Iterative Improvement of an Additively Regularized Topic Model

Alex Gorbulev[1], **Vasiliy Alekseev**[1], Konstantin Vorontsov[2]

[1] Moscow Institute of Physics and Technology
[2] Lomonosov Moscow State University

AIST 2024: The 12th International Conference on Analysis of Images, Social Networks and Texts
17 October, 2024

TL;DR

TL;DR

Knowledge preservation in the model

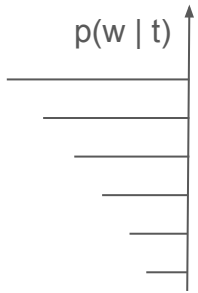
via *objective function modification*.

…The open country in the suburbs was quiet and deserted. Moreover, few would venture out into the snow at this time of the night. After leaving the house, Zhu Zhen looked back and saw no footprints. He then wended his way to Miss Zhou's grave. …Unfortunately for him, the grave keepers had a dog. At this point, it emerged from its straw kennel to bark at the intruding stranger. Earlier in the day, Zhu Zhen had prepared a piece of fried dough and stuffed some drug in it. He now tossed the dough to the barking dog. The dog sniffed at it and, liking the aroma, ate it up. The very next moment, the dog gave a bark and collapsed to the ground. Zhu Zhen drew near the grave…

The Fan Tower Restaurant as Witness to the Love of Zhou Shengxian,
from Stories to Awaken the World (Xingshi Hengyan); translated by Shuhui Yang and Yunqin Yang

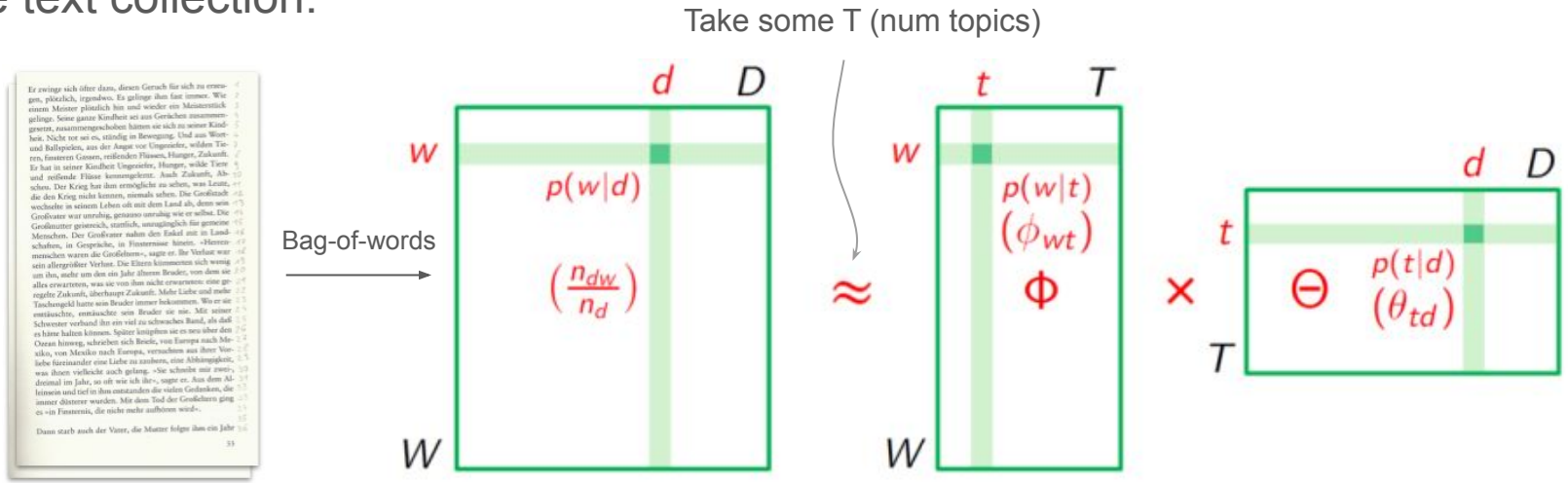
…The open country in the suburbs was quiet and deserted. Moreover, few would venture out into the snow at this time of the night. After leaving the house, Zhu Zhen looked back and saw no footprints. He then wended his way to Miss Zhou's grave. …Unfortunately for him, the grave keepers had a dog. At this point, it emerged from its straw kennel to bark at the intruding stranger. Earlier in the day, Zhu Zhen had prepared a piece of fried dough and stuffed some drug in it. He now tossed the dough to the barking dog. The dog sniffed at it and, liking the aroma, ate it up. The very next moment, the dog gave a bark and collapsed to the ground. Zhu Zhen drew near the grave…

Nature	Winter night	Adventure	Illegal entry	Cemetery	Dogs	Food	Poison
forest	snow	venture	thief	grave	dog	dough	drug
sky	night	danger	house	grave keeper	bark	fried dough	antidote
grass	frost	risk	intrude	tombstone	barking dog	eat	sick
straw	snowflake	stranger	steal	coffin	friend	aroma	suffer
open country	quiet	footprint	money	crypt	kennel	rice	collapse
suburbs	deserted	escape	danger	night	collar	bacalhau	snake



Topic Modeling

Topic modelling assumes that there are a number of *latent topics* which explain the text collection.



Konstantin Vorontsov, Probabilistic Topic Modeling (in Russian).

Topic Modeling

Given:

- D — text collection
- W — set of words found in texts (vocabulary)
- n_{wd} — frequency of the word ‘w’ in the document ‘d’

Find:

- set of *hidden* topics T as distributions p(w | t)
- distributions of topics in documents p(t | d)

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t} \phi_{wt} \theta_{td}$$

Topic Modeling

Criterion: maximization of regularized log-likelihood:

