Поиск полного набора тем с помощью обучения нескольких моделей

Алексеев Василий

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

МФТИ, группа 874

3 июня 2020

Введение в тематическое моделирование

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

$$\mathcal{L}(\Phi, \Theta \mid D) \equiv \ln p(\Phi, \Theta \mid D)$$
$$= \sum_{d \in D} \sum_{w \in W_d} n_{wd} \log \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

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

Содержание

Проблема

- 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

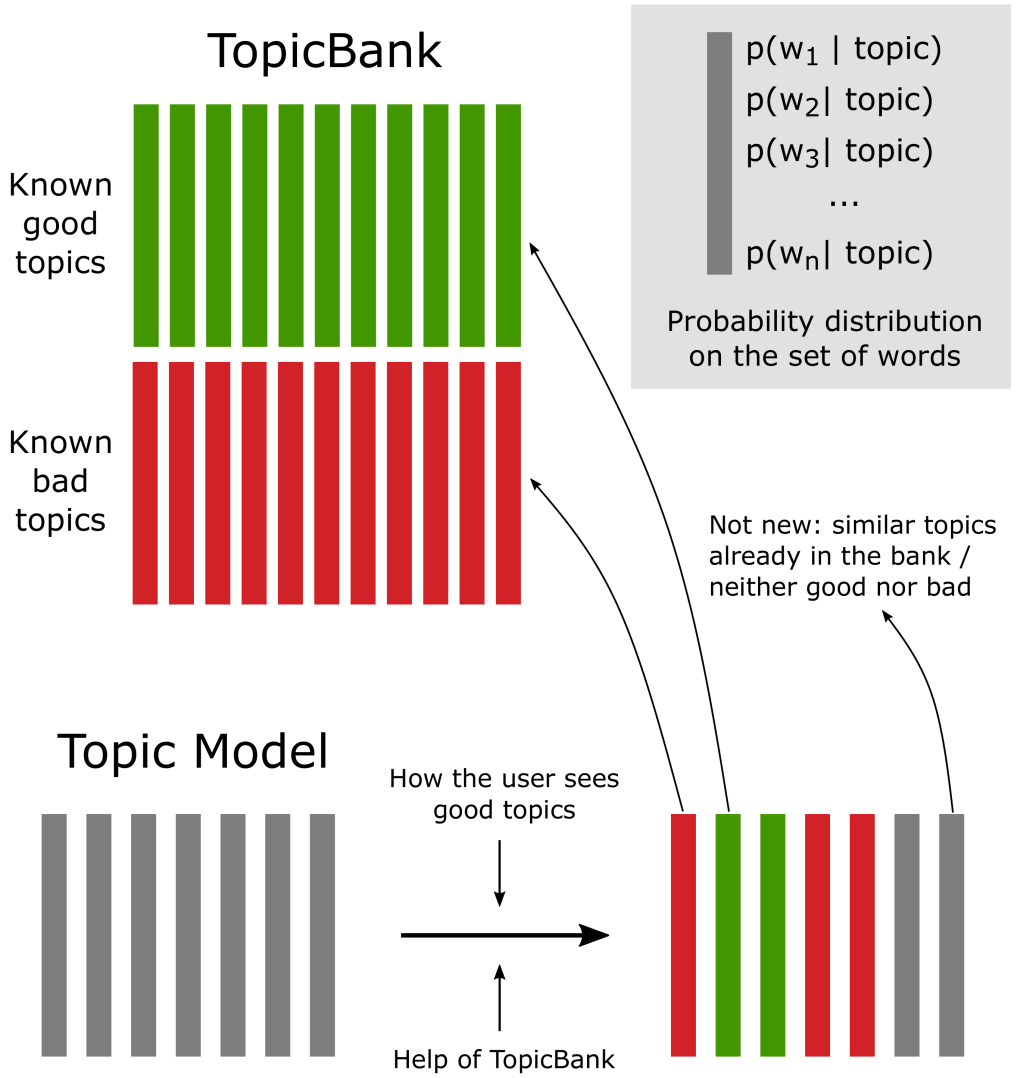
На нескольких датасетах оценить качество ряда тематических моделей с помощью банка тем.

