Вероятностные тематические модели
Лекция 11.
Модели локального контекста

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ – ФИЦ ИУ РАН • 30 ноября 2023

Содержание

1. **Тематические модели сегментации текста**
   - Задачи сегментации и метод TopicTiling
   - Измерение и оптимизация качества сегментации
   - Тематические модели предложений
2. **Линейная тематизация текста**
   - Однопроходный E-шаг
   - Локализованный E-шаг
   - Модели внимания
3. **Регуляризация E-шага**
   - Постобработка E-шага
   - Регуляризация E-шага
   - Примеры регуляризаторов E-шага

Цели и прикладные задачи сегментации текстов

Цель: разделение текста на семантически однородные *сегменты* для поиска, классификации, суммаризации.

Примеры текстов, обладающих сегментной структурой

- научные статьи
- патенты
- учебные курсы
- юридические документы
- новостные дайджесты
- тексты резюме
- обсуждения в социальных медиа
- мультиязычные документы

M.A.Hearst. TextTiling: A Quantitative Approach to Discourse Segmentation. 1993.
I.Pak, P.L.Teh. Text Segmentation Techniques: A Critical Review. 2018.

Задача k-сегментации последовательности (k-segmentation)

Дано:

последовательность векторов $X = (x_i)_{i=1}^n$, $x_i \in \mathbb{R}^T$

Для текстов x_i — эмбеддинги слов / предложений / абзацев

Найти:

разбиение на k непересекающихся подпоследовательностей $S_1 \sqcup \cdots \sqcup S_k = X$ и систему их представителей $\mu_1, \ldots, \mu_k \in \mathbb{R}^T$

Критерий:

$$\sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|^2 \;\to\; \min_{\{S_j, \mu_j\}}$$

Оптимальное решение: динамическое программирование, $O(n^2 k)$

На практике используются приближённые эвристики, $O(nk)$

Richard Bellman. On the approximation of curves by line segments using dynamic programming. 1961.

Метод тематической сегментации TopicTiling

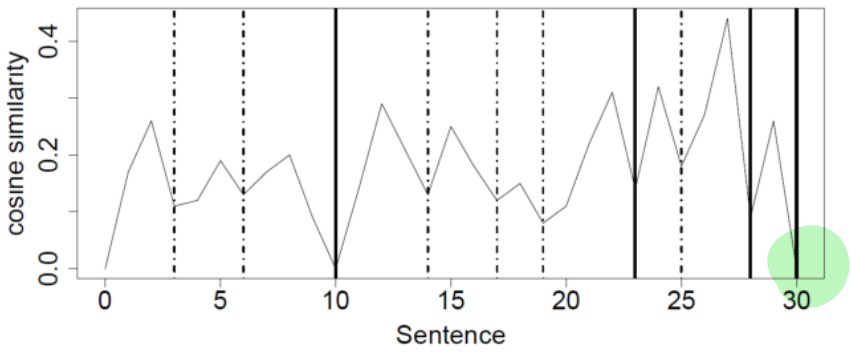
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

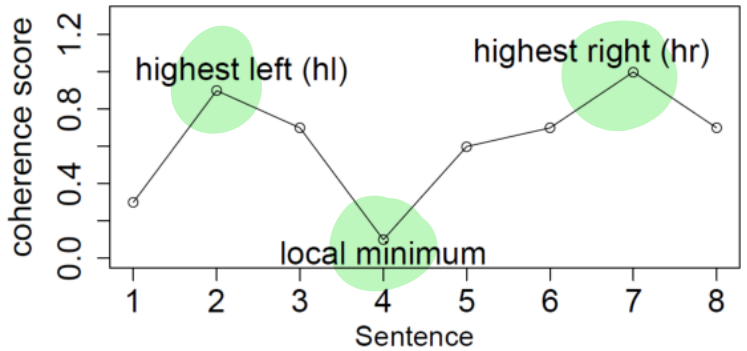
$p(t|d,s) = \frac{1}{|s|}\sum_{w\in s} p(t|d,w)$ — тематика предложения s

$p_j = \big(p(t|d,s_j)\big)_{t\in T}$ — тематический вектор предложения s_j

$c_j = \cos(p_{j-1}, p_j)$ — *coherence score*, оценка близости соседних предложений (чем глубже провал, тем чётче граница)

$d_j = \frac{1}{2}(\mathrm{hl}_j + \mathrm{hr}_j - 2c_j)$ — *depth score*, оценка глубины провала





Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Эвристики для TopicTiling

Эвристики для определения числа сегментов:

- заданное число провалов с наибольшей глубиной d_j
- провалы с глубиной более $\text{avr}\{d_j\} + \delta\,\text{stdev}\{d_j\}$, $\delta = 0{,}5..1{,}2$

Дополнительные эвристики и параметры:

- filter: игнорировать короткие предложения (менее 5 слов)
- игнорировать стоп-слова
- подбирать число предложений слева и справа от j

Эвристики для тематической сегментации:

- использовать фоновые темы и игнорировать их в p_j
- использовать $p(t|d,w)$ или $\arg\max_t p(t|d,w)$
- подбирать число итераций
- подбирать параметры $|T|$, α, β в модели LDA

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Измерение качества сегментации

Базовые методы сегментации по векторам $p(w|s_j)$ и $p(t|s_j)$

- TT и TT-LDA — TextTiling (Hearst, 1997)
- C99 и C99-LDA — кластеризация предложений (Choi, 2000)

Коллекции для сравнения методов сегментации:

- *Choi dataset*: синтетический корпус, 700 документов по 10 сегментов, нарезанных из «Brown corpus»
- *Galley dataset*: синтетический корпус, 500 документов по 4–22 сегментов, нарезанных из «WSJ corpus»

Метрики для сравнения методов сегментации:

- Precision/Recall не учитывают границы между сегментами
- P_k (Beeferman et al., 1997)
- WD, WindowDiff (Pevzner and Hearst, 2002)

