DOUBLE DEGREE IN COMPUTER SCIENCE:
DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

MASTER THESIS

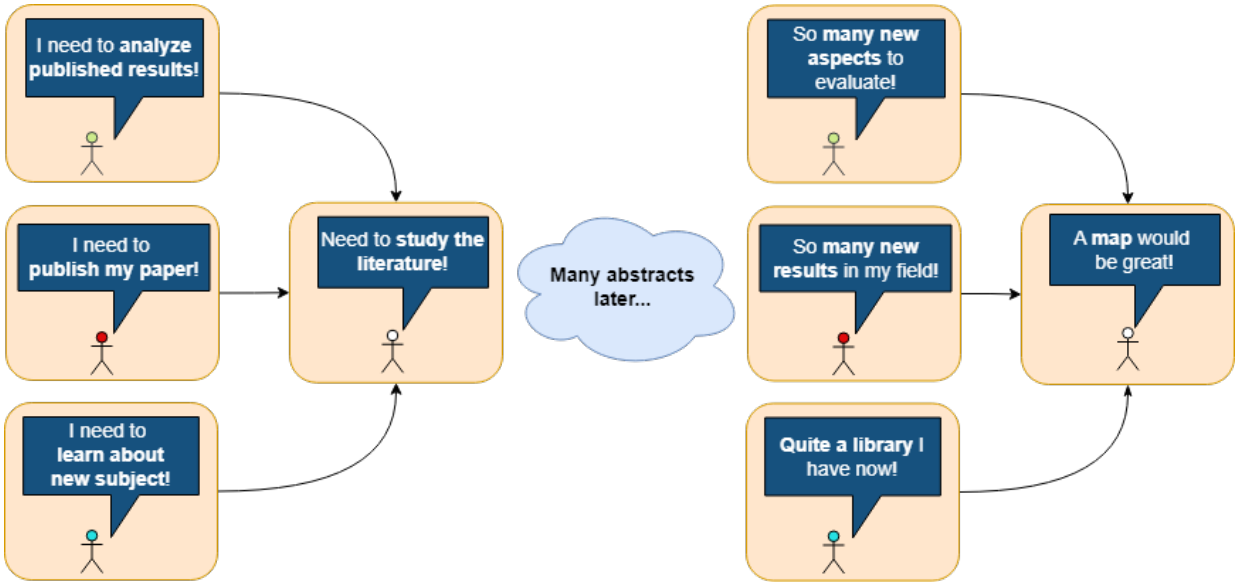
Unsupervised machine learning approach for hierarchical graph-based representation of natural language text collections

Jevgenijs Bodrenko
st81238
4203MDA

Supervisor: *Professor, Dr. sc. ing.* Irina Jackiva
Consultant: *Professor, Dr. sc. ing.* Dmitry Pavlyuk

Actuality & Motivation

- According to (Hassani *et al.*, 2020)
  text mining is a vast area with wide range of application, including:
  - Strategic management
  - Scientific research
  - Cybersecurity
  - Market research
  - Marketing
  - Politics
- (Hanson *et al.*, 2023) argued that annual growth rate of number of papers has some negative effects on science quality.



Hassani, H., Beneki, C., Unger, S., Mazinani, M.T. and Yeganegi, M.R. (2020) Text Mining in Big Data Analytics. Big Data and Cognitive Computing [online]. 4 (1), p. 1.
Hanson, M.A., Barreiro, P.G., Crosetto, P. and Brockington, D. (2023) *The strain on scientific publishing* [online]. Available from: http://arxiv.org/abs/2309.15884 [Accessed 11 May 2024].

- **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 LLMs.

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

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

- **Research aim**

  To develop a prototype of the machine learning pipeline for analyzing structures of document collections with focus on visual representation under condition of restricted computational resources.

- **Research objectives**
  - Implement a machine learning pipeline to detect document similarities in the context of topical hierarchy.
  - Develop visualization approaches for analyzing structure of a document collection and highlighting document similarity.
  - Ensure the possibility to run the solution in an environment where resources are not sufficient for LLM fine-tuning.

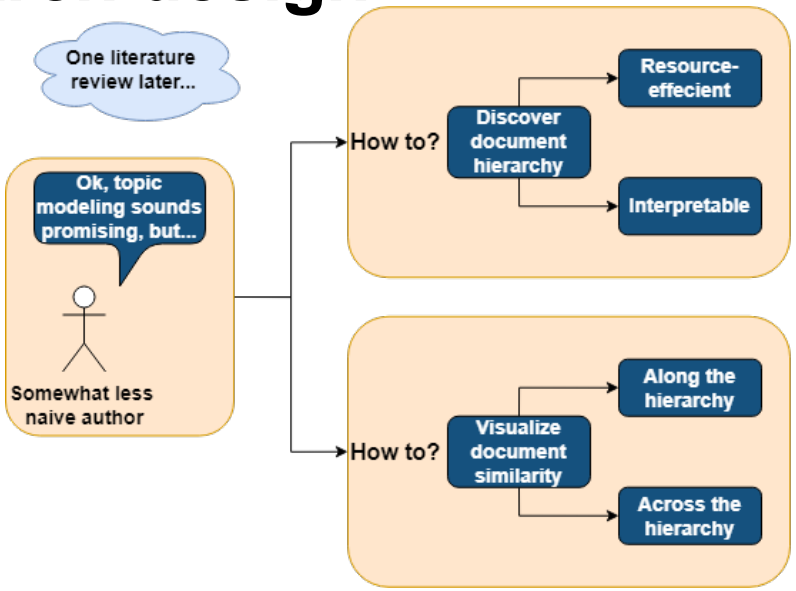
Research questions

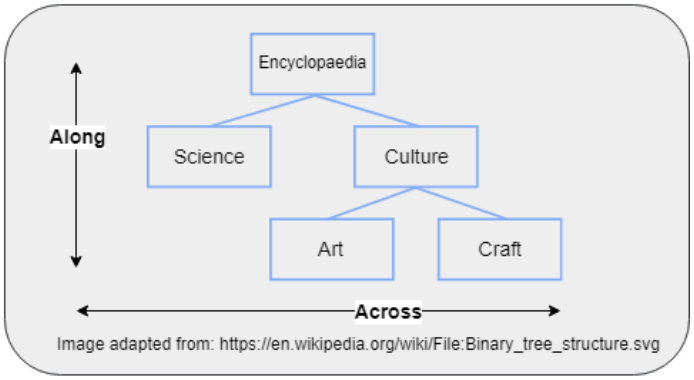
1. How to discover the topic hierarchy in a collection of English texts using unsupervised machine learning methods, given that it can be discovered by a human?

2. What model and quality metrics allow to ensure that the resulting hierarchy is understandable, insightful about the structure of the collection, and include interpretable connections between its different levels?

