Ансамблевые алгоритмы обучения устойчивых и полных тематических моделей

Алексеев Василий Антонович

Специальность 1.2.1 — «Искусственный интеллект и машинное обучение»
Степень — кандидат технических наук

Научный руководитель:
д-р физ.-мат. наук
Воронцов Константин Вячеславович

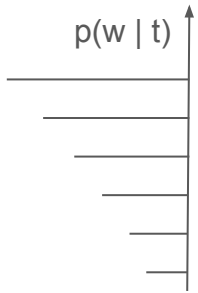
10 октября 2024

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.

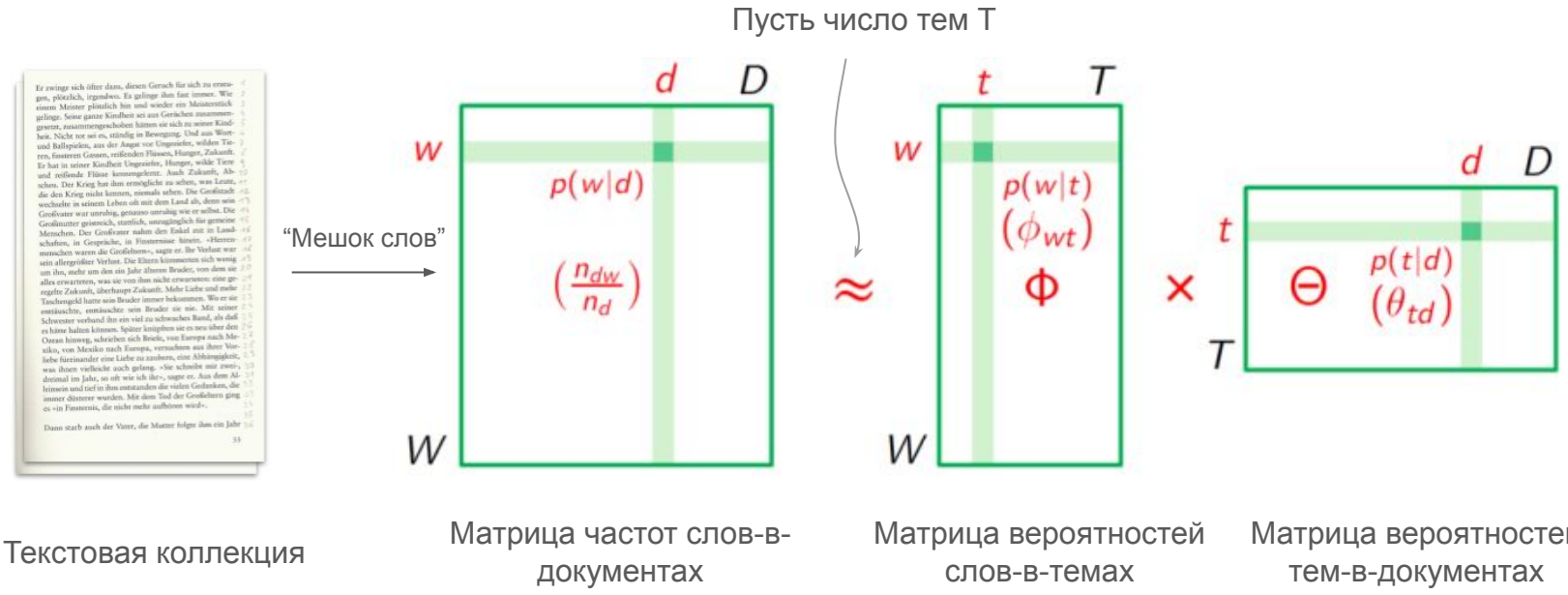
Ещё не кончилась вторая стража и на улицах было людно, но за городской стеной, на открытом месте, стояла мёртвая тишина, снег падал всё гуще, и прогулка за воротами никого не соблазняла. Пройдя несколько шагов, Чжу Чжэнь оглянулся: всё хорошо, следов не видно. То и дело озираясь, прокрался он на кладбище 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*.



