TopicBank: Collection of Coherent Topics Using Multiple Model Training with Their Further Use for Topic Model Validation

Vasiliy Alekseev[1], Konstantin Vorontsov[2] et al.

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

MLIS 2023: The 5th International Conference on Machine Learning and Intelligent Systems

18 November, 2023

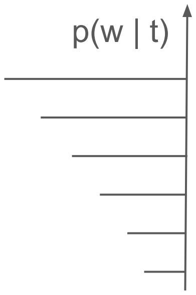
https://www.sciencedirect.com/science/article/abs/pii/S0169023X21000483

…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

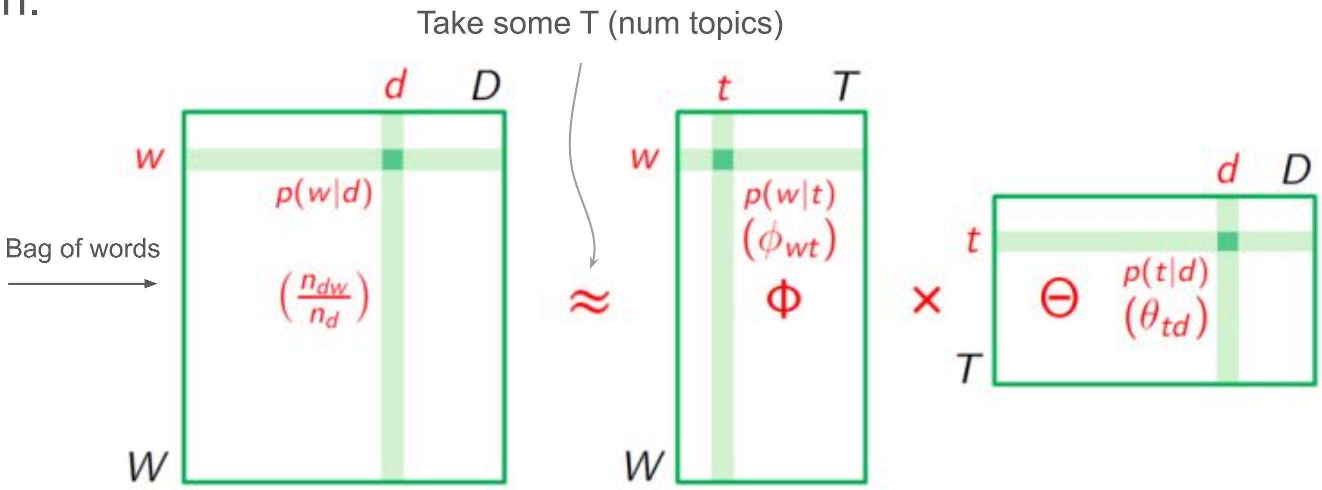
…The open country in the suburbs was quiet and deserted. Moreover, few would venture out into the snow at this time of the night. After leaving the house, Zhu Zhen looked back and saw no footprints. He then wended his way to Miss Zhou’s grave. …Unfortunately for him, the grave keepers had a dog. At this 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.



Text collection D (num docs), W (vocab size)	Matrix of word-in-document relative frequencies	Matrix of word-in-topic probabilities	Matrix of topic-in-document probabilities
Input		Output	

Konstantin Vorontsov, [Probabilistic Topic Modeling](#) (in Russian).

