Итеративное улучшение аддитивно регуляризованной тематической модели

Алексей Горбулев[1], **Василий Алексеев**[1], Константин Воронцов[2]

[1] Московский физико-технический институт,
[2] Московский государственный университет имени М. В. Ломоносова

66-я Всероссийская научная конференция МФТИ

6 апреля 2024

…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…

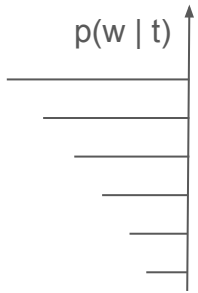
The Fan Tower Restaurant as Witness to the Love of Zhou Shengxian,
from Stories to Awaken the World (Xingshi Hengyan); translated by Shuhui Yang and Yunqin Yang

Ещё не кончилась вторая стража и на улицах было людно, но за городской стеной, на открытом месте, стояла мёртвая тишина, снег падал всё гуще, и прогулка за воротами никого не соблазняла. Пройдя несколько шагов, Чжу Чжэнь оглянулся: всё хорошо, следов не видно. То и дело озираясь, прокрался он на кладбище и перелез через ограду вокруг могилы девицы Чжоу. Но вот беда – смотрители кладбища держали собаку. Когда Чжу Чжэнь перелезал через ограду, собака его учуяла и, выскочив из конуры, залилась истошным лаем. Чжу Чжэнь, однако же, предусмотрительно запасся лепёшкою, начинённой ядом. Как только раздался лай, он бросил за ограду лепёшку. Собака подбежала, понюхала лепёшку и мигом проглотила. Ещё миг – и она взвизгнула, опрокинулась на спину и околела. Чжу Чжэнь подступил к могиле…

«Дважды умершая», повесть из сборника «*Слово вечное, мир пробуждающее*».
Перевод с китайского: Дмитрий Воскресенский.

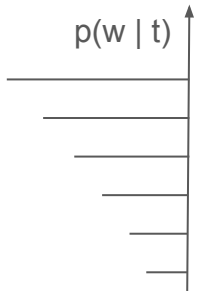
…The open country in the suburbs was quiet and deserted. Moreover, few would venture out into the snow at this time of the night. After leaving the house, Zhu Zhen looked back and saw no footprints. He then wended his way to Miss Zhou's grave. …Unfortunately for him, the grave keepers had a dog. At this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



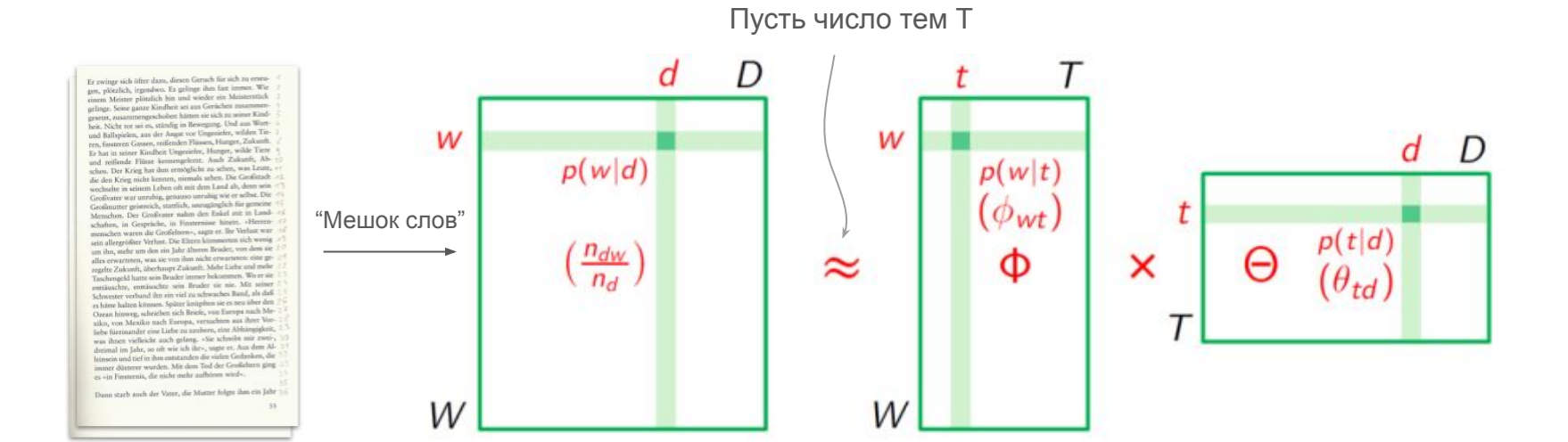
Ещё не кончилась вторая стража и на улицах было людно, но за городской стеной, на открытом месте, стояла мёртвая тишина, снег падал всё гуще, и прогулка за воротами никого не соблазняла. Пройдя несколько шагов, Чжу Чжэнь оглянулся: всё хорошо, следов не видно. То и дело озираясь, прокрался он на кладбище и перелез через ограду вокруг могилы девицы Чжоу. Но вот беда – смотрители кладбища держали собаку. Когда Чжу Чжэнь перелезал через ограду, собака его учуяла и, выскочив из конуры, залилась истошным лаем. Чжу Чжэнь, однако же, предусмотрительно запасся лепёшкою, начинённой ядом. Как только раздался лай, он бросил за ограду лепёшку. Собака подбежала, понюхала лепёшку и мигом проглотила. Ещё миг – и она взвизгнула, опрокинулась на спину и околела. Чжу Чжэнь подступил к могиле…

