

# Поиск полного набора тем с помощью обучения нескольких моделей

Алексеев Василий

Научный руководитель:  
д.ф.-м.н. Воронцов К. В.

МФТИ, группа 874

17 июня 2020

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

$$\begin{aligned} \mathcal{L}(\Phi, \Theta | D) &\equiv \ln p(\Phi, \Theta | D) \\ &= \sum_{d \in D} \sum_{w \in W_d} n_{wd} \log \sum_{t \in T} \varphi_{wt}\theta_{td} \rightarrow \max_{\Phi, \Theta} \end{aligned}$$

$$\begin{cases} \mathcal{L} + \sum_{i=1}^n \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \\ \tau_i \geq 0, i = 1, \dots, n \end{cases}$$

- 1 Проблема
- 2 Банк тем
- 3 Создание банка тем
- 4 Использование банка тем для оценки качества моделей

- Тематические модели *неполны и неустойчивы*.
- Получение хорошей тематической модели, как правило, требует больших затрат времени.
- Не существует идеального автоматического способа оценки качества тематических моделей.

## Решение

Банк тем — инструмент для сохранения найденных интерпретируемых тем с целью последующего их использования для оценки качества моделей.

## Цели

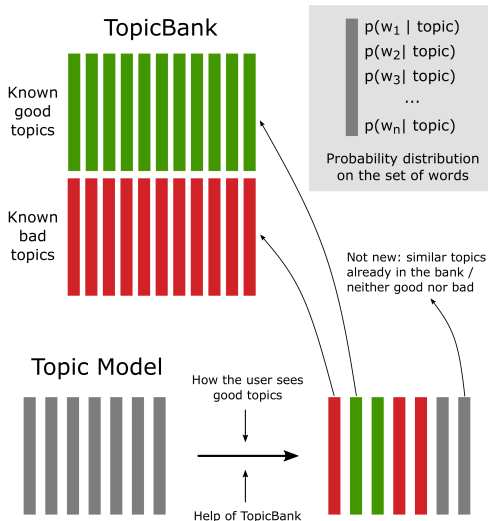
На нескольких датасетах оценить качество ряда тематических моделей с помощью банка тем.

- 1 Проблема
- 2 **Банк тем**
- 3 Создание банка тем
- 4 Использование банка тем для оценки качества моделей

# Банк тем: сохранение интерпретируемых тем

Банк тем — модель  
полного набора тем:  
таких тем, которые

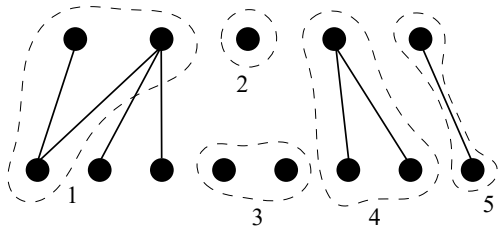
- 1) интерпретируемы
- 2) различны
- 3) обеспечивают  
высокое  
правдоподобие  
коллекции  
 $p(\Phi, \Theta | D)$



# Банк тем: обеспечение различности тем, отличия от иерархической модели

$$\underbrace{p(w | t)}_{\varphi_{wt}^{\text{parent}}} = \sum_{s \in S} \underbrace{p(w | s)}_{\varphi_{ws}^{\text{child}}} \underbrace{p(s | t)}_{\psi_{st}} \quad \text{Hierarchy}$$

$$\underbrace{p(w | t)}_{\varphi_{wt}^{\text{bank}}} = \sum_{s \in S} \underbrace{p(w | s)}_{\varphi_{ws}^{\text{new}}} \underbrace{p(s | t)}_{\psi_{st}} \quad \text{TopicBank}$$



| № | Hierarchy | TopicBank |
|---|-----------|-----------|
| 1 | ok        | no        |
| 2 | ok        | ok        |
| 3 | no        | ok        |
| 4 | ok        | maybe     |
| 5 | ok        | maybe     |

- 1 Проблема
- 2 Банк тем
- 3 Создание банка тем**
- 4 Использование банка тем для оценки качества моделей



- Множественное обучение тематических моделей
- Бинарная классификация тем на интерпретируемые и нет на основании одного признака – когерентности<sup>1</sup>