3. How can the output be visualized and used to explore the hierarchy structure and document similarity within the topic hierarchy?

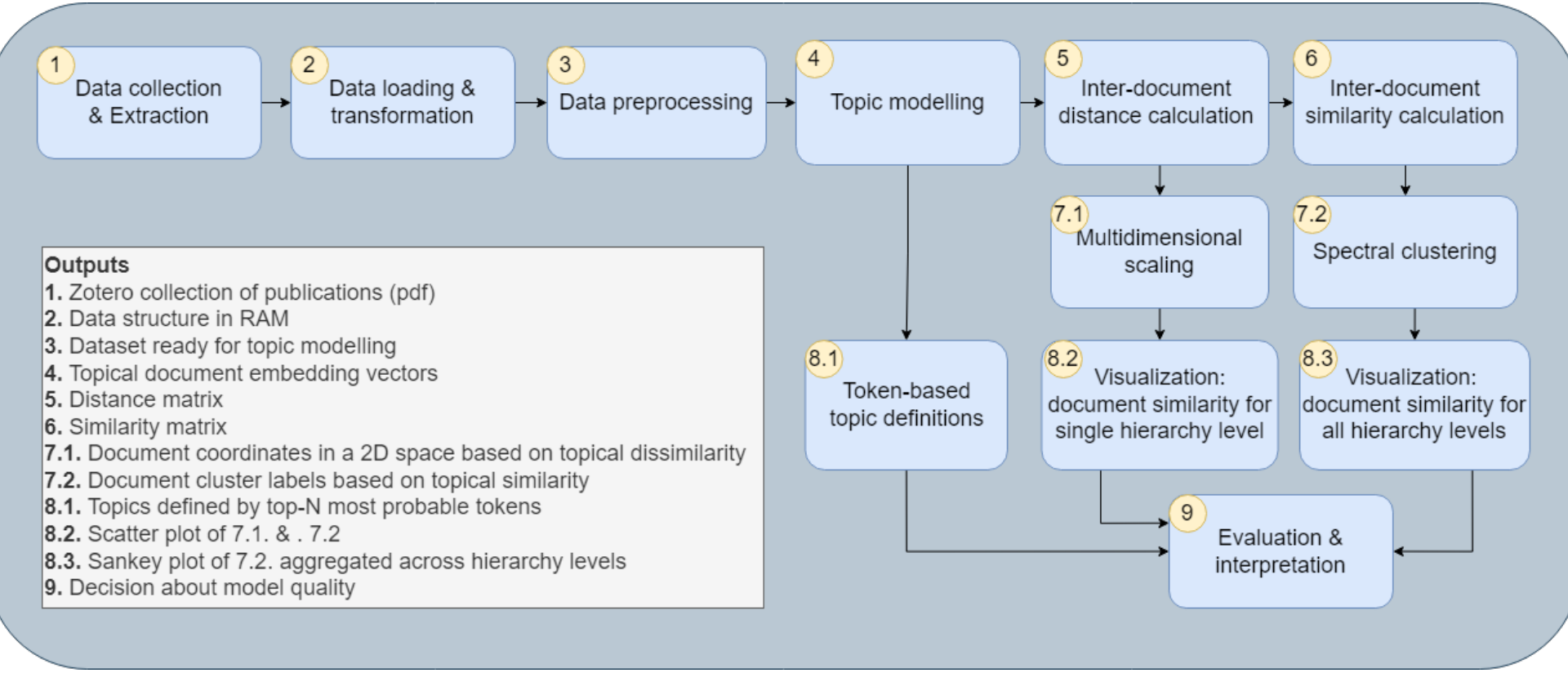
Research design





- **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 document clustering based on similarity.
- **Visualizations: Sankey plot** based on **Spectral Clustering** and **Scatter plot** based on **Multidimensional Scaling** to evaluate document similarity at each hierarchy level and across all levels based on minimum number of visalizations.
- **Efficiency:** Unsupervised topic modelling approach allows to relax computational resource requirements compared to Large Language Model fine-tuning.
- **Applications:** Potential use in academia and industry for self-study, reviews, and meta-analyses.