Problems of Topic Models

Topic model

Topics

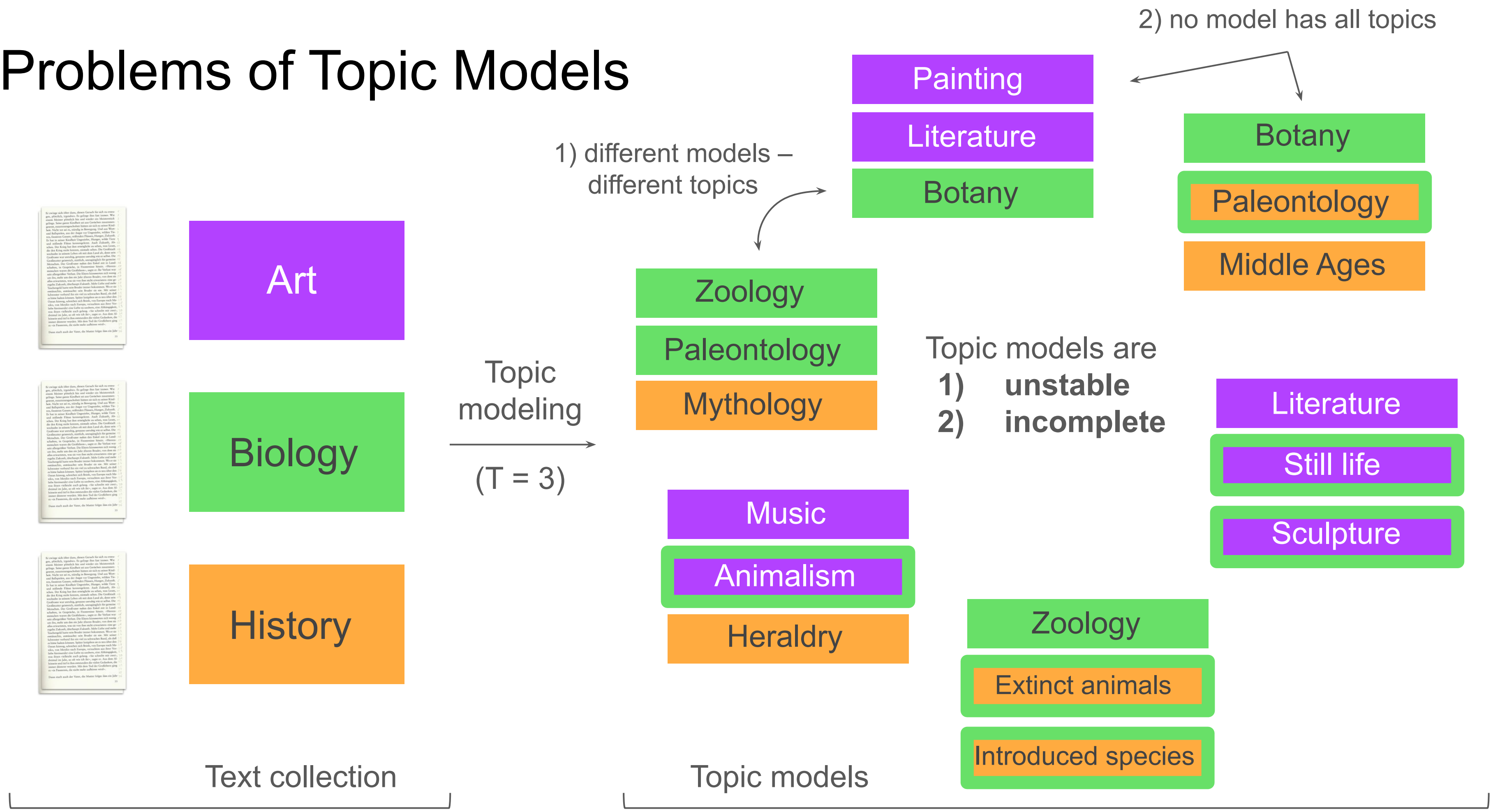
Interpretable

- china, portugal, casino, pataca, st. paul, serradura
- machine learning, intelligent systems, model, recognition, prediction, analysis
- autumn, yellow leaves, cool weather, wind, rain, school

Uninterpretable

- dinosaur, maths, sun, suspicion, quick, small
- i, she, go, to, take, with, call, say
- teacher, teach, school, taught, teachers, lesson