Природа	Зимняя ночь	Приключение	Воровство	Кладбище	Собаки	Еда	Яды
небо	снег	опасность	вор	могила	собака	рис	яд
улица	ночь	риск	ограда	мёртвый	лай	лепёшка	лекарство
трава	холод	след	красть	смотритель	друг	проглотить	больной
воздух	снежинка	соблазнить	деньги	гроб	учуять	нюхать	страдать
прогулка	тишина	стража	прокрасться	склеп	конура	котлета	околеть
городской	пустынно	девица	опасность	ночь	ошейник	гречка	змея

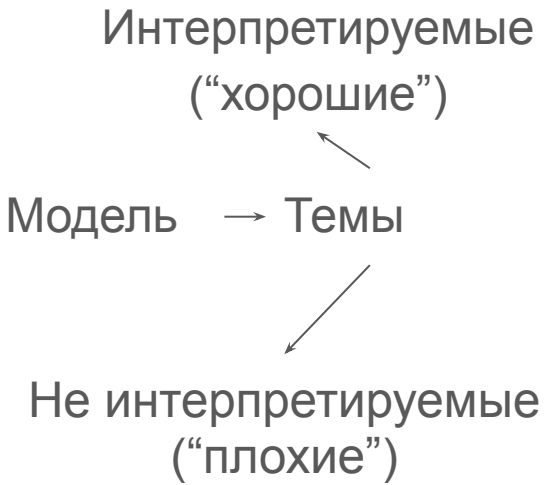


Тематическое моделирование

В текстовой коллекции содержится набор *скрытых тем*.



Проблема тематических моделей



- мфти, конференция, секция, тезисы, рецензирование, льгота, магистратура
- машинное обучение, интеллектуальные системы, модель, распознавание, предсказание
- весна, птица, солнце, тепло, температура, рекорд

- динозавр, математика, луна, подозрение, быстрый
- я, она, идти, в, взять, с, позвать, говорить
- учитель, учить, школа, учил, учителя, урок

Схема типичного эксперимента

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

Схема типичного эксперимента

```
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)
```

Схема типичного эксперимента

```
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)
```

Схема типичного эксперимента

```
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)
```