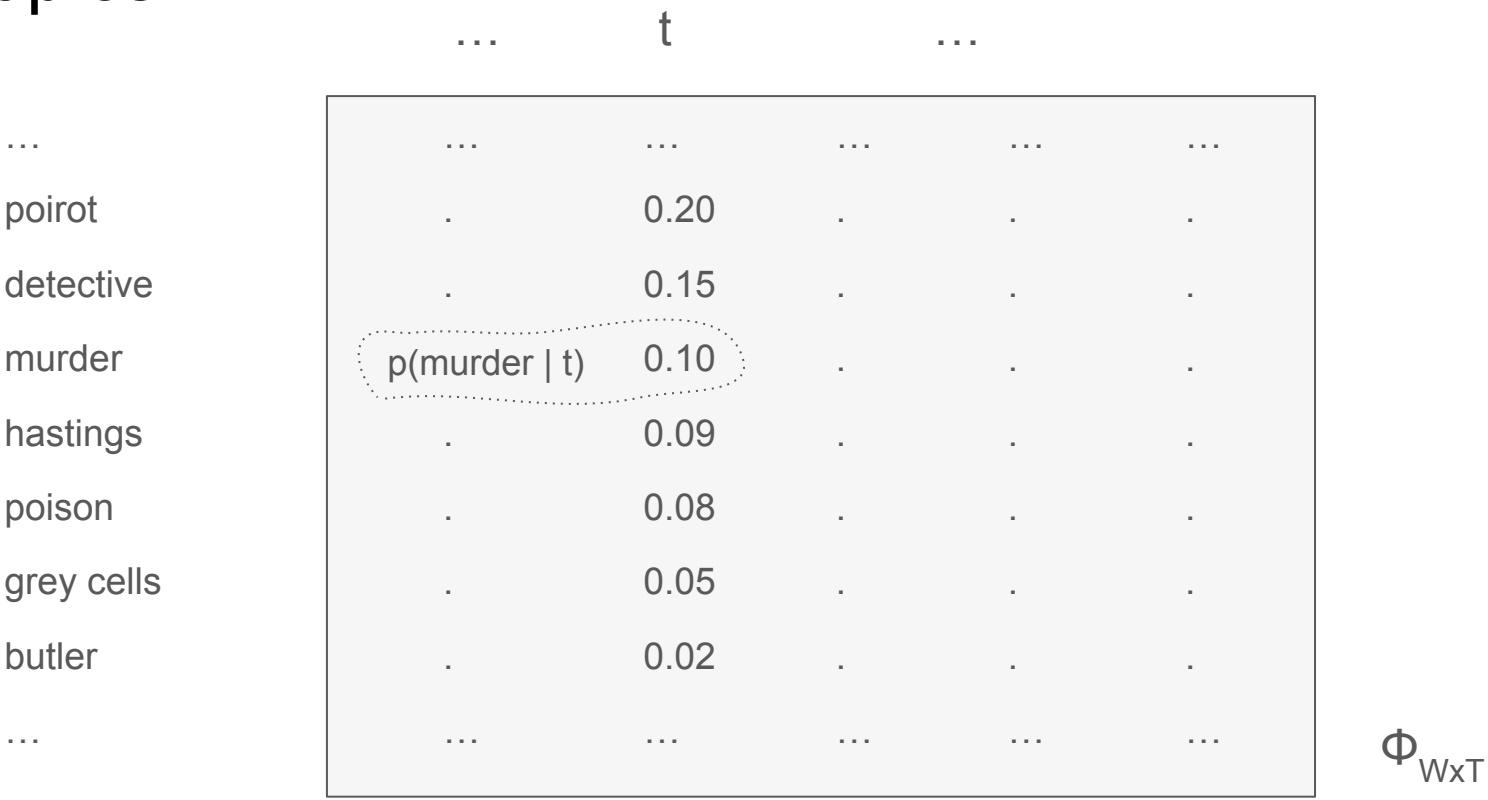
$$\underbrace{\ln p(\Phi,\Theta)}_{\mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n}\tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

$$\sum_{w\in W}\phi_{wt}=1,\ \phi_{wt}\geq 0 \qquad \sum_{t\in T}\theta_{td}=1,\ \theta_{td}\geq 0$$

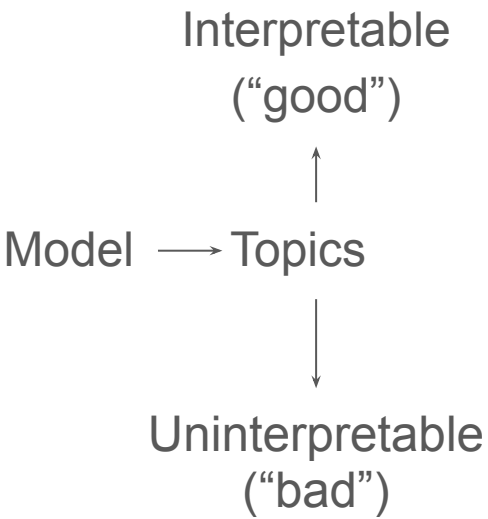
Solution: fixed-point iteration (Vorontsov, 2014):

E-step	$p_{tdw}=\underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td})$
M-step	$\phi_{wt}=\underset{w\in W}{\text{norm}}\left(n_{wt}+\phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$
	$\theta_{td}=\underset{t\in T}{\text{norm}}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$

Topics


… t …
… … … … … …
poirot . 0.20 . . .
detective . 0.15 . . .
murder p(murder | t) 0.10 . . .
hastings . 0.09 . . .
poison . 0.08 . . .
grey cells . 0.05 . . .
butler . 0.02 . . .
… … … … … …
Φ WxT

Problem of Topic Models



- conference, aist, recognition, prediction
- bishkek, oak park, mountains, som, jansak, manas
- autumn, cold, rain, puddles, leaf palette

- dinosaur, math, moon, suspicion, quick
- I, she, go, in, take, with, call, talk
- teacher, teach, school, taught, teachers, lesson