Содержание

Банк тем: сохранение интерпретируемых тем

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

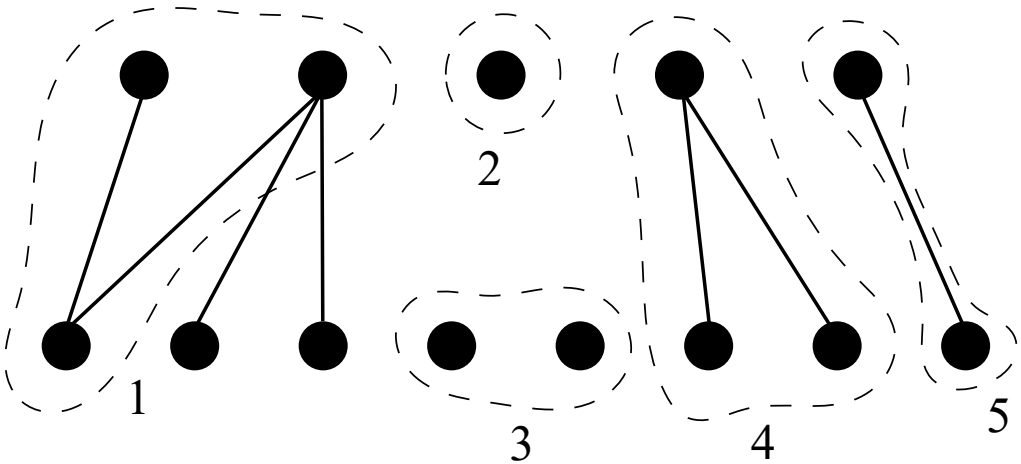
1) интерпретируемы
2) различны
3) обеспечивают высокое правдоподобие коллекции $p(\Phi, \Theta \mid D)$



Банк тем: обеспечение различности тем, отличия от иерархической модели

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

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



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

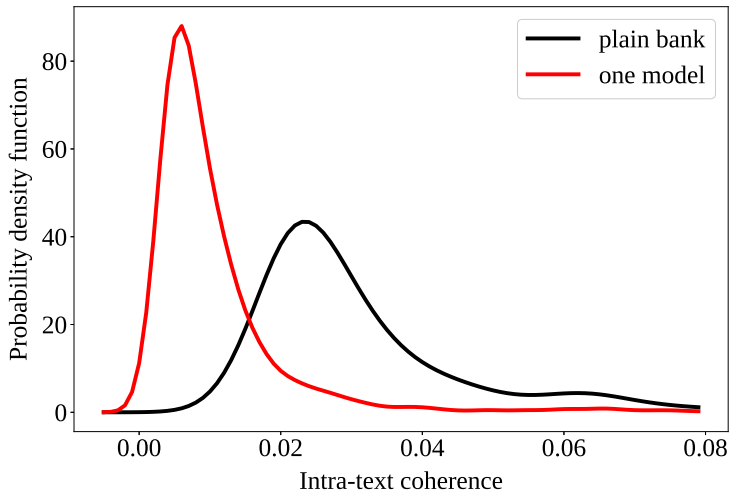
Содержание

Создание банка тем

- Множественное обучение тематических моделей
- Бинарная классификация тем на интерпретируемые и нет на основании одного признака — когерентности[1]