Задача тематического моделирования

Дано:

- D — коллекция текстов
- W — множество слов, встречающихся в текстах
- n_{dw} — абсолютная частота слова w в документе d

Найти:

- множество скрытых тем T как распределения p(w | t)
- распределения тем в документах p(t | d)

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

Задача тематического моделирования

Критерий: максимум лог-правдоподобия с регуляризаторами:

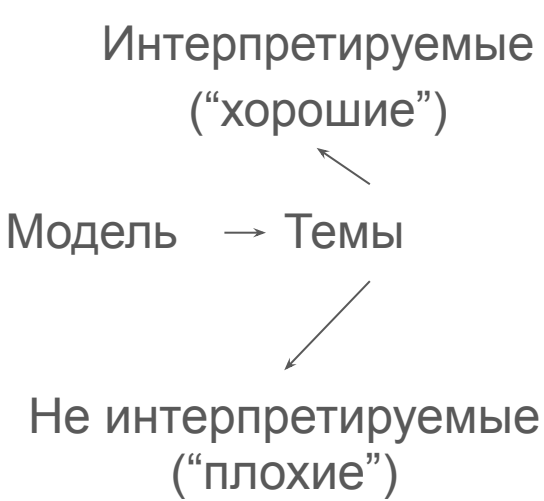
$$\underbrace{\ln p(\Phi,\Theta)}_{\mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n}\tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

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

Решение: методом простой итерации (Воронцов, 2014):

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

Проблемы тематических моделей



- диссер, автореферат, отзыв, рецензия, антиплагиат
- машинное обучение, интеллектуальные системы, модель, распознавание, предсказание
- осень, холод, дождь, лужи, палитра листьев

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

Проблемы тематических моделей

Искусство

Биология

История

Текстовая коллекция

TM

(|T| = 3)

Зоология

Бактериология

Мифология

Музыка

Анималистика

Геральдика

1) разные модели – разные темы

Рисование

Литература

Ботаника

2) ни одна модель не содержит всех тем

Ботаника

Палеонтология

Средние века

Тематические модели:

1) **неустойчивы**
2) **неполны**

Литература

Натюрморт

Скульптура

Зоология

Вымершие животные

Интродукция

Тематические модели

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):
```

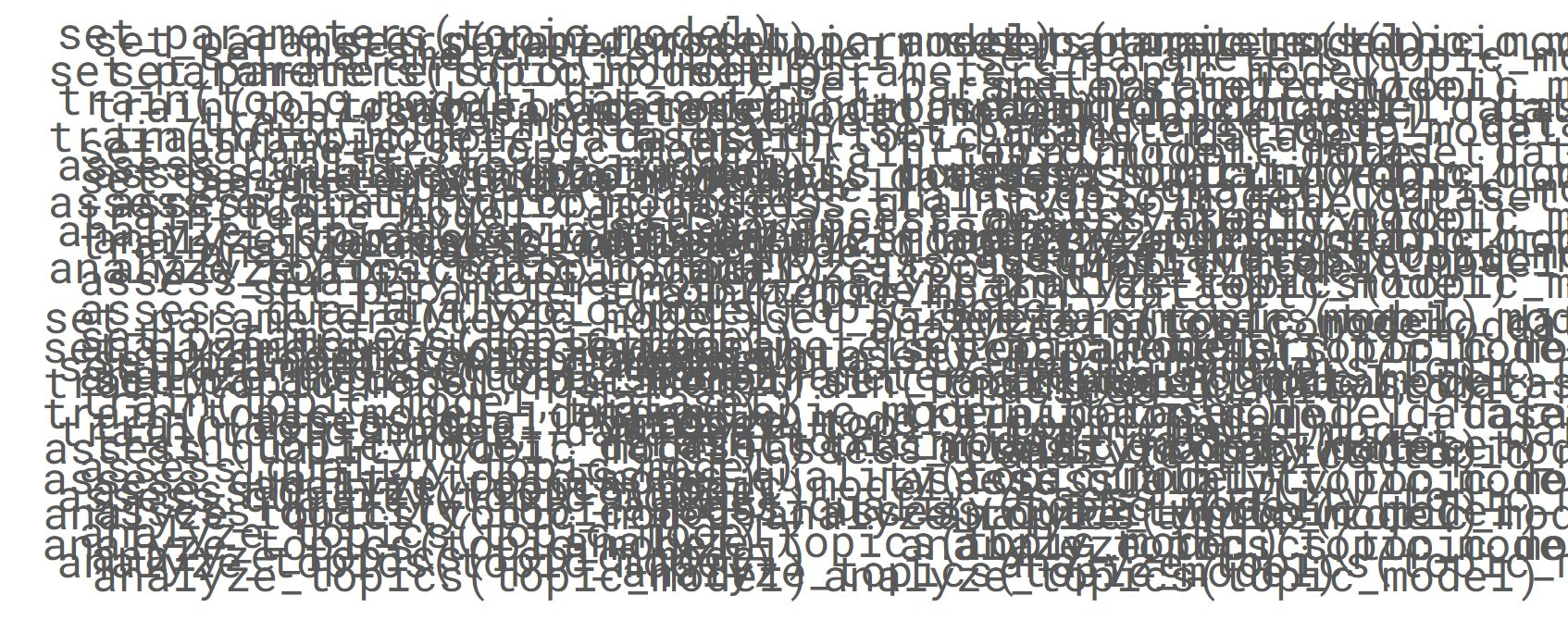


Схема типичного эксперимента

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

Автоматическая оценка качества тем

+

Итеративное улучшение модели

Цель и задачи диссертационного исследования

Цель: разработка комплекса программ, позволяющего с помощью множественного обучения тематических моделей получать полные и устойчивые тематические модели.

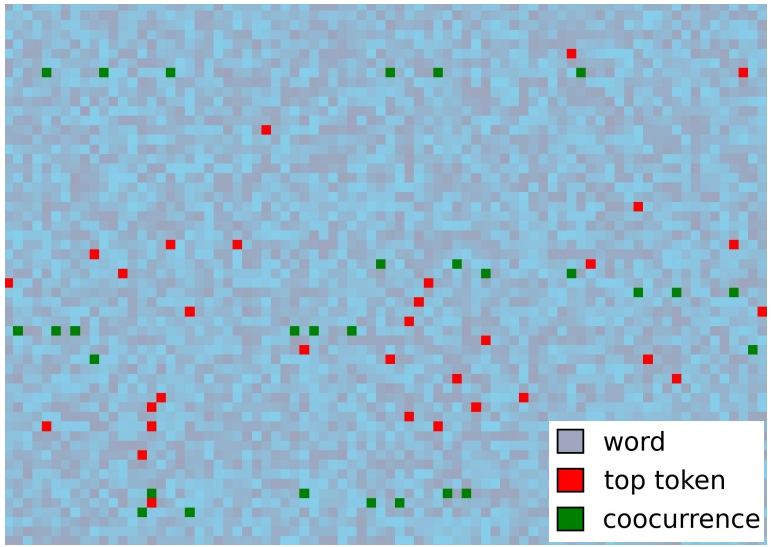
Задачи:

1. Разработать автоматический способ оценки интерпретируемости темы, учитывающий распределение всей темы в тексте.
2. Разработать комплекс программ для автоматической оценки качества тематических моделей по большому числу внутренних критериев.
3. Исследовать возможность поиска оптимального значения числа тем в модели с помощью внутренних критериев качества.
4. Разработать автоматический способ валидации тематических моделей с учётом их неустойчивости и неполноты.
5. Предложить и реализовать регуляризаторы в рамках АРТМ, предназначенные для улучшения тематической модели в процессе множественного обучения.

Задача

1. Разработать автоматический способ оценки интерпретируемости темы, учитывающий распределение всей темы в тексте

Проблема классической когерентности (Newman, 2010)



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

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

k топ-слов темы

вероятность встретить в тексте два слова рядом

- Десять самых частых слов темы занимают малую долю текста.
- Их совстречаемости — ещё меньше.

Проблема классической когерентности (Newman, 2010)

Во фрагменте текста виден *лишь один* из 10 топ-токенов (“частиц”). Все другие слова темы будут проигнорированы классической когерентностью.

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

Эксперимент 1

Цель: найти когерентность, лучше всего согласующуюся с гипотезой о сегментной структуре текста.[1]