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
    set_parameters(topic_model)
    train(topic_model, dataset)
    assess_quality(topic_model)
    analyze_topics(topic_model)
```

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
    set_parameters(topic_model)
    train(topic_model, dataset)
    assess_quality(topic_model)
    analyze_topics(topic_model)
```

```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
    set_parameters(topic_model)
    train(topic_model, dataset)
    assess_quality(topic_model)
    analyze_topics(topic_model)
```

```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
    set_parameters(topic_model)
    train(topic_model, dataset)
    assess_quality(topic_model)
    analyze_topics(topic_model)
```

```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

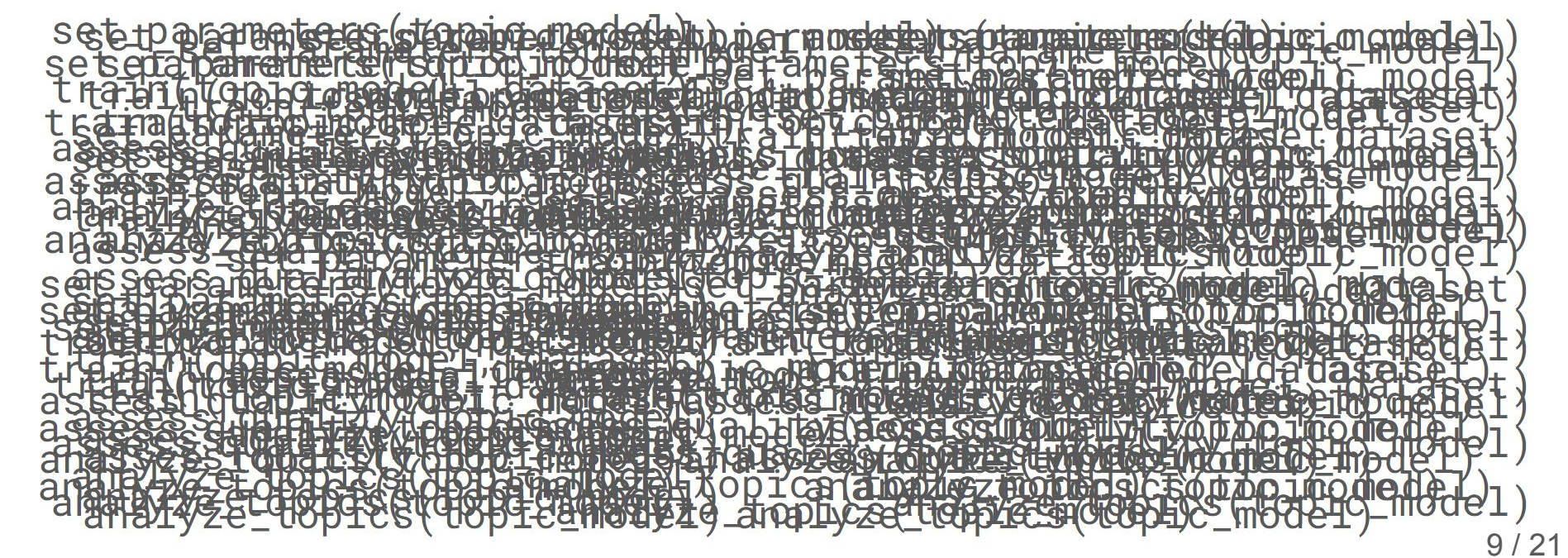
```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

```
set_parameters(topic_model)
train(topic_model, dataset)
assess_quality(topic_model)
analyze_topics(topic_model)
```

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
```



Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):
```

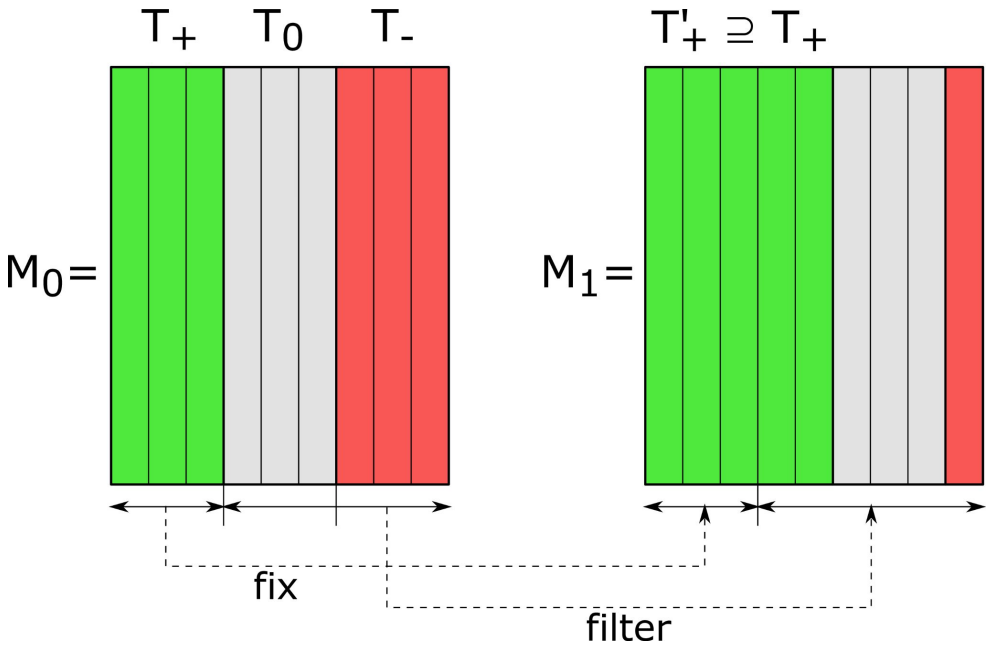
Iterative Improvement of a Topic Model

Problem:

- *A lot of* experiments to find a good model.
- The good topics found in the process are *lost*.

Solution:

- *Fix* good topics.
- Train the remaining free topics to be *unlike* the bad ones.



Additive Regularization

Maximization of the regularized log-likelihood:

ARTM: $$L(\Phi, \Theta) + R_{\text{sparse}}(\Phi) + R_{\text{decorr}}(\Phi) \to \max_{\Phi,\Theta}$$

$$R_{\text{sparse}}(\Phi)|_{\tau>0} = -\tau \sum_{t\in T} \sum_{w\in W} \beta_w \ln \phi_{wt} \to \max_{\Phi}$$

$$R_{\text{decorr}}(\Phi)|_{\tau>0} = -\tau \sum_{t\in T} \sum_{s\in T\setminus t} \sum_{w\in W} \phi_{wt}\phi_{ws} \to \max_{\Phi}$$

Vorontsov K. et al. BigARTM: Open source library for regularized multimodal topic modeling, 2015.

Additive Regularization for Iterative Improvement

Maximization of the regularized log-likelihood:

ITAR:

collected topics

$$L(\Phi,\Theta) + R_{\text{sparse}}(\Phi) + R_{\text{decorr}}(\Phi) + R_{\text{fix}}(\Phi,\widetilde{\Phi}) + R_{\substack{\text{decorr}\\\text{bad}}}(\Phi,\widetilde{\Phi}) + R_{\substack{\text{decorr}\\\text{good}}}(\Phi,\widetilde{\Phi}) \to \max_{\Phi,\Theta}$$

$$R_{\text{fix}}(\Phi,\widetilde{\Phi})|_{\tau\gg 1} = \tau \sum_{t\in T_+}\sum_{w\in W} \widetilde{\phi}_{wt} \ln \phi_{wt} \to \max_{\Phi}$$

$$R_{\substack{\text{decorr}\\\text{bad/good}}}(\Phi,\widetilde{\Phi})|_{\tau>0} = -\tau \sum_{t\in T\setminus T_+}\sum_{s\in T_-/T_+}\sum_{w\in W} \phi_{wt}\widetilde{\phi}_{ws} \to \max_{\Phi}$$

Experiment

Purpose:

- Verify that the number of good topics iteratively increases.
- Compare ITAR by the number of good topics with other topic models.

