

# Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций

Василий Алексеев

Научный руководитель  
д. ф.-м. н. Воронцов Константин Вячеславович

Защита бакалаврской работы

28 июня 2018



- $W$  – множество слов
- $D$  – множество документов
- $W_d$  – упорядоченное мультимножество слов, из которых состоит материальный аналог документа  $d \in D$
- $T$  – множество тем
- $n_{dw}$  – количество вхождений слова  $w$  в  $W_d$
- $\nu_{wd}$  – частота появления слова  $w$  в  $W_d$

- *Гипотеза*: слово в документе связано с некоторой темой
- $D \times W \times T$  – дискретное вероятностное пространство
- Элемент коллекции  $(d_i, w_j, t_k) \sim p(d, w, t)$ , при этом  $d_i$  и  $w_j$  – наблюдаемые, а  $t_k$  – скрытые
- *Гипотеза*: для определения тем не важен порядок документов в  $D$
- *Гипотеза (мешка слов)*: для определения тем не важен порядок слов в  $W_d, d \in D$
- *Гипотеза (условной независимости)*:  $p(w | d, t) = p(w | t)$

# Задача тематического моделирования

- $\varphi_{wt} \equiv p(w | t)$  – вероятность встретить слово  $w$  в теме  $t$
- $\theta_{td} \equiv p(t | d)$  – вероятность найти тему  $t$  в документе  $d$

## Задача матричного разложения

$$r_{wd} \approx p(w | d) = \sum_T p(w | t)p(t | d)$$

## Решение

Максимизация регуляризованного логарифма правдоподобия<sup>1</sup>

$$\sum_{d \in D} \sum_{w \in W_d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

<sup>1</sup>Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections, 2015

## Плохие темы

Тема  $t \in T$ , найденная тематической моделью, может характеризоваться словами, которые

- вместе ни с чем не ассоциируются у человека,
- разбросаны по тексту в случайном порядке.

*Интерпретируемость* означает, может ли человек по самым частым словам темы дать ей подходящее название.

## Хорошо интерпретируемая тема (самые частые слова)

актёр, пьеса, музыкальный, премьера, партер, зритель, продюсер, аудитория, занавес, оркестр

## Недостатки подхода

Для оценки интерпретируемости темы необходимо привлекать экспертов. Также учитывается не вся информация о теме.

## Когерентность

Оценка неслучайности того, что топ-слова темы встречаются недалеко друг от друга в тексте.

$$\text{coh}(D, W, \varphi.t) = \text{Average}_{w_i, w_j \text{ from } 10 \text{ top-words}} \text{PMI}(w_i, w_j)$$

Хорошо коррелирует с экспертными оценками интерпретируемости.

## Когерентность тематической модели

Среднее значение когерентности по темам  $T$  модели.

- Newman<sup>2</sup>: 
$$\text{PMI}(w_i, w_j) = \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

- Mimno<sup>3</sup> : 
$$\text{PMI}(w_i, w_j) = \ln \frac{D(w_i, w_j) + 1}{D(w_i)}$$

- $p(w_i)$ ,  $p(w_i, w_j)$  – вероятность встретить слово  $w_i$  и два слова  $w_i, w_j$  в одном окне заданного размера в тексте
- $D(w_i)$ ,  $D(w_i, w_j)$  – количество документов, содержащих слово  $w_i$  и два слова  $w_i, w_j$  в одном окне