Problems of Topic Models



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

Typical Topic Modeling Experiment Pipeline

```
while not is_good(topic_model):

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

Typical Topic Modeling Experiment Pipeline

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

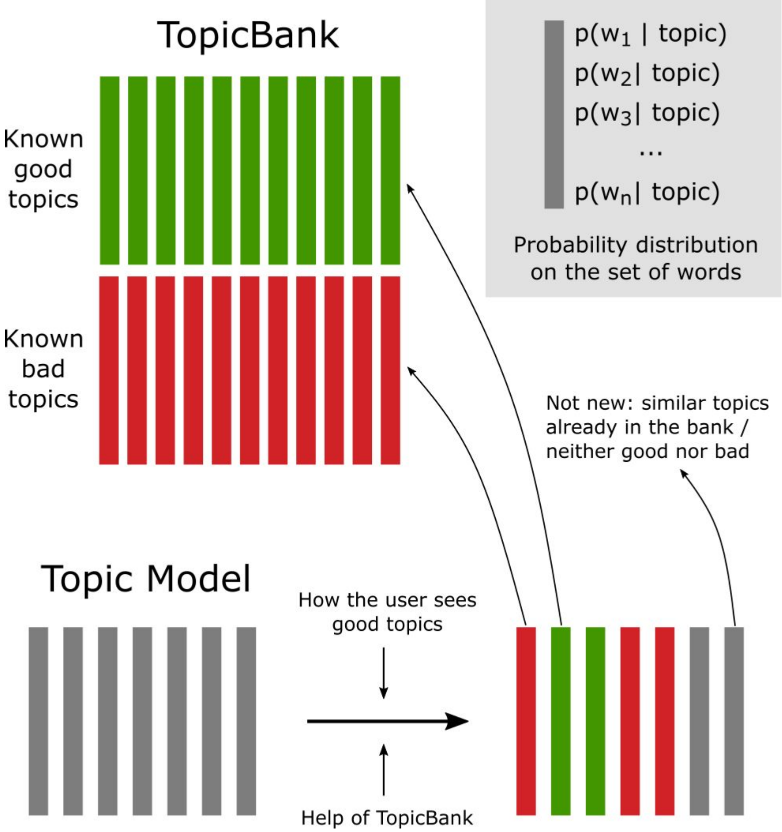
TopicBank: Collection of Coherent Topics

Problem:

- *Huge* number of experiments to find best topic model.
- Found good topics may be *lost*.

Solution:

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



Proposed Methodology

```
for i in range(N):
    set_parameters(topic_model)
    train(topic_model, dataset)
    good_topics = analyze_topics(topic_model)
    add_topics(topic_bank, good_topics)
```

TopicBank creation:

- Input: dataset
- Output: topic bank

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

TopicBank application:

- Input: dataset, topic bank
- Output: topic model (best)

```
assess_quality(best_topic_model, human)
analyze_topics(best_topic_model, human)
```

Proposed Methodology