Key points:

- “Iteration” is one model training.
- Good topics are topics with high coherence values.
- Several topic models. Several text collections.
- Several iterations of training for each topic model (a series of 20 topic models).
- The final iterative model is the *last* model in the series.
- The final non-iterative model is the *best* one in the series (in terms of the number of good topics).

Topic Models

- **PLSA**: model with a single hyperparameter T.
- **LDA**: model where the Φ and Θ columns are generated by Dirichlet distributions.
- **ARTM**: model with additive regularization
- **TLESS**: model without Θ matrix, with sparse topics.
- **BERTopic**: neural network topic model.
- **TopicBank**: iteratively updated model without regularizers.

Hofmann, T. Probabilistic latent semantic analysis, 1999.
Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation, 2003.
Vorontsov K. et al. BigARTM: Open source library for regularized multimodal topic modeling, 2015.
Irkhin I., Bulatov V., Vorontsov K. Additive regularizarion of topic models with fast text vectorizartion, 2020.
Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022.
Alekseev V., Vorontsov K. et al. TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation, 2021.

Data

Dataset	D	Len	W	Lang
PostNauka	3404	421,2	19186	Ru
$20Newsgroups_{train}$	11301	93,9	52744	En
RuWiki-Good	8603	1934,6	61688	Ru
RTL-Wiki-Person	1201	1600,1	37739	En
ICD-10	2036	550,0	22608	Ru

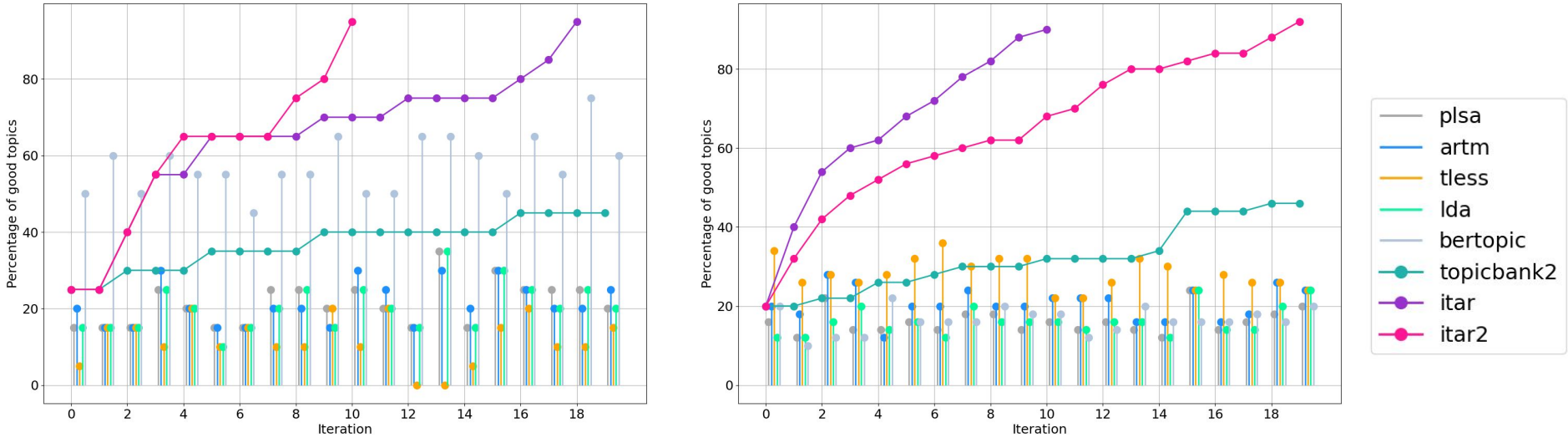
Datasets used in the experiments (D — number of documents, Len — average document length).
Pre-processing: lemmatization, stop word removal, "bag-of-words".

(Source: https://huggingface.co/TopicNet.)

Alekseev V., Bulatov V., Vorontsov K. Intra-text coherence as a measure of topic models' interpretability, 2018.
Bulatov V., Alekseev V., Vorontsov K. et al. TopicNet: Making additive regularisation for topic modelling accessible, 2020.
Chang J. et al. Reading tea leaves: How humans interpret topic models, 2009.

Results

- ITAR model contains the largest number of good[1] topics
- With good over 80% of the model's topics



Percentage of good model topics depending on the iteration (↑).
RuWiki-Good, models for 20 topics (left); 20Newsgroups, models for 50 topics (right).

[1]Newman D. et al. Automatic evaluation of topic coherence, 2010.

Results

- Highest % of good topics
- Topics are diverse
- Perplexity is moderate

Model	PostNauka (20 topics)				RuWiki-Good (50 topics)			
	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)
plsa	**2,99**	0,74	20	0,60	**3,46**	0,81	26	0,66
artm	3,15	0,79	40	0,61	3,62	0,86	30	0,67
tless	3,65	0,75	30	**0,75**	4,98	0,71	24	0,72
lda	**2,99**	0,73	25	0,58	**3,48**	0,83	24	0,65
bertopic	4,26/5,93	**1,16**	75	0,67	3,17/5,06	**1,34**	70	0,67
topicbank	4,22/6,11	0,98	30	0,60	7,39/12,94	**1,33**	20	0,68
topicbank2	4,12/8,11	**1,10**	70	0,67	6,09/11,30	1,16	44	0,69
itar	3,79	1,02	**90**	**0,76**	4,62	1,12	**86**	**0,77**
itar2	3,75	1,00	**90**	0,74	4,53	1,23	**96**	**0,77**

Some properties of the final models: perplexity (PPL), average topic coherence (Coh), percentage of good topics (Good T), diversity of topics (Div) as Jensen–Shannon divergence. PostNauka, models for 20 topics (left); RuWiki-Good, models for 50 topics (right).

Results

- Highest % of good topics
- Topics are diverse
- Perplexity is moderate

Model	PostNauka (20 topics)				RuWiki-Good (50 topics)			
	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)
plsa	**2,99**	0,74	20	0,60	**3,46**	0,81	26	0,66
artm	3,15	0,79	40	0,61	3,62	0,86	30	0,67
tless	3,65	0,75	30	**0,75**	4,98	0,71	24	0,72
lda	**2,99**	0,73	25	0,58	**3,48**	0,83	24	0,65
bertopic	4,26/5,93	**1,16**	75	0,67	3,17/5,06	**1,34**	70	0,67
topicbank	4,22/6,11	0,98	30	0,60	7,39/12,94	**1,33**	20	0,68
topicbank2	4,12/8,11	**1,10**	70	0,67	6,09/11,30	1,16	44	0,69
itar	3,79	1,02	**90**	**0,76**	4,62	1,12	**86**	**0,77**
itar2	3,75	1,00	**90**	0,74	4,53	1,23	**96**	**0,77**

Some properties of the final models: perplexity (PPL), average topic coherence (Coh), percentage of good topics (Good T), diversity of topics (Div) as Jensen–Shannon divergence. PostNauka, models for 20 topics (left); RuWiki-Good, models for 50 topics (right).