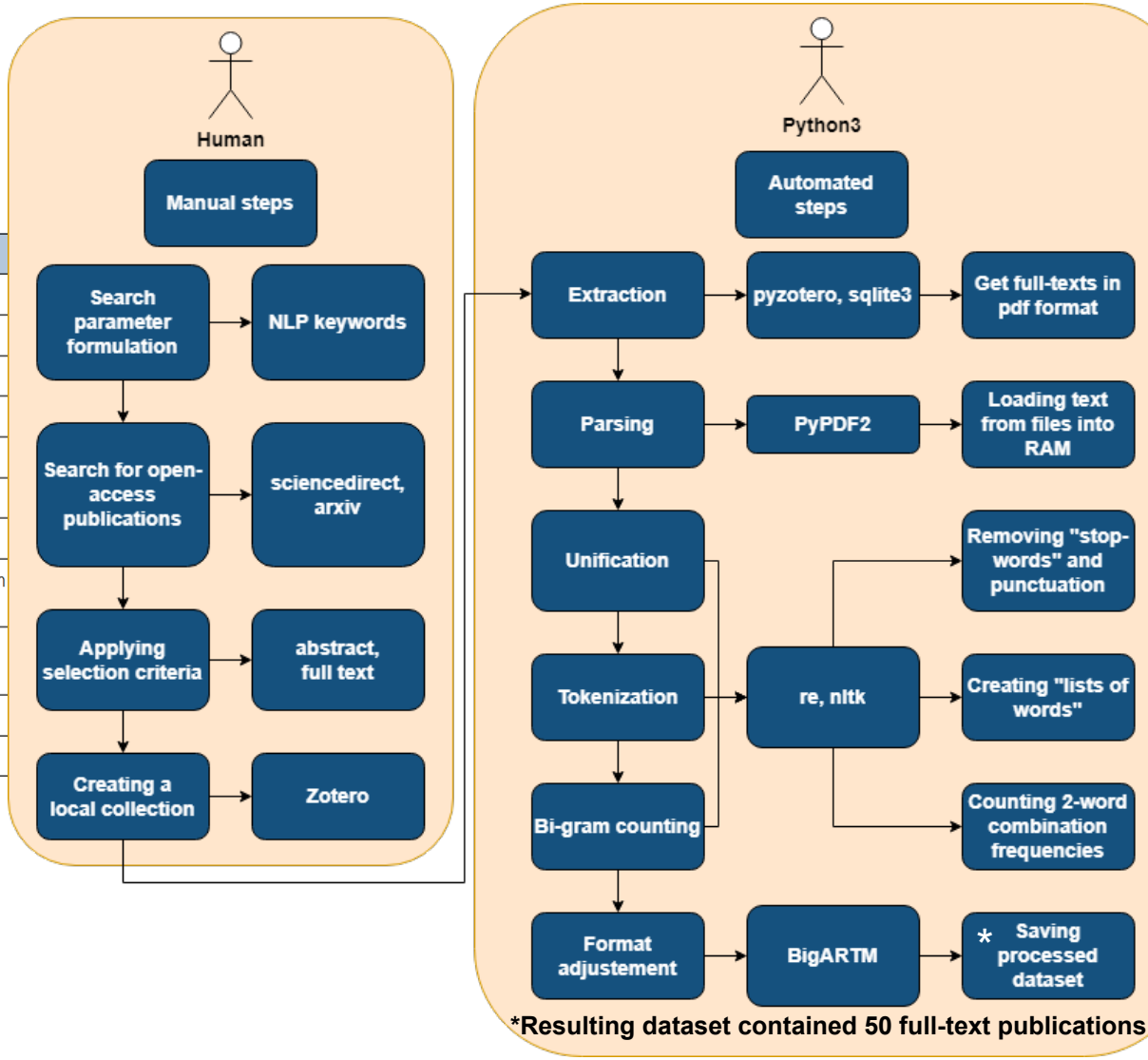
Pipeline diagram



Dataset preparation and preprocessing

Table 1: Terms used to construct queries for publication search

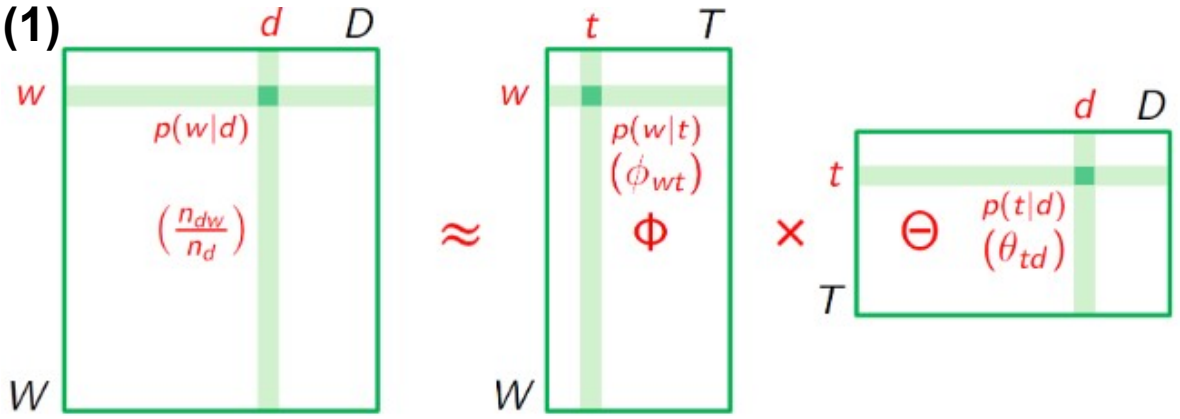
Search term	Synonyms		
Unsupervised	Without labels	Independent	Self-guided
NLP	Text mining		
Machine learning	ML		
Pipeline	Workflow		
Semantic	Meaning-based	Contextual	Context-based
Feature engineering	Feature extraction	Attribute engineering	Feature creation
Clustering	Grouping	Categorization	Partitioning
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)

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

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

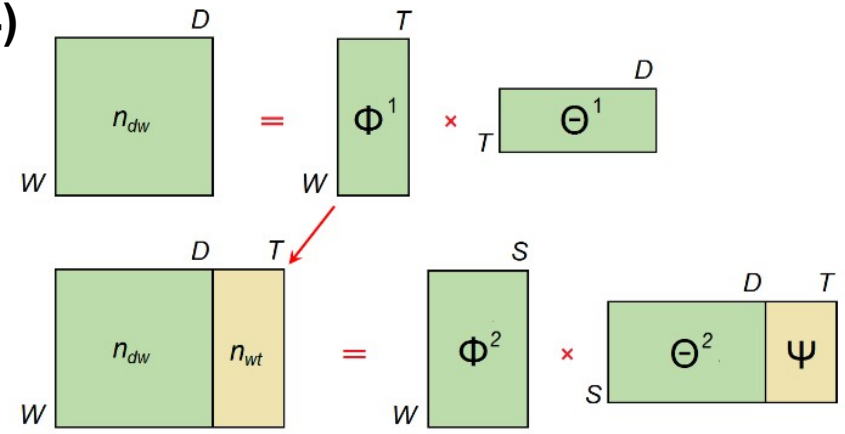
$$\sum_{w \in W} \phi_{wt} = 1;\ \phi_{wt} \geqslant 0; \sum_{t \in T} \theta_{td} = 1;\ \theta_{td} \geqslant 0;$$

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.

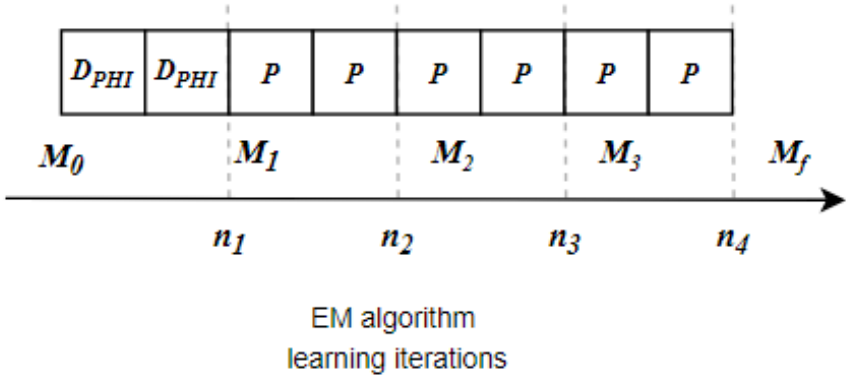
Methods: hARTM*

Hierarchical additive-regularization topic model

(4)



(5)



(6)

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

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

(7)

