Unsupervised machine learning approach for hierarchical graph-based representation of natural language text collections

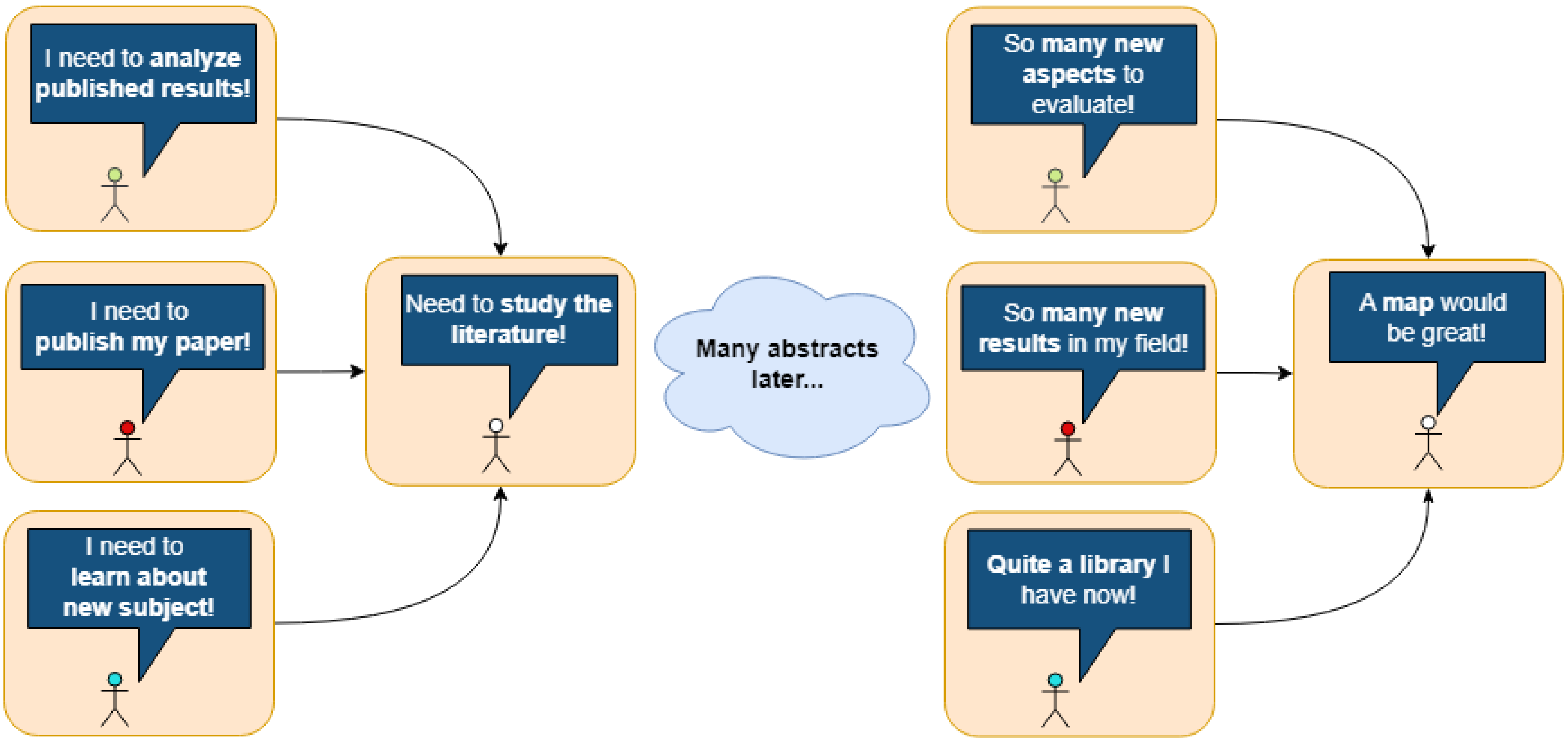
Jevgenijs Bodrenko [1]

[1] Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
eugene.bodrenko@gmail.com

Outline

- Research motivation
- Research problem & objectives
- Research questions & methodology
- Methods
- Results & Conclusions

Motivation


I need to analyze published results!
I need to publish my paper!
I need to learn about new subject!
Need to study the literature!
Many abstracts later...
So many new aspects to evaluate!
So many new results in my field!
Quite a library I have now!
A map would be great!

- **Research problem**
  - **Shortage of visual open-source tools** for analysis of document collections, with a focus on document similarities across topical hierarchy, **without the computational demands typical for Large Language Models** (LLMs).

- **Research objectives**
  1. Implement a machine learning pipeline to detect document similarities based on a topical hierarchy.
  2. Develop visualization approaches for analyzing document collection structures and highlighting document similarity.
  3. Optimize the solution for efficiency to reduce computational resource requirements compared to LLM fine-tuning.

- **Research aim**

  To develop a machine learning component of a visual open-source tool for analyzing structures of document collection with focus on document similarities across topical hierarchy, and minimization of computational resources.

- **Research object**
  - Hierarchical semantic structures in collections of English language texts.

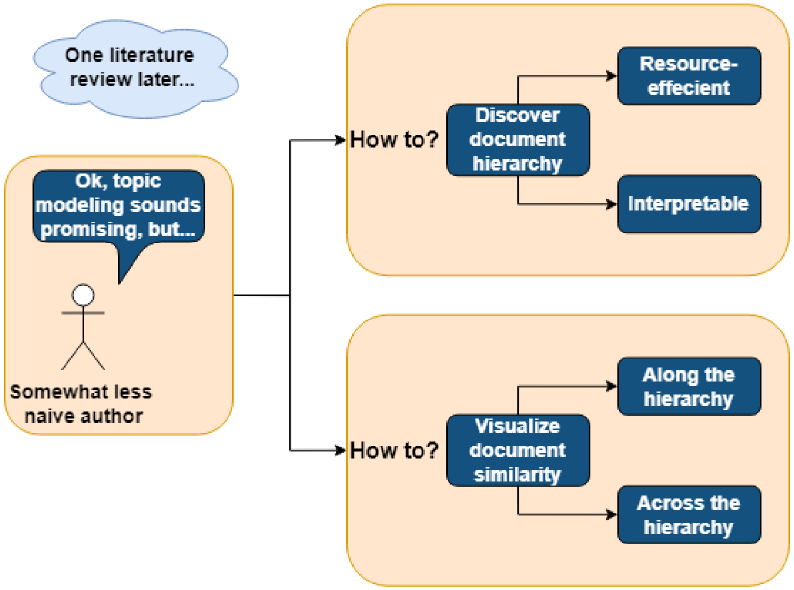
- **Research subject**
  - Topic modeling methods for detecting and visualizing hierarchical semantic stucture in collections of scientific English language texts.