Схема типичного эксперимента

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

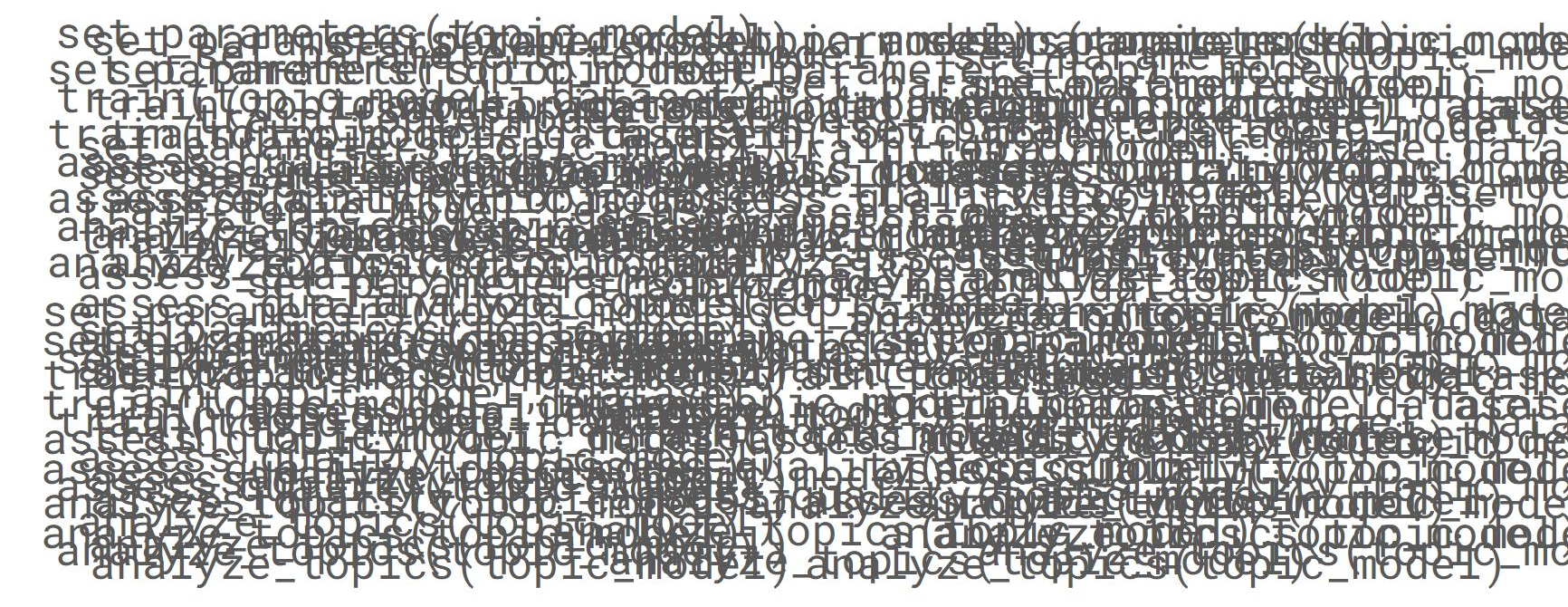


Схема типичного эксперимента

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

Итеративное улучшение

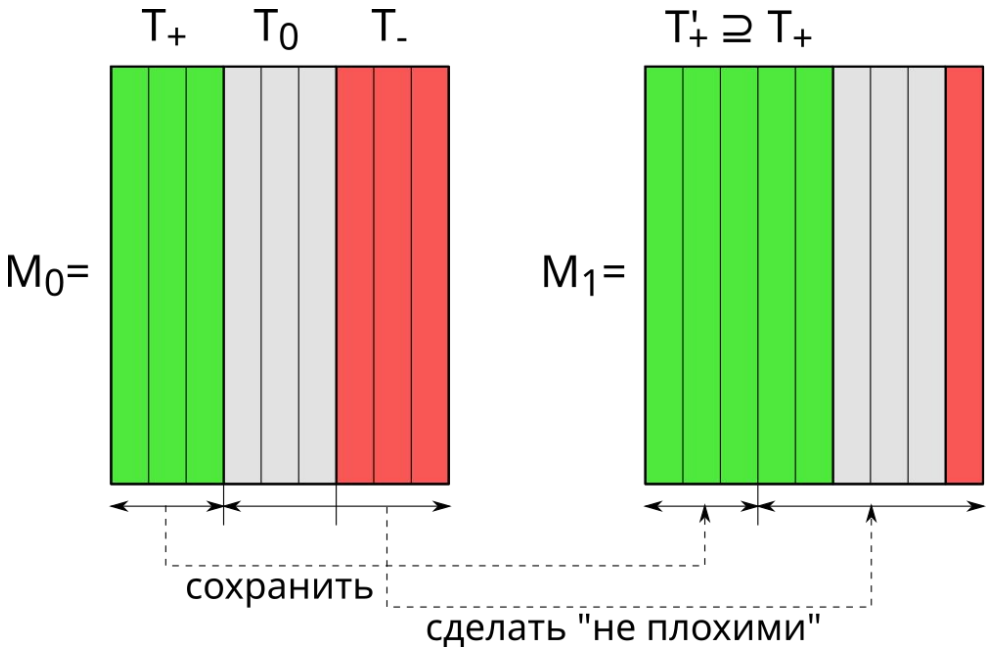
Итеративное улучшение тематической модели

Проблема:

- *Много* экспериментов, чтобы найти хорошую модель.
- Найденные хорошие темы *теряются*.