```
for i in range(N):
    set_parameters(topic_model)
    train(topic_model, dataset)
    good_topics = analyze_topics(topic_model)
    add_topics(topic_bank, good_topics)
```

TopicBank creation:

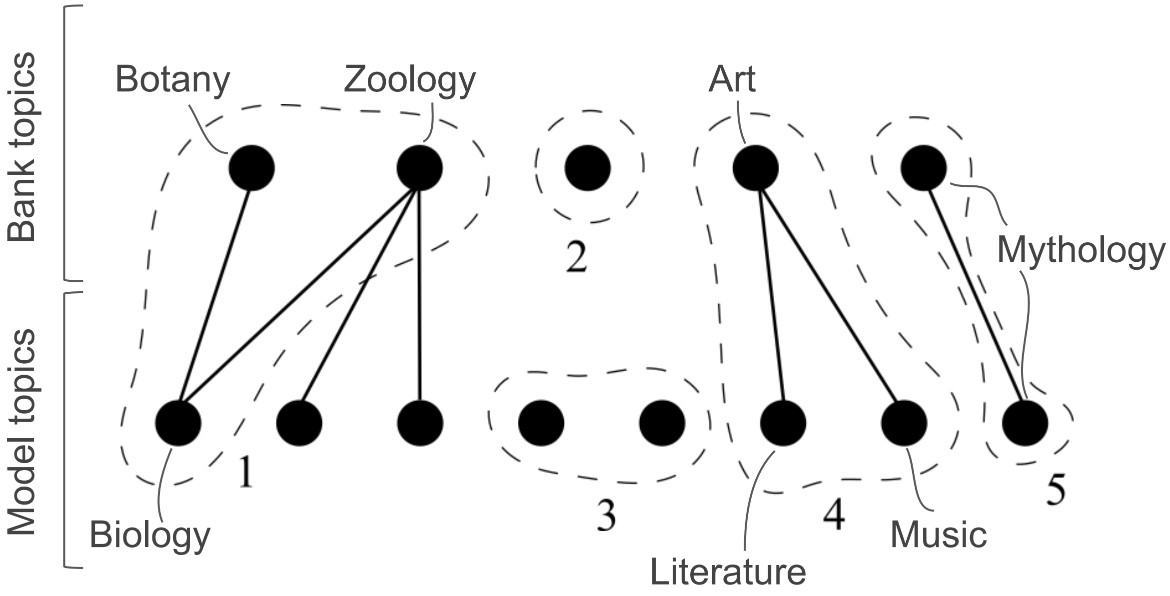
- Input: dataset
- Output: topic bank

- Automatic or semi-automatic evaluation of the quality of new topics (*topic coherence*).
- Evaluation of the *dependencies* between new topics and the topics of the topic bank (*two-level hierarchical topic model*).
- Good topics *can be added* to the topic bank if the topics of the topic bank remain *different*.

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



Proposed Methodology