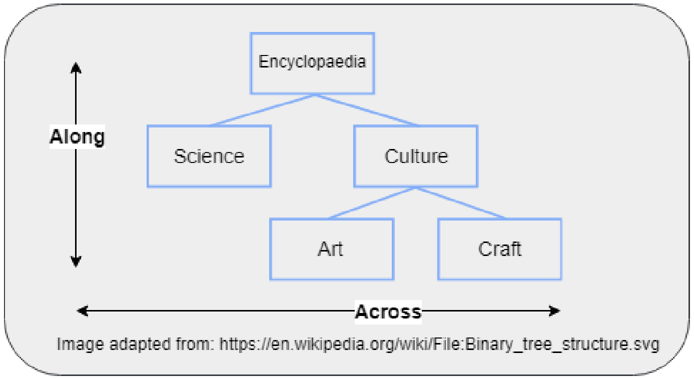
Research questions

1. How to discover the topic hierarchy in a collection of English texts using unsupervised machine learning methods, given that it was discovered by a human?

2. How can the output be visualized and used to explore the hierarchy structure and document similarity along and across the topic hierarchy?

Research design



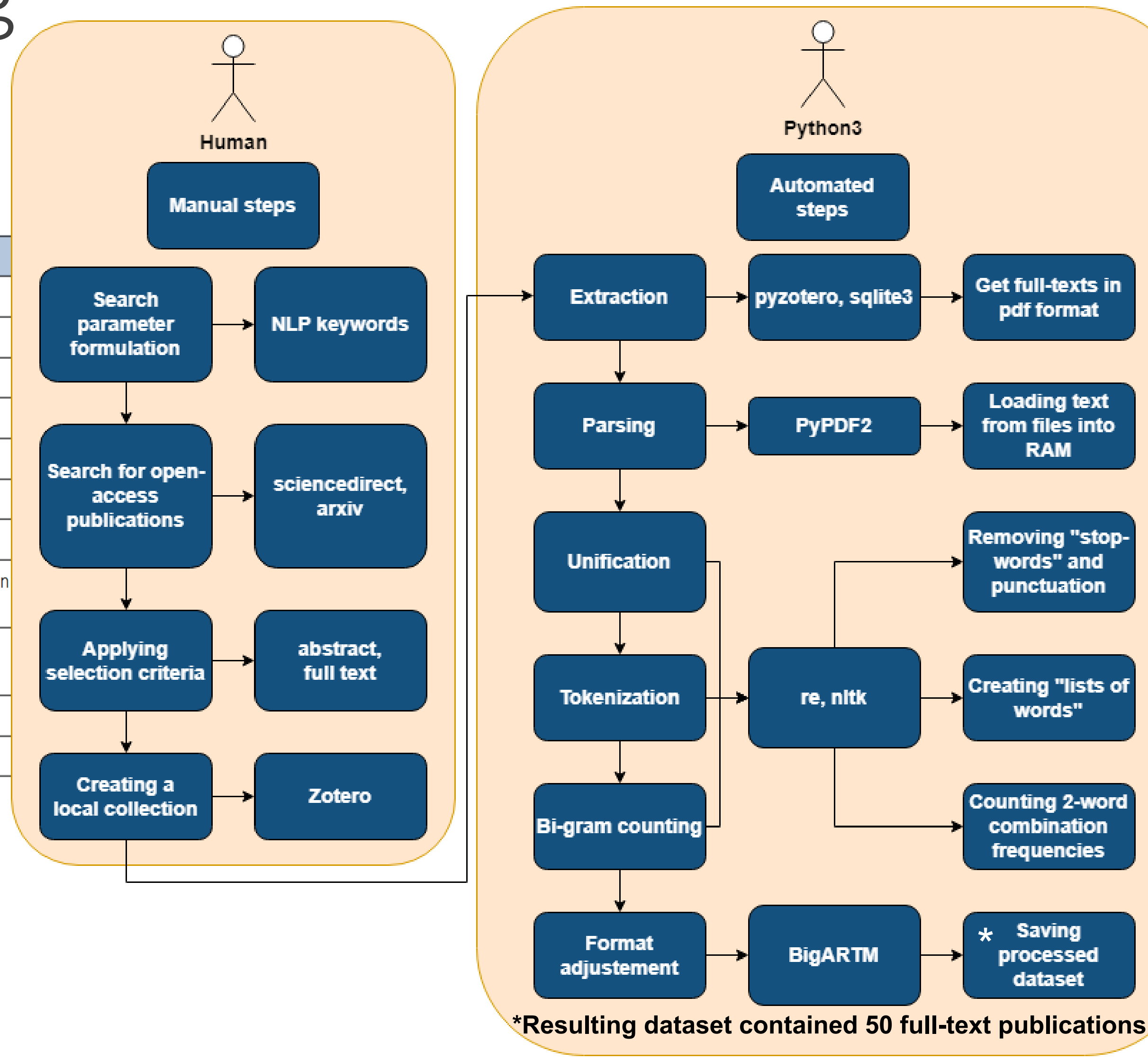


- **Approach:** Natural Language Processing and Machine Learning.
- **Data:** Full texts of scientific publications in English.
- **Pipeline:** Hierarchical Additive Regularization Topic Model augmented by Spectral Clustering for clustering based on document similarity.
- **Visualizations:** Sankey plot and Multidimensional Scaling to evaluate document similarity at each hierarchy level and across all levels simultaneously.
- **Efficiency:** Unsupervised topic modeling approach allows to reduce computational resource requirements compared to Large Language Model fine-tuning.
- **Applications:** Potential use in academia and industry for self-study, reviews, and analyses.

Dataset preparation and preprocessing

Table 1: Terms used to construct queries for publication search

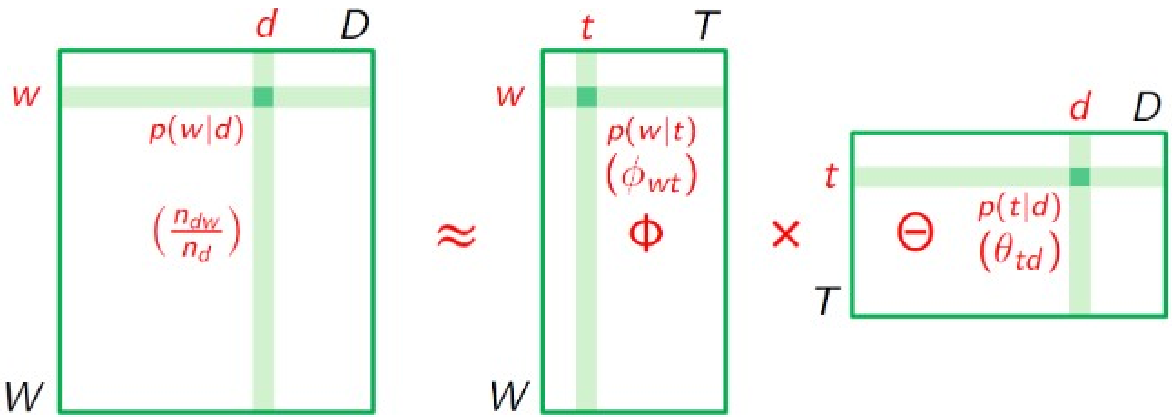
Search term	Synonyms		
Unsupervised	Without labels	Independent	Self-guided
NLP	Text mining		
Machine learning	ML		
Pipeline	Workflow		
Semantic	Meaning-based	Contextual	Context-based
Feature engineering	Feature extraction	Attribute engineering	Feature creation
Clustering	Grouping	Categorization	Partitioning
Community detection	Subgraph identification	Group discovery	Subnetwork detection
Graph representation	Text graph embedding	Text network representation	Document-graph embedding
Embedding	Encoding	Vectorization	
Document	Text		