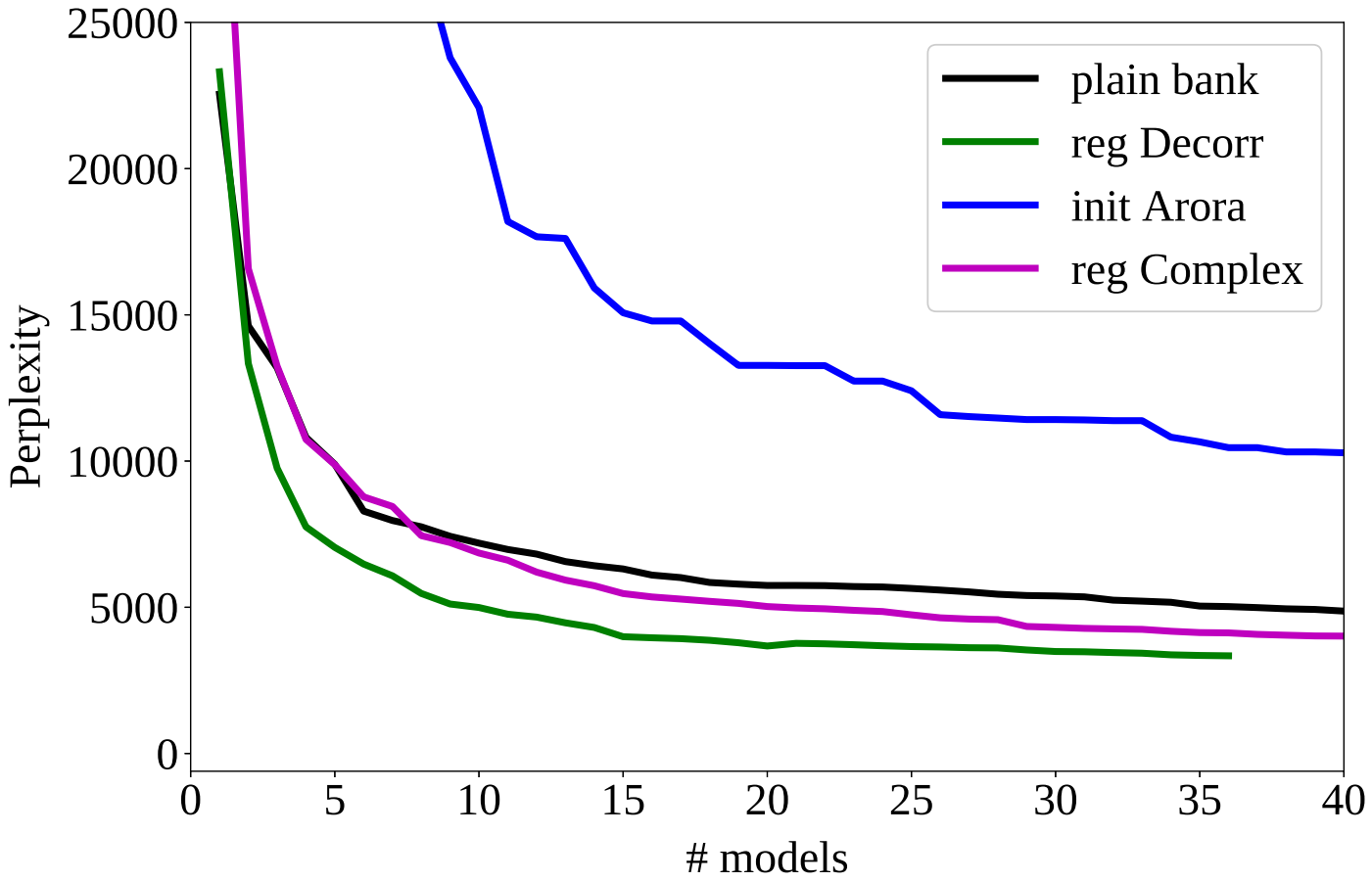
```
# B: topic bank
# N: number of model trainings

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)
```



[1] 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.

Содержание

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.

Функция качества модели

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

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):

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

Где $\text{Ker}_i \equiv \text{Ker}(t_i)$, $\text{Ker}_{12} \equiv \text{Ker}(t_1) \cap \text{Ker}(t_2)$
и $\text{Ker}(t) = \left\{w \in t : p(w \mid t) > 1/|W|\right\}$ — ядро темы t