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

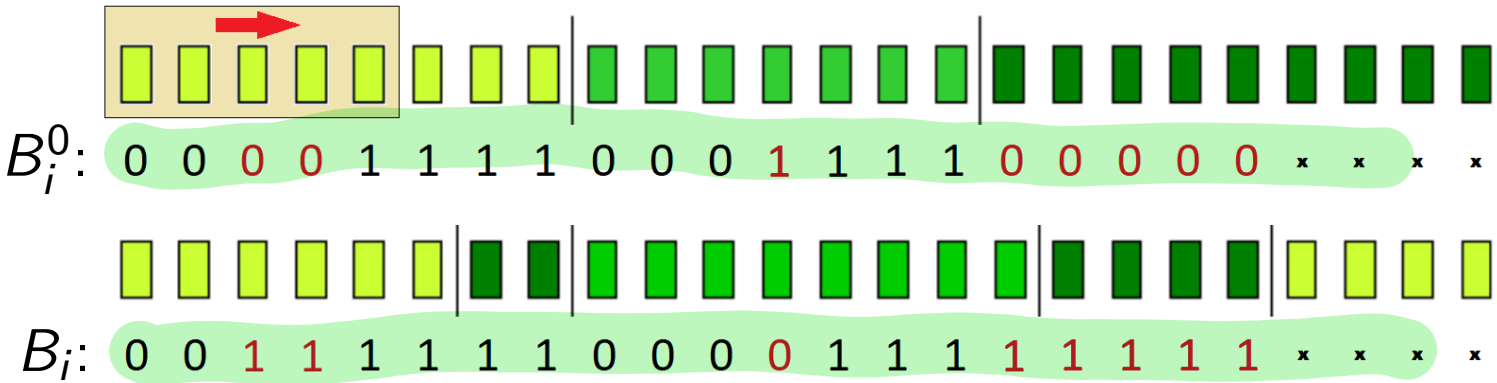
Метрики для сравнения методов сегментации

Сравнение с идеальной сегментацией (gold standard).

Метрика P_k — чем меньше, тем лучше:

- $B_i = [\text{словопозиции } i \text{ и } i+k-1 \text{ лежат в разных сегментах}]$
- B_i^0 — то же самое для идеальной сегментации
- P_k — доля позиций (в %), для которых $B_i \neq B_i^0$

Пример: $k = 5, \quad P_k = \frac{8}{20} = 40\%$



Doug Beeferman, Adam Berger, John Lafferty. Statistical models for text segmentation. 1999.

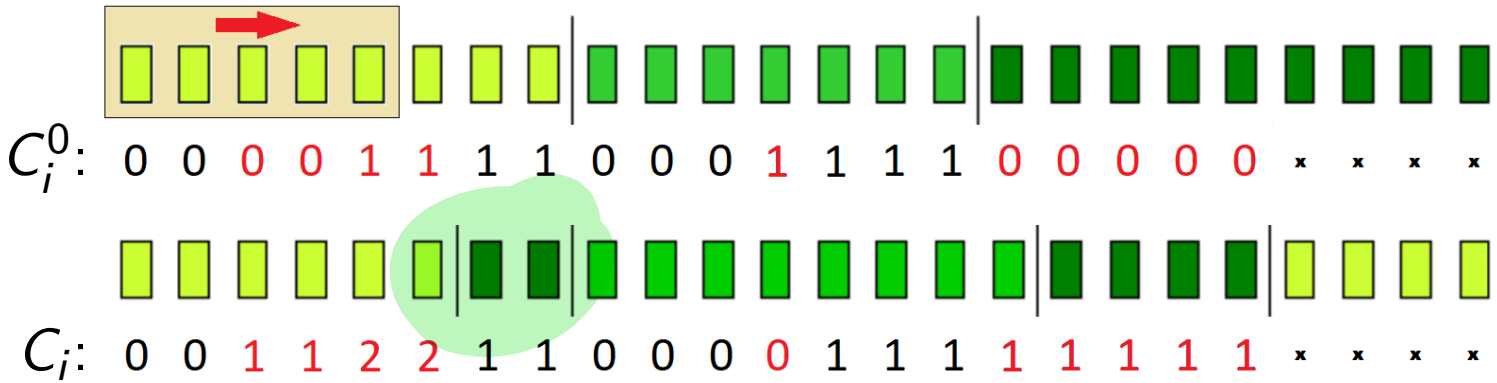
Метрики для сравнения методов сегментации

Сравнение с идеальной сегментацией (gold standard).

Метрика WD, WindowDiff — чем меньше, тем лучше:

- $C_i = (\text{число сегментов между позициями } i \text{ и } i+k-1)$
- C_i^0 — то же самое для идеальной сегментации
- WD — доля позиций (в %), для которых $C_i \neq C_i^0$

Пример: $k = 5$, $\text{WD} = \frac{10}{20} = 50\%$,



WD сильнее, чем P_k, штрафует короткие ложные сегменты

Lev Pevzner, Marti Hearst. A critique and improvement of an evaluation metric for text segmentation. 2002.

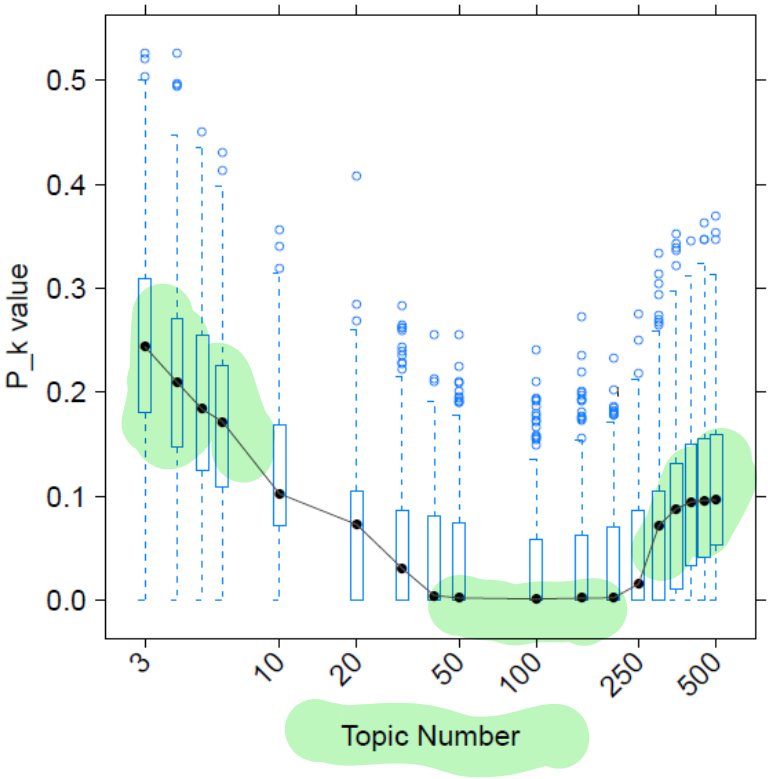
Результаты сравнения методов сегментации (Choi dataset)