Methods: AR*-based topic modeling

*additive-regularization

(1)



(2)

Input: document collection D, number of topics $|T|$;
Output: Φ, Θ;

initialize vectors ϕ_t, θ_d randomly;
repeat
  zeroize n_{wt}, n_{td}, n_t, n_d for all $d \in D$, $w \in W$, $t \in T$;
  forall $d \in D$, $w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    forall $t \in T$: $\phi_{wt}\theta_{td} > 0$ **do**
      increase n_{wt}, n_{td}, n_t, n_d by $\delta = n_{dw}\phi_{wt}\theta_{td}/Z$;
  $\phi_{wt} := n_{wt}/n_t$ for all $w \in W$, $t \in T$;
  $\theta_{td} := n_{td}/n_d$ for all $d \in D$, $t \in T$;
until Φ and Θ converge;

(3)

$$\sum_{d \in D} \sum_{w \in d} n_{td} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \; + \; R(\Phi, \; \Theta) \; \to \; \max(\Phi, \; \Theta);$$

$$R(\Phi, \; \Theta) = \sum_{1}^{k} \tau_i R_i(\Phi, \; \Theta)$$

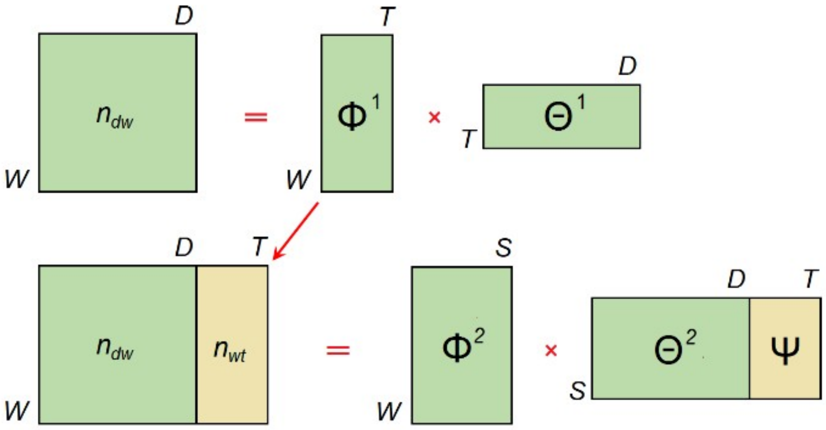
$$\sum_{w \in W} \phi_{wt} = 1; \;\; \phi_{wt} \; \geqslant \; 0; \sum_{t \in T} \theta_{td} = 1; \;\; \theta_{td} \; \geqslant \; 0;$$

Images adapted from:
Vorontsov, K. and Potapenko, A. (2015) ‘Additive regularization of topic models’, Machine Learning, 101(1), pp. 303–323. Available at: https://doi.org/10.1007/s10994-014-5476-6.

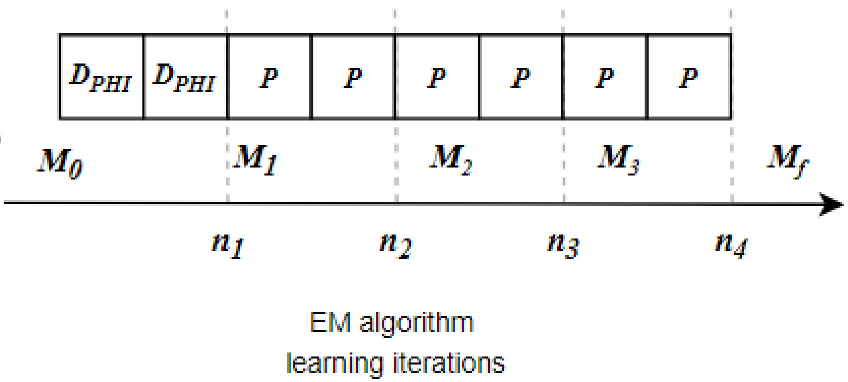
Methods: hARTM*

*Hierarchical additive-regularization topic model

(4)



(5)



(6)

$$R(\Phi) = -\gamma \sum_{t\in T} \sum_{s\in T\setminus t} \sum_{w\in W} \phi_{ws}\phi_{wt} \to max(\Phi)$$

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D} \sum_{t\in T} \alpha_t \ln \theta_{td} \to max(\Phi,\ \Theta)$$

(7)

$$PMI(u,v) = \ln \frac{|D|N_{uv}}{N_u N_v}$$

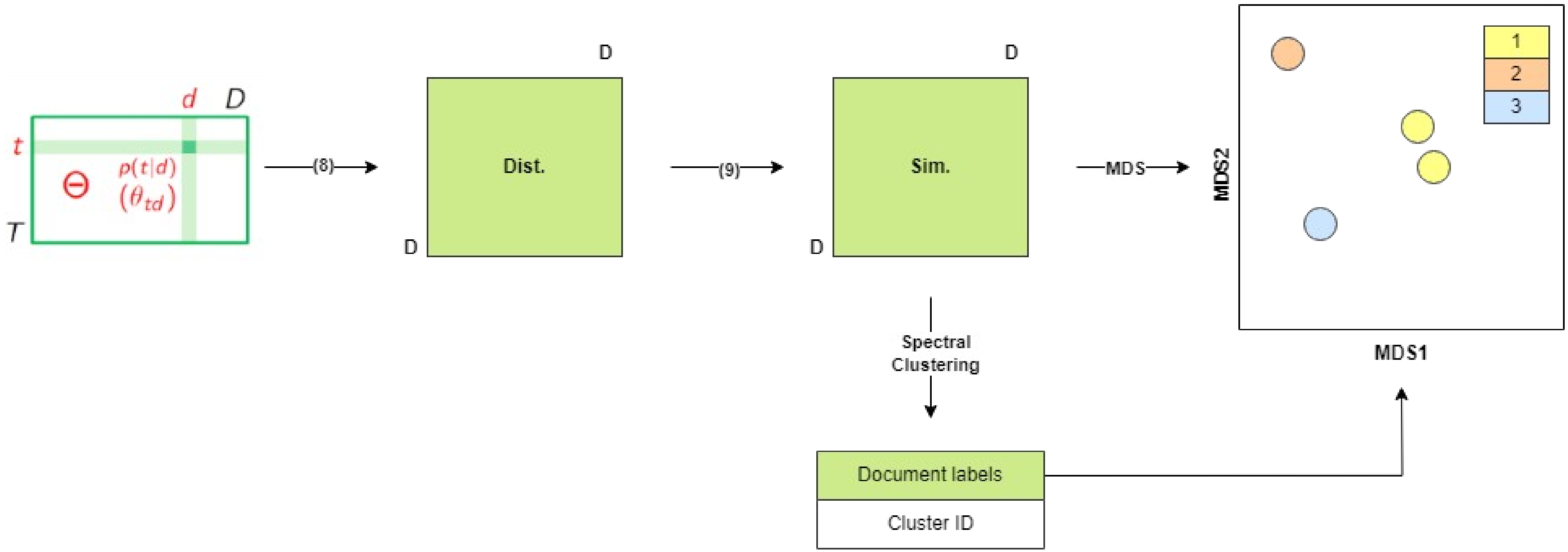
$$C_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=1}^{k} PMI(w_i, w_j)$$