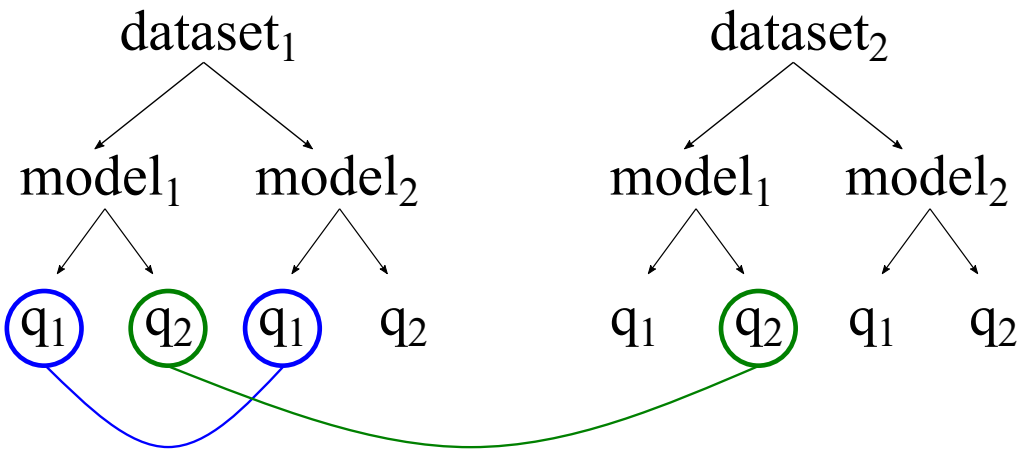
Усреднение

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

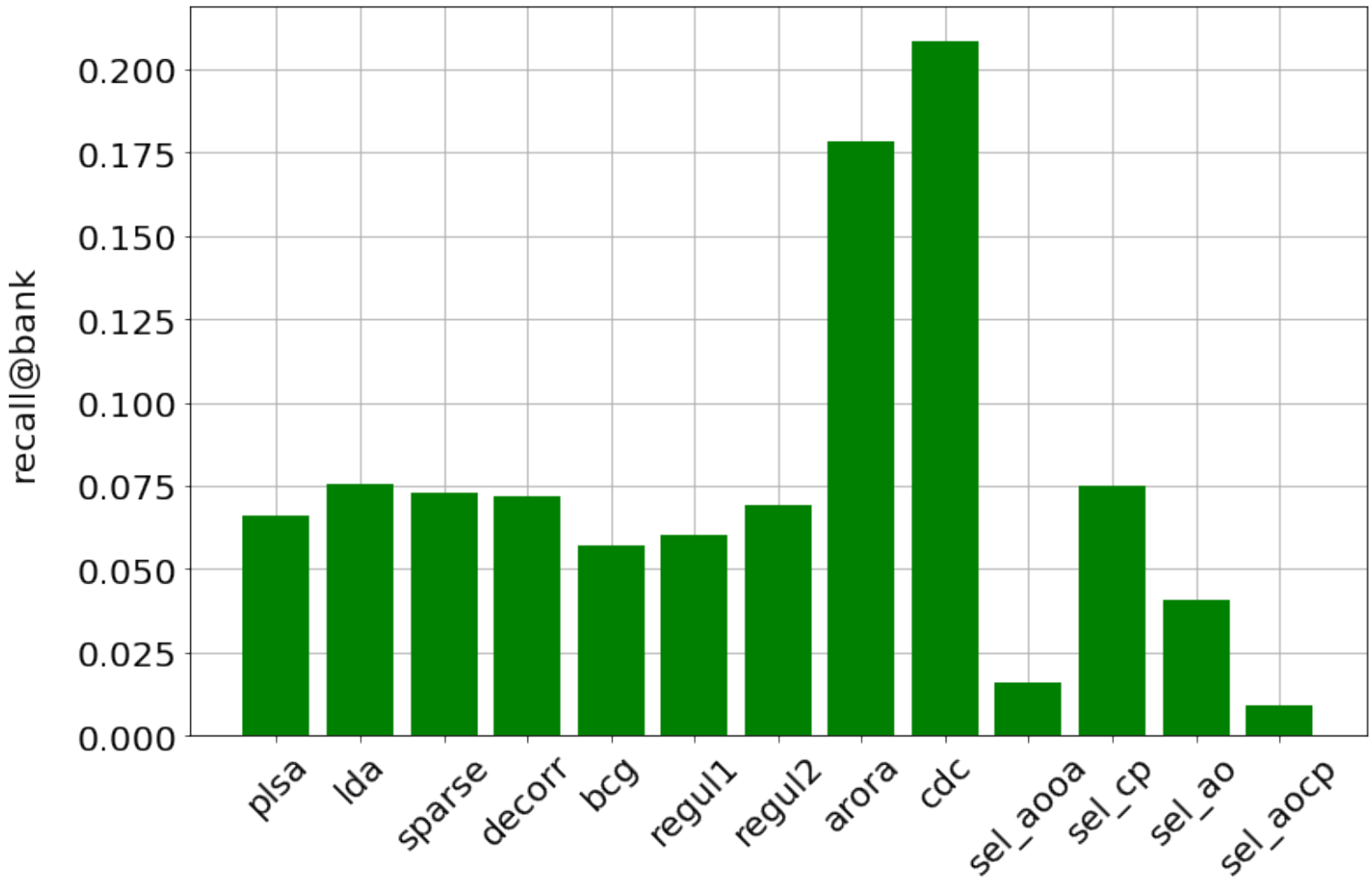
$$\text{recall@bank} = f(d, m, h),\ d \in \mathcal{D},\ m \in \mathcal{M},\ h \in H$$

$$\big\langle \text{recall@bank}(m, d, h) \big\rangle_h = \frac{\sum\limits_{\eta \in H} w(\eta) \cdot \text{recall@bank}(m, d, \eta)}{\sum\limits_{\eta \in H} w(\eta)},\ w(\eta) \geqslant 0$$