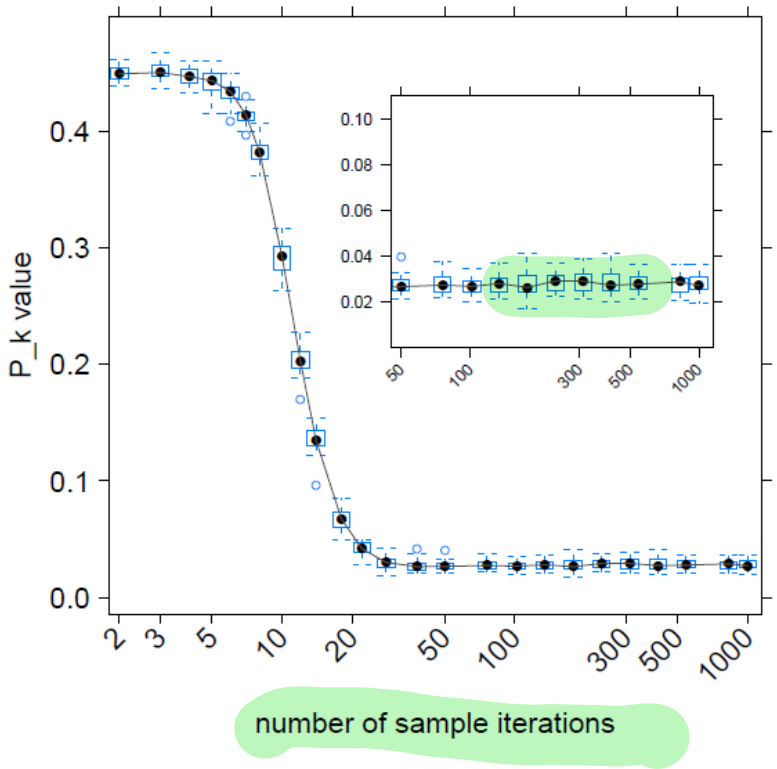
Method	Segments provided		Segments unprovided	
	P_k	WD	P_k	WD
C99	11.20	12.07	12.73	14.57
C99LDA	4.16	4.89	8.69	10.52
TT	44.48	47.11	49.51	66.16
TTLDA	1.85	2.10	16.41	21.40
TopicTiling	2.65	3.02	4.12	5.75
TopicTiling (filtered)	**1.50**	**1.72**	**3.24**	**4.58**

- Тематические модели лучше
- Лидирует TopicTiling с фильтрацией коротких предложений
- «Segments provided» — число сегментов известно (на реальных данных это нереалистичное предположение)

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимости P_k ($k = 6$) от параметров модели

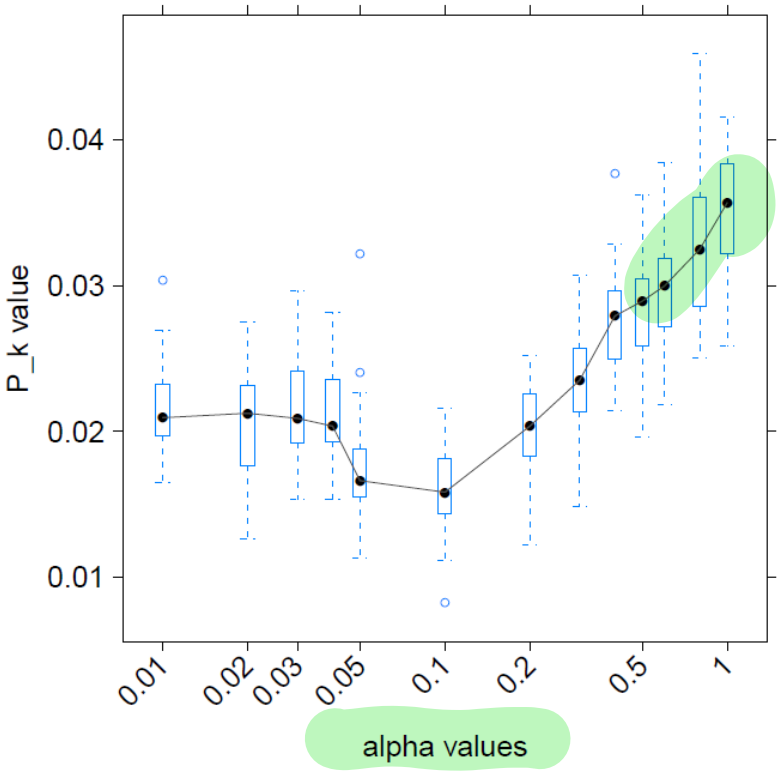


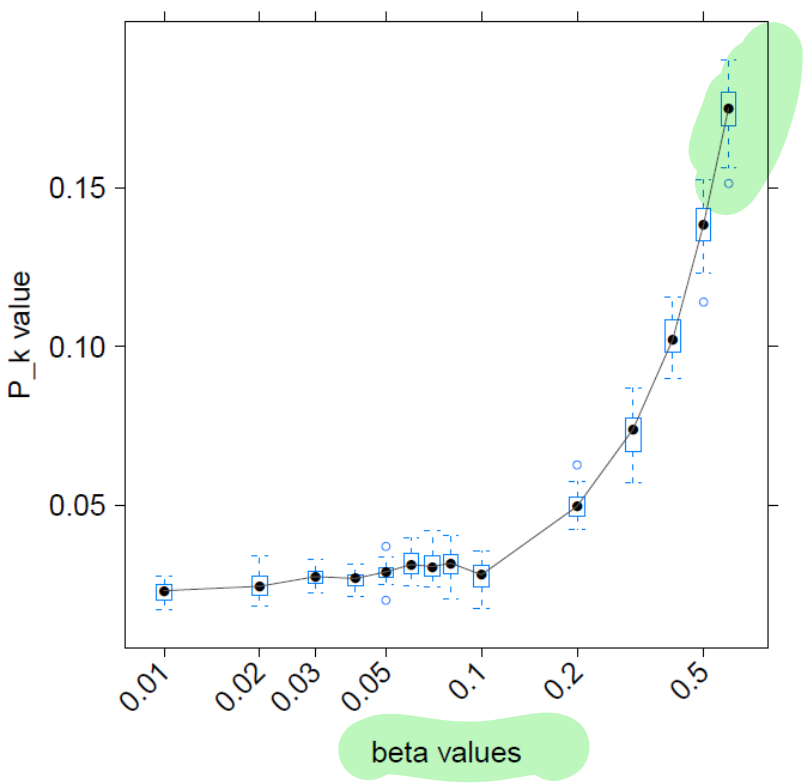


- Качество сегментации сильно зависит от $|T|$
- оптимальный диапазон $|T| = 50..150$ достаточно широк
- при $|T| = 100$ сходимость за 20–30 итераций

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимости P_k ($k = 6$) от параметров α, β модели LDA