$$W_t = \left\{ w \in W | \phi_{wt} > \frac{1}{|W|} \right\}$$

$$T_d = \left\{ t \in T | \theta_{td} > \frac{1}{|T|} \right\}$$

Images adapted from:
(4,6,7) Chirkova, N.A., JSC Antiplagiat, and Lomonosov Moscow State University (2016) ‘Additive Regularization for Hierarchical Multimodal Topic Modeling’, Machine Learning and Data Analysis, 2(2), pp. 187–200. Available at: https://doi.org/10.21469/22233792.2.2.05.
(5) Khodorchenko, M. et al. (2020) ‘Optimization of Learning Strategies for ARTM-Based Topic Models’, in. Available at: https://doi.org/10.1007/978-3-030-61705-9_24.

Methods: Cross-sectional view

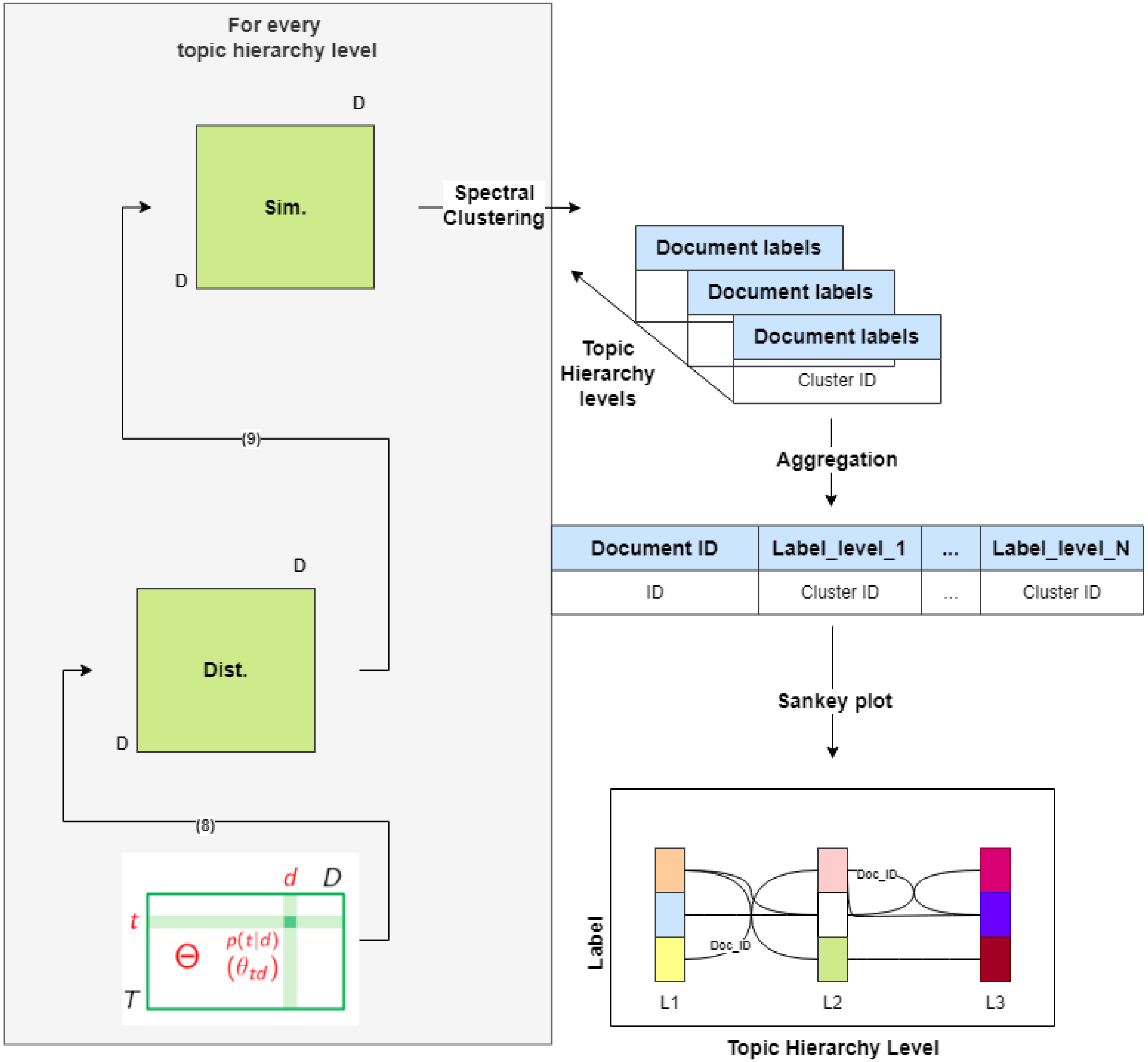


(8) $$H(A,B) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{a_{ij}} - \sqrt{b_{ij}}\right)^2}$$