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

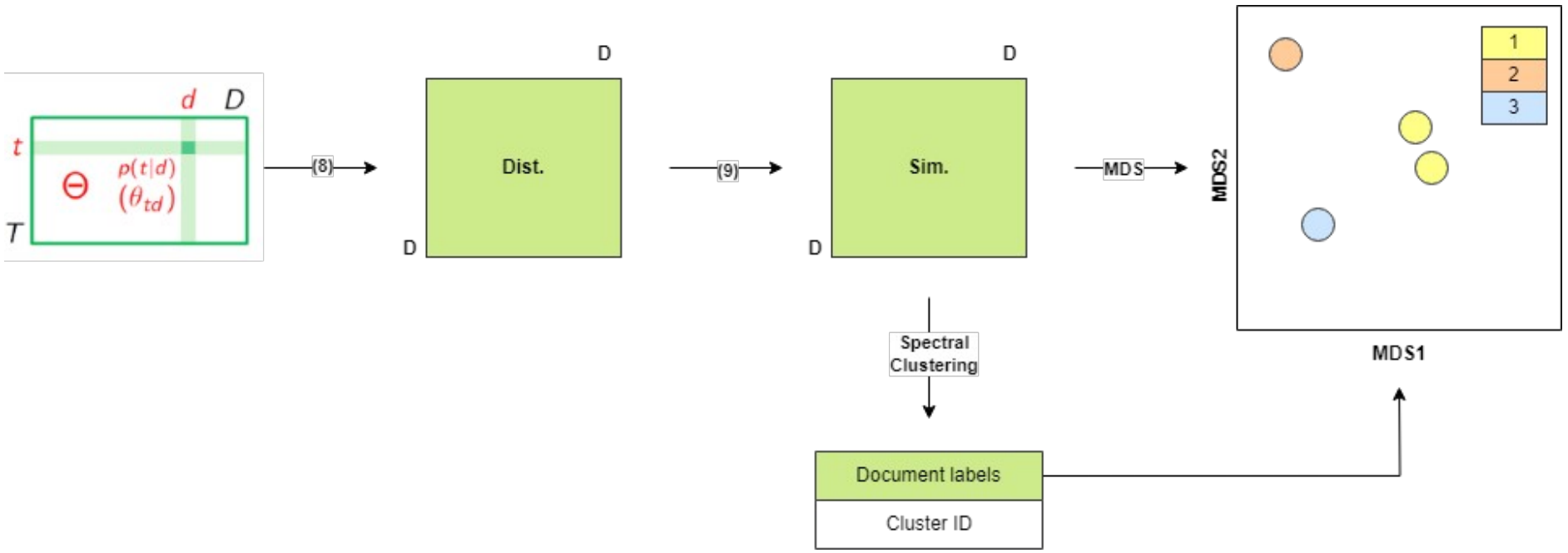
$$C_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=1}^{k} PMI(w_i, w_j)$$

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

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

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) $$H(A,B) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{a_i} - \sqrt{b_i}\right)^2}$$

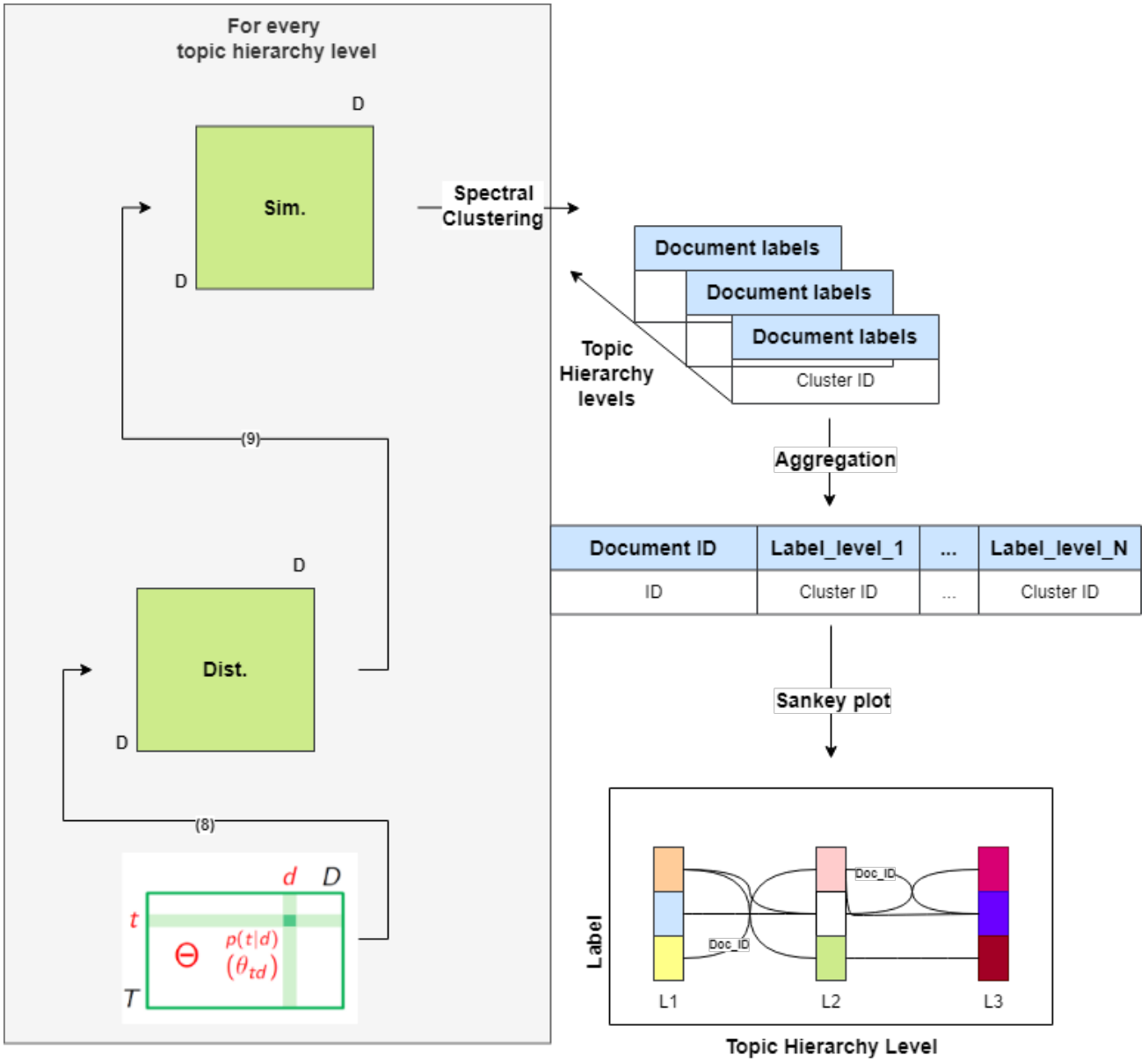
(9) $$B(A,B) = 1 - H(A,B)^2$$

Images adapted from:

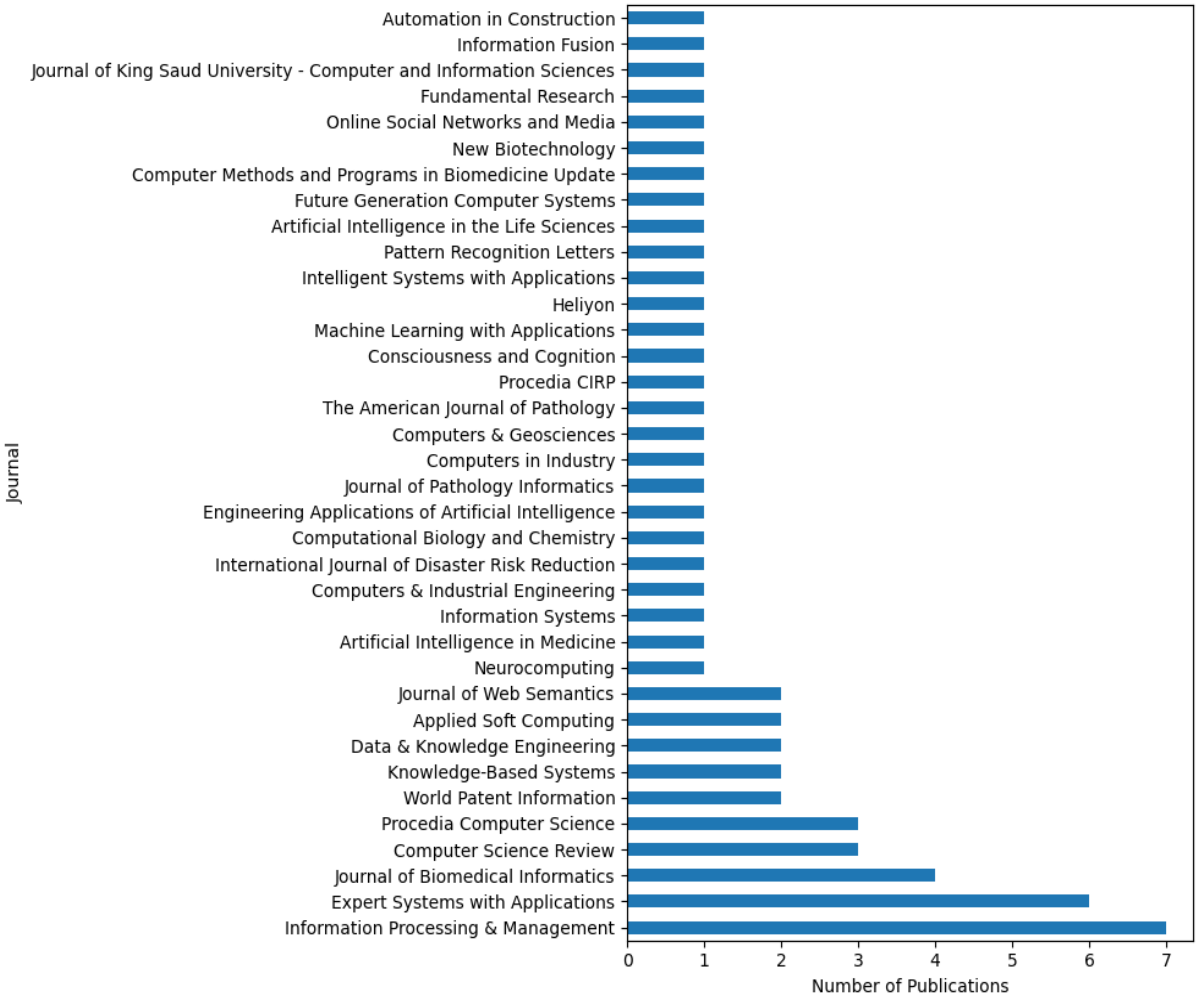
(8) Chirkova, N.A., JSC Antiplagiat, and Lomonosov Moscow State University (2016) 'Additive Regularization for Hierarchical Multimodal Topic Modeling', Machine Learning and Data Analysis, 2(2), pp. 187–200. Available at: https://doi.org/10.21469/22233792.2.2.05.

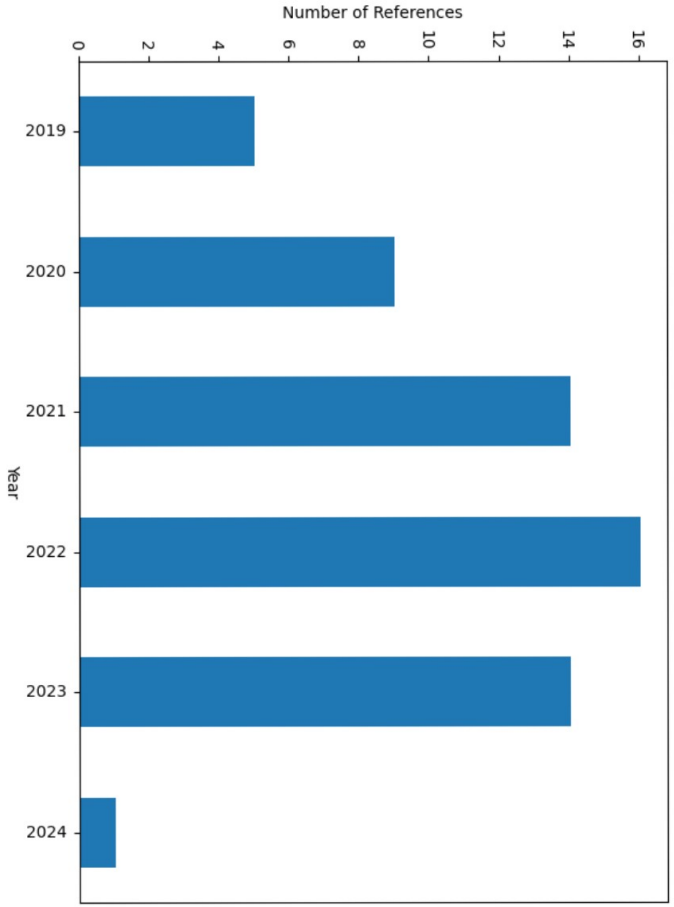
(9) Kitsos, C.P. and Nisiotis, C.-S. (2022) 'Considering distance measures in Statistics', Biometrical Letters, 59(1), pp. 65–75. Available at: https://doi.org/10.2478/bile-2022-0006.