Results

- Highest % of good topics
- Topics are diverse
- Perplexity is moderate

Model	PostNauka (20 topics)				RuWiki-Good (50 topics)			
	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)
plsa	**2,99**	0,74	20	0,60	**3,46**	0,81	26	0,66
artm	3,15	0,79	40	0,61	3,62	0,86	30	0,67
tless	3,65	0,75	30	**0,75**	4,98	0,71	24	0,72
lda	**2,99**	0,73	25	0,58	**3,48**	0,83	24	0,65
bertopic	4,26/5,93	**1,16**	75	0,67	3,17/5,06	**1,34**	70	0,67
topicbank	4,22/6,11	0,98	30	0,60	7,39/12,94	**1,33**	20	0,68
topicbank2	4,12/8,11	**1,10**	70	0,67	6,09/11,30	1,16	44	0,69
itar	3,79	1,02	**90**	**0,76**	4,62	1,12	**86**	**0,77**
itar2	3,75	1,00	**90**	0,74	4,53	1,23	**96**	**0,77**

Some properties of the final models: perplexity (PPL), average topic coherence (Coh), percentage of good topics (Good T), diversity of topics (Div) as Jensen–Shannon divergence. PostNauka, models for 20 topics (left); RuWiki-Good, models for 50 topics (right).

Results

- Highest % of good topics
- Topics are diverse
- Perplexity is moderate

Model	PostNauka (20 topics)				RuWiki-Good (50 topics)			
	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Div (↑)
plsa	**2,99**	0,74	20	0,60	**3,46**	0,81	26	0,66
artm	3,15	0,79	40	0,61	3,62	0,86	30	0,67
tless	3,65	0,75	30	**0,75**	4,98	0,71	24	0,72
lda	**2,99**	0,73	25	0,58	**3,48**	0,83	24	0,65
bertopic	4,26/5,93	**1,16**	75	0,67	3,17/5,06	**1,34**	70	0,67
topicbank	4,22/6,11	0,98	30	0,60	7,39/12,94	**1,33**	20	0,68
topicbank2	4,12/8,11	**1,10**	70	0,67	6,09/11,30	1,16	44	0,69
itar	3,79	1,02	**90**	**0,76**	4,62	1,12	**86**	**0,77**
itar2	3,75	1,00	**90**	0,74	4,53	1,23	**96**	**0,77**

Some properties of the final models: perplexity (PPL), average topic coherence (Coh), percentage of good topics (Good T), diversity of topics (Div) as Jensen–Shannon divergence. PostNauka, models for 20 topics (left); RuWiki-Good, models for 50 topics (right).

Ablation study

- Fixing good topics increases the proportion of good topics in the model
- + decorrelation with bad ones reduces the frequency of bad topics
- + decorrelation with good ones results in more diverse topics

Model	PostNauka (20 topics)					
	Train iters, % (↓)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Seen bad T, % (↓)	Div (↑)
itar	**50**	3,79	**1,02**	**90**	**100**	**0,76**
itar_0-0-1	85	**3,30**	0,81	35	275	0,66
itar_0-1-0	**60**	**3,31**	0,86	50	350	0,71
itar_0-1-1	85	3,31	0,93	50	325	0,71
itar_1-0-0	70	3,56	0,90	60	230	0,69
itar_1-0-1	90	3,65	0,95	75	200	0,72
itar_1-1-0	90	3,75	**1,05**	**95**	**95**	**0,75**

The effect of different parts of the ITAR model on the final result. Name format: “itar_[is there fixation of good topics]-[is there decorrelation with bad topics]-[is there decorrelation with good topics]”. “Train iters” is how many iterations the training took (as a percentage of the maximum number of iterations).

Ablation study

- Fixing good topics increases the proportion of good topics in the model
- + decorrelation with bad ones reduces the frequency of bad topics
- + decorrelation with good ones results in more diverse topics

Model	PostNauka (20 topics)					
	Train iters, % (↓)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Seen bad T, % (↓)	Div (↑)
itar	**50**	3,79	**1,02**	**90**	**100**	**0,76**
itar_0-0-1	85	**3,30**	0,81	35	275	0,66
itar_0-1-0	**60**	**3,31**	0,86	50	350	0,71
itar_0-1-1	85	3,31	0,93	50	325	0,71
itar_1-0-0	70	3,56	0,90	60	230	0,69
itar_1-0-1	90	3,65	0,95	75	200	0,72
itar_1-1-0	90	3,75	**1,05**	**95**	**95**	**0,75**

The effect of different parts of the ITAR model on the final result. Name format: “itar_[is there fixation of good topics]-[is there decorrelation with bad topics]-[is there decorrelation with good topics]”. “Train iters” is how many iterations the training took (as a percentage of the maximum number of iterations).

Ablation study

- Fixing good topics increases the proportion of good topics in the model
- + decorrelation with bad ones reduces the frequency of bad topics
- + decorrelation with good ones results in more diverse topics

Model	PostNauka (20 topics)					
	Train iters, % (↓)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Seen bad T, % (↓)	Div (↑)
itar	**50**	3,79	**1,02**	**90**	**100**	**0,76**
itar_0-0-1	85	**3,30**	0,81	35	275	0,66
itar_0-1-0	**60**	**3,31**	0,86	50	350	0,71
itar_0-1-1	85	3,31	0,93	50	325	0,71
itar_1-0-0	70	3,56	0,90	60	230	0,69
itar_1-0-1	90	3,65	0,95	75	200	0,72
itar_1-1-0	90	3,75	**1,05**	**95**	**95**	**0,75**

The effect of different parts of the ITAR model on the final result. Name format: “itar_[is there fixation of good topics]-[is there decorrelation with bad topics]-[is there decorrelation with good topics]”. “Train iters” is how many iterations the training took (as a percentage of the maximum number of iterations).

