Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций

Василий Алексеев

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

Защита бакалаврской работы

27 июня 2018

Обозначения

- 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
- ν_{wd} — частота появления слова w в W_d

Гипотезы

- *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)*: $p(w \mid d, t) = p(w \mid t)$

Тематическое моделирование

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

Задача матричного разложения

$$\nu_{wd} \approx p(w \mid d) = \sum_{T} p(w \mid t) p(t \mid d)$$

Решение

Максимизация регуляризованного логарифма правдоподобия[1]

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

[1] Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections, 2015

Темы, затрагиваемые в коллекции документов

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.

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

Качество темы $t \in T$

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

Функция качества темы $q(t)$

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

Проблема

$q(t)$ не известна

Интерпретируемость темы

Топ-слова темы — её самые частые слова.

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

Хорошо интерпретируемая тема (самые частые слова)

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

Плохо интерпретируемая тема (самые частые слова)

экспресс, эпиграф, туманный, результат, образ, право, заём, иероглиф, лак, футбол

Недостатки

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

Когерентность по топ-словам $\mathrm{coh}\left(D, W, \varphi_{\cdot t}\right)$

Когерентность

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

$$\mathrm{coh}\left(D, W, \boldsymbol{\varphi}_{\cdot t}\right) = \underset{w_i, w_j \text{ from } \boldsymbol{k} \text{ top-words}}{\text{Average}} \mathrm{PMI}(w_i, w_j)$$

Когерентность тематической модели

Среднее значение когерентности по темам T модели.

Недостаток

Частые совстречаемости *определённого числа* слов темы есть лишь *косвенный признак* того, что тема представлена в тексте в виде однородных сегментов.

Когерентности по топ-словам: Ньюман, Мимно

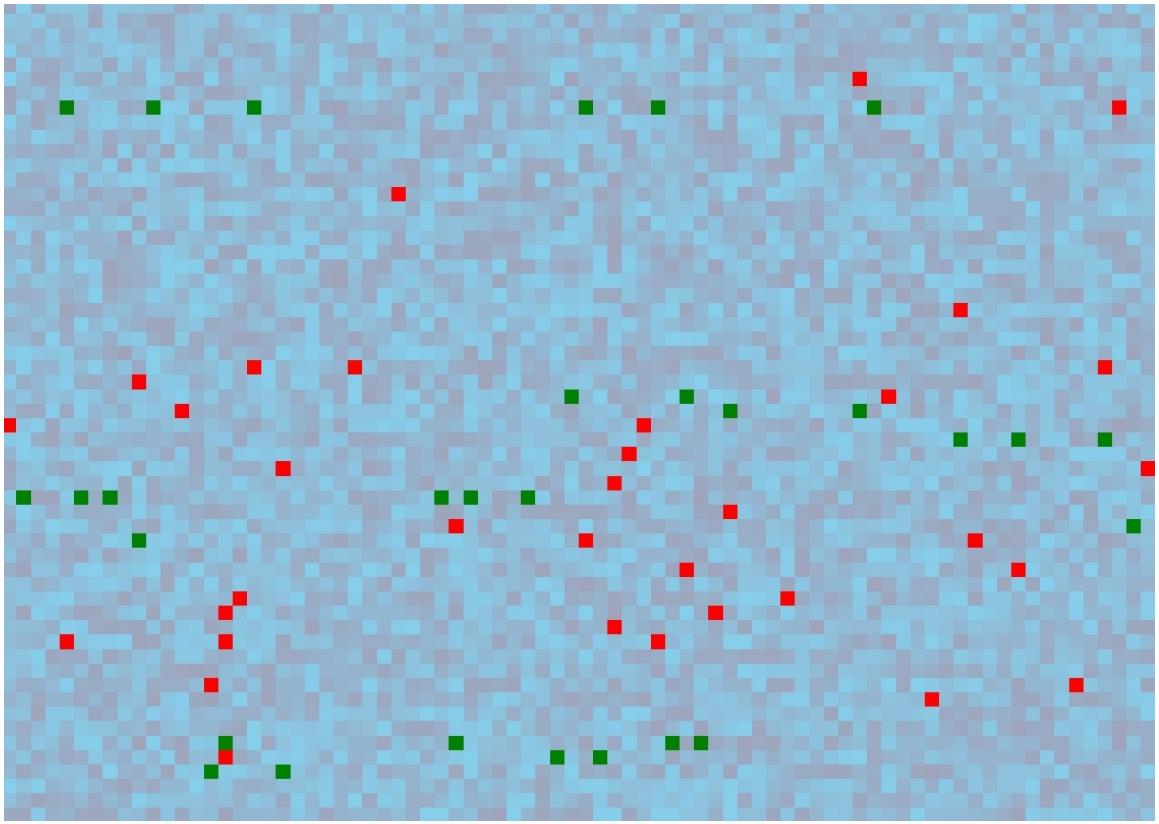
- Newman [2]: $\mathrm{PMI}(w_i, w_j) = \ln \dfrac{p(w_i, w_j)}{p(w_i)p(w_j)}$
- Mimno [3]: $\mathrm{PMI}(w_i, w_j) = \ln \dfrac{D(w_i, w_j) + 1}{D(w_i)}$

- $p(w_i)$, $p(w_i, w_j)$ — 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 $d \in D$.