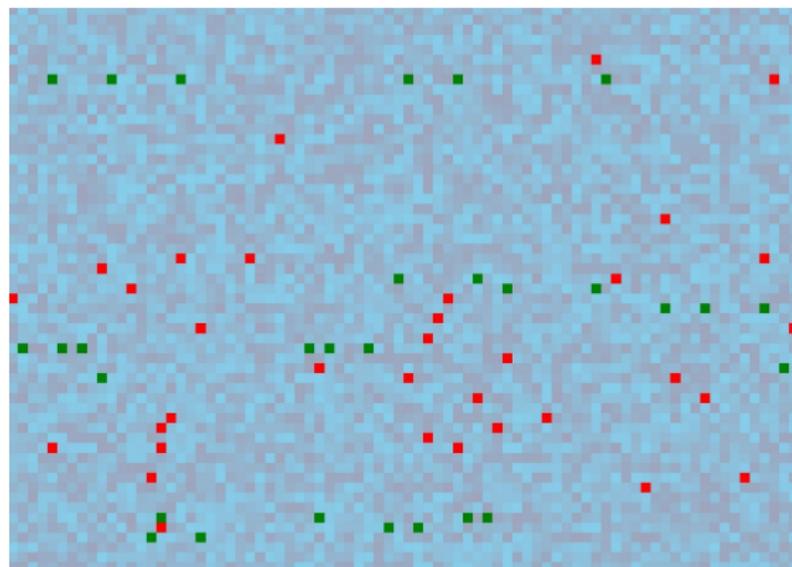
*Совстречаемость слов  $U \subseteq W$  – факт нахождения двух слов  $w_i, w_j \in U$  в одном текстовом окне в  $W_d$  для некоторого  $d \in D$ .*

---

<sup>2</sup>Newman et al. Automatic Evaluation of Topic Coherence, 2010

<sup>3</sup>Mimno et al. Optimizing Semantic Coherence in Topic Models, 2011

# Проблема когерентностей по топ-словам



- слова
- топ-слова
- совстречаемости

Совстречаемости 10 топовых слов тем датасета из статей «ПостНауки» занимают лишь **1.2%** от всего текста  $\bigcup_{d \in D} W_d$ .

Во фрагменте текста есть *только одно* слово «частиц» из списка 10 топ-слов темы «физика», и *ни одной* совстречаемости.

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.

Первые топ-слова темы «физика», включая первые **топ-10-слов**, с вероятностями (%):

**частица** (2.7), **электрон** (1.5), **кварк** (1.5), **атом** (1.3), **энергия** (1.2), **вселенная** (1.1), **фотон** (1.0), **физика** (0.9), **физик** (0.9), **эксперимент** (0.9), **масса** (0.7), **теория** (0.7), **свет** (0.7), **симметрия** (0.7), **протон** (0.7), **эйнштейн** (0.5), **нейтрино** (0.5), **вещество** (0.5), **квантовый** (0.5), **ускоритель** (0.5), **детектор** (0.4), **волна** (0.4), **эффект** (0.4), **свойство** (0.4), **спин** (0.4), **гравитация** (0.4), **материя** (0.4), **адрон** (0.4), **поль** (0.4), **частота** (0.4)

## Проблема

Когерентности по топ-словам учитывают не всю информацию о тематической модели

## Решение

Смотреть распределение темы по *всем* словам текста

## Задачи

- Предложить новые методы подсчёта когерентности
- Сравнить с существующими, основанными на встречаемости топ-слов

## Гипотеза о сегментной структуре текста

Тексты естественного языка сегментированы, состоят из сегментов разных тем.

## Следствие гипотезы

Слова каждой темы расположены группами, а не разбросаны по тексту беспорядочно.

*Качество темы* – абстрактное понятие, отражающее то, насколько хорошо распределение темы в тексте соответствует гипотезе о сегментной структуре текста.

Функция качества темы  $q(t)$ ,  $t \in T$

$q(t_1) < q(t_2) \leftrightarrow$  тема  $t_1$  менее качественная, чем тема  $t_2$

**Проблема**

$q(t)$  не известна

## Semantic Closeness (SemantiC)

$$\text{SemantiC}_{l_2} \Big|_t = \left\langle [0 < j - i < \text{window}] - \|\mathbf{w}_i - \mathbf{w}_j\|_2 \right\rangle_{\substack{w_i, w_j \in \bigcup_d W_d \\ \arg \max_s \mathbf{w}_i(s) = t \\ \arg \max_s \mathbf{w}_j(s) = t}}$$

$$\text{SemantiC}_{\text{Var}} \Big|_t = \left\langle -\text{Var}(\mathbf{w}_i(t), \mathbf{w}_{i+1}(t), \dots, \mathbf{w}_{i+\text{window}}(t)) \right\rangle_{w_i \in \bigcup_d W_d}$$