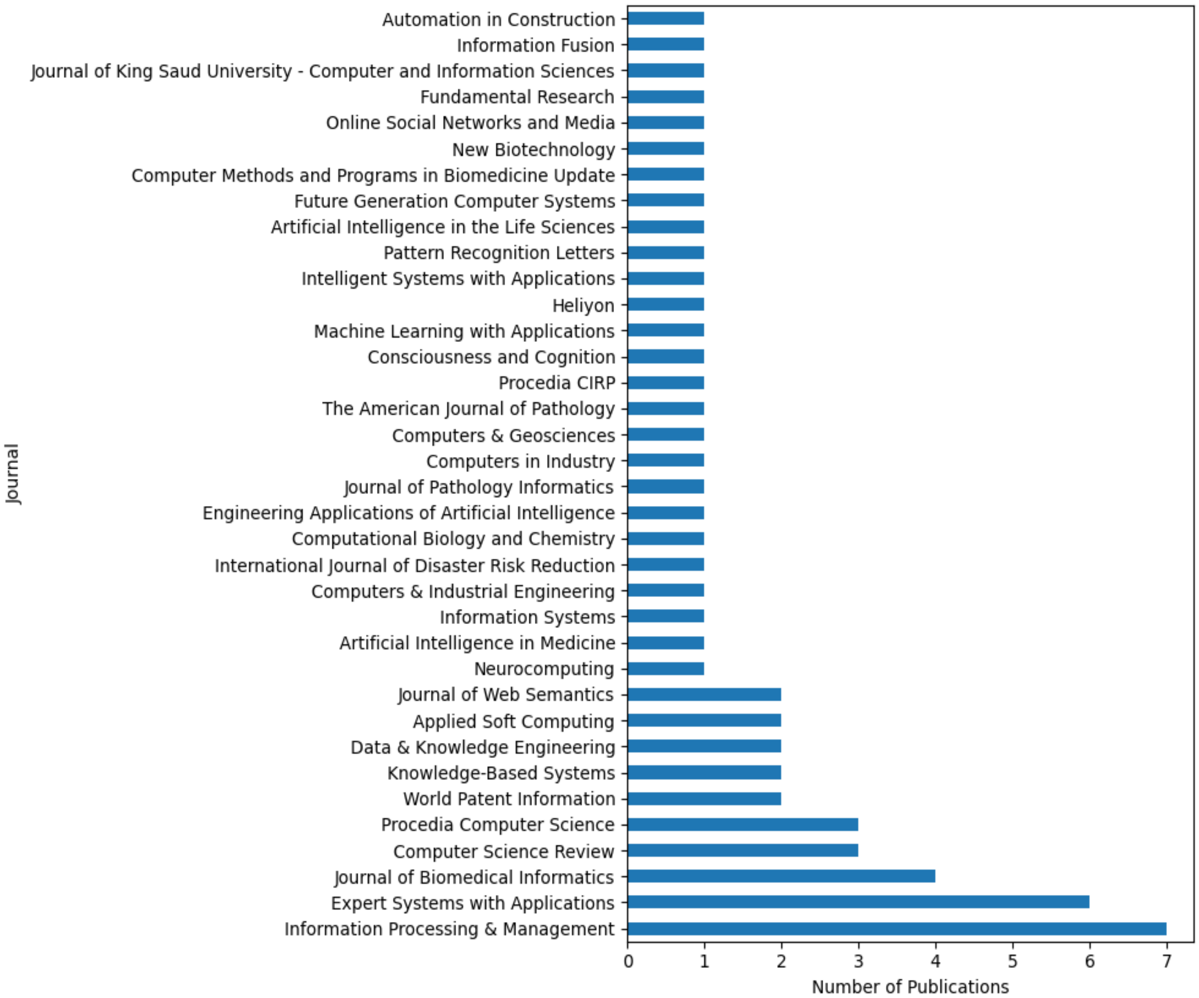
(9) $$B(A,B) = 1 - H(A,B)^2$$

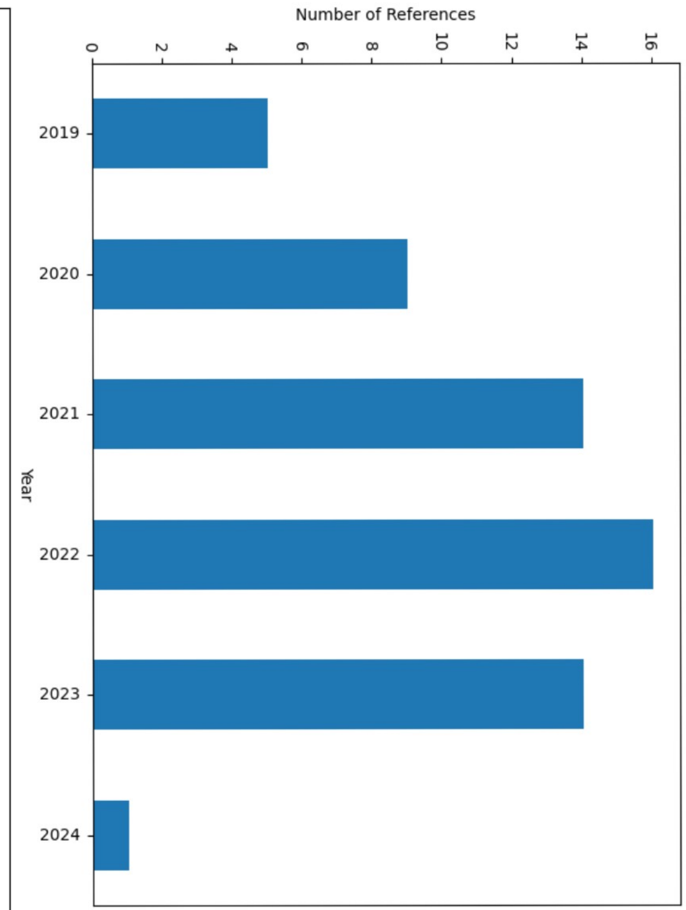
Images adapted from:
(8) Chirkova, N.A., JSC Antiplagiat, and Lomonosov Moscow State University (2016) 'Additive Regularization for Hierarchical Multimodal Topic Modeling', Machine Learning and Data Analysis, 2(2), pp. 187–200. Available at: https://doi.org/10.21469/22233792.2.2.05.
(9) Kitsos, C.P. and Nisiotis, C.-S. (2022) 'Considering distance measures in Statistics', Biometrical Letters, 59(1), pp. 65–75. Available at: https://doi.org/10.2478/bile-2022-0006.