Критерий: корреляция Спирмена между значениями когерентности и качества сегментации моделями текста с известной разметкой.

Данные:

- Коллекция статей ПостНауки.
- Полусинтетический датасет, составленный из сегментов.

[1]*Гипотеза о сегментной структуре текста*: слова темы распределены в тексте не случайно, а группами, сегментами.

Результаты 1

- Предложен автоматический метод сравнения когерентностей — по корреляции с качеством сегментации размеченного текста.
- Предложена внутритекстовая когерентность, учитывающая распределение всей темы в тексте.
- Внутритекстовая когерентность превосходит когерентность по совстречаемостям том-слов.

	Coh	Corr
	Newman	0.80
	Mimno	0.94
Внутритекстовые	SC L2	0.70
	SC Cos	−0.97
	SC Var	**1.00**
	TopLen	**1.00**
	FoCon	**1.00**

Корреляции Спирмена между когерентностями и качеством сегментации.

Vasiliy Alekseev, Victor Bulatov, Konstantin Vorontsov. Intra-Text Coherence as a Measure of Topic Models' Interpretability, 2018.

Задачи

2. Разработать комплекс программ для автоматической оценки качества тематических моделей по большому числу внутренних критериев.
3. Исследовать возможность поиска оптимального значения числа тем в модели с помощью внутренних критериев качества.

Число тем

Искусство

Биология

История

T

Число тем?

Искусство

Биология

История

T

Рисование

Литература

Музыка

Ботаника

Зоология

Бактериология

Мифология

Средние века

Геральдика

Анималистика

Натюрморт

Скульптура

Палеонтология

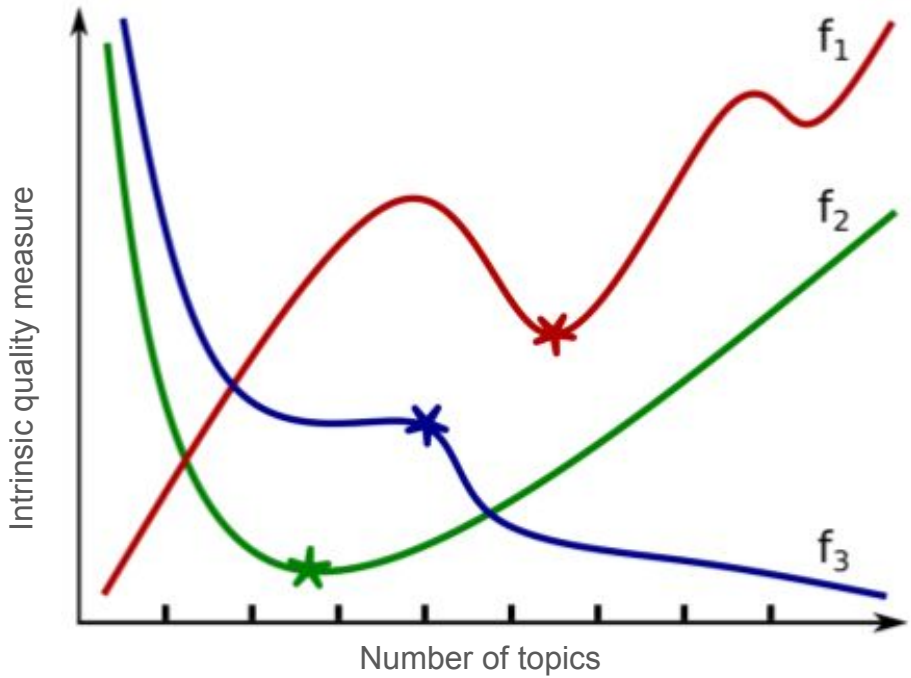
Вымершие животные

Интродукция

T’ > T

Определение числа тем по внутреннему критерию

Идея: если есть оптимальное число тем, то его можно найти по графику зависимости критерия качества модели от числа тем.



Ожидаемые возможные зависимости внутреннего критерия качества модели от числа тем.

Эксперимент 2

Цель: исследовать возможность определения числа тем в датасета с по внутренним критериям качества тематических моделей.