Решение:

- *Фиксировать* хорошие темы.
- Свободные темы учить *непохожими* на плохие.



Аддитивная регуляризация

Максимизация регуляризованного логарифма правдоподобия:

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.

Итеративная аддитивная регуляризация

Максимизация регуляризованного логарифма правдоподобия:

ITAR:

отложенные темы

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

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

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

Эксперимент

Цели:

- Проверить, что число хороших тем итеративно увеличивается.
- Сравнить по числу хороших тем с другими тематическими моделями.

Основные моменты:

- “Итерация” — обучение одной модели.
- Хорошие темы — темы с высокой когерентностью.
- Несколько тематических моделей.
- Несколько текстовых коллекций.
- Несколько итеративных обучений для каждой модели.
- Итоговая итеративная модель — последняя модель, итоговая не итеративная модель — лучшая по числу хороших тем из всей серии.

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

Модели

- **PLSA**: модель с единственным гиперпараметром T.
- **LDA**: модель, где столбцы Φ и Θ порождаются распределениями Дирихле.
- **ARTM**: модель с аддитивной регуляризацией.
- **TLESS**: модель без матрицы Θ, с разреженными темами.
- **BERTopic**: нейросетевая тематическая модель.
- **TopicBank**: итеративно обновляемая модель без регуляризаторов.

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 (2021). TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation, 2021.

Данные

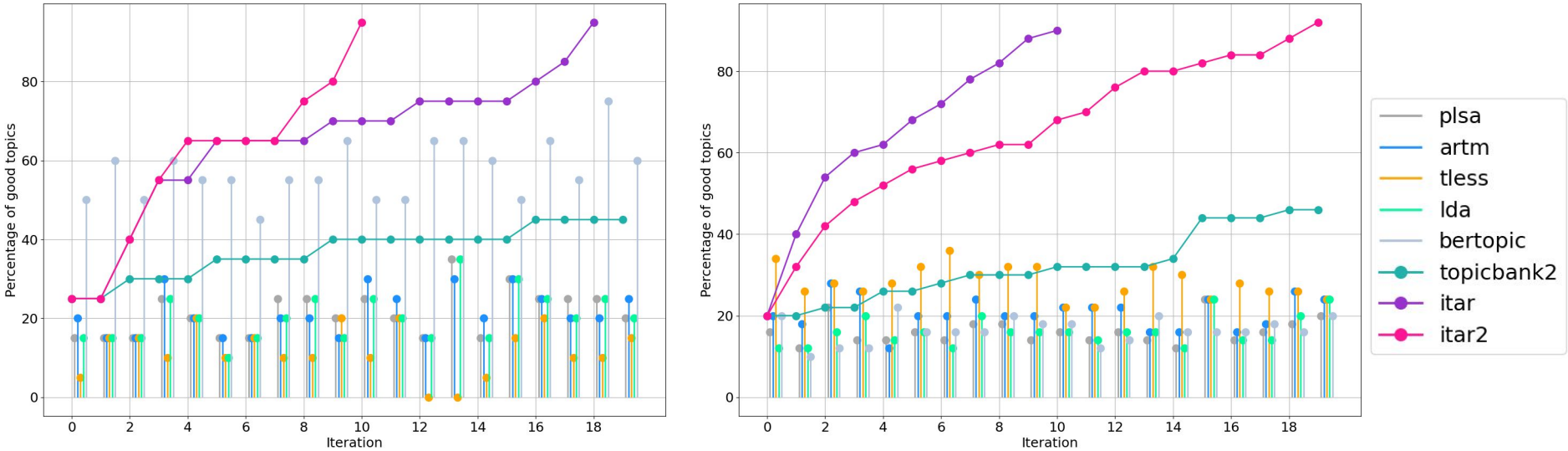
Dataset	D	Len	W	Lang
PostNauka	3404	421,2	19186	Ru
$20Newsgroups_{train}$	11301	93,9	52744	En
RuWiki-Good	8603	1934,6	61688	Ru
RTL-Wiki-Person	1201	1600,1	37739	En
ICD-10	2036	550,0	22608	Ru

Датасеты, которые использовались в экспериментах (D — количество документов, Len — средняя длина документа). Предобработка: лемматизация, удаление стоп-слов, “мешок слов”.