$$W_d \ni w \mapsto \mathbf{w} \equiv (p(t | d, w))_{t \in T}$$

## Topic Length (TopLen)

$$\text{TopLen} \Big|_t = \left\langle \overbrace{l(0)}^{l_1}, \overbrace{l(l_1)}^{l_2}, l(l_1 + l_2), \dots, l\left(\sum_{r=0}^{k-1} l_r\right), \dots \right\rangle_{l_j > 0}$$

$$l(\cdot) : \begin{cases} l(i) = \max \left\{ L : \text{threshold} + \sum_{j=i}^{i+L} \left( \mathbf{w}_j(t) - \max_{\substack{1 \leq \tau \leq |T| \\ \tau \neq t}} \mathbf{w}_j(\tau) \right) \geq 0 \right\} \\ \arg \max_s \mathbf{w}_i(s) = t \\ l(i) = 0 \\ \arg \max_s \mathbf{w}_i(s) \neq t \end{cases}$$

$$W_d \ni w \mapsto \mathbf{w} \equiv (p(t | d, w))_{t \in T}$$

## Focus Consistency (FoCon)

$$\text{FoCon} = - \sum_{d \in D} \sum_{\substack{w_i, w_j \in W_d \\ |j-i|=1}} |w_i(t) - w_j(t)| + |w_i(\tau) - w_j(\tau)|$$

$$\begin{cases} t = \arg \max_s w_j(s) \\ \tau = \arg \max_s w_j(s) \end{cases}$$

$$W_d \ni w \mapsto \mathbf{w} \equiv (p(t | d, w))_{t \in T}$$

Метод не привязан к теме, а даёт значение когерентности для *тематической модели* как целого.

## Гипотеза о сегментной структуре текста

Тексты естественного языка сегментированы, состоят из сегментов разных тем.

## Следствие

Чем лучше функция когерентности, тем лучше она должна описывать способность тематической модели угадывать сегментную структуру текста.

## Проблема

Позиции сегментов не известны.

## Решение: полусинтетический датасет

2000 *монотематических* статей «ПостНауки» разрезаются на сегменты одинаковой длины, которые потом сшиваются в новые документы  $D'$ .

## Segmentation quality ( $sq$ )

$$sq(D', W, \Phi', \Theta') \Big|_{t'} = \sum_{d' \in D'} \sum_{\substack{w' \in W'_{d'} \\ \arg \max_s \mathbf{w}(s) = t'}} p(t' | d', w')$$

$$\begin{cases} W_d \ni w \mapsto w' \in W'_{d'} \\ W_d \ni w \mapsto \mathbf{w} \equiv (p(t | d, w))_{t \in T} \\ T \ni t \leftrightarrow t' \in T' \end{cases}$$

## Гипотеза о сохранении порядка по качеству

Функция  $sq(\cdot)$  задаёт тот же порядок на образах тем  $T'$ , что и неизвестная функция качества тем  $q(\cdot)$  на исходных темах  $T$

$$sq(t'_1) < sq(t'_2) \leftrightarrow q(t_1) < q(t_2)$$

# Корреляция между когерентностями и качеством сегментации

## Ряд моделей

$$\Phi'(\alpha) = \alpha \cdot \Phi_{bad} + (1 - \alpha) \cdot \Phi_{good} \quad | \quad \alpha \in [0, 1]$$

$\Phi_{good}$  – модель исходной коллекции «ПостНауки»

$\Phi_{bad}$  – случайная, столбцы из Dirichlet(**0.01**<sub>|W|</sub>)