# B: topic bank

# N: number of model trainings

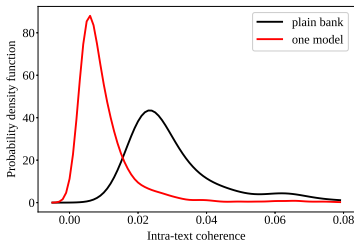
for i = 1 to N:

  model ← train\_model(i)

  new\_topics ← get\_new(model, B)

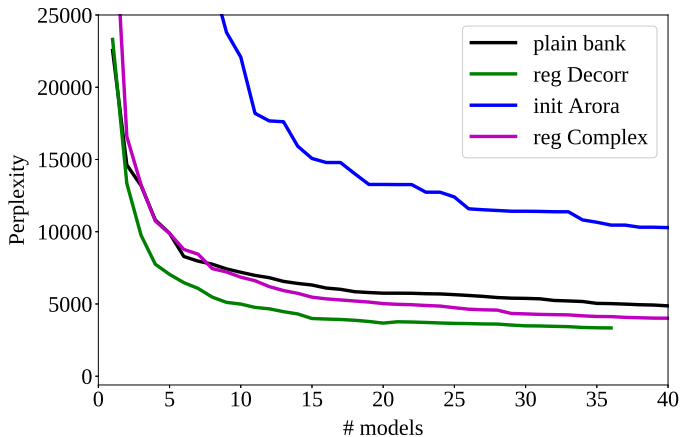
  good\_topics ← get\_good(model)

  B ← update(B,  
    new\_topics ∩ good\_topics)



<sup>1</sup>Alekseev V., Bulatov V., and Vorontsov K. *Intra-text coherence as a measure of topic models' interpretability*. Dialogue, 2018

# Зависимость банка тем от обучаемых моделей



Vorontsov K. et al. *Additive regularization of topic models*, 2015.

Arora S. et al. *Learning topic models—going beyond SVD*, 2012.

Hofmann T. *Probabilistic latent semantic indexing*, 1999.

- 1 Проблема
- 2 Банк тем
- 3 Создание банка тем
- 4 Использование банка тем для оценки качества моделей

# План эксперимента по валидированию моделей

- $B$  – темы в банке тем
- $\mathcal{D}$  – датасеты,  $|\mathcal{D}| < \infty$
- $\mathcal{M}$  – модели,  $|\mathcal{M}| < \infty$

Модель по набору данных даёт множество тем

$$\begin{cases} m : d \mapsto T \\ m \in \mathcal{M}, d \in \mathcal{D} \end{cases}$$

Качество модели можно оценить с помощью банка тем по тому, насколько темы модели  $T$  похожи на темы банка  $B$ .

$$\text{recall@bank} = \frac{|t \in B \mid \exists \tau \in T : \rho(t, \tau) < \text{threshold}|}{|B|}$$

$B$  – множество тем в банке тем, а  $T$  – множество тем во вновь обученной модели

Где  $\rho(t_1, t_2)$  – расстояние между темами (по мере Жаккара):

$$\rho(t_1, t_2) = 1 - \frac{\sum_{w \in \text{Ker}_{12}} \min_{i \in \{1,2\}} (p(w \mid t_i))}{\left( \sum_{i=1}^2 \sum_{w \in \text{Ker}_i \setminus \text{Ker}_{12}} p(w \mid t_i) + \sum_{w \in \text{Ker}_{12}} \max_{i \in \{1,2\}} (p(w \mid t_i)) \right)}$$

Где  $\text{Ker}_i \equiv \text{Ker}(t_i)$ ,  $\text{Ker}_{12} \equiv \text{Ker}(t_1) \cap \text{Ker}(t_2)$

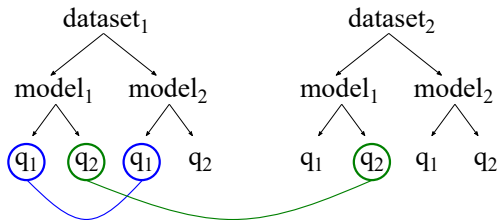
и  $\text{Ker}(t) = \{w \in t : p(w \mid t) > 1/|W|\}$  – ядро темы  $t$