Ablation study

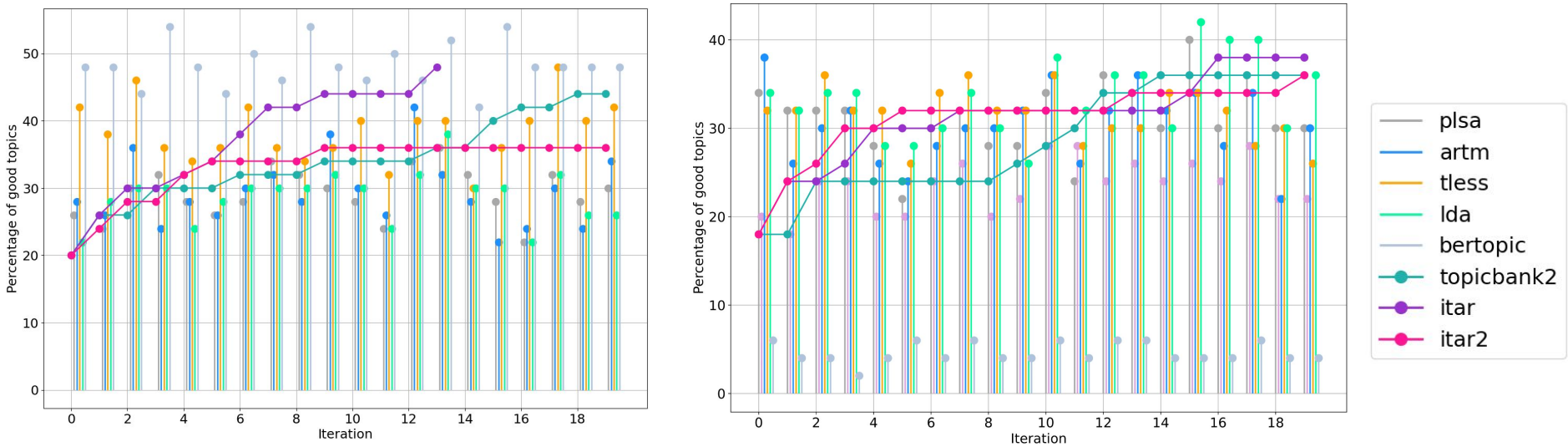
- Fixing good topics increases the proportion of good topics in the model
- + decorrelation with bad ones reduces the frequency of bad topics
- + decorrelation with good ones results in more diverse topics

Model	PostNauka (20 topics)					
	Train iters, % (↓)	PPL / 1000 (↓)	Coh (↑)	Good T, % (↑)	Seen bad T, % (↓)	Div (↑)
itar	**50**	3,79	**1,02**	**90**	**100**	**0,76**
itar_0-0-1	85	**3,30**	0,81	35	275	0,66
itar_0-1-0	60	**3,31**	0,86	50	350	0,71
itar_0-1-1	85	3,31	0,93	50	325	0,71
itar_1-0-0	70	3,56	0,90	60	230	0,69
itar_1-0-1	90	3,65	0,95	75	200	0,72
itar_1-1-0	90	3,75	**1,05**	**95**	**95**	**0,75**

The effect of different parts of the ITAR model on the final result. Name format: “itar_[is there fixation of good topics]-[is there decorrelation with bad topics]-[is there decorrelation with good topics]”. “Train iters” is how many iterations the training took (as a percentage of the maximum number of iterations).

Other topic goodness criterion

ITAR model may contain the comparable number of good[1] topics
if the quality of one topic *depends* on other topics



Percentage of good model topics depending on the iteration (↑).
ICD-10, models for 50 topics (left); 20Newsgroups, models for 50 topics (right).

[1] *Alekseev V.* Intra-Text Coherence as a Measure of Topic Models' Interpretability, 2018.

Conclusion

- An iteratively updated additively regularized topic model is presented (ITAR).
- It accumulates (fixes) good topics and filters out the bad ones.
- It outperforms all other models on several text collections in terms of good topics, with its topics being diverse and its perplexity moderate.
- Its learning process convergence depends on the criterion by which the goodness of topics is determined.

Possible directions for further research:

- Accelerate ITAR model training (ideally in a single iteration).
- Selecting good topics not by coherence, but somehow else (LLM-as-a-judge).
- Investigating whether it is possible to get 100% good topics.

Assets

- Paper draft (see the final version in the conference proceedings 🙂):

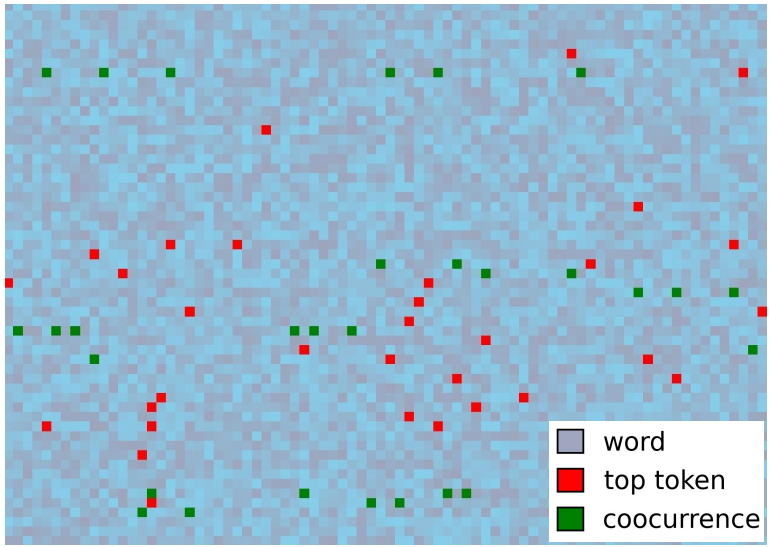
  https://arxiv.org/abs/2408.05840
- Code (will be here in a while 😅):

  https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics
- Datasets (all used in the experiments):

  https://huggingface.co/TopicNet

Appendix 1
Intra-Text Coherence

Problem of Classical Coherence (Newman, 2010)



$$\text{coh}(t) = \underset{w_i, w_j \in \text{top}_k(t)}{\text{mean}} \text{PMI}(w_i, w_j)$$

$$\text{PMI}(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

k topwords of the topic

probability of finding two words close to each other in text

- Ten most frequent words of the topic occupy a small proportion of the text.
- Their co-occurrences are even smaller.

Problem of Classical Coherence (Newman, 2010)

Only one of the top 10 tokens (“частиц”) of the topic is visible in the text fragment. All other words of the topic will be ignored by classical coherence.