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.

Результаты

- Итеративная модель содержит больше всего хороших тем
- При этом хорошие больше 80% тем модели



Процент хороших тем модели в зависимости от итерации (↑).
RuWiki-Good, модели на 20 тем (слева); 20Newsgroups, модели на 50 тем (справа).

Результаты

- Наибольший % хороших тем
- Темы различны
- Умеренная перплексия

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**

Некоторые свойства итоговых моделей: перплексия (PPL), средняя когерентность тем (Coh), процент интерпретируемых тем (Good T), различность тем (Div).
PostNauka, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа).

Результаты

- Наибольший % хороших тем • Темы различны • Умеренная перплексия

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**

Некоторые свойства итоговых моделей: перплексия (PPL), средняя когерентность тем (Coh), процент интерпретируемых тем (Good T), различность тем (Div).
PostNauka, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа).

Результаты

- Наибольший % хороших тем
- Темы различны
- Умеренная перплексия

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**

Некоторые свойства итоговых моделей: перплексия (PPL), средняя когерентность тем (Coh), процент интерпретируемых тем (Good T), различность тем (Div).
PostNauka, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа).

Результаты

- Наибольший % хороших тем
- Темы различны
- Умеренная перплексия

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**

Некоторые свойства итоговых моделей: перплексия (PPL), средняя когерентность тем (Coh), процент интерпретируемых тем (Good T), различность тем (Div).
PostNauka, модели на 20 тем (слева); RuWiki-Good, модели на 50 тем (справа).

Ablation study

- Фиксация хороших тем увеличивает перплексию
- + декорреляция с плохими снижает частоту появления плохих тем
- + декорреляция с хорошими приводит в более различным темам

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**

Влияние разных частей ITAR модели на итоговый результат. Формат имени: “itar_[есть ли фиксация хороших]-[есть ли декорреляция с плохими]-[есть ли декорреляция с хорошими]”. Train iters — сколько итераций заняло обучение (в процентах от максимального числа в 20 итераций).

Ablation study

- Фиксация хороших тем увеличивает долю хороших тем в модели
- + декорреляция с плохими снижает частоту появления плохих тем
- + декорреляция с хорошими приводит в более различным темам

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**

Влияние разных частей ITAR модели на итоговый результат. Формат имени: “itar_[есть ли фиксация хороших]-[есть ли декорреляция с плохими]-[есть ли декорреляция с хорошими]”. Train iters — сколько итераций заняло обучение (в процентах от максимального числа в 20 итераций).

Ablation study

- Фиксация хороших тем увеличивает долю хороших тем в модели
- + декорреляция с плохими снижает частоту появления плохих тем
- + декорреляция с хорошими приводит в более различным темам

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**

Влияние разных частей ITAR модели на итоговый результат. Формат имени: “itar_[есть ли фиксация хороших]-[есть ли декорреляция с плохими]-[есть ли декорреляция с хорошими]”. Train iters — сколько итераций заняло обучение (в процентах от максимального числа в 20 итераций).

Ablation study

- Фиксация хороших тем увеличивает долю хороших тем в модели
- + декорреляция с плохими снижает частоту появления плохих тем
- + декорреляция с хорошими приводит в более различным темам

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**

Влияние разных частей ITAR модели на итоговый результат. Формат имени: “itar_[есть ли фиксация хороших]-[есть ли декорреляция с плохими]-[есть ли декорреляция с хорошими]”. Train iters — сколько итераций заняло обучение (в процентах от максимального числа в 20 итераций).

Заключение

- Представлена итеративно обновляемая аддитивно регуляризованная тематическая модель, накапливающая (фиксирование тем) и “ищущая” (декоррелирование с отложенными темами) хорошие темы.
- Проведены эксперименты по сравнению с другими тематическими моделями, на нескольких коллекциях текстов естественного языка.
- По количеству хороших тем новая модель превосходит все остальные, при этом её темы различны, а перплексия умеренная.

Возможные направления дальнейших исследований:

- Ускорение обучения модели (в идеале — за одну итерацию).
- Увеличение средней когерентности накапливаемых в модели тем.
- Отбор хороших тем не по когерентности, а как-то ещё.
- Снижение перплексии / исследование вопроса о возможности получить 100% хороших тем.