Methods: Longitudinal view

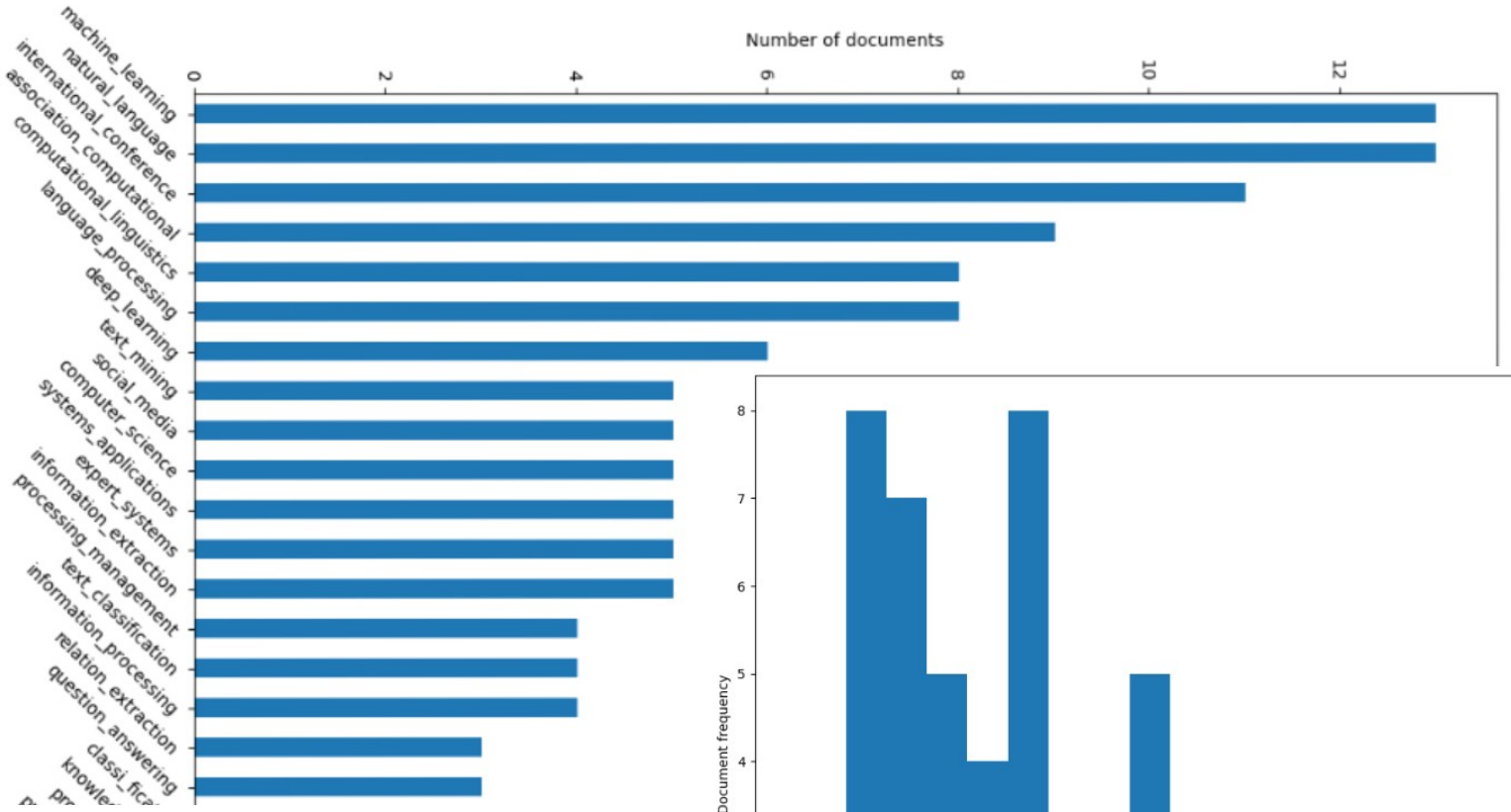


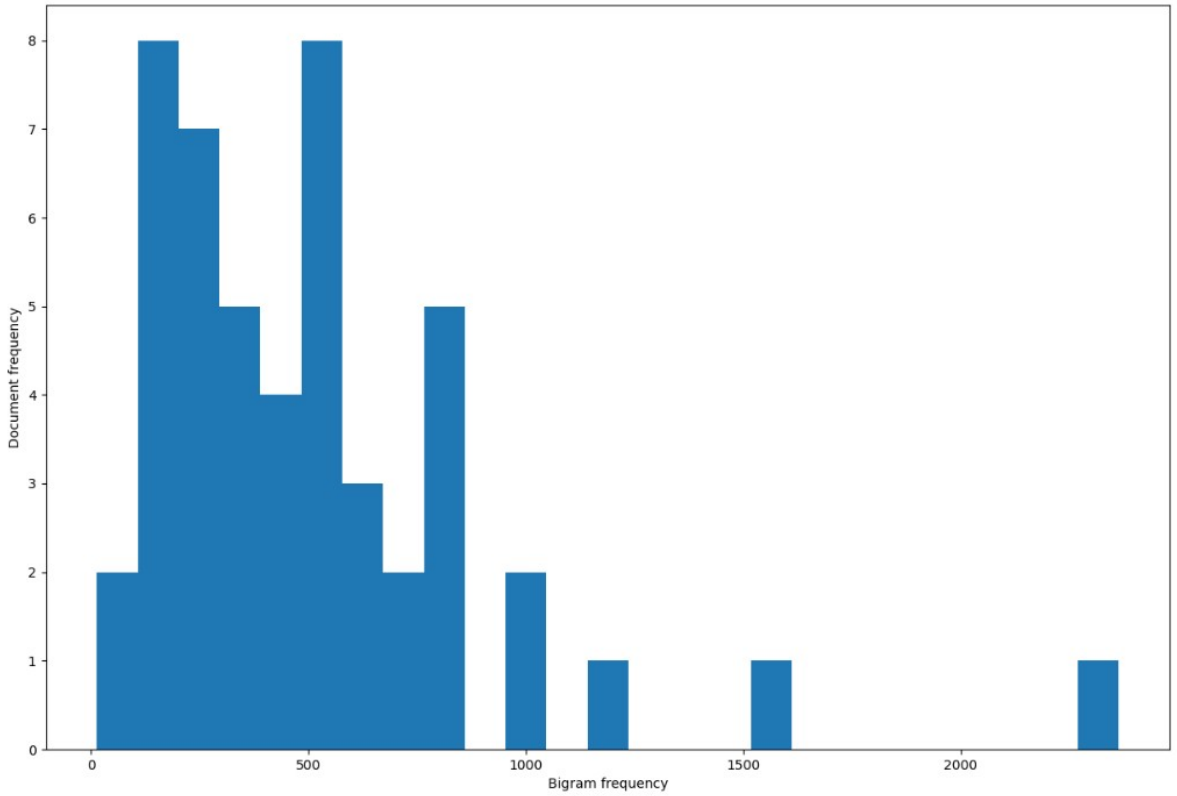
Results: Dataset metadata





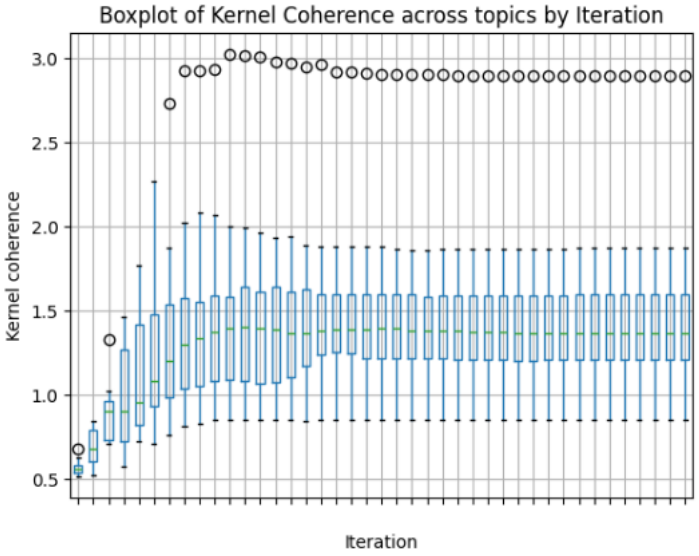
Results: Bigram-based dataset view



