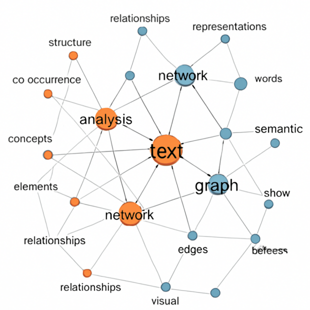
- Connects text analysis and network (graph) theory
- Graph $G = (V, E)$, where V is the set of vertices (nodes) and E is the set of edges (connections between vertices)
- Models relationships between documents, concepts, words, or topics

Goals

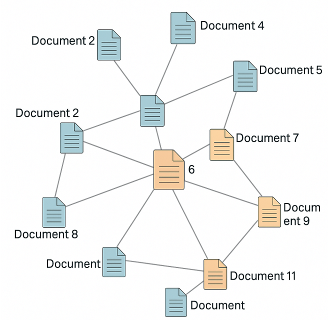
- Discover hidden semantic structures in texts
- Visualize relationships and communities
- Quantify the influence or centrality of key documents or concepts

Network Text Analysis

- Nodes are keywords, concepts
- Edges are relationships between them, e.g., co-occurrence in the document



- Nodes are documents
- Edges are relationships between documents, e.g., common topic, citation



Example - Text Analysis of Scientific Articles

Assignment

- Map the area of Neural Architecture Search (NAS)
- NAS is a relatively narrow subfield of automated machine learning, but in the center great interest
- Deals with the automatic selection of neural network architecture

Approach

- Application of text analysis and network text analysis
- Articles available on ArXiv along with metadata from the OpenAlex database
- In the initial study, only abstracts of the downloaded articles were used

Data Sources

arXiv

- Allows access via API
- Automatic download of metadata, abstracts, and full texts via the application interface
- Allows searching, e.g., by topic
- We downloaded a large number of articles on the topic and performed filtering using LLM $\rightarrow$ a set of approx. 2,500 articles

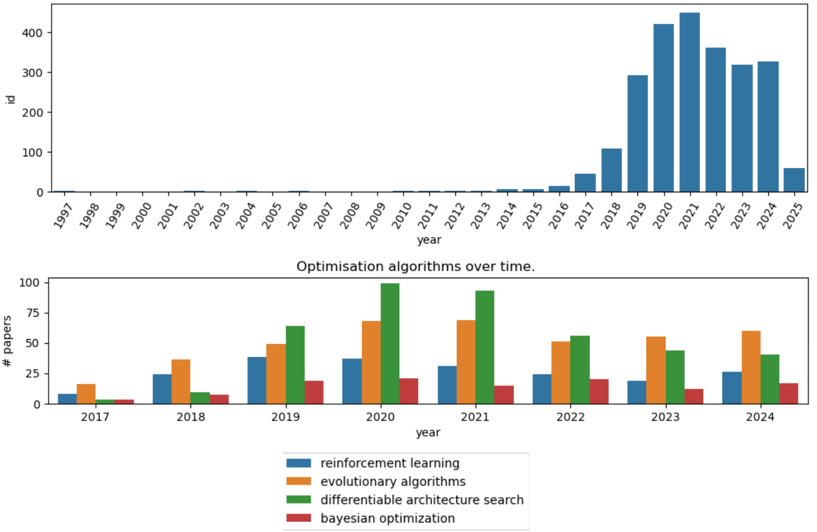
OpenAlex

- Huge database of scientific articles, www.openalex.org, provides access via API
- Source of information about citations and references