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

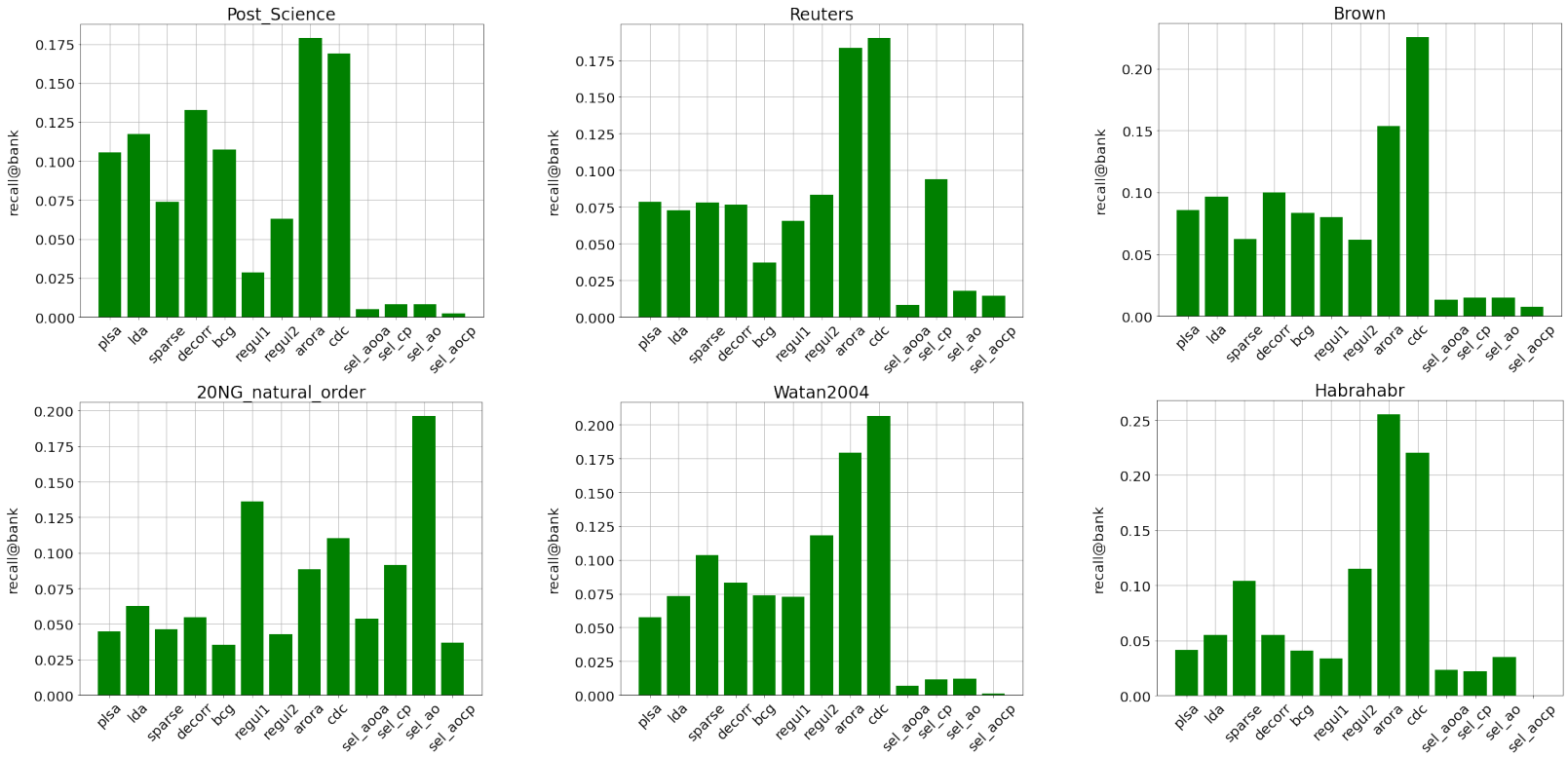


Результат, усреднённый по датасетам



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

Результаты

- 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[1]

Публикации

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

[1] github.com/machine-intelligence-laboratory/OptimalNumberOfTopics
[2] aclweb.org/anthology/2020.lrec-1.833
[3] ieeexplore.ieee.org/abstract/document/9007329