```
for i in range(N):
    set_parameters(topic_model)
    train(topic_model, dataset)
    good_topics = analyze_topics(topic_model)
    add_topics(topic_bank, good_topics)
```

TopicBank creation:

- Input: dataset
- Output: topic bank

Topic model topics are compared with the topics stored in the topic bank.

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

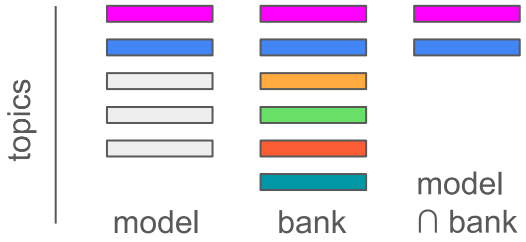
TopicBank application:

- Input: dataset, topic bank
- Output: topic model (best)

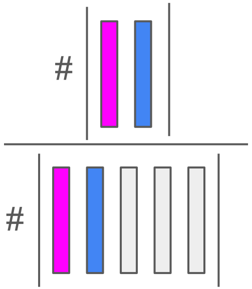
```
assess_quality(best_topic_model, human)
analyze_topics(best_topic_model, human)
```

TopicBank as Intrinsic Quality Measure

- The more the model managed to find good topics, the better.
- The distance between topics is calculated as jaccard distance.



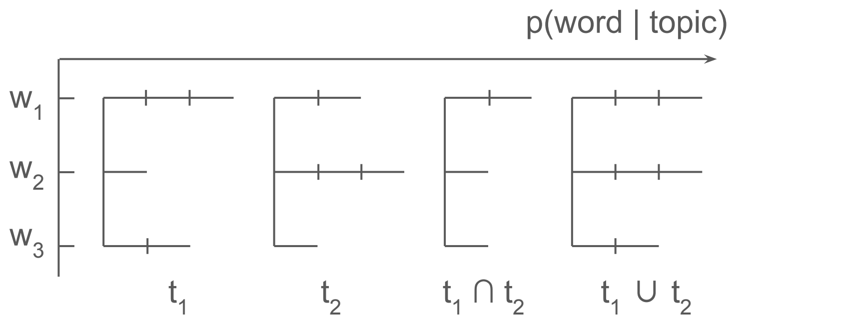
$$\text{quality(model)} = \frac{|\text{model} \cap \text{bank}|}{|\text{model}|} =$$



$$\text{sim}(t_1, t_2) = \frac{|t_1 \cap t_2|}{|t_1 \cup t_2|} =$$

TopicBank as Intrinsic Quality Measure

- The more the model managed to find good topics, the better.
- The distance between topics is calculated as jaccard distance.



$$\mathrm{sim}(t_1, t_2) = \frac{|\, t_1 \cap t_2 \,|}{|\, t_1 \cup t_2 \,|} =$$

Experiment

Goal:
Understand if the topic bank can be used to *assess the quality* of topic models.

Task:
Check if the topic bank allows to *find the best model* from a fixed set of models.

Plan:

- Take several text collections.
- Create a topic bank for each text collection.
- Take a set of topic models.
- Evaluate the quality of topic models on all datasets (using topic banks).

Models

- **PLSA**: a simple topic model without any hyperparameters aside from T.
- **LDA**: a well-known topic model, having priors for Φ and ϴ distributions.
- **ARTM**: a PLSA extension which can obtain topics with desired qualities.
- **Arora**, **CDC**: topic models with specific topic distributions initialization.

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.
Arora S. et al. Computing a nonnegative matrix factorization – provably, 2012.