[2] Newman et al. Automatic Evaluation of Topic Coherence, 2010
[3] Mimno et al. Optimizing Semantic Coherence in Topic Models, 2011

Проблема когерентностей по топ-словам



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

Проблема когерентностей по топ-словам

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)

Цель исследования

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 совстречаемостях топ-слов

Semantic Closeness (SemantiC)

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

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

$$W_d \ni w \mapsto \boldsymbol{w} \equiv \left(p(t \mid d, w)\right)_{t \in T}$$

Topic Length (TopLen)

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

$$l(\cdot): \begin{cases} l(i) = \max\left\{L : \mathrm{threshold} + \sum_{j=i}^{i+L}\left(\boldsymbol{w}_j(t) - \max\limits_{\substack{1\leqslant\tau\leqslant|T|\\ \tau\neq t}} \boldsymbol{w}_j(\tau)\right) \geqslant 0\right\} \\ \arg\max_s \boldsymbol{w}_i(s) = t \end{cases}$$

$$\begin{cases} l(i) = 0 \\ \arg\max_s \boldsymbol{w}_i(s) \neq t \end{cases}$$

$$W_d \ni w \mapsto \boldsymbol{w} \equiv \big(p(t \mid d, w)\big)_{t\in T}$$

Внутритекстовая когерентность $\mathrm{coh}\big(D, W, \Phi, \Theta\big)$

Focus Consistency (FoCon)

$$\mathrm{FoCon} = -\sum_{d \in D} \sum_{\substack{w_i, w_j \in W_d \\ j-i=1}} \big|\boldsymbol{w}_i(t) - \boldsymbol{w}_j(t)\big| + \big|\boldsymbol{w}_i(\tau) - \boldsymbol{w}_j(\tau)\big|$$

$$\begin{cases} t = \arg\max_s \boldsymbol{w}_i(s) \\ \tau = \arg\max_s \boldsymbol{w}_j(s) \end{cases}$$

$$W_d \ni w \mapsto \boldsymbol{w} \equiv \big(p(t \mid d, w)\big)_{t \in T}$$

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

Качество функций когерентности

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'.

Методы оценки качества сегментации текста

Segmentation quality (*sq*)

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

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

Гипотеза

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

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

Φ_{good} – модель исходной коллекции «ПостНауки»

Φ_{bad} – случайная, столбцы из $\text{Dirichlet}\left(\mathbf{0.01}_{|W|}\right)$

Корреляция (по Спирмену)

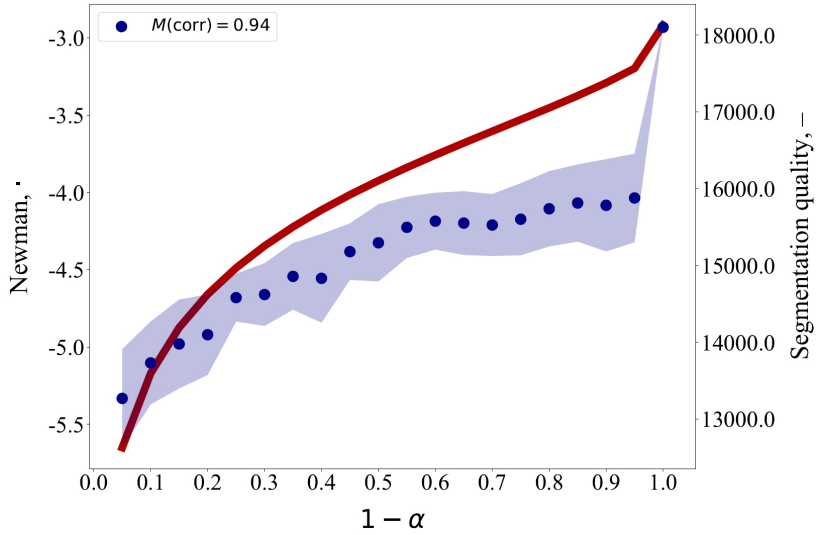
$$\text{Corr}\left\{ \left(\text{coh}(\Phi'(\alpha))\right)_{\alpha}, \left(sq(\Phi'(\alpha))\right)_{\alpha} \right\}$$