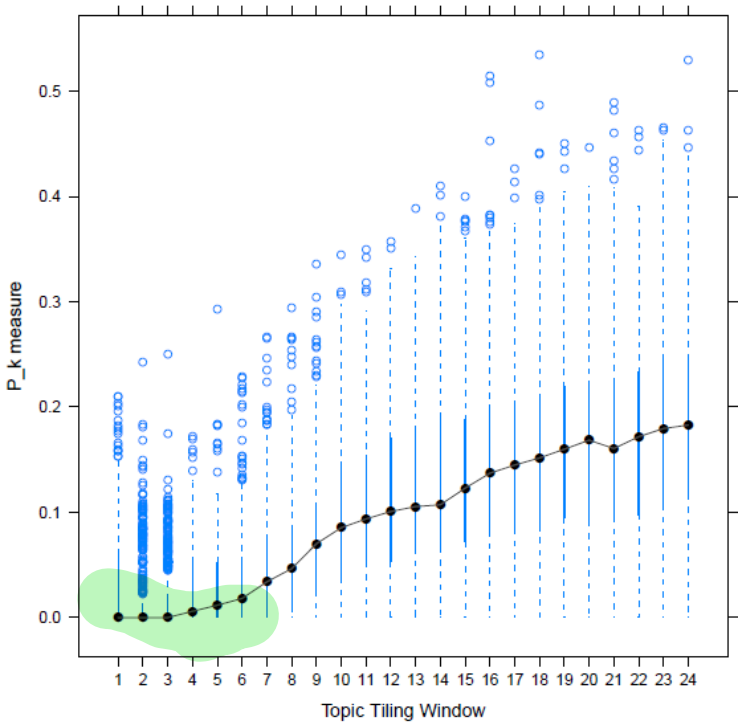




- Разреживать надо, но матрицу Θ — не слишком сильно
- параметры α, β менее критичны, чем число тем

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимость P_k ($k = 6$) от ширины окна w (window)



фиксированное число сегментов:

seg. size	3-5		6-8		9-11		3-11	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
d=false,w=1	2.71	3.00	3.64	4.14	5.90	7.05	3.81	4.32
d=true,w=1	3.71	4.16	1.97	2.23	2.42	2.92	2.00	2.30
d=false,w=2	1.46	1.51	1.05	1.20	1.13	1.31	1.00	1.15
d=true,w=2	**1.24**	**1.27**	**0.76**	**0.85**	**0.56**	**0.71**	**0.95**	**1.08**
d=false,w=5	2.78	3.04	1.71	2.11	4.47	4.76	3.80	4.46
d=true,w=5	2.34	2.65	1.17	1.35	4.39	4.56	3.20	3.54

определяемое число сегментов:

seg. size	3-5		6-8		9-11		3-11	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
d=false,w=1	**2.39**	**2.45**	4.09	5.85	9.20	15.44	4.87	6.74
d=true,w=1	3.54	3.59	1.98	2.57	3.01	5.15	2.04	2.62
d=false,w=2	15.53	15.55	0.79	0.88	1.98	3.23	1.03	1.36
d=true,w=2	14.65	14.69	**0.62**	**0.62**	**0.67**	**0.88**	**0.66**	**0.78**
d=false,w=5	21.47	21.62	16.30	16.30	6.01	6.14	14.31	14.65
d=true,w=5	21.57	21.67	17.24	17.24	6.44	6.44	15.51	15.74

- Оптимальная ширина окна $w = 2$–3 предложения
- «d=true»: усреднение $\arg\max_t p(t|d,w)$ по каждому w
- Почему они не догадались использовать $p(t|d,w)$?

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Эксперименты на более реалистичных данных Galley's WSJ

фиксированное число сегментов:

Parameters	All words		Filtered	
	P_k	WD	P_k	WD
d=false,w=1	37.31	43.20	37.01	43.26
d=true,w=1	35.31	41.27	33.52	39.86
d=false,w=2	22.76	28.69	21.35	27.28
d=true,w=2	21.79	27.35	19.75	25.42
d=false,w=5	14.29	19.89	12.90	18.87
d=true,w=5	**13.59**	**19.61**	**11.89**	**17.41**
d=false,w=10	14.08	22.60	14.09	22.22
d=true,w=10	13.61	21.00	13.48	20.59

определяемое число сегментов:

Parameters	All words		Filtered	
	P_k	WD	P_k	WD
d=false,w=1	53.07	72.78	52.63	72.66
d=true,w=1	53.42	74.12	51.84	72.57
d=false,w=2	46.68	65.01	44.81	63.09
d=true,w=2	46.08	64.41	43.54	61.18
d=false,w=5	30.68	43.73	28.31	40.36
d=true,w=5	28.29	38.90	26.96	36.98
d=false,w=10	19.93	32.98	18.29	29.29
d=true,w=10	**17.50**	**26.36**	**16.32**	**24.75**

- Качество сегментации сильно зависит от коллекции
- Определять число сегментов стало труднее
- Окно пришлось расширить до $w = 5$–10 предложений
- Здесь «filtered» — учитывать только существительные, прилагательные и глаголы — помогает, но не сильно
- Возможно ли критерием качества сегментации повлиять на саму тематическую модель?

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Тематические модели предложений (или коротких текстов)

Примеры *коротких текстов* (short text):

- твиты одного автора
- комментарии в одном блоге
- заголовки новостей за один день
- заголовки статей в одном журнале
- реплики в одном диалоге клиента и оператора
- предложения в одном документе

Основные предположения о коротких текстах:

- границы короткого текста (сегмента) известны
- слов не хватает для надёжного определения тематики
- короткий текст относится только к одной теме
- текст может содержать фоновые слова общей лексики

Тематическая модель Twitter-LDA

Предположения:

1. Каждый автор $a \in A$ написал множество сообщений $d \in D_a$.
2. Каждое сообщение d относится к одной теме $p(t|d) \in \{0, 1\}$.
3. Есть фоновая тема $b \in T$ с распределением $p(w|b)$.
4. Вероятность фона одинакова для сообщений, $p(b|d) = \pi$.

Порождающий процесс:

Вход: распределения $p(w|t)$, $p(t|a)$
для всех авторов $a \in A$
  для всех сообщений $d \in D_a$ автора a
    выбрать тему t из $p(t|a)$, кроме фоновой, $t \neq b$;
    для всех позиций слов $i = 1, \ldots, n_d$ в сообщении d
      выбрать слово w_i из $(1 - \pi)p(w|t) + \pi p(w|b)$;

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models // ECIR 2011.

Тематическая модель предложений senLDA и её обобщение

S_d — множество сегментов, на которые разбит документ d;

n_s — длина сегмента s;

n_{sw} — число вхождений терма w в сегмент s.