Критерий: наличие оптимума на графике зависимости критерия от числа тем и совпадение этого числа тем с ожидаемым.

Данные: WikiRef220, 20NewsGroups, Reuters, Brown, StackOverflow, PostNauka, RuWiki-Good.[1]

Модели: PLSA, LDA, ARTM (decorrelated), ARTM (sparse), ARTM (sparse decorrelated).[1]

[1] *Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov, Evgeny Egorov*. TopicNet: Making Additive Regularisation for Topic Modelling Accessible, 2020.

Результаты 2

Число тем не является свойством одного лишь корпуса текстов.

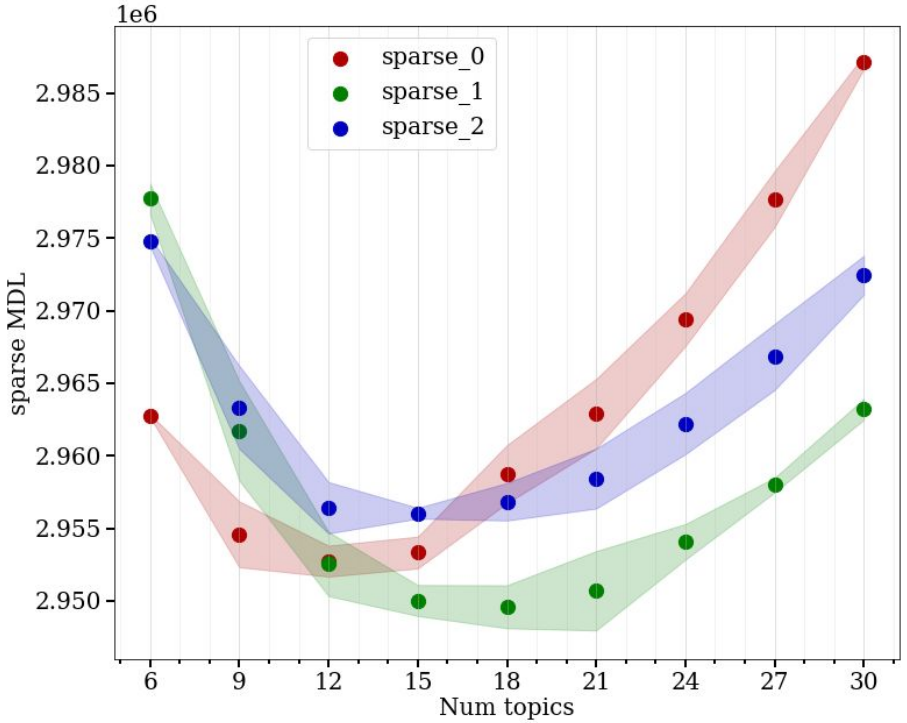
Три характеристики для каждого критерия качества:

- Jaccard: независимость результата от случайной инициализации (↓)
- Informativity: читаемость получаемых графиков (↑)
- Expected: точность предсказания числа тем по графику (↑)

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
Diversity	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
	D-avg-JH	0.302	0.053	0.022
Clustering	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
	average coherence	0.780	0.472	0.208
	uni-theta-divergence	0.470	0.197	0.047

Результаты 2

- Оптимальное число тем зависит от модели.
- Случайность в инициализации порождает дисперсию.



Sparse MDL критерий для разреженных моделей с разными значениями гиперпараметра (WikiRef220).

Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov. Determination of the Number of Topics Intrinsically: Is It Possible?, 2023.

Задача

4. Разработать автоматический способ валидации тематических моделей с учётом их неустойчивости и неполноты.

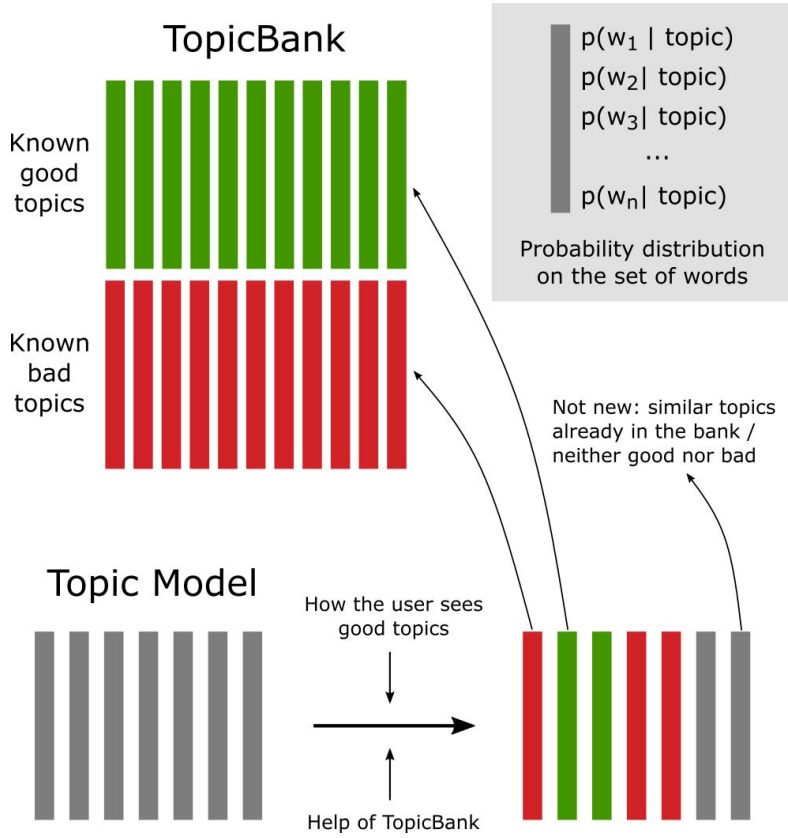
TopicBank: Сохранение хороших тем

Проблема:

- *Большое* число экспериментов по поиску лучшей модели.
- Найденные в процессе хорошие темы *теряются*.

Решение:

- *Сохранять* найденные темы в банк тем.
- Использовать банк тем для *валидации* новых моделей.

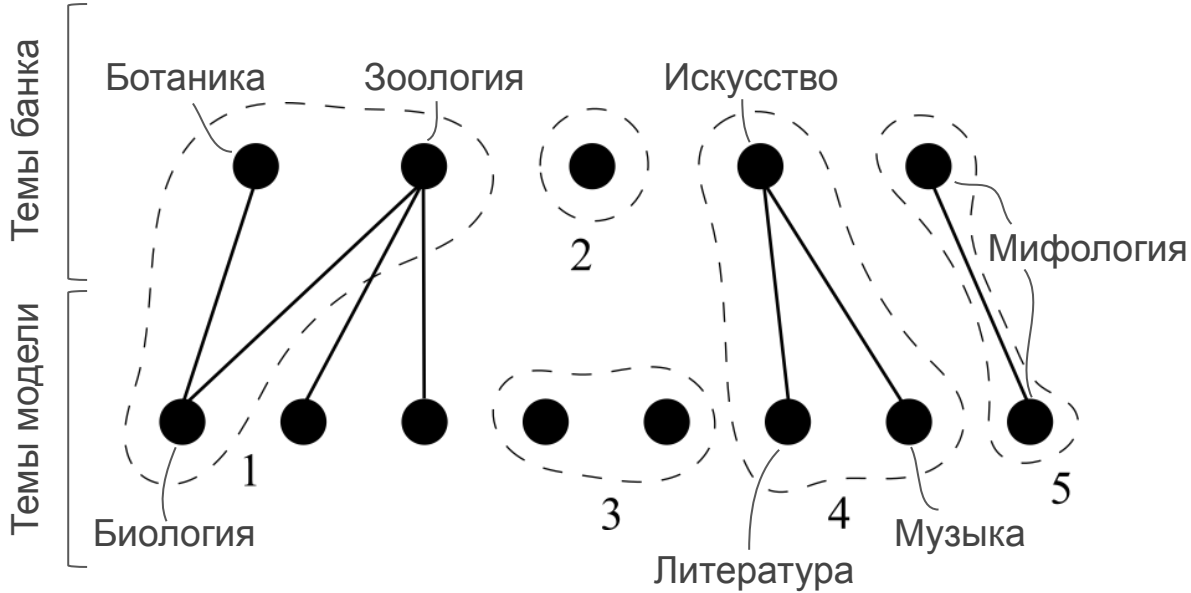


Связи между темами банка и модели

- Банк тем хранит хорошие и *различные* темы.
- Решение о добавлении темы в банк — из анализа *связей* между темами.

Возможные типы связей:

1) слияние тем
2) нет дочерних
3) нет родительских
4) расщепление темы
5) сохранение темы



Эксперимент 3

Цель: убедиться в возможности использования Банка тем для оценки качества тематических моделей.

Критерий: для большинства датасетов Банком тем в качестве лучшей определена *одна и та же модель*.[1]

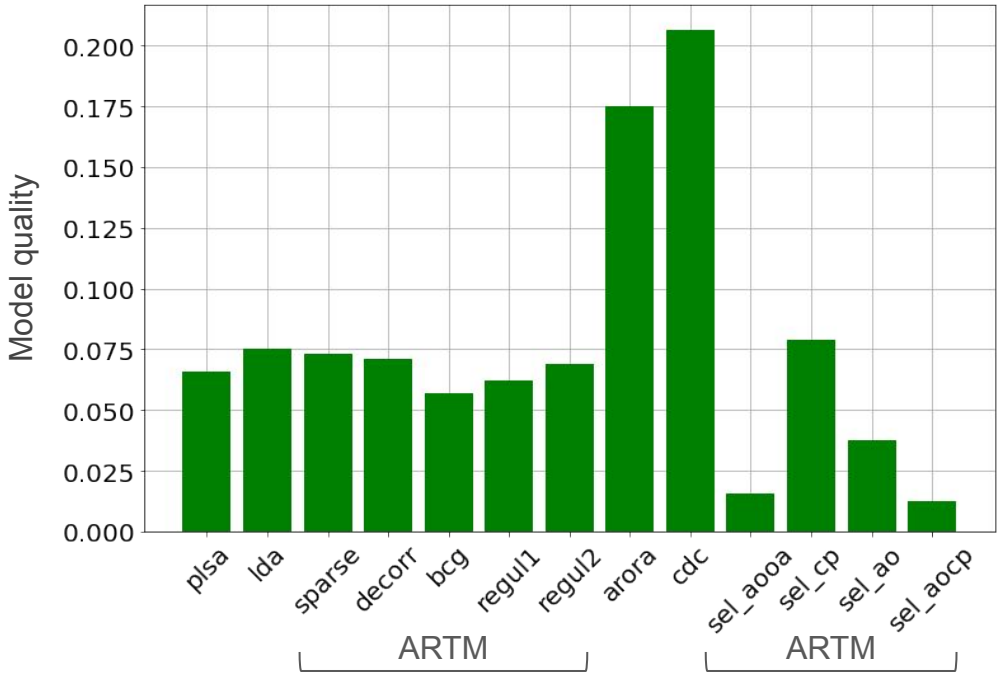
Данные: PostNauka, Reuters, Brown, 20Newsgroups, AG News, Watan2004, Habrahabr.

Модели: PLSA, LDA; ARTM (несколько видов); Arora, CDC.

[1]*Гипотеза*: если текстовые коллекции по структуре похожи, то и качество разных тематических моделей на этих коллекциях будет схоже, поэтому в фиксированном множестве моделей должна найтись лучшая.

Результаты 3

- TopicBank выявил модели с наибольшим числом хороших тем.
- Лучшими моделями оказались модели с неслучайной инициализацией (Arora, CDC)



Усреднённое качество моделей, рассчитанное с помощью Банка тем.

Vasiliy Alekseev, Konstantin Vorontsov et al. TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation, 2021.

Задача

5. Предложить и реализовать регуляризаторы в рамках АРТМ, предназначенные для улучшения тематической модели в процессе множественного обучения.

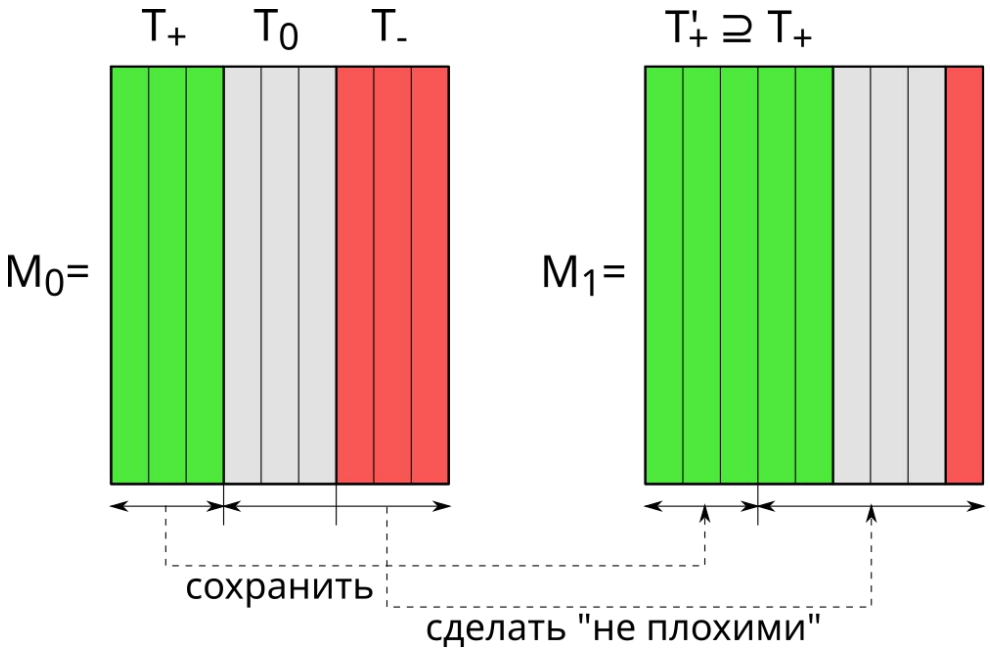
Итеративное улучшение тематической модели

Проблема:

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

Решение:

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



Эксперимент 4

Цели:

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

Критерии:

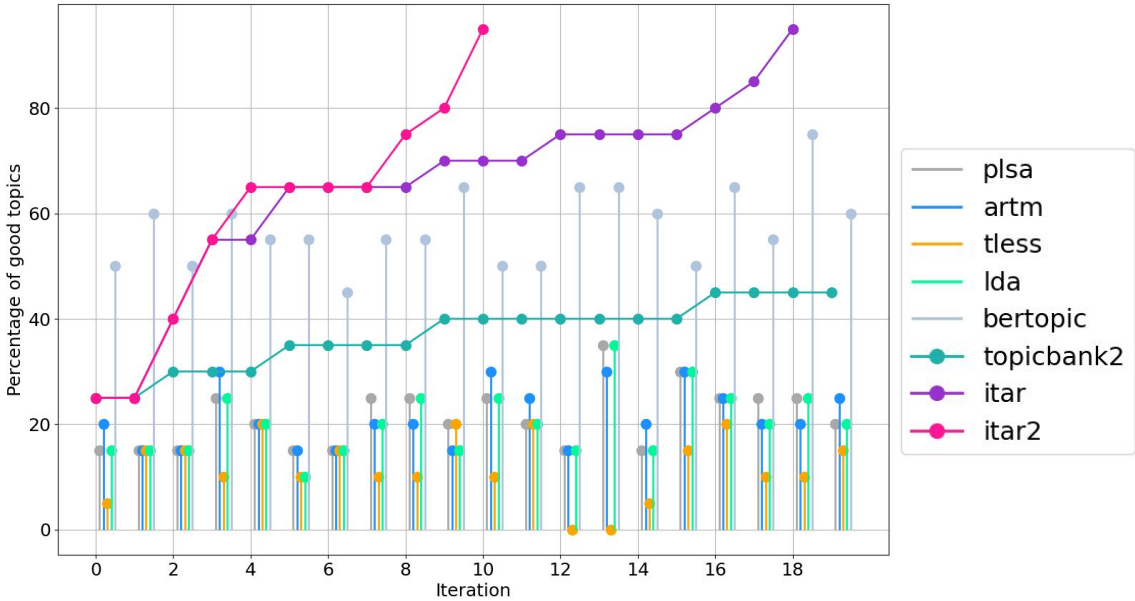
- Доля хороших тем в модели (хорошие — с высокой когерентностью).
- Различность тем.
- Перплексия.

Данные: PostNauka, 20Newsgroups, RuWiki-Good, RTL-Wiki-Person, ICD-10.

Модели: PLSA, LDA, ARTM, TLESS, BERTopic, TopicBank.

Результаты 4

- Предлагаемая модель монотонно улучшается
- Предлагаемая модель содержит больше всего хороших тем.
- Темы модели различны.
- Перплексия модели умеренная.
- Показана связь между когерентностями (по совстречаемостям и внутритекстовой)