## Корреляция (по Спирмену)

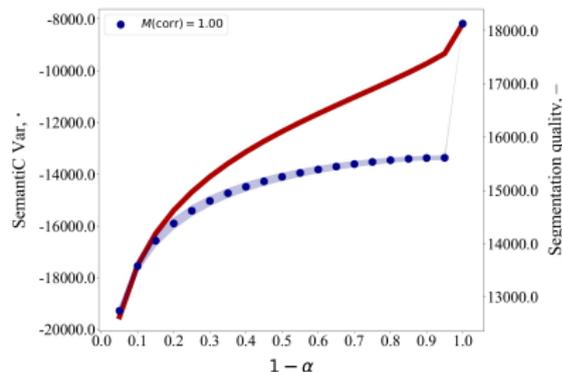
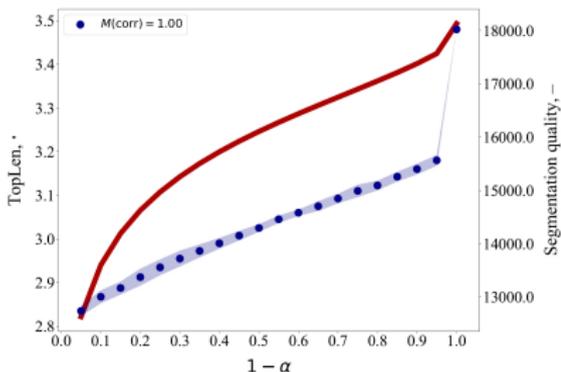
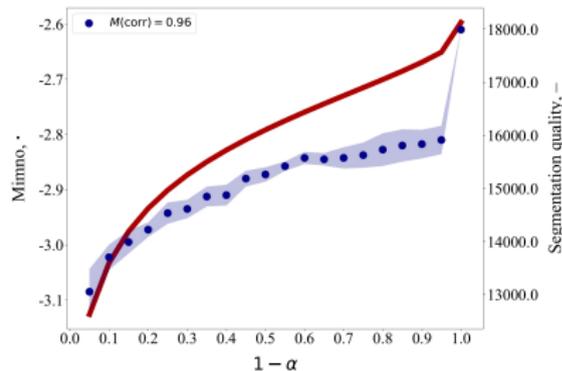
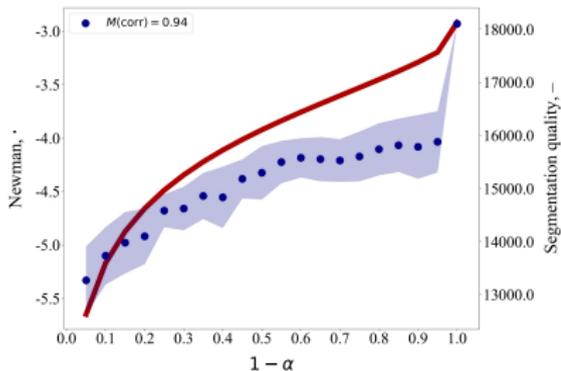
$$\text{Corr} \left\{ (\text{coh}(\Phi'(\alpha)))_{\alpha}, (\text{sq}(\Phi'(\alpha)))_{\alpha} \right\}$$

Чем ближе корреляция к 1, тем лучше когерентность

Coh	Corr
Newman	0.80
Mimno	0.94
SemantiC <sub>l<sub>2</sub></sub>	0.70
SemantiC <sub>Var</sub>	<b>1.00</b>
TopLen	<b>1.00</b>
FoCon	<b>1.00</b>

Корреляции при размере сегментов 200 слов  
и при 5 темах в каждом сшитом из сегментов документе

# Когерентности (синие) и качество сегментации (красное) как функции качества тематической модели



- Проиллюстрирован недостаток когерентностей по топ-словам: покрытие лишь малой части текстовой коллекции.
- Предложен полуавтоматический метод оценки качества функций когерентности: по корреляции с качеством сегментации полусинтетического текста тематическими моделями.
- Представлены методы *внутритекстовой* когерентности. По предложенной функции качества некоторые внутритекстовые методы лучше, чем когерентности по топ-словам.

## Публикация

Alekseev V. A., Bulatov V. G., Vorontsov K. V.  
Intra-Text Coherence as a Measure of Topic Models'  
Interpretability // Computational Linguistics and Intellectual  
Technologies. Dialogue 2018. Pp. 1-13.