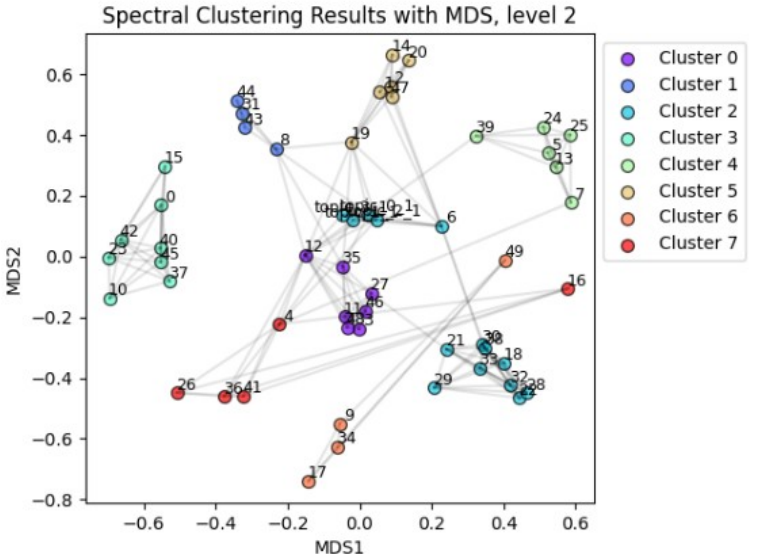


Results: Topic Modelling

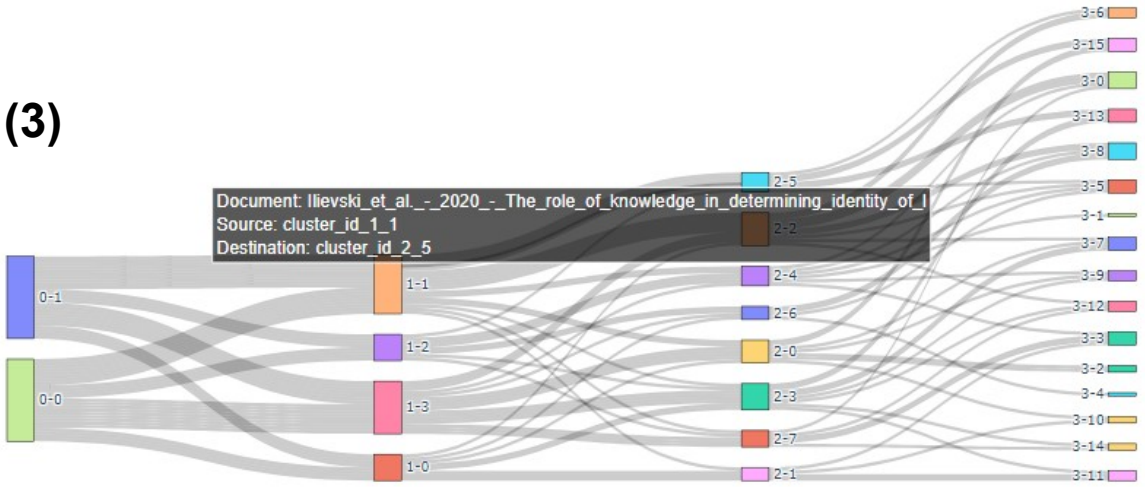
(1)



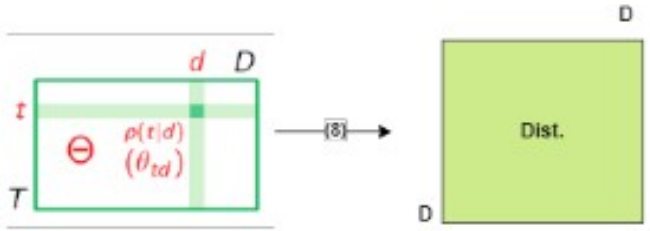
(2)



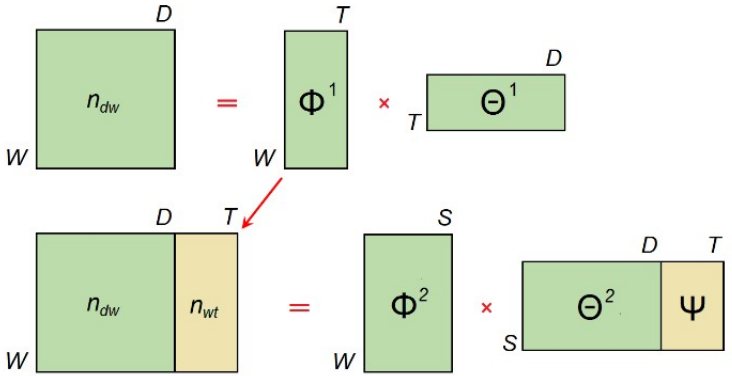
(3)