Тематическая модель монотематичного сегмента:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}$$

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

В senLDA регуляризатор $R(\Phi,\Theta)$ — распределения Дирихле.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Тематическая модель предложений в ARTM

Это модель гиперграфа (вершины — слова, сегменты — рёбра):

$$\sum_{d\in D}\sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$p_{tds} \equiv p(t|d,s) = \operatorname*{norm}_{t\in T}\Big(\theta_{td} \prod_{w\in s} \phi_{wt}^{n_{sw}}\Big);$$

M-шаг:
$$\begin{cases} \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\Big); \quad n_{wt} = \sum\limits_{d\in D}\sum\limits_{s\in S_d} n_{sw} p_{tds} \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big); \quad n_{td} = \sum\limits_{s\in S_d} n_s p_{tds} \end{cases}$$

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Тематическая модель предложений с фоновой темой

Слова сегмента порождаются либо темой $p(w|t) = \phi_{wt}$, либо фоновым распределением $p(w) = \psi_w$ слов общей лексики:

$$\sum_{d\in D}\sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \big(x_{dsw}\phi_{wt} + (1-x_{dsw})\psi_w\big)^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

Варианты модели (что лучше? — открытая проблема):

- $x_{dsw} = x$ — доля тематических слов в коллекции
- $x_{dsw} = x_d$ — доля тематических слов в документе
- $x_{dsw} = [\phi_{wt} > \psi_w]$ — результат аналитической оптимизации для каждого слова $\langle d, s, w\rangle$ (возможно переобучение?)
- $\psi_w = \frac{n_w}{n}$ — фиксированное распределение
- ψ_w обучается по коллекции

Гиперграфовые тематические модели языка

Ребро гиперграфа формализует гипотезу, что термы связаны друг с другом по смыслу и порождаются одной общей темой.

Что может быть ребром гиперграфа:

- предложение
- синтагма, ветка синтаксического дерева
- именная группа
- факт «объект, субъект, действие»
- связанные термы в одном или соседних предложениях: два синонима, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст сообщения и его автор
- финансовая транзакция с текстом платёжного поручения

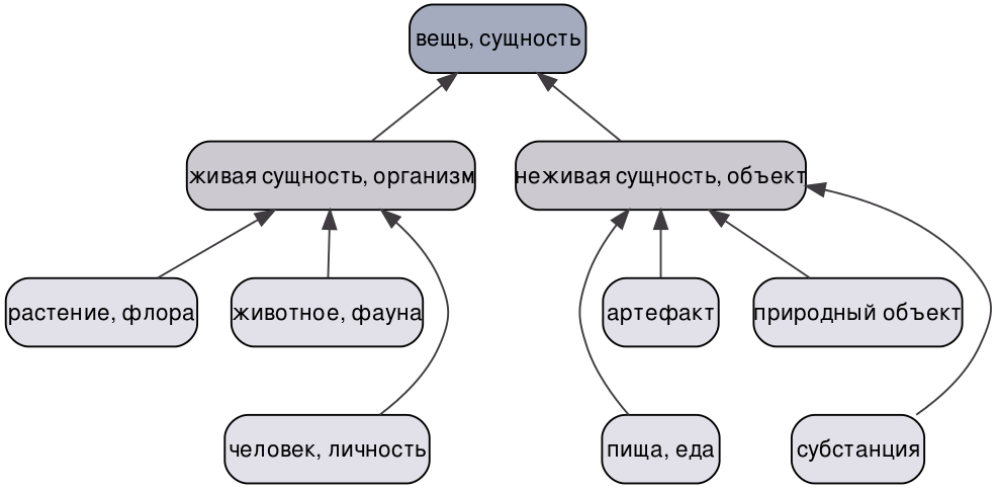
Семантическая сеть WordNet

117K *наборов синонимов* (synset), 155K слов, с определениями и примерами, связанных семантическими отношениями:

- *гипероним* — более общее (родовое) понятие
- *гипоним* — частное (видовое) понятие
- *холоним* — объемлющее целое
- *мероним* — составная часть

Словари разделены по частям речи:

- существительные
- глаголы
- прилагательные
- наречия



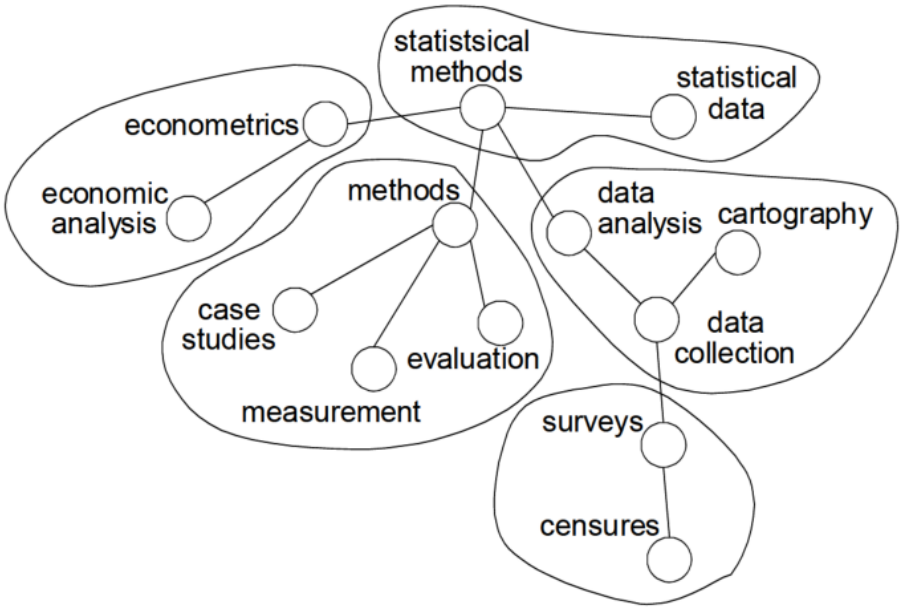
http://wordnet.princeton.edu

Метод лексических цепочек (Lexical Chains)

Лексическая цепочка — множество терминов:

- пары терминов связаны тезаурусными связями
- соседние термины на расстоянии не более 2 предложений
- возможна транзитивная связь через третий термин

Сильная цепочка — (почти) все слова связаны (клика)



Jane Morris, Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. 1991.

Пример выделения лексических цепочек

Пример использования русскоязычного тезауруса РуТез

О порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы.

Во исполнение Закона Российской Федерации "О статусе военнослужащих" и в целях обеспечения прав на жилище военнослужащих и граждан, уволенных с военной службы, Правительство Российской Федерации постановляет:
1. Утвердить прилагаемое Положение о порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы.
2. Министерству обороны Российской Федерации и иным федеральным органам исполнительной власти, в которых предусмотрена военная служба:
в месячный срок разработать и утвердить формы и перечень документов, необходимых для принятия решения об оказании военнослужащим безвозмездной

Возможна ли тематизация текста за один проход?