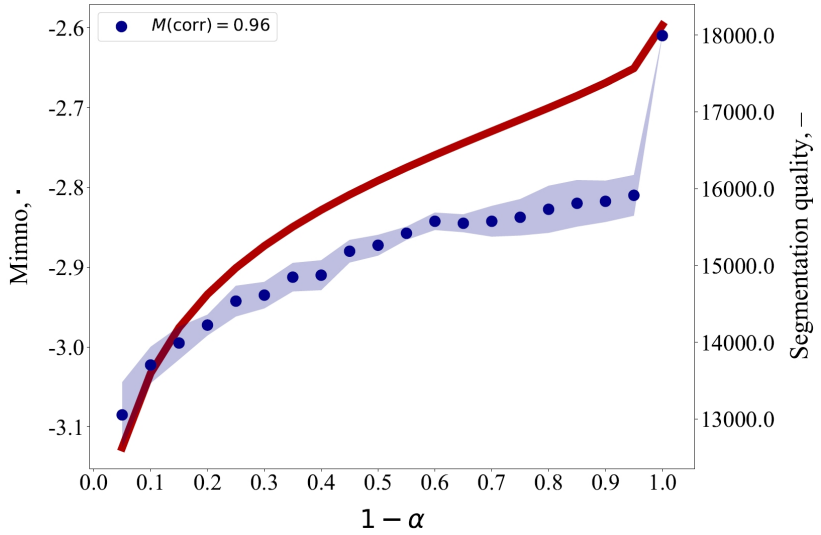
Результаты корреляций

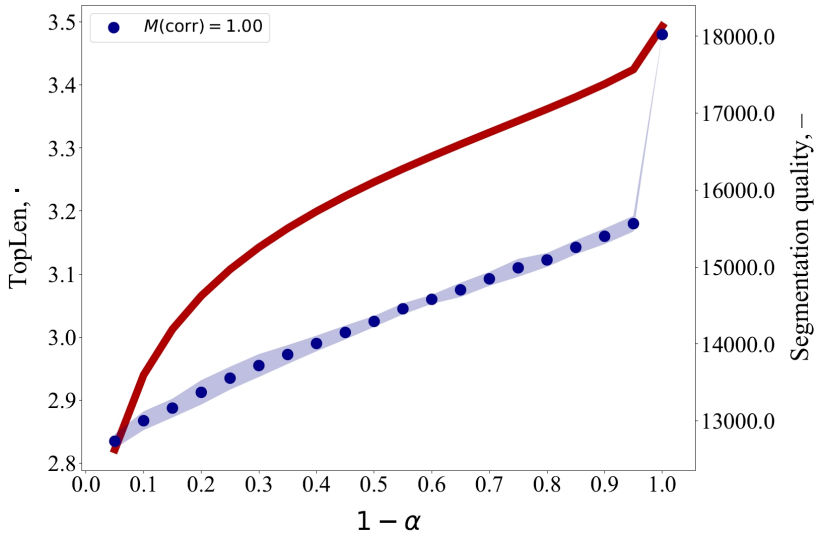
Coh	Corr
Newman	0.80
Mimno	0.94
SemantiC$_{l_2}$	0.70
SemantiC$_{\text{Var}}$	**1.00**
TopLen	**1.00**
FoCon	**1.00**

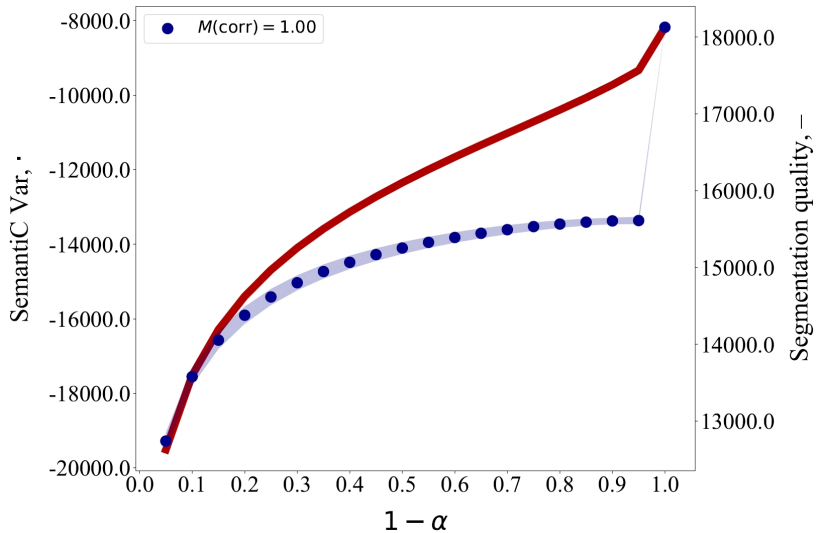
Корреляции при размере сегментов 200 слов
и при 5 темах в каждом сшитом из сегментов документе

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









Результаты

- 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.