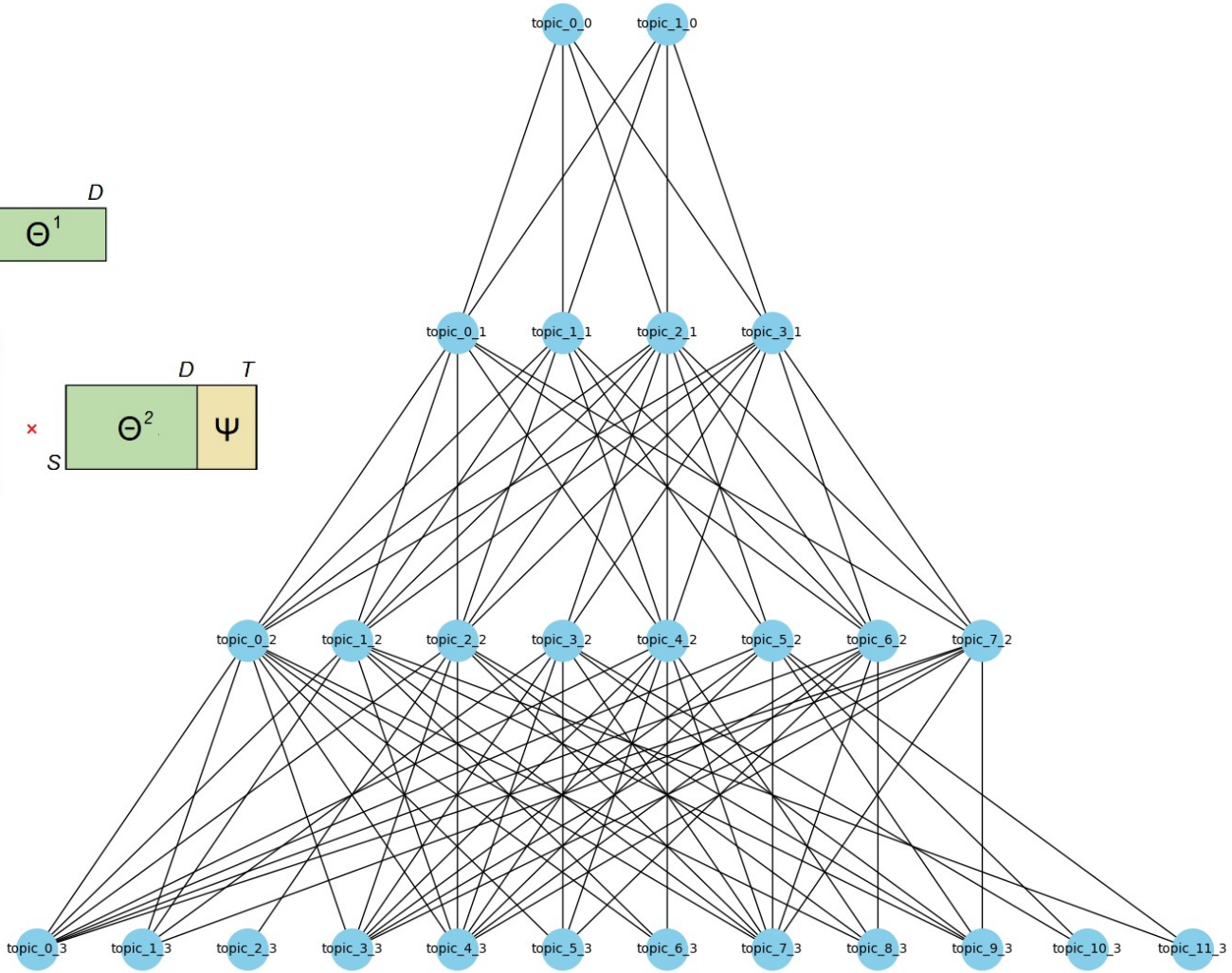


(4)



Results: Topic connectivity





Conclusions

- Literature review conducted and the State of the art formulated based on 59 discovered research papers and review articles.
- NLP pipeline prototype was developed using hARTM approach allowing to model the datasets as hierarchies of interpretable topics.
  - Tested on two collections of scientific publication full texts
  - Worked in the environment with limited computational resources
  - Allowed to evaluate document and topic similarity using developed visualizations
  - Provides a basis for potentially valuable text mining tools

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 definitions

```
level0

topic_0:  ['natural_language', 'language_processing', 'computational_linguistics', 'association_computational', 'international_
topic_1:  ['machine_learning', 'international_conference', 'network_embedding', 'computer_science', 'deep_learning', 'natural_l

level1

topic_0:  ['text_mining', 'spam_detection', 'clinical_trial', 'social_spam', 'dream_reports', 'clinical_trials', 'argument_mini
topic_1:  ['natural_language', 'language_processing', 'computational_linguistics', 'concept_extraction', 'association_computati
topic_2:  ['social_media', 'data_set', 'information_processing', 'processing_management', 'text_classification', 'stance_detect
topic_3:  ['machine_learning', 'network_embedding', 'deep_learning', 'relation_extraction', 'representation_learning', 'interna

level2

topic_0:  ['clinical_trial', 'clinical_trials', 'computer_science', 'seed_words', 'seed_vocabulary', 'stance_detection', 'label
topic_1:  ['patent_text', 'online_news', 'atomic_changes', 'quality_control', 'question_retrieval', 'atomic_change', 'data_set'
topic_2:  ['information_processing', 'electronic_health', 'word_embeddings', 'twitter_data', 'health_records', 'neural_networks
topic_3:  ['social_media', 'classi_fication', 'learning_methods', 'processing_management', 'piskorski_information', 'haneczok_pi
topic_4:  ['spam_detection', 'social_spam', 'dream_reports', 'expert_systems', 'problems_solutions', 'argument_mining', 'system
topic_5:  ['social_distancing', 'jain_borah', 'spectral_clustering', 'borah_biswas', 'text_mining', 'distancing_index', 'biomat
topic_6:  ['concept_extraction', 'proceedings_conference', 'relation_extraction', 'named_entity', 'anaphora_resolution', 'subje
topic_7:  ['network_embedding', 'representation_learning', 'core_competency', 'thematic_areas', 'institute_technology', 'themat

level3

topic_0:  ['expert_systems', 'systems_applications', 'problems_solutions', 'core_competency', 'prefiltering_model', 'advantaged
topic_1:  ['network_embedding', 'representation_learning', 'machine_learning', 'international_conference', 'computer_science',
topic_2:  ['dream_reports', 'twitter_data', 'data_set', 'data_sets', 'document_clustering', 'contextual_groups', 'reddit_data',
topic_3:  ['natural_language', 'deep_learning', 'language_processing', 'machine_learning', 'named_entity', 'concept_extraction'
topic_4:  ['natural_language', 'computational_linguistics', 'language_processing', 'anaphora_resolution', 'association_computat
topic_5:  ['spam_detection', 'social_spam', 'machine_learning', 'social_networks', 'deep_learning', 'international_conference',
topic_6:  ['patent_text', 'machine_learning', 'stance_detection', 'deep_learning', 'modeling_combinations', 'international_conf
topic_7:  ['computational_linguistics', 'association_computational', 'natural_language', 'text_classification', 'international_
topic_8:  ['clinical_trial', 'natural_language', 'language_processing', 'clinical_trials', 'text_mining', 'argument_mining', 't
topic_9:  ['knowledge_graph', 'international_conference', 'classi_fication', 'pathology_reports', 'quality_control', 'natural_la
topic_10:  ['social_media', 'text_mining', 'cjeu_vat', 'computer_science', 'procedia_computer', 'tax_rulings', 'biomaterials_ar
topic_11:  ['social_media', 'online_news', 'social_distancing', 'atomic_changes', 'data_set', 'atomic_change', 'news_accuracy',
```