Дано: s — фрагмент текста, Φ — готовая тематическая модель
Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- как не переобучить вектор $p(t|s)$, если текст короткий?
- как согласовать $p(t|s)$ с объемлющим контекстом?
- как согласовать $p(t|s)$ с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ термов w?

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in W} n_{dw} p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\underset{t\in T}{\operatorname{norm}}(\phi_{wt}\theta_{td}^0)$$

- формула полной вероятности:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\,p(t|w) = \sum_{w\in d}\frac{n_{dw}}{n_d}\underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}p_t\big)$$

EM-алгоритм для ARTM без матрицы Θ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}};$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}};$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по Лемме о максимизации на симплексах)

Оптимизационная задача М-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}\ln\big(\phi_{us}\theta_{sd}(\Phi)\big) + R\big(\Phi,\Theta(\Phi)\big) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,s,u} n_{du}p_{sdu}\frac{\phi_{wt}}{\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{\phi_{wt}}{\theta_{sd}}\underbrace{\left(\sum_{u\in d} n_{du}p_{sdu} + \theta_{sd}\frac{\partial R}{\partial\theta_{sd}}\right)}_{n_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right) + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\underbrace{\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{n_{sd}}{\theta_{sd}}\phi_{wt}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right)}_{p'_{tdw}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}.$$

■

EM-алгоритм для линейной тематизации документов

$$\theta_{td}(\Phi) = \sum_{w \in d} p_{wd} \underset{t \in T}{\operatorname{norm}}(\phi_{wt} p_t) \quad \Rightarrow \quad \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = p_{wd} \phi'_{tw} \left(\delta_{st} - \phi'_{sw}\right)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w \in d} p_{wd} \phi'_{tw};$$

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw};$$

$$n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}};$$

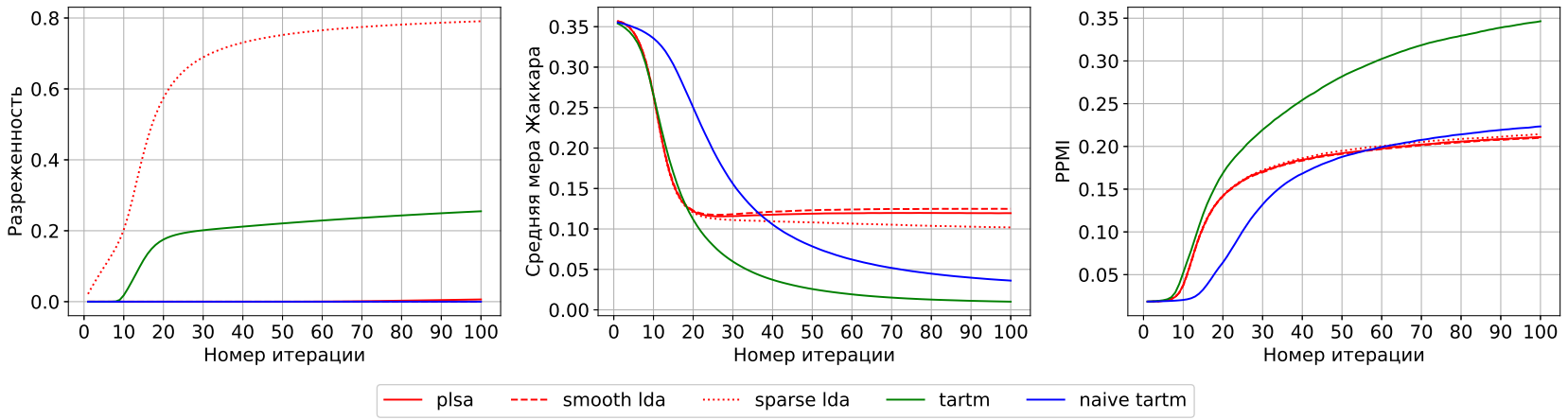
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right);$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

Упрощение EM-алгоритма для линейной тематизации

- Нет регуляризации по Θ, следовательно, $\frac{\partial R}{\partial \theta_{td}} = 0$
- Подстановка несмещённых оценок $\theta_{td} = \frac{n_{td}}{n_d}$, $\theta_{sd} = \frac{n_{sd}}{n_d}$ в формулу M-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w \in d} p_{wd} \phi'_{tw};$$

$$p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw};$$

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right).$$

Это обычный EM-алгоритм, только с однопроходным E-шагом!
ОГО! И ТАК МОЖНО БЫЛО?!

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \ldots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\ldots, w_i, \ldots)$ терма w_i:

$$\theta_{ti}(\Phi) \equiv p(t|i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Тематизация локального контекста с распределением весов:

$$\theta_{ti}(\Phi) \equiv p(t|i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i), \quad \sum_{u \in C_i} \alpha(u|i) = 1, \quad \alpha(u|i) \geqslant 0$$

Локализованная тематическая модель (похожа на BitermTM):

$$p(w|d,i) = \sum_{t \in T} p(w|t)\, p(t|i) = \sum_{t \in T} \phi_{wt} \sum_{u \in C_i} \phi'_{tu} \alpha(u|i)$$

EM-алгоритм с локализованным E-шагом

$w_1, \ldots, w_n$ — сквозная нумерация термов во всей коллекции

C_i — локальный контекст (окружение) терма w_i

$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для терма w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{ti} = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i);$$

$$p_{ti} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{w_i t} \theta_{ti}\right); \qquad n_t = \sum_{i=1}^{n} p_{ti};$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\sum_{i=1}^{n} [w_i = w] p_{ti} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right).$$

Быстрое вычисление двунаправленных векторов контекста

Два прохода по тексту — «слева направо» и «справа налево» для вычисления *экспоненциальных скользящих средних* (ЭСС):

$$\vec{p}(t|i) = \gamma_i\, p(t|w_i) + (1-\gamma_i)\, \vec{p}(t|i-1), \quad i = 1,\ldots,n, \quad \gamma_1 = 1$$
$$\overleftarrow{p}(t|i) = \gamma_i\, p(t|w_i) + (1-\gamma_i)\, \overleftarrow{p}(t|i+1), \quad i = n,\ldots,1, \quad \gamma_n = 1$$

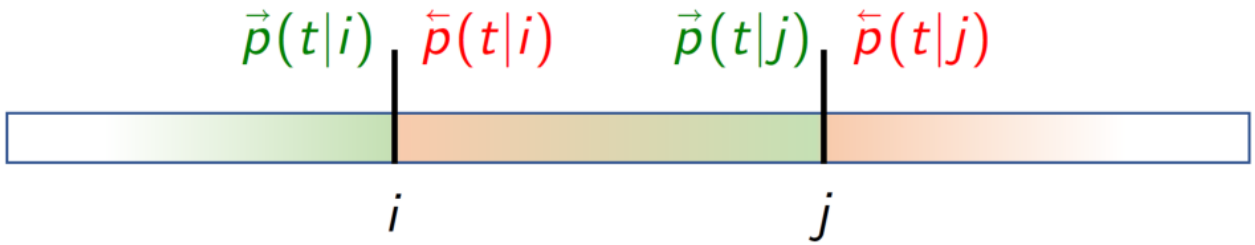
где γ_i — коэффициент сглаживания в позиции i