Dobrynin V., Patterson D., Rooney N. Contextual document clustering, 2004.

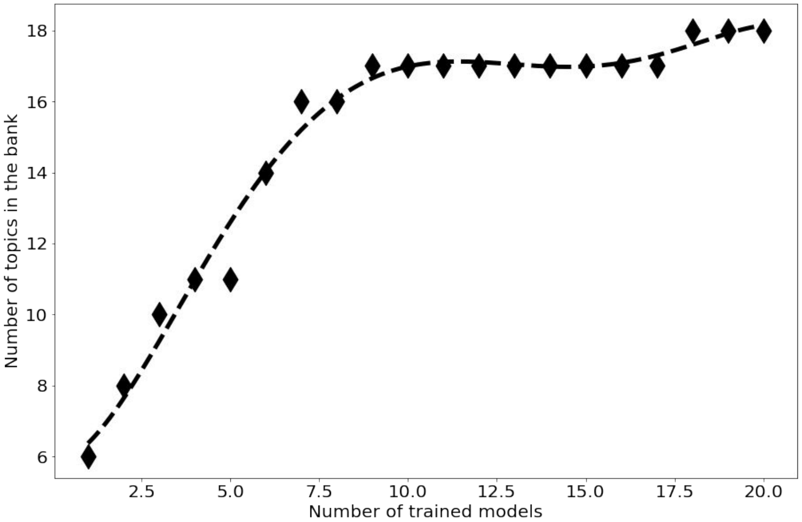
Datasets

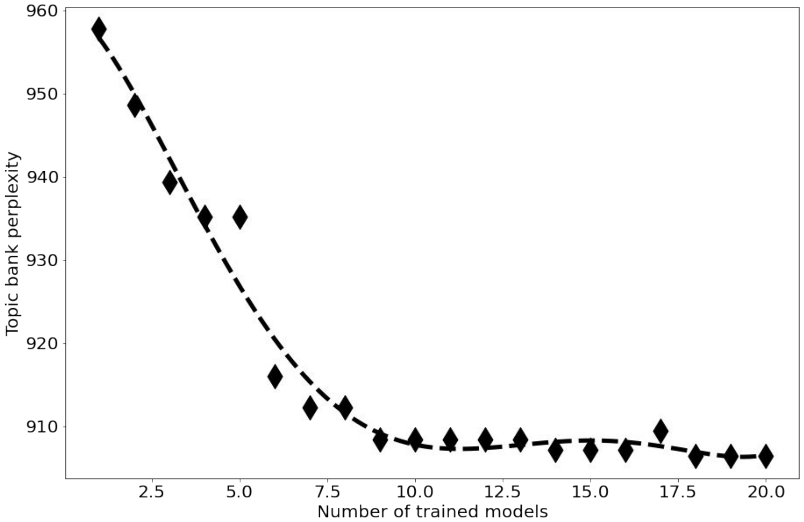
Name	$\|D\|$	Language
PostNauka	3 446	Russian
Reuters	10 788	English
Brown	500	English
20 NG	18 846	English
AG News	127 600	English
Watan2004	20 291	Arabic
Habrahabr	133 978	Russian

Datasets used in the experiments (|D| is the number of documents in a dataset).
Preprocessing: lemmatization, stop-words removal.

Results

The process of bank creation *reaches saturation*: no more new topics are added.



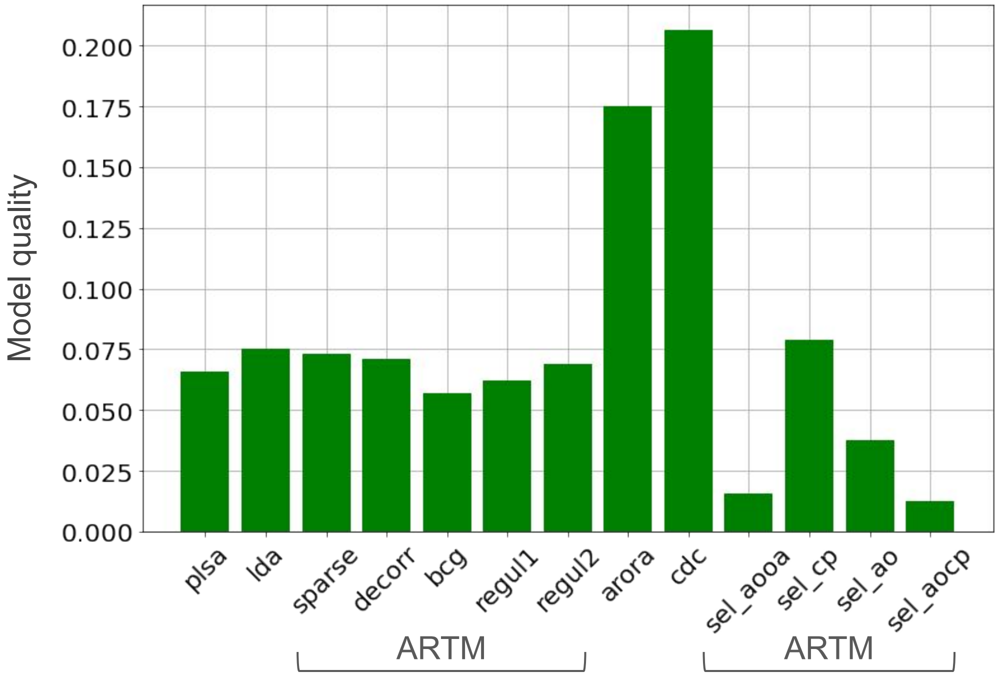


Some characteristics of the topic bank depending on the number of trained topic models: number of topics in the topic bank (left); perplexity of the topic bank as a topic model (right; the lower the better).

Results

TopicBank managed to find the topic models with the *largest number of interpretable topics* (Arora and CDC).

Averaged over datasets model quality estimates calculated using topic banks. Horizontal axis is topic model. Vertical axis is an average proportion of model's good topics calculated with the help of topic banks (the higher, the better).



Conclusion

- TopicBank is introduced which is a “wrapper” over topic modeling that should *accelerate the validation* of newly trained topic models.
- Algorithm for *automatically creating* a topic bank for a given text collection is proposed.
- Experiment was conducted on real data, confirming the possibility of using TopicBank to *assess the quality of topic models*.

Possible future directions:

- Validate neural topic models with TopicBank.
- Investigate the possibility for faster TopicBank creation.
- Require that TopicBank itself should be a good topic model (low perplexity).

Publication: Alekseev V. et al. TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation //*Data & Knowledge Engineering*. – 2021. – Vol. 135. – p. 101921. – https://doi.org/10.1016/j.datak.2021.101921.

Code: https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.