Methods: Longitudinal view

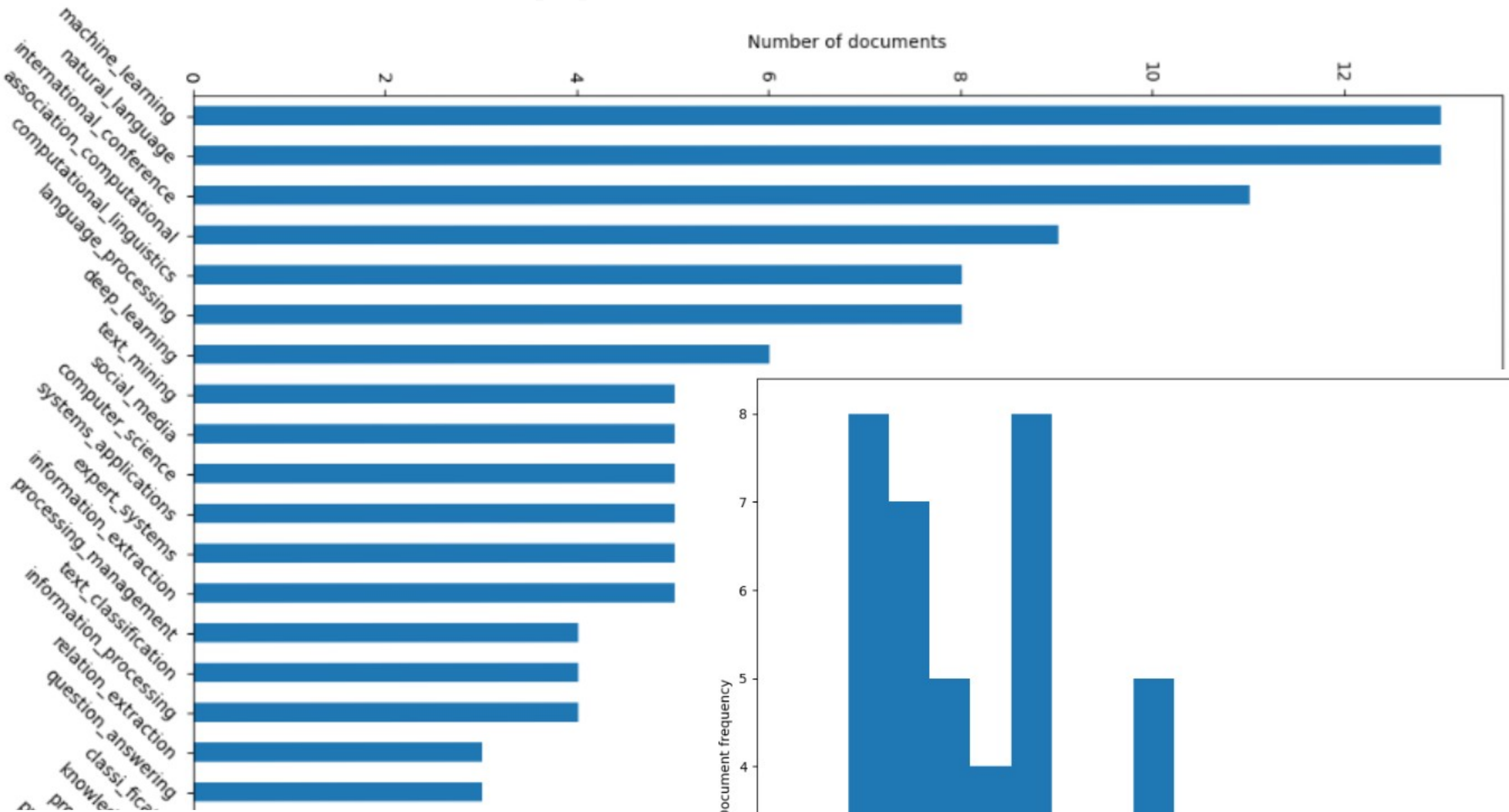


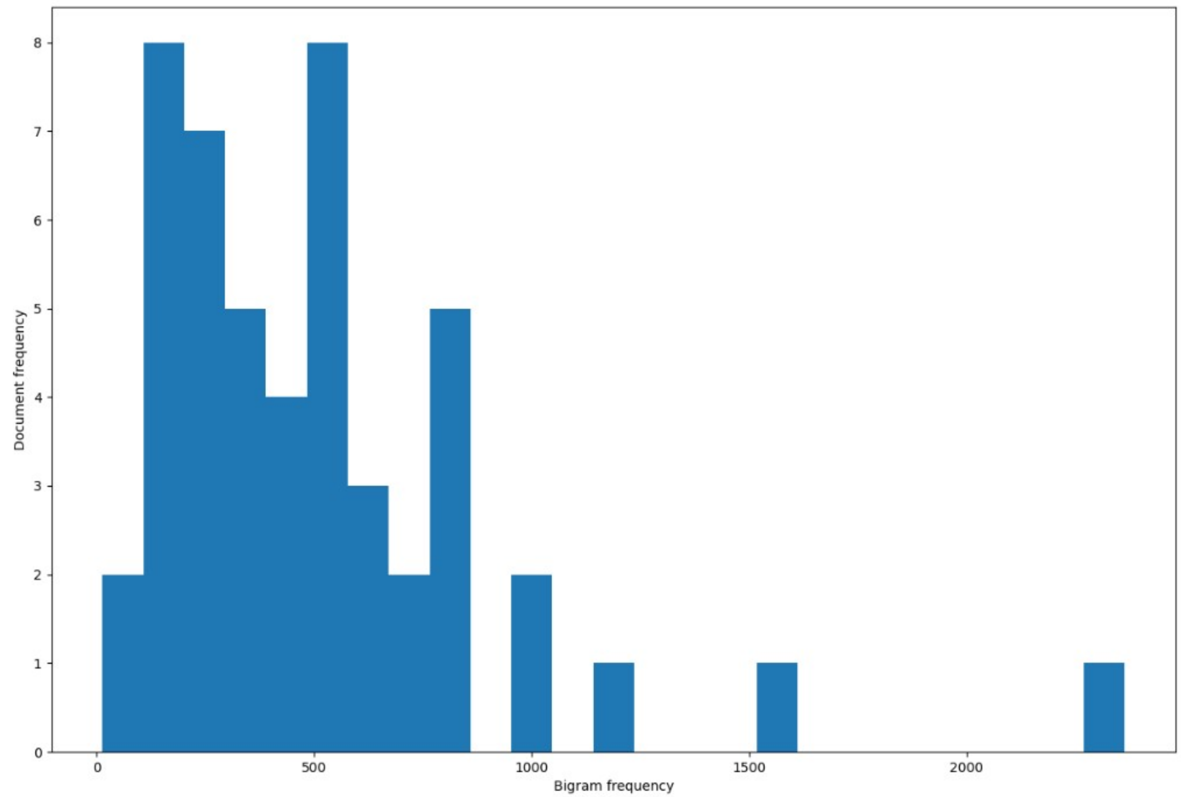
Results: Dataset metadata





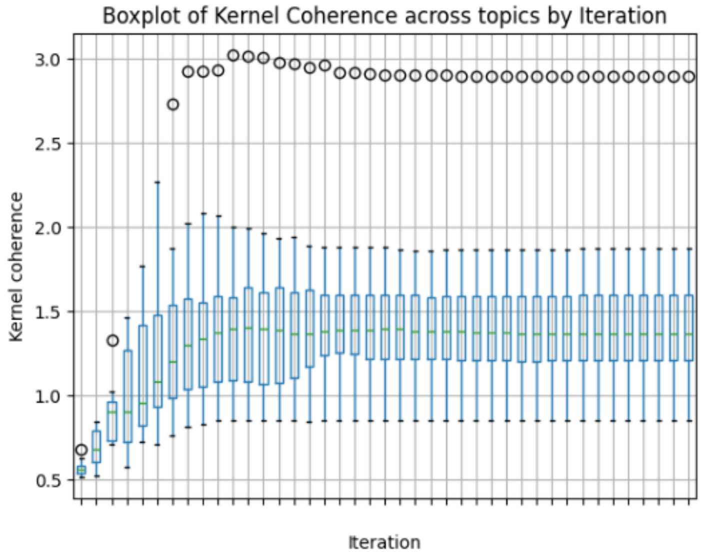
Results: Bigram-based dataset view