Основное свойство: если $\gamma_i = \gamma$, то $\alpha(w_k|i) = \gamma(1-\gamma)^{|i-k|}$

Несколько соображений, как распоряжаться выбором γ_i:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм
- γ_i можно умножать на оценку важности терма

Использование двунаправленных векторов контекста



Двунаправленные тематические векторы определяют:

- $\vec{p}(t|i)$ — тематику левого контекста терма w_i
- $\overleftarrow{p}(t|i)$ — тематику правого контекста терма w_i
- $\frac{1}{2}\left(\vec{p}(t|i) + \overleftarrow{p}(t|i)\right)$ — тематику двустороннего контекста w_i
- $p(t|i\ldots j) = \frac{1}{2}\left(\overleftarrow{p}(t|i) + \vec{p}(t|j)\right)$ — тематику сегмента $[i\ldots j]$
- тематическую однородность сегмента $[i\ldots j]$:
  насколько распределения $\overleftarrow{p}(t|i)$ и $\vec{p}(t|j)$ схожи
- позиции i границ между сегментами:
  насколько распределения $\vec{p}(t|i)$ и $\overleftarrow{p}(t|i)$ не схожи
- короткие и длинные контексты при различных γ_i

Модель внимания Query–Key–Value

q — *вектор-запрос* для трансформации в вектор контекста z

$(k_1, \ldots, k_n)$ — *векторы-ключи*, чтобы сравнивать с q

$(v_1, \ldots, v_n)$ — *векторы-значения*, составляющие контекст

Модель внимания — это выпуклая комбинация векторов v_u:

$$z = \sum_u v_u \operatorname*{SoftMax}_u \langle k_u, q \rangle,$$

где $\langle k_u, q \rangle$ — *оценка релевантности* ключа k_u запросу q

Модель внутреннего внимания (самовнимания, self-attention):

$$z_i = \sum_u Vx_u \operatorname*{SoftMax}_u \langle Kx_u, Qx_i \rangle$$

трансформирует последовательность векторов $(x_1, \ldots, x_n)$
в выходную последовательность векторов контекста $(z_1, \ldots, z_n)$

Vaswani et al. Attention is all you need. 2017.
Dichao Hu. An Introductory Survey on Attention Mechanisms in NLP Problems. 2018.

Сравнение локализованного E-шага с моделью self-attention

Тематический вектор локального контекста на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\left(\phi_{w_i t}\theta_{ti}\right) = \underset{t\in T}{\text{norm}}\Big(\sum_{u\in C_i} \phi'_{tu}\phi_{w_i t}\alpha(u|i)\Big)$$

Вектор контекста (эмбединг) на выходе модели внимания:

$$z_i = \sum_{u\in C_i} Vx_u\,\alpha(u|i) = \sum_{u\in C_i} Vx_u\,\underset{u\in C_i}{\text{SoftMax}}\langle Qx_i, Kx_u\rangle.$$

Сходство:

- вектор терма w_i трансформируется в вектор его контекста
- путём усреднения векторов ϕ'_u из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i.

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i};
- нет обучаемых параметров Q, K, V как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

Регуляризатор E-шага: $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$:

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

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$\begin{cases} p_{tdw} = \underset{t\in T}{\operatorname{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \end{cases}$$

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Набросок доказательства: три шага

1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

2. Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw};\quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

3. Подставляем это в формулы M-шага:

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

Шаг 1. Замечательное тождество

Для функции $p_{tdw}(\Phi,\Theta) = \frac{\phi_{wt}\theta_{td}}{\sum_z \phi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Воспользуемся определением функции $p_{tdw}(\Phi,\Theta)$:

$$\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}} = \phi_{wt}\frac{[z=t]\theta_{td}\sum_u \phi_{wu}\theta_{ud} - \theta_{td}\phi_{wz}\theta_{zd}}{(\sum_u \phi_{wu}\theta_{ud})^2} =$$

$$= p_{tdw}[z=t] - p_{tdw}p_{zdw} = p_{tdw}\big([z=t] - p_{zdw}\big);$$

$$\theta_{td}\frac{\partial p_{zdw}}{\partial \phi_{td}} = \theta_{td}\frac{[z=t]\phi_{wt}\sum_u \phi_{wu}\theta_{ud} - \phi_{wt}\phi_{wz}\theta_{zd}}{(\sum_u \phi_{wu}\theta_{ud})^2} =$$

$$= p_{tdw}[z=t] - p_{tdw}p_{zdw} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

Шаг 2. Дифференцирование суперпозиции $R(\Pi(\Phi,\Theta),\Phi,\Theta)$

Пусть $R(\Pi)$ не зависит от переменных p_{tdw} при $w \notin d$. Тогда

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_d p_{tdw} Q_{tdw}; \quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_w p_{tdw} Q_{tdw}$$

где $Q_{tdw} = \dfrac{\partial R}{\partial p_{tdw}} - \sum\limits_{z\in T} p_{zdw}\dfrac{\partial R}{\partial p_{zdw}}$.

Заметим: $\dfrac{\partial p_{zdw'}}{\partial \phi_{wt}} = 0,\ w \neq w'$; $\quad \dfrac{\partial p_{zd'w}}{\partial \theta_{td}} = 0,\ d \neq d'$; $\quad \dfrac{\partial R}{\partial p_{tdw}} = 0,\ w \notin d$.

$$\phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}} = \phi_{wt}\left(\frac{\partial R}{\partial \phi_{wt}} + \sum_{z,d,w'} \frac{\partial R}{\partial p_{zdw'}}\frac{\partial p_{zdw'}}{\partial \phi_{wt}}\right) = \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \sum_{d,z} \frac{\partial R}{\partial p_{zdw}}\phi_{wt}\frac{\partial p_{zdw}}{\partial \phi_{wt}}$$

$$\theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\left(\frac{\partial R}{\partial \theta_{td}} + \sum_{z,d',w} \frac{\partial R}{\partial p_{zd'w}}\frac{\partial p_{zd'w}}{\partial \theta_{td}}\right) = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w,z} \frac{\partial R}{\partial p_{zdw}}\theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}}$$

В силу «замечательного тождества» шага 1

$$\sum_{z\in T} \frac{\partial R}{\partial p_{zdw}} p_{tdw}\big([z=t] - p_{zdw}\big) = p_{tdw} Q_{tdw}.$$