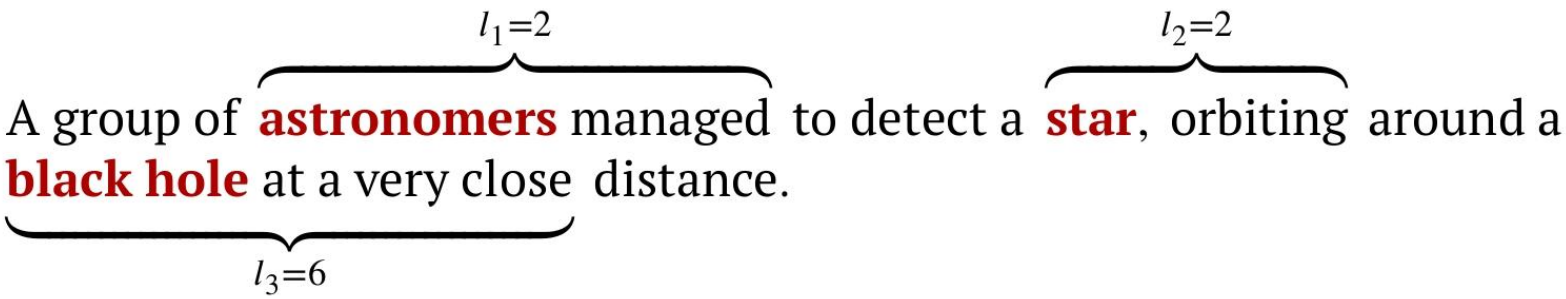
Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of *topic 3: физика* with top 10 in bold: **частица**, **электрон**, **кварк**, **атом**, **энергия**, **вселенная**, **фотон**, **физика**, **физик**, **эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

Intra-Text Coherence

Hypothesis about the segment structure of text: Words of a topic are distributed in the text not randomly, but in groups, *segments*.

Idea: Count words of the topic, penalizing when a word of another topic is encountered (thus estimating the *length* of the topic in the text).



Example of a text segment connected with topic about black holes.

Appendix 2
TopicBank

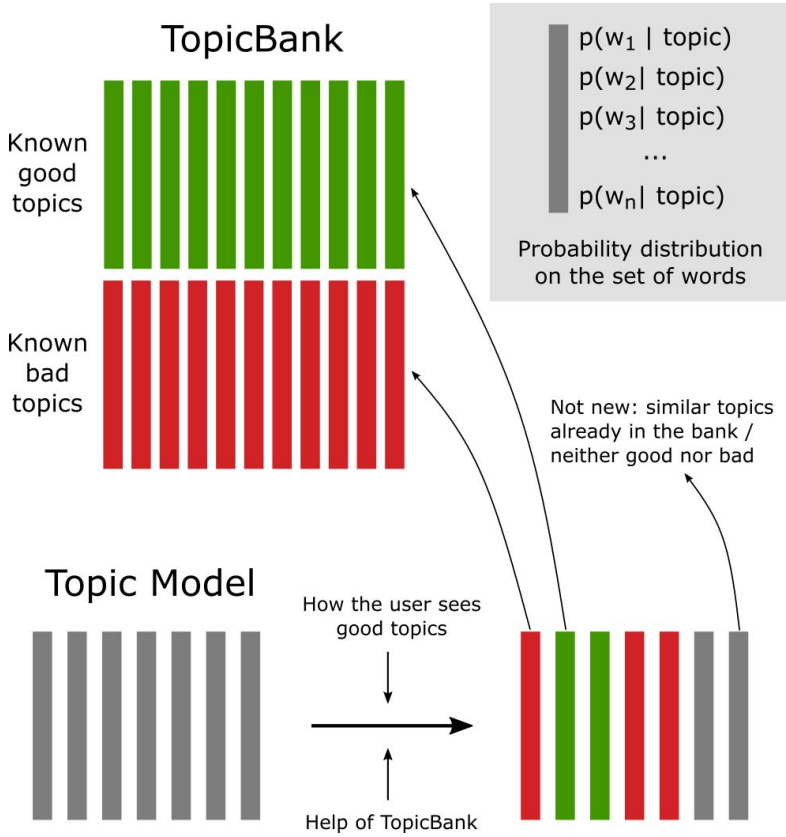
TopicBank: Collection of Coherent Topics

Idea:

- *Save* found good (and, optionally, bad) topics in the topic bank.
- Use topic bank to *validate* newly trained topic models.

New topic is added to the bank if it is:

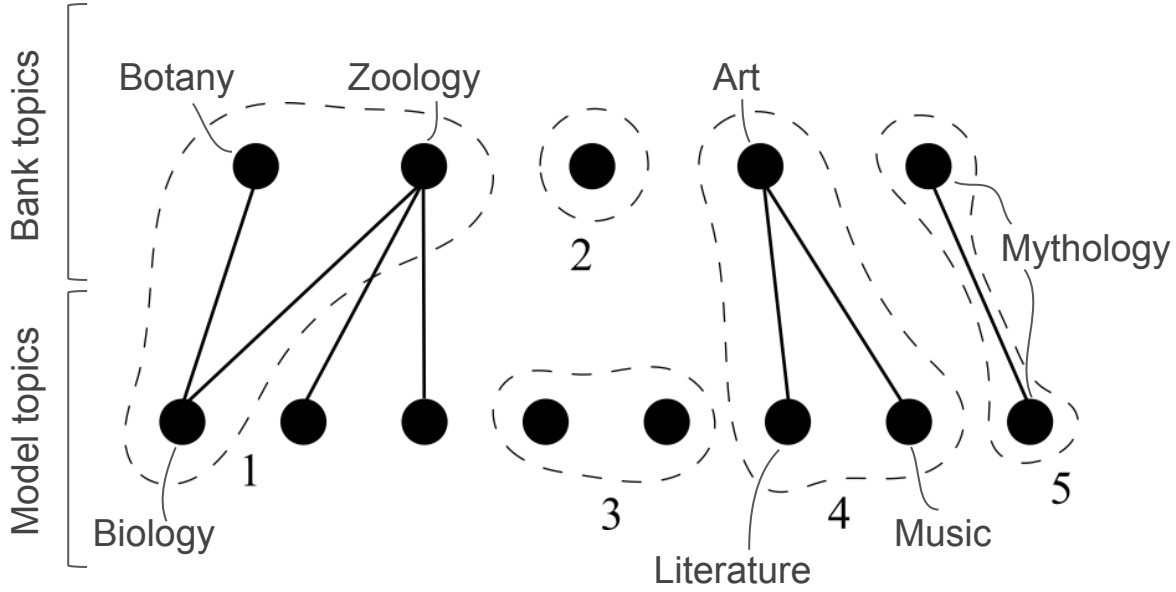
- *good* (in top of the model's topics in terms of coherence)
- *different* (high Jaccard distance to the nearest bank topic)