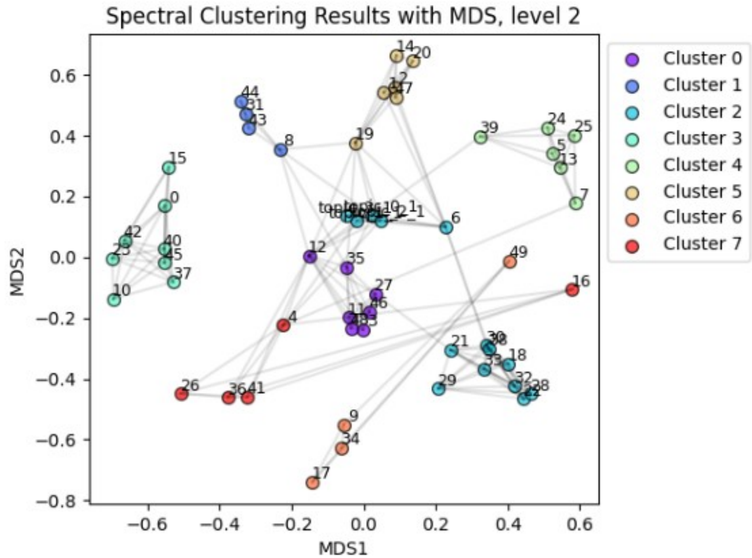


Results: Topic Modeling

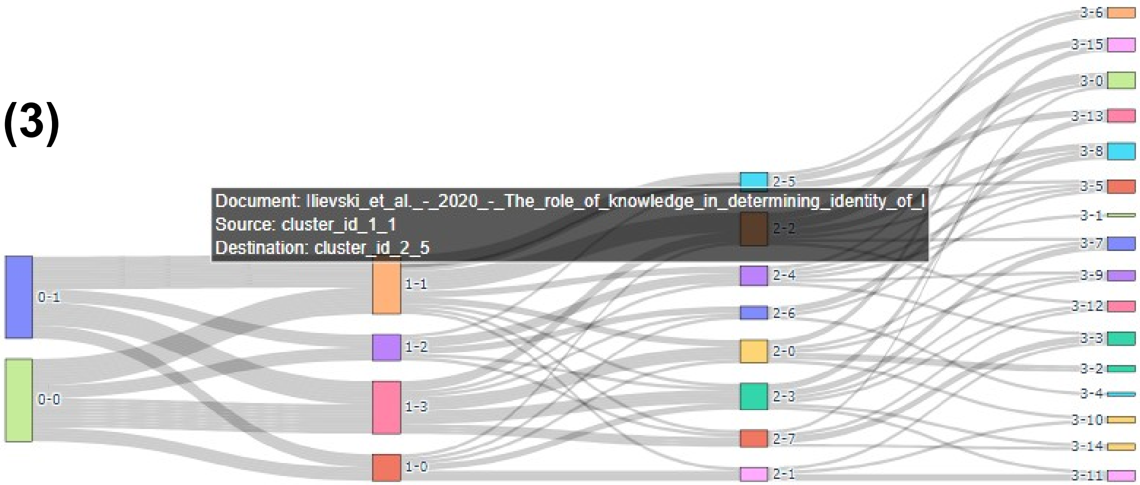
(1)



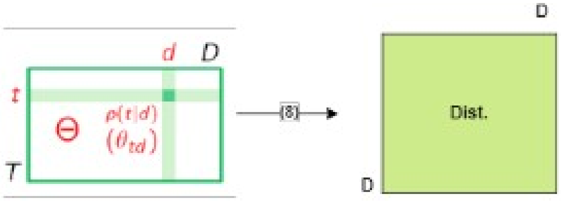
(2)



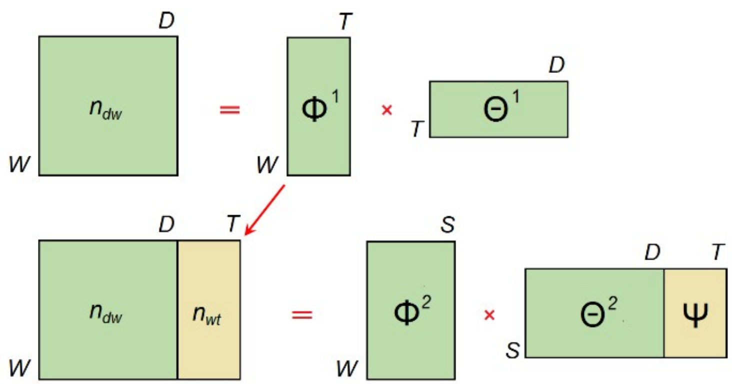
(3)

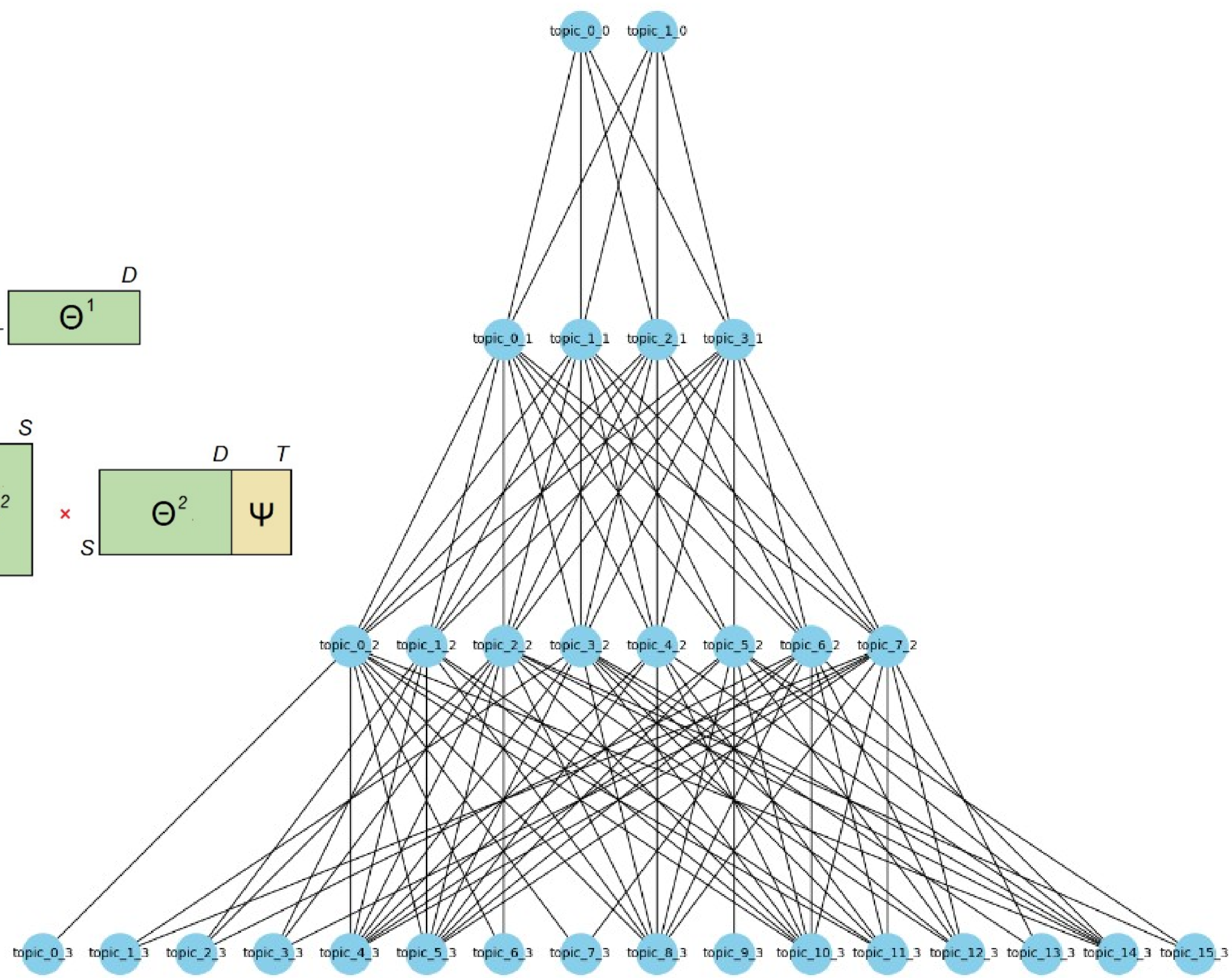


(4)



Results: Topic connectivity





Conclusions

- NLP pipeline was developed to effectively detect topic hierarchy in collections of scientific publication texts.
- hARTM coupled with Spectral Clustering and Multidimensional Scaling allowed to identify interpretable topics and evaluate document similarity in the context of topic hierarchy.
- hARTM approach allowed to achieve reduced computational requirements compared to LLM fine-tuning.
- The resulting approach provides a basis for development of valuable tools for self-study, literature reviews, and meta-analyses.

Acknowledgements

I am deeply grateful to Dr. sc. ing., Professor Jackiva Irina, and Dr. sc. Ing., Professor Dmitry Pavlyuk, for their unwavering support, expert guidance, and invaluable mentorship throughout this thesis.

Thank you for your attention!
Are there any questions?

Bigram-based topic interpretation

```
level0

topic_0:  ['natural_language', 'language_processing', 'computational_linguistics', 'association_computational', 'international_conference', 'machine_learning', 'social_m
topic_1:  ['machine_learning', 'international_conference', 'network_embedding', 'computer_science', 'deep_learning', 'natural_language', 'association_computational', 'co

level1

topic_0:  ['text_mining', 'spam_detection', 'clinical_trial', 'social_spam', 'dream_reports', 'clinical_trials', 'argument_mining', 'natural_language', 'social_networks'
topic_1:  ['natural_language', 'language_processing', 'computational_linguistics', 'concept_extraction', 'association_computational', 'international_conference', 'procee
topic_2:  ['social_media', 'data_set', 'information_processing', 'processing_management', 'text_classification', 'stance_detection', 'computational_linguistics', 'comput
topic_3:  ['machine_learning', 'network_embedding', 'deep_learning', 'relation_extraction', 'representation_learning', 'international_conference', 'patent_text', 'associ

level2

topic_0:  ['clinical_trial', 'clinical_trials', 'computer_science', 'seed_words', 'seed_vocabulary', 'stance_detection', 'label_names', 'cjeu_vat', 'eligibility_criteria
topic_1:  ['patent_text', 'online_news', 'atomic_changes', 'quality_control', 'question_retrieval', 'atomic_change', 'data_set', 'news_articles', 'knowledge_sources', 'k
topic_2:  ['information_processing', 'electronic_health', 'word_embeddings', 'twitter_data', 'health_records', 'neural_networks', 'data_sets', 'data_set', 'clinical_note
topic_3:  ['social_media', 'classi_fication', 'learning_methods', 'processing_management', 'piskorski_information', 'haneczok_piskorski', 'event_templates', 'digital_path
topic_4:  ['spam_detection', 'social_spam', 'dream_reports', 'expert_systems', 'problems_solutions', 'argument_mining', 'systems_applications', 'advantageous_effects', '
topic_5:  ['social_distancing', 'jain_borah', 'spectral_clustering', 'borah_biswas', 'text_mining', 'distancing_index', 'biomaterials_annotator', 'accessed_april', 'mone
topic_6:  ['concept_extraction', 'proceedings_conference', 'relation_extraction', 'named_entity', 'anaphora_resolution', 'subjectivity_detection', 'methods_natural', 'em
topic_7:  ['network_embedding', 'representation_learning', 'core_competency', 'thematic_areas', 'institute_technology', 'thematic_area', 'network_representation', 'recor

level3

topic_0:  ['expert_systems', 'systems_applications', 'problems_solutions', 'core_competency', 'prefiltering_model', 'advantageous_effects', 'technical_problem', 'trainin
topic_1:  ['network_embedding', 'representation_learning', 'machine_learning', 'international_conference', 'computer_science', 'question_retrieval', 'word_embeddings', '
topic_2:  ['dream_reports', 'twitter_data', 'data_set', 'data_sets', 'document_clustering', 'contextual_groups', 'reddit_data', 'topic_modelling', 'consciousness_cogniti
topic_3:  ['natural_language', 'deep_learning', 'language_processing', 'machine_learning', 'named_entity', 'concept_extraction', 'computational_linguistics', 'electronic
topic_4:  ['natural_language', 'computational_linguistics', 'language_processing', 'anaphora_resolution', 'association_computational', 'concept_extraction', 'proceedings
topic_5:  ['spam_detection', 'social_spam', 'machine_learning', 'social_networks', 'deep_learning', 'international_conference', 'learning_methods', 'neural_network', 'pr
topic_6:  ['patent_text', 'machine_learning', 'stance_detection', 'deep_learning', 'modeling_combinations', 'international_conference', 'disease_networks', 'patent_texts
topic_7:  ['computational_linguistics', 'association_computational', 'natural_language', 'text_classification', 'international_conference', 'language_processing', 'relat
topic_8:  ['clinical_trial', 'natural_language', 'language_processing', 'clinical_trials', 'text_mining', 'argument_mining', 'text_data', 'machine_learning', 'spectral_c
topic_9:  ['knowledge_graph', 'international_conference', 'classi_fication', 'pathology_reports', 'quality_control', 'natural_language', 'arxiv_preprint', 'jain_borah', '
topic_10:  ['social_media', 'text_mining', 'cjeu_vat', 'computer_science', 'procedia_computer', 'tax_rulings', 'biomaterials_annotator', 'vat_cases', 'natural_language', '
topic_11:  ['social_media', 'online_news', 'social_distancing', 'atomic_changes', 'data_set', 'atomic_change', 'news_accuracy', 'distancing_index', 'linguistic_errors',
```