Шаг 3. Подстановка производных $\tilde{R}(\Phi,\Theta)$ в формулы M-шага

Точка максимума (Φ,Θ) регуляризованного log-правдоподобия

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

удовлетворяет системе уравнений относительно ϕ_{wt}, θ_{td}, p_{tdw}:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big);$$

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

Общий член в формулах M-шага переносится в E-шаг, если ввести новую переменную $\tilde{p}_{tdw} = p_{tdw}\big(1 + \frac{1}{n_{dw}}Q_{tdw}\big)$. ■

Любая пост-обработка E-шага — это регуляризатор $R(\Pi)$

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует преобразование $p_{tdw} \to \tilde{p}_{tdw}$.
Верно и обратное:

Теорема. Если на k-й итерации EM-алгоритма для каждого (d, w): $n_{dw} > 0$ в формулах M-шага вместо вектора $(p^k_{tdw})_{t \in T}$ подставить вектор $(\tilde{p}^k_{tdw})_{t \in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}^k_{tdw} = 1$, то это эквивалентно добавлению регуляризатора сглаживания–разреживания

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} (\tilde{p}^k_{tdw} - p^k_{tdw}) \ln p_{tdw}.$$

$p(t|d, w)$ можно подвергать любой разумной пост-обработке!
ОГО! И ТАК МОЖНО БЫЛО?!

Пример 1. Кросс-энтропийное разреживание $p(t|d,w)$

Путь каждый терм относится к небольшому числу тем:

$$\mathrm{KL}\Big(\tfrac{1}{|T|} \,\Big\|\, p(t|d,w)\Big) \to \max.$$

Суммируем по всем термам всех документов:

$$R(\Pi) = -\frac{\tau}{|T|}\sum_{d\in D}\sum_{w\in d} n_{dw}\sum_{t\in T}\ln p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} - \tau\Big(\tfrac{1}{|T|} - p_{tdw}\Big).$$

Интерпретация: Если $p_{tdw} < \frac{1}{|T|}$, то p_{tdw} станет ещё меньше.
Тематика терма концентрируется в небольшом числе тем.

Недостаток: Тематика соседних термов разреживается независимо.

Пример 2. Тематическая модель сегментированного текста

S_d — множество микро-сегментов документа d

n_{sw} — число вхождений терма w в сегмент s длины n_s

Тематика сегмента $s \in S_d$ — среднее по всем его термам:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

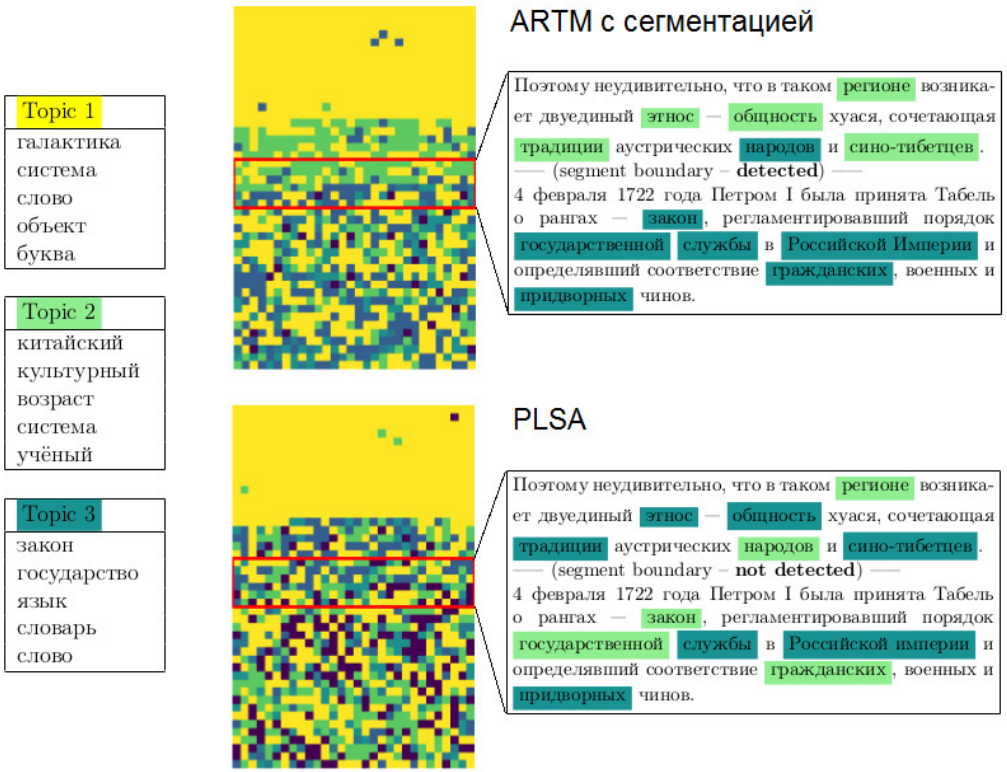
$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.

Тематика сегмента концентрируется в небольшом числе тем.

Пример 2. Эксперимент на полусинтетической коллекции

Сегментация текстов, склеенных из сегментов монотематичных статей научно-просветительского портала postnauka.ru



N.Skachkov, K.Vorontsov. Improving topic models with segmental structure of texts. Dialogue, 2018.

Теоретическое задание №6

1. Для тематической модели предложений с фоновой темой (слайд 19) выведите решение для переменной x_{dsw}.

2. Для тематической модели предложений с фоновой темой выведите формулы M-шага в случаях: $x_{dsw} = x$, $x_{dsw} = x_d$.

3*. Выведите EM-алгоритм с локализованным E-шагом (слайд 31) для локализованной тематической модели. Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw}?

4.** Предложите параметризацию для тематической модели внимания (слайд 35). Используя «основную лемму», получите уравнения для новых параметров модели.

Резюме

- Механизмы учёта порядка слов в ARTM:
  — модели сочетаемости пар слов: BitermTM, WNTM;
  — гиперграфовые модели, в том числе
  модели предложений: Twitter-LDA, senLDA;
  — тематическая сегментация (TopicTiling и др.);
  — регуляризация E-шага;
  — линейная однопроходная тематизация документов,
  в том числе линейная локальная тематизация и
  двунаправленные тематические модели контекста.
- Возможны ли тематические модели внимания?
  (есть много попыток объединять тематические модели
  с нейросетевыми моделями языка attention, transformer)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022
He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021
Liang Yang et al. Graph Attention Topic Modeling Network. 2020
Tian Tian et al. Attention-based Autoencoder Topic Model for Short Texts. 2019
Shuangyin Li et al. Recurrent Attentional Topic Model. 2017