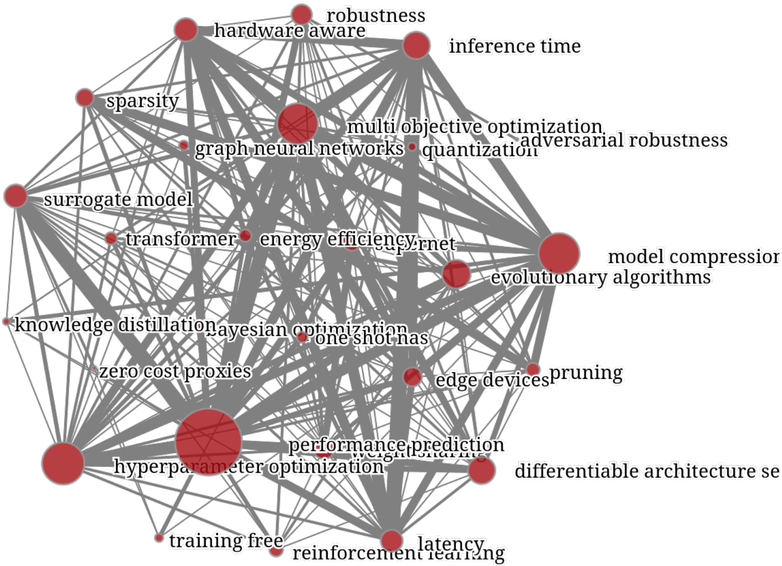
Keyword Extraction

- We assign keywords to each document
- Determining keywords using LLM
- Approach 1: instruction for LLM - select up to 5 keywords
- Approach 2: a pre-selected fixed set of keywords based on knowledge of the topic, we query for each keyword whether it relates to the document
- Combination of both: we obtain a set of keywords for the whole topic, the most frequent keywords are used in Approach 2
- For each document, we have a vector of boolean values of keyword membership