TopicBank Creation: Dependencies Between Topics

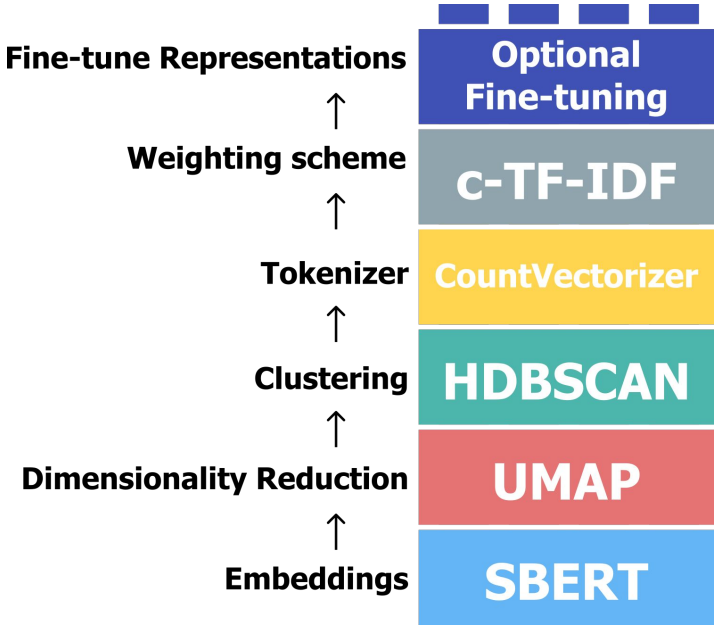
Possible relationship types between model topics and topics in the topic bank:

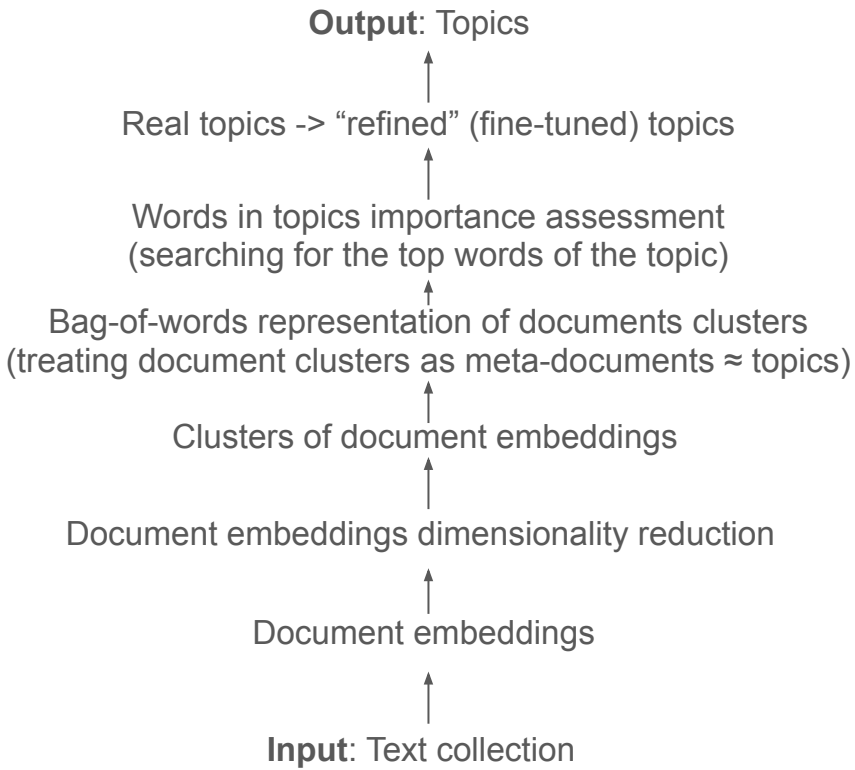
1) merging topics
2) no child topics
3) no parent topics
4) splitting topic
5) remaining topic



Appendix 3
BERTopic

BERTopic Model Architecture

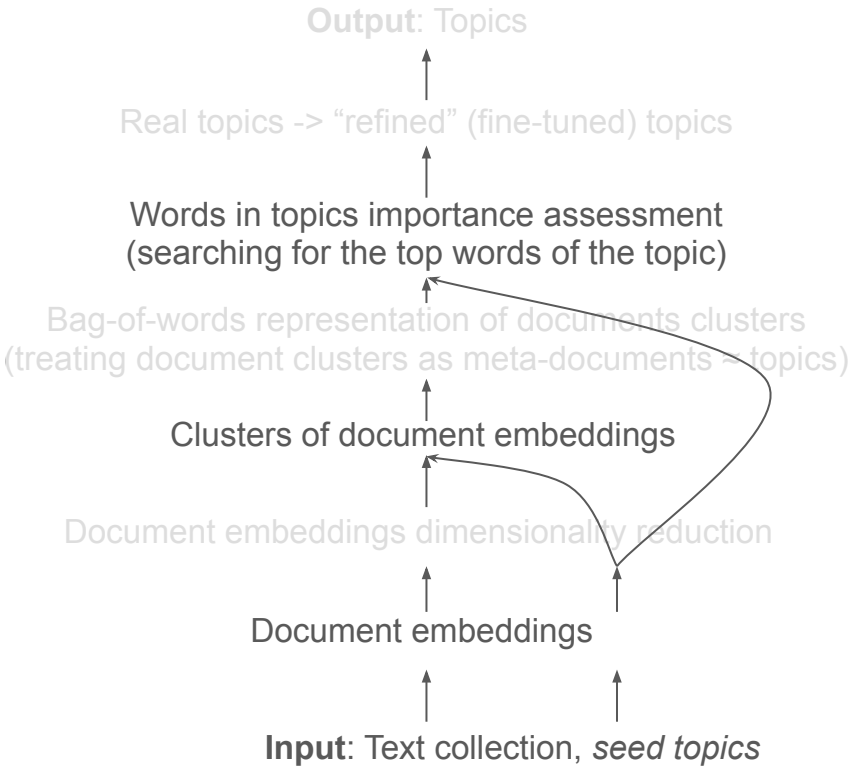




https://maartengr.github.io/BERTopic/algorithm/algorithm.html.

Guided Topic Modeling

«…BERTopic is more likely to model the defined seeded topics. *However, BERTopic is merely nudged towards creating those topics.* In practice, if the seeded topics do not exist or might be divided into smaller topics, *then they will not be modeled*. Thus, seed topics need to be accurate to accurately converge towards them.»



https://maartengr.github.io/BERTopic/getting_started/guided/guided.html.