Процент хороших тем модели в зависимости от итерации (↑).
RuWiki-Good, модели на 20 тем.

Alex Gorbulev, Vasiliy Alekseev, Konstantin Vorontsov. Iterative Improvement of an Additively Regularized Topic Model //arXiv preprint arXiv:2408.05840. – 2024.

Эксперимент 5

Цель: выяснить роль трёх определяющих предлагаемую модель регуляризаторов (сохранение хороших тем, декорреляция свободных тем с плохими, декорреляция свободных тем с хорошими).

Данные: PostNauka, 20Newsgroups, RuWiki-Good, RTL-Wiki-Person, ICD-10.

Критерии:

- Когерентность тем.
- Различность тем.
- Перплексия.
- Общее число просмотренных плохих тем.

Alex Gorbulev, Vasiliy Alekseev, Konstantin Vorontsov. Iterative Improvement of an Additively Regularized Topic Model //arXiv preprint arXiv:2408.05840. – 2024.

Результаты 5

- 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).

Положения, выносимые на защиту

- Предложена внутритекстовая когерентность как метод оценки интерпретируемости темы по распределению её слов в тексте.
- Реализованы когерентность и алгоритмы обучения интерпретируемых тематических моделей в рамках библиотеки TopicNet[1].
- Разработана библиотека OptimalNumberOfTopics[2] для оценки качества тематических моделей по внутренним критериям.
- Представлен метод TopicBank[2] оценки качества тематических моделей с учётом их неустойчивости и неполноты.
- Предложен многопроходной алгоритм обучения тематической модели ITAR[2], приводящий к более устойчивой и полной модели.

[1] https://github.com/machine-intelligence-laboratory/TopicNet.
[2] https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics.

Доклады

- Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций — 60-я Научная конференция МФТИ, 2017.
- Intra-Text Coherence as a Measure of Topic Models' Interpretability — 24-я Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», 2018.
- Topic Modelling for Extracting Behavioral Patterns from Transactions Data — IC-AIAI 2019: International Conference on Artificial Intelligence: Applications and Innovations, 2019.
- Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей — 64-я научная конференция МФТИ, 2021.
- Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей — Математические методы распознавания образов (ММРО-2021).
- Determination of the Number of Topics Intrinsically: Is It Possible? — The 11th International Conference on Analysis of Images, Social Networks and Texts (AIST 2023).
- TopicBank: Collection of Coherent Topics Using Multiple Model Training with Their Further Use for Topic Model Validation — The 5th International Conference on Machine Learning and Intelligent Systems (MLIS 2023).
- Determination of the Number of Topics Intrinsically: Is It Possible? — The 66th MIPT All-Russian Scientific Conference, 2024.
- Итеративное улучшение аддитивно регуляризованной тематической модели — 66-я Всероссийская научная конференция МФТИ, 2024.

Публикации

- *Alekseev V., Bulatov V., Vorontsov K.* Intra-text coherence as a measure of topic models' interpretability // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. – 2018. – Pp. 1-13.
- *Egorov E., Nikitin F., Alekseev V., Goncharov A., Vorontsov K.* Topic Modelling for Extracting Behavioral Patterns from Transactions Data // IEEE: 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI). – 2019. – Pp. 44-49.
- *Bulatov V., Alekseev V., Vorontsov K., Polyudova D., Veselova E., Goncharov A., Egorov E.* TopicNet: Making Additive Regularisation for Topic Modelling Accessible //Proceedings of The 12th Language Resources and Evaluation Conference. – 2020. – Pp. 6745-6752.
- *Alekseev V., Vorontsov K. 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.
- *Bulatov V., Alekseev V., Vorontsov K.* Determination of the Number of Topics Intrinsically: Is It Possible? // International Conference on Analysis of Images, Social Networks and Texts. – Cham : Springer Nature Switzerland. – 2023. – Pp. 3-17.

- *Gorbulev A., Alekseev V., Vorontsov K.* Iterative Improvement of an Additively Regularized Topic Model // arXiv preprint arXiv:2408.05840. – 2024. (Accepted for AIST 2024)