Пусть  $H \subseteq \text{Im}(\rho)$  – пороги,  $|H| < \infty$ . Тогда

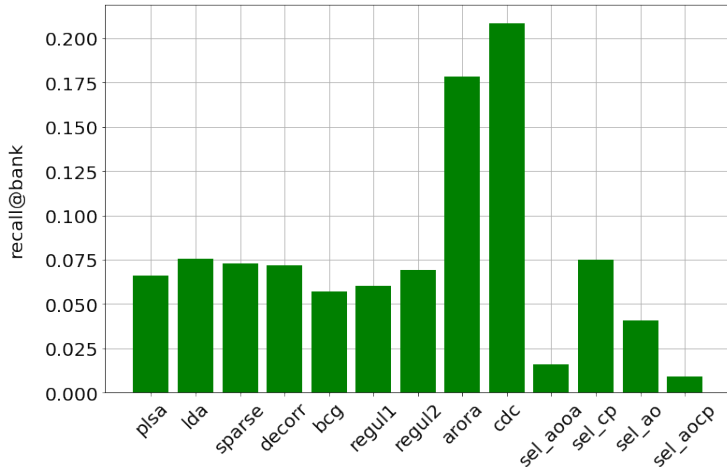
$$\text{recall@bank} = f(d, m, h), \quad d \in \mathcal{D}, \quad m \in \mathcal{M}, \quad h \in H$$

$$\langle \text{recall@bank}(m, d, h) \rangle_h = \frac{\sum_{\eta \in H} w(\eta) \cdot \text{recall@bank}(m, d, \eta)}{\sum_{\eta \in H} w(\eta)}, \quad w(\eta) \geq 0$$

$$\langle \text{recall@bank}(m, d, h) \rangle_{d,h} = \frac{1}{|\mathcal{D}|} \sum_{\delta \in \mathcal{D}} \frac{\langle \text{recall@bank}(m, \delta, h) \rangle_h}{\sum_{\mu \in \mathcal{M}} \langle \text{recall@bank}(\mu, \delta, h) \rangle_h}$$



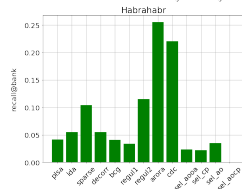
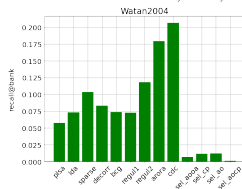
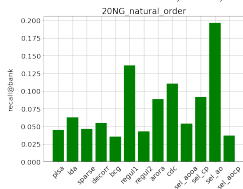
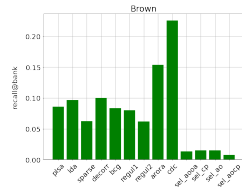
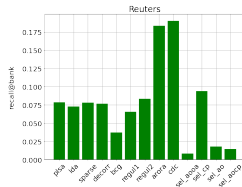
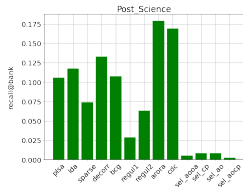
# Результат, усреднённый по датасетам



Dobrynin V. et al. *Contextual document clustering*, 2004.

Blei D. et al. *Latent dirichlet allocation*, 2003.

# Результат по разным датасетам



habr.com

postnauka.ru

nltk.org/book/ch02.html

sites.google.com/site/mouradabbas9/corpora

scikit-learn.org/0.19/datasets/twenty\_newsgroups.html



- Предложен алгоритм создания банка тем с использованием множественного обучения моделей
- Предложена методика оценки качества моделей с помощью банка тем
- Разработана система для использования банка тем<sup>1</sup>

## Публикации

- Alekseev V. et al. *TopicNet: Making Additive Regularisation for Topic Modelling Accessible*. LREC, 2020.<sup>2</sup>
- Alekseev V. et al. *Topic Modelling for Extracting Behavioral Patterns from Transactions Data*. IEEE, 2019.<sup>3</sup>

---

<sup>1</sup>[github.com/machine-intelligence-laboratory/OptimalNumberOfTopics](https://github.com/machine-intelligence-laboratory/OptimalNumberOfTopics)

<sup>2</sup>[aclweb.org/anthology/2020.lrec-1.833](https://aclweb.org/anthology/2020.lrec-1.833)

<sup>3</sup>[ieeexplore.ieee.org/abstract/document/9007329](https://ieeexplore.ieee.org/abstract/document/9007329)