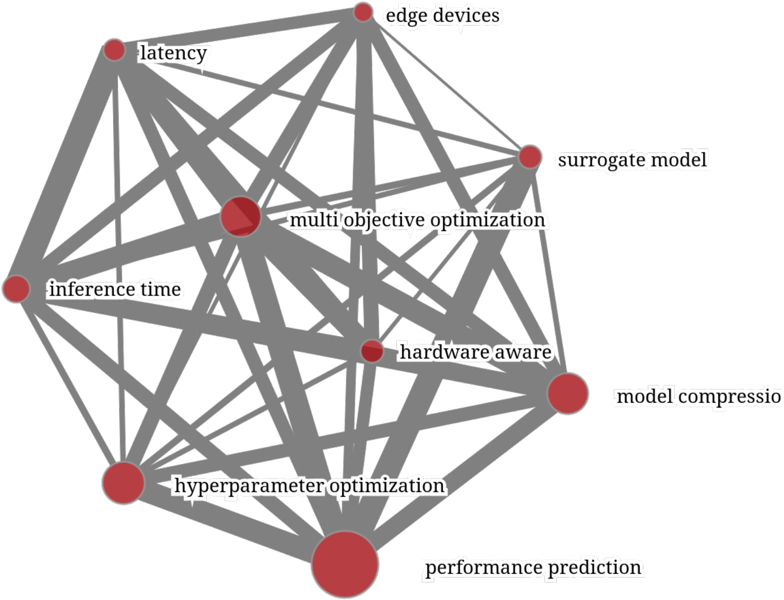
Evolution over Time



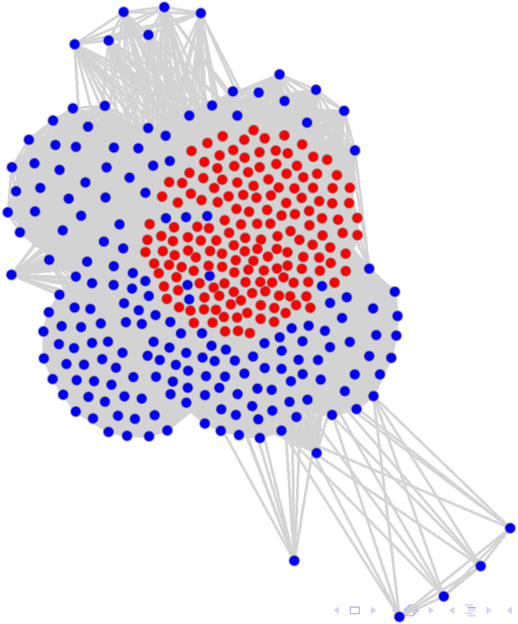
Relationships between Keywords



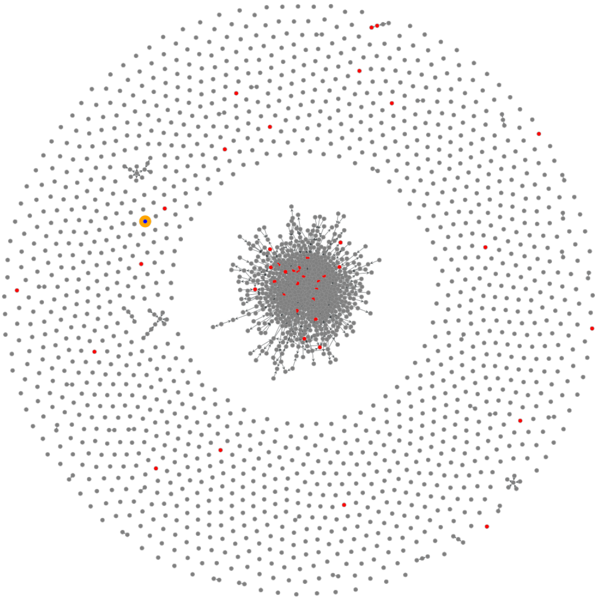
Relationships between Keywords



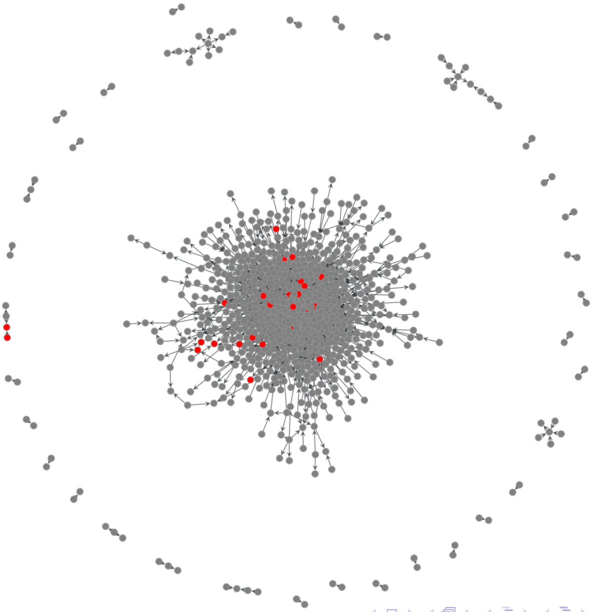
Relationships between Documents



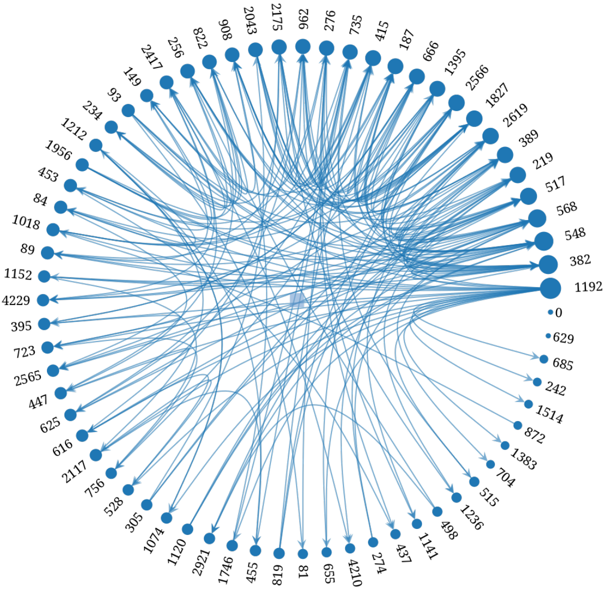
Citation Networks



Citation Networks



Citation Networks



Using Citation Networks

- The network carries more information than just the number of citations and references (one node vs. whole network)
- Node properties in the network can be utilized – centralities

- Pagerank - not all citations are equal, a more important citation is from a significant article than a marginal one

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M_i} PR(j)/L(j)$$

N	total number of articles
d	damping factor
M_i	set of articles that cite i
L_j	set of references of article j

Using Citation Networks

Betweenness centrality

- How many times a node lies on the shortest paths between nodes in the network
- Reflects the importance of the node as a mediator of information flow between parts of the network
- Identification of articles connecting different research directions

- High value – the article connects different directions, a review article that summarizes different approaches, an interdisciplinary article
- Low value - the article is part of a closed cluster where there is high mutual citation

Novelty Detection

- Novelty, breakthrough nature of the article
- Attempts to detect breakthrough articles

Metric - Disruption index

- Again carries more information than just the citation counts
- Disruption index
    - only article i is cited – it replaced older works, a disruptive article
    - both i and its predecessors are cited – rather a developing article
    - only predecessors of i are cited – the article does not have a large impact

$$DI = \frac{N_i - N_j}{N_i + N_j + N_k}$$

N_i citations of the article, N_j citations of both the article and its sources,

N_k citations of sources

Conclusion

Future Research Directions

- Opportunities are opening up for research into novelty detection by analyzing full texts
- The question of whether novelty or disruptiveness can be measured immediately, without waiting for citations and future impact of the article
- Citation classification – replacing mere citation counts with information about why the article cites

Thank you for your attention.

This work has been funded by a grant from the Programme Johannes Amos Comenius under the Ministry of Education, Youth and Sports of the Czech Republic, CZ.02.01.01/00/23_025/0008711.

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY