Вероятностные тематические модели
Лекция 8. Модели связного текста и тематическая сегментация

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 • 28 октября 2020

Содержание

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 известны
  (задачу сегментации пока не ставим)
- 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

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

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

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

Гиперграфовые тематические модели языка

Что ещё может быть ребром гиперграфа?

Любое подмножество термов, связанных друг с другом по смыслу, и порождаемых одной общей темой.

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

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 d, s, w (возможно переобучение)
- ψ_w обучается по коллекции
- $\psi_w = \frac{n_w}{n}$ — фиксированное распределение

Контекстная документная кластеризация (CDC)

n_{uw} — частота сочетания пары слов u, w в некотором окне

$p(u|w) = \frac{n_{uw}}{n_w}$ — контекст слова w

$H(w) = -\sum_{u \in W} p(u|w) \log p(u|w)$ — энтропия контекста слова w

Узкий контекст — контекст с низкой энтропией, аналог темы, слова u, неслучайно часто встречающиеся рядом со словом w

Метод CDC — Contextual Document Clustering:

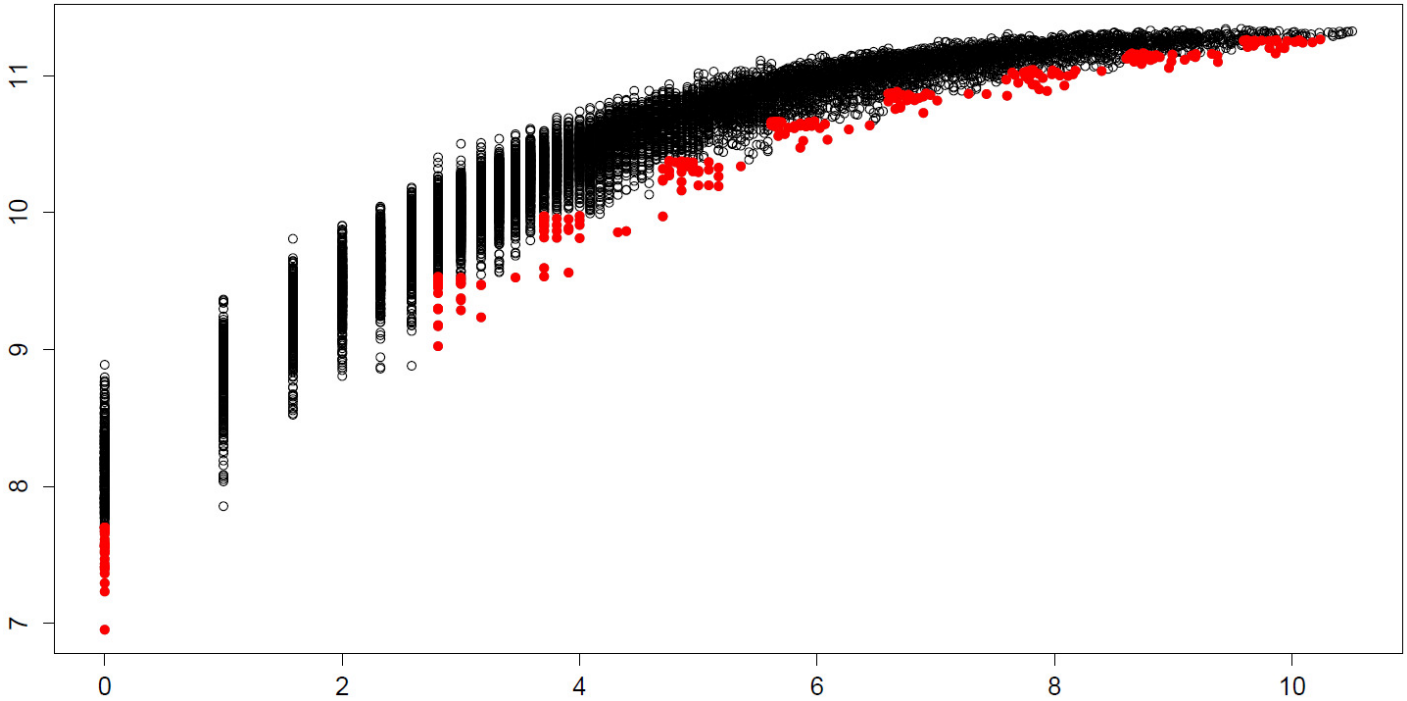
1. выделить «тематичные» слова с узкими контекстами
2. кластеризовать узкие контексты (найти темы)
3. разбить документы на однородные сегменты (абзацы)
4. отнести каждый сегмент к ближайшей теме

V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.

Выделение слов, имеющих узкие контексты

Оригинальный CDC: диапазон $\log_2 N_w$ разбивается на интервалы, в каждом интервале отбираются слова с наименьшими $H(w)$:

Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$

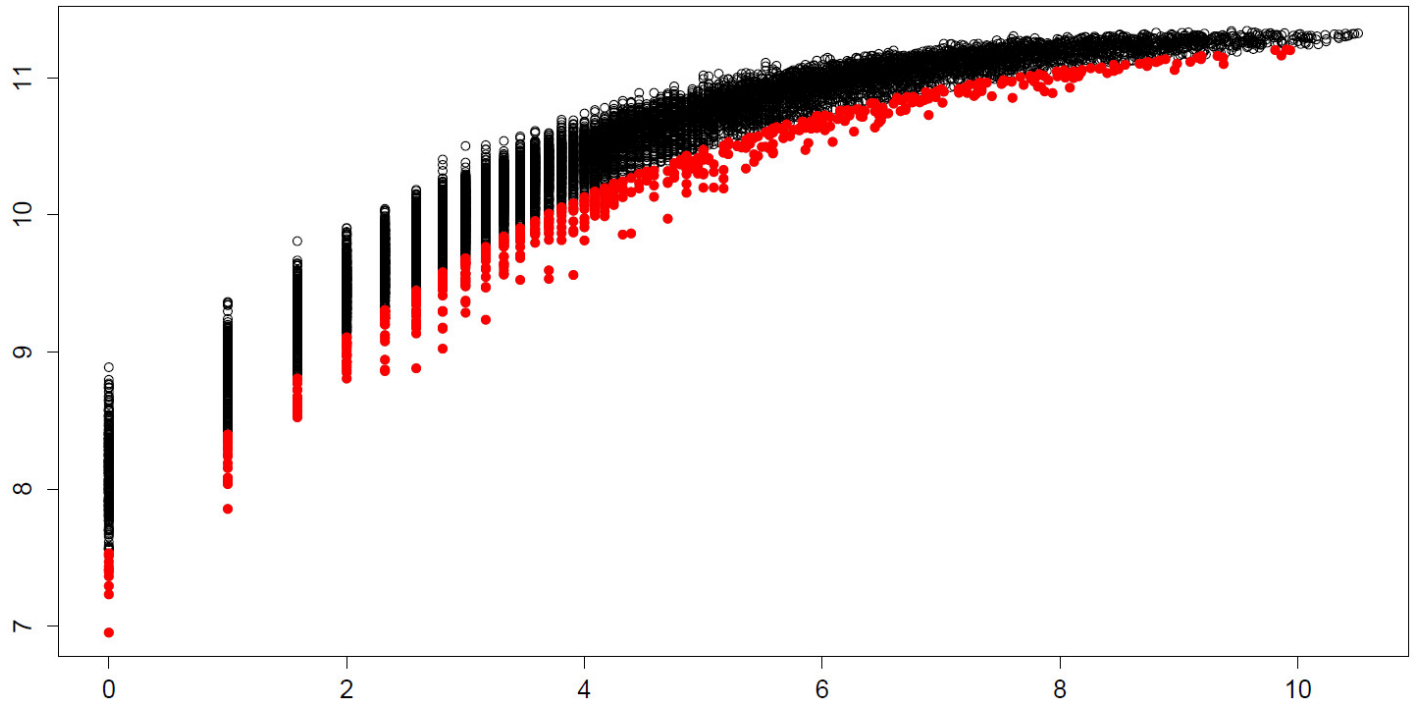


V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.

Выделение слов, имеющих узкие контексты

Более аккуратный отбор локальных контекстов
с помощью квантильной регрессии (отсекаем 5% снизу).

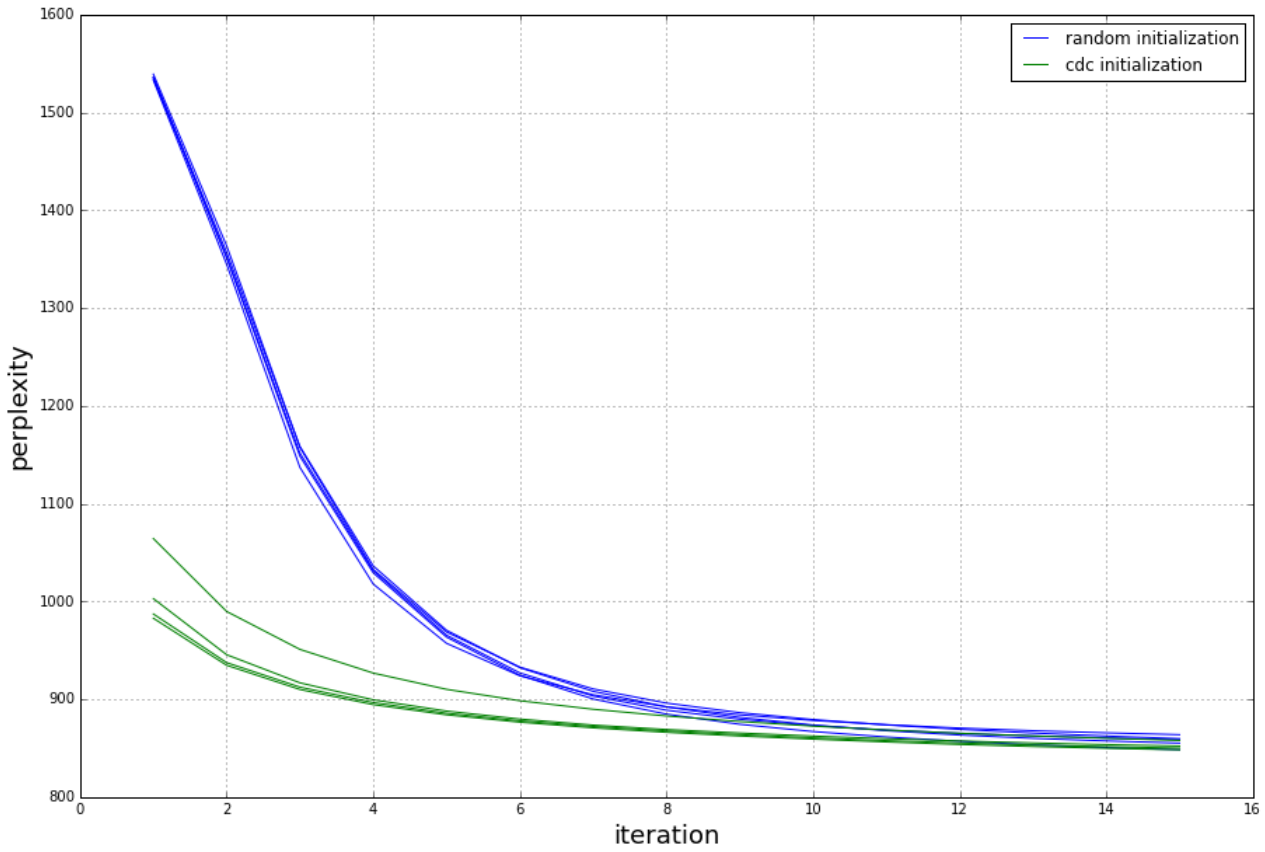
Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



А.Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей. 2015. МФТИ.

Инициализация тематической модели с помощью CDC

Зависимость перплексии от числа итераций (коллекция MMRO)



А.Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей. 2015. МФТИ.

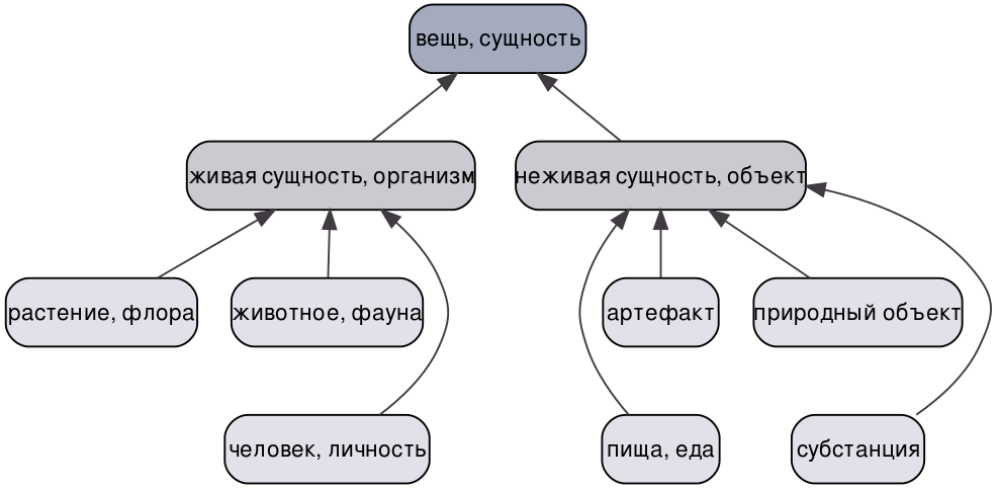
Семантическая сеть WordNet

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



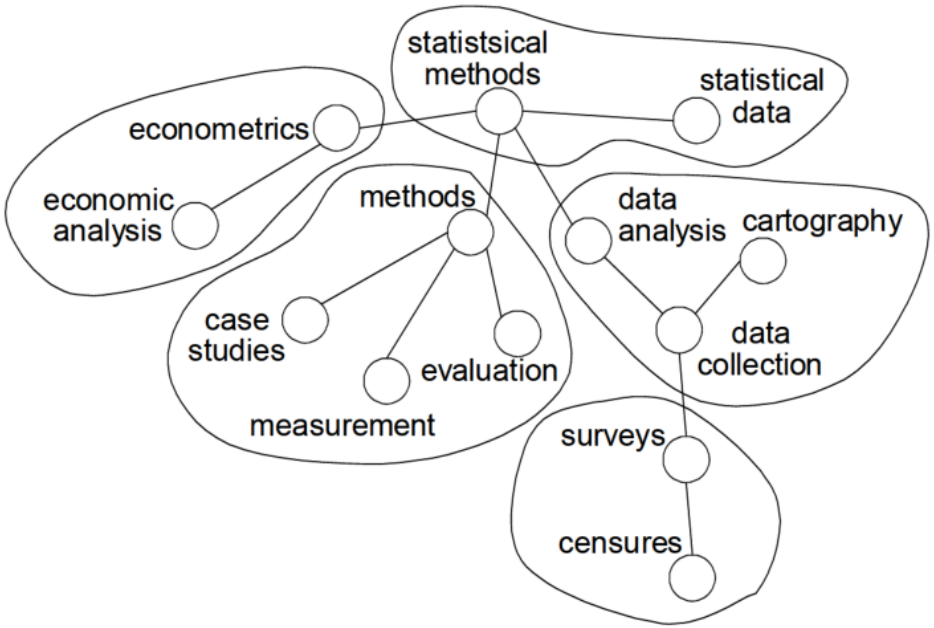
http://wordnet.princeton.edu

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

Применение TM для построения LC без тезауруса

LDA Mode Method (LDA-MM):

- тема каждого термина: $t(w) = \arg\max_t p(t|d,w)$
- термины с одинаковыми $t(w)$ образуют цепочку
- возможен учёт второй темы t' при $p(t'|d,w) > \varepsilon$

LDA Graph Method (LDA-GM):

- граф близостей всех терминов документа по $p(t|d,w)$
- максимальные клики этого графа образуют цепочки

LDA Top-N Method (LDA-TM):

- для каждого d выбираем top-N тем из $p(t|d)$
- для каждой t выбираем top-M терминов из $p(w|t)$
- все такие термины из d образуют цепочку

Steffen Remus. Automatically Identifying Lexical Chains by Means of Statistical Methods — A Knowledge-Free Approach. 2012.

Измерение качества построения лексических цепочек

Эксперты выделяли термины и лексические цепочки:

- по принципу однородности тематики
- повторения терминов, синонимы, коллокации, меронимы, гиперонимы, антонимы

	LDA-MM	LDA-GM	LDA-TM	S&M	G&M	Anno A	Anno B
avg. num. of lexical items per doc.	38.20	29.32	30.82	14.40	15.29	38.66	38.96
avg. num. of chains per doc.	13.80	9.12	7.32	5.83	5.71	11.25	7.38
avg. num. of links per doc.	8.60	2.06	1.44	–	–	5.47	2.41
avg. size lexical chains	2.82	3.41	4.61	2.48	2.68	3.69	5.57
avg. num. of merged lexical chains	5.76	7.06	5.98	–	–	6.10	4.99
avg. size merged lexical chains	8.29	4.45	5.57	–	–	7.60	8.91

Результаты:

- тематические модели сравнимы с экспертами
- тематические модели лучше семантических сетей

Steffen Remus. Automatically Identifying Lexical Chains by Means of Statistical Methods — A Knowledge-Free Approach. 2012.

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:

k непересекающихся подпоследовательностей $S_1 \sqcup \cdots \sqcup S_k = X$,

и (опционально) систему 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

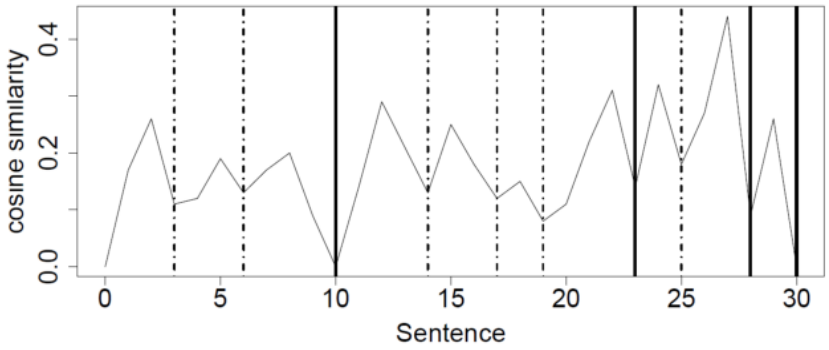
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

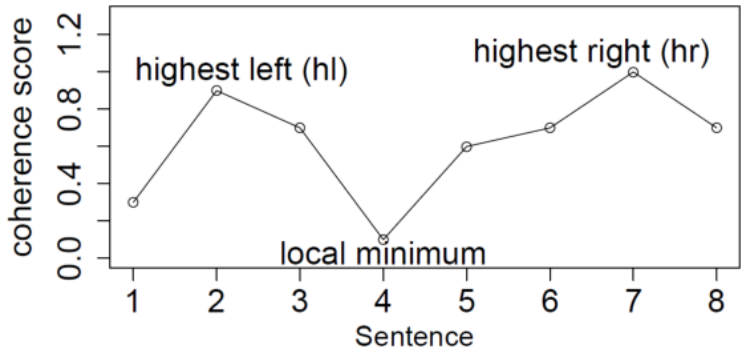
$p(t|d,s) = \frac{1}{|s|}\sum_{w\in s} p(t|d,w)$ — тематика предложения s

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

$d_j = \frac{1}{2}\big(\mathrm{hl}_j + \mathrm{hr}_j - 2c_j\big)$ — *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)

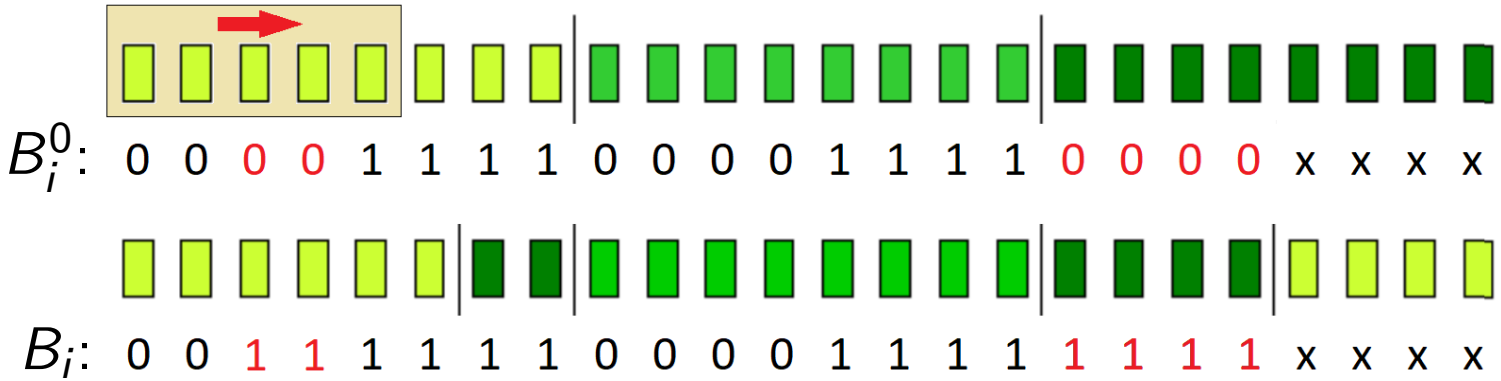
Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Метрики для сравнения методов сегментации

Сравнение с идеальной сегментацией (gold standard).

Метрика P_k — чем меньше, тем лучше:

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



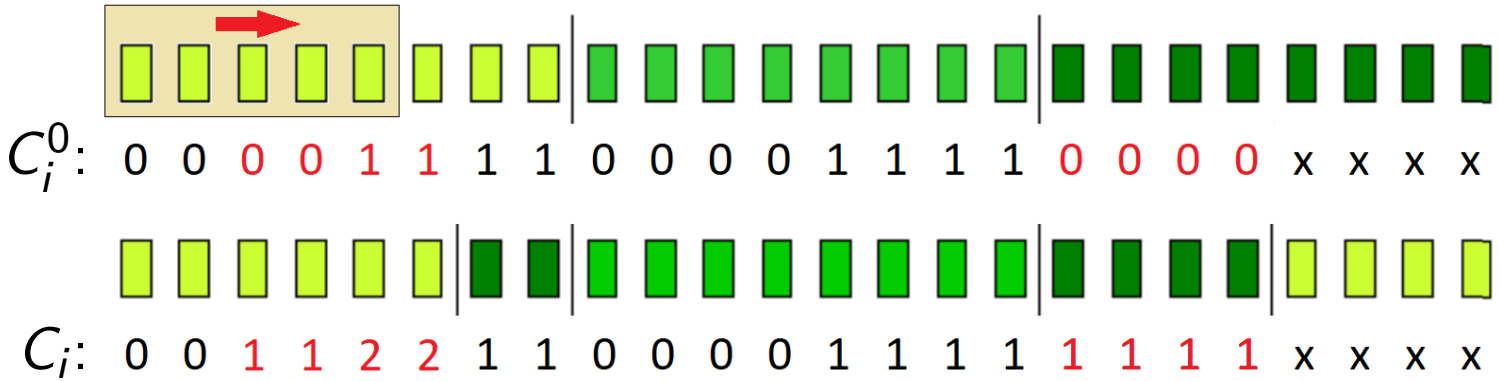
В данном примере $k = 5$, $P_k = \frac{6}{20}$

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:

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



WD $= \frac{8}{20}$ — сильнее штраф за короткие ложные сегменты

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

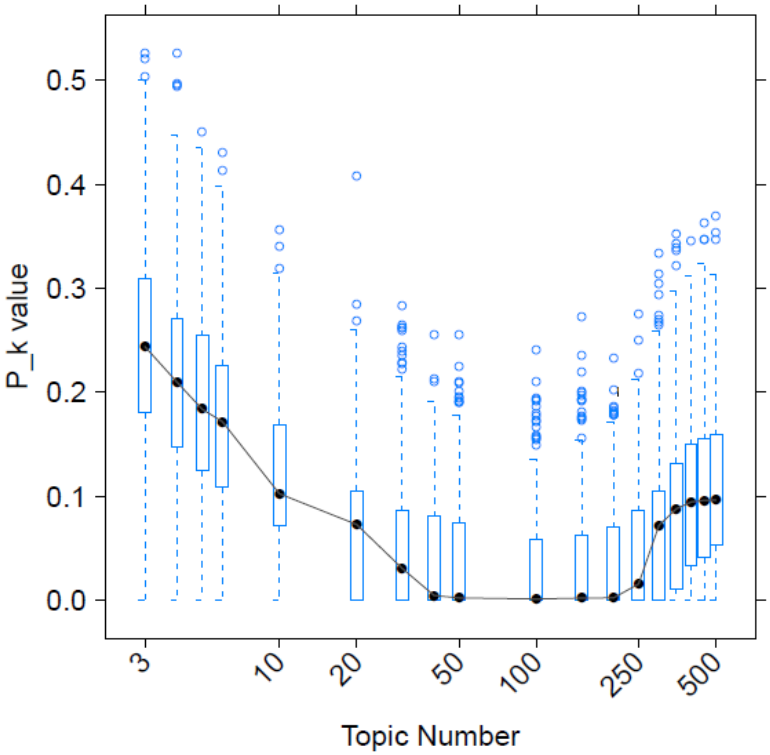
Результаты сравнения методов сегментации (Choi dataset)

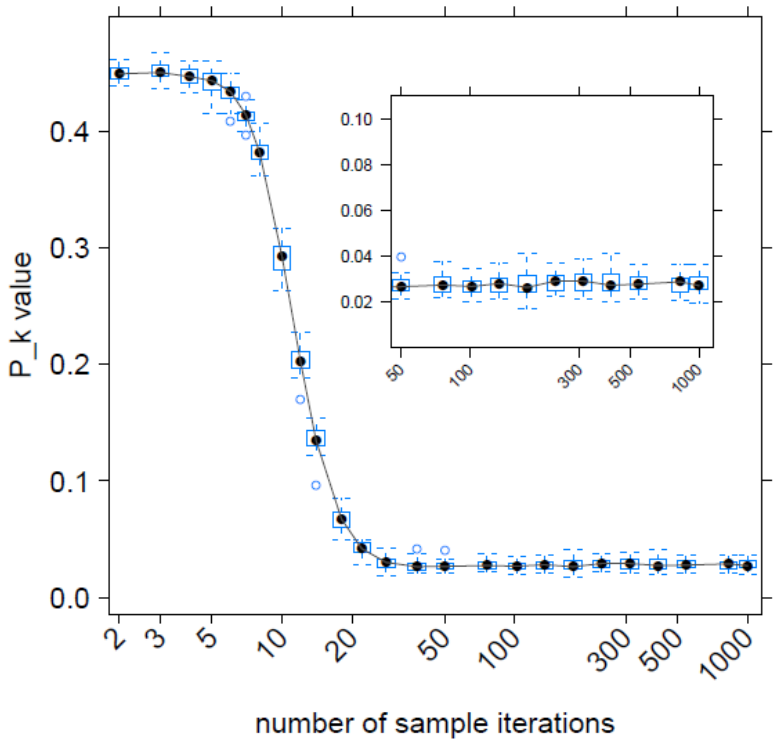
Method	Segments provided		Segments unprovided	
	P_k	WD	P_k	WD
C99	11.20	12.07	12.73	14.57
C99LDA	4.16	4.89	8.69	10.52
TT	44.48	47.11	49.51	66.16
TTLDA	1.85	2.10	16.41	21.40
TopicTiling	2.65	3.02	4.12	5.75
TopicTiling (filtered)	**1.50**	**1.72**	**3.24**	**4.58**

- 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

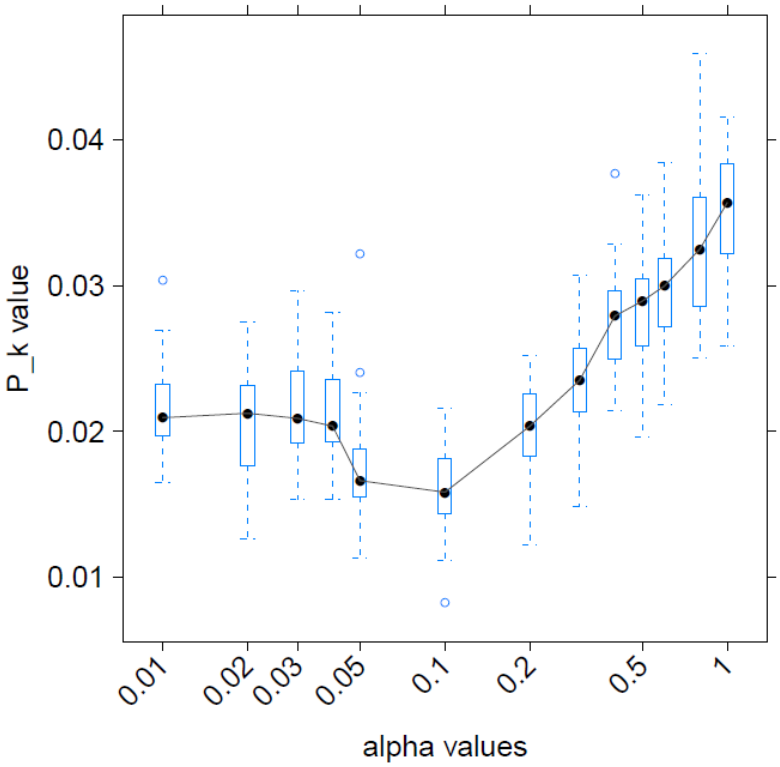


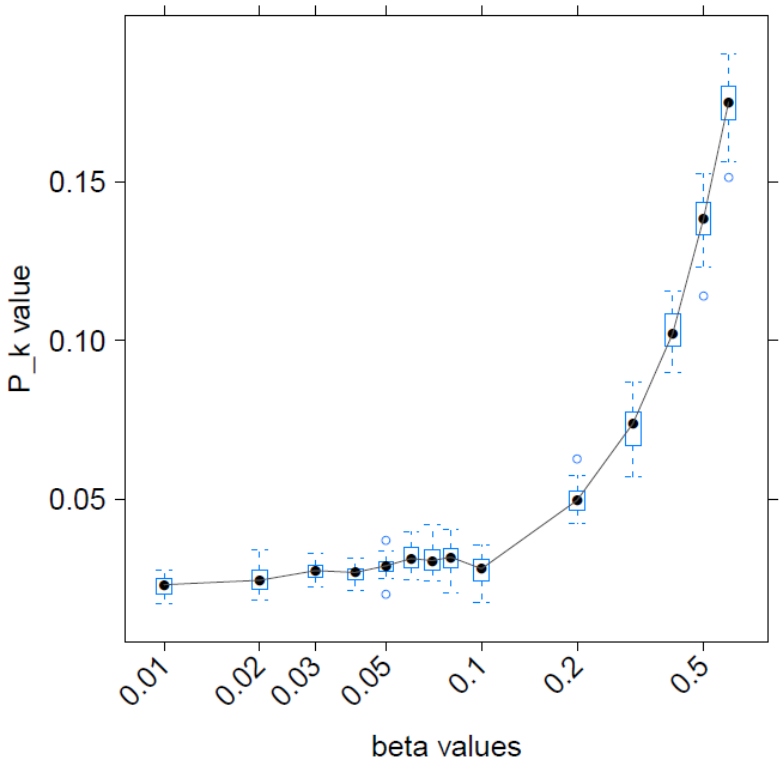


- Качество сегментации сильно зависит от $|T|$
- оптимальный диапазон $|T| = 50..150$ достаточно широк
- при $|T| = 100$ сходимость за 20–30 итераций

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

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

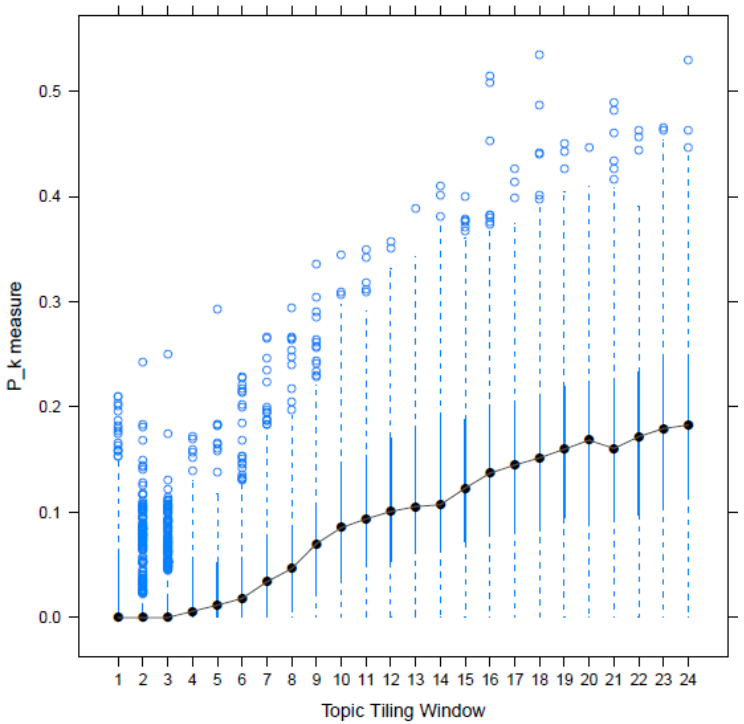




- Разреживать надо, но матрицу Θ — не слишком сильно
- параметры α, β менее критичны, чем число тем

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

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

Напоминания. ARTM — аддитивная регуляризация

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

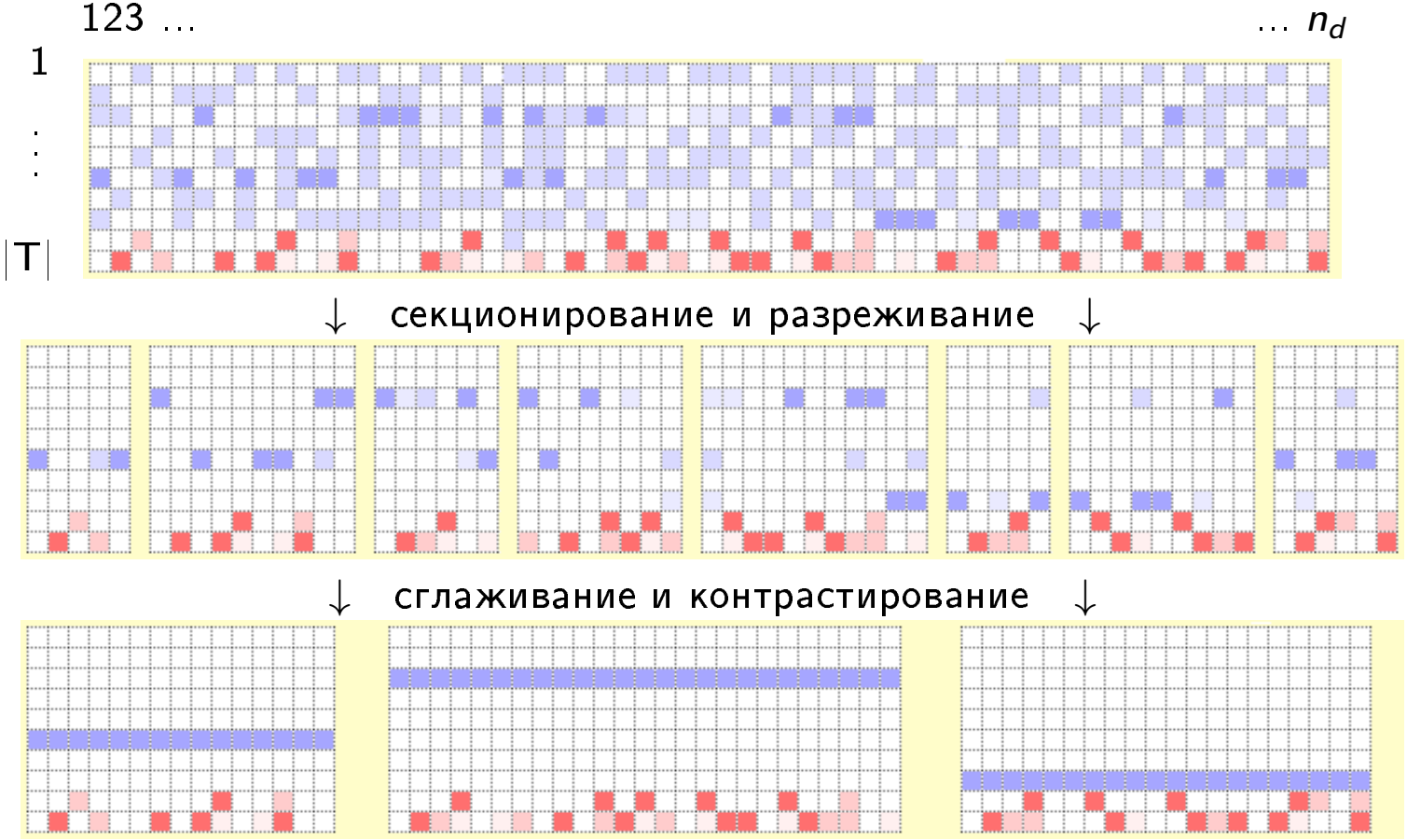
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum\limits_{d\in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum\limits_{w\in d} n_{dw} p_{tdw} \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

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: $\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Набросок доказательства: три леммы

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

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

Введём вспомогательную функцию от переменных Π, Φ, Θ:

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

Лемма 2. Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

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

Лемма 3. Формулы M-шага:

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

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

Лемма 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 $p_{tdw}(\Phi,\Theta)$:

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

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

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

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

Лемма 2. Дифференцируем суперпозицию $R(\Pi(\Phi,\Theta),\Phi,\Theta)$

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

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

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

В силу леммы 1 справедливо тождество

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

Лемма 3. Подставляем производные $\tilde{R}(\Phi, \Theta)$ в M-шаг

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 $\tilde{p}_{tdw} = p_{tdw}\big(1 + \frac{1}{n_{dw}} Q_{tdw}\big)$.

Гипотеза о пост-обработке E-шага

Между E- и M-шагом добавляется обработка матрицы (p_{tdw}) тематических векторов последовательности термов документа:

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \qquad (1)$$

Пост-обработка E-шага позволяет учитывать порядок термов в документе в обход гипотезы «мешка слов».

Гипотеза

Любое «разумное» преобразование $p_{tdw} \to \tilde{p}_{tdw}$ эквивалентно некоторому регуляризатору $R(\Pi(\Phi, \Theta))$.

Открытый вопрос: при каких условиях по заданным p_{tdw} и $\tilde{p}_{tdw}$ возможно подобрать функцию $R(\Pi)$ так, чтобы выполнялось уравнение пост-обработки (1)?

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:

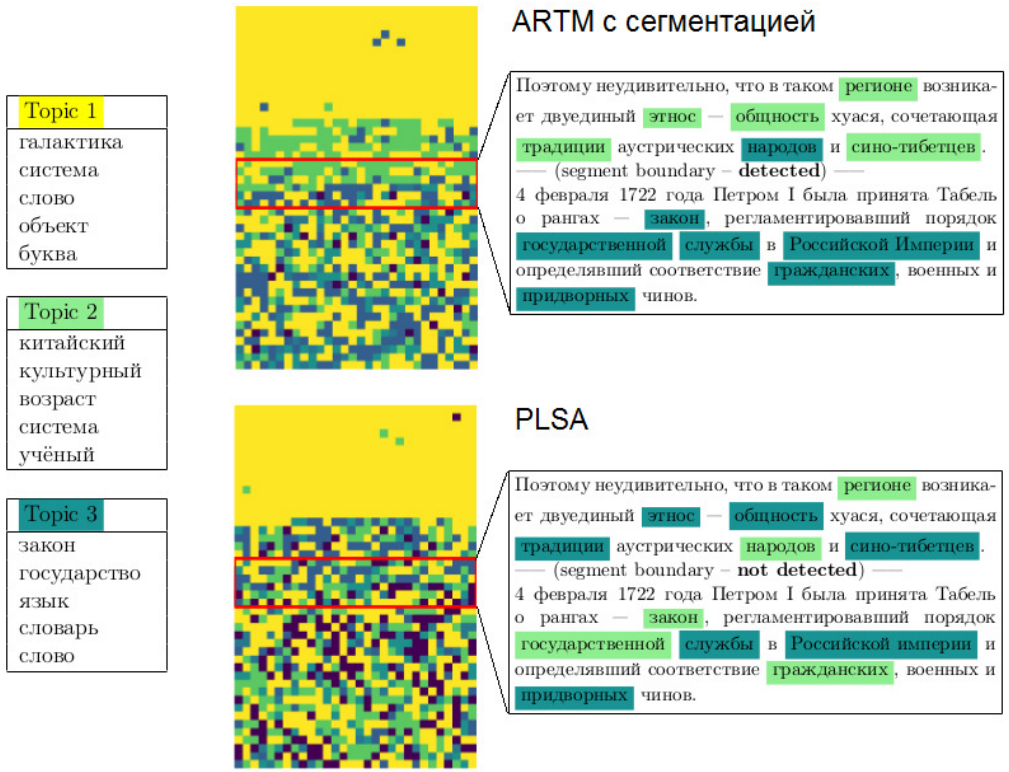
$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}}\right)\right).$$

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.

Резюме

- Гиперграфы и регуляризация E-шага — общие механизмы ARTM для учёта связности текста в обход «мешка слов»
- Twitter-LDA, senLDA — тематические модели предложений, могут быть реализованы через гиперграфовую ARTM
- TopicTiling — метод тематической сегментации
- CDC, LC — невероятностные подходы к выделению тематики текстов и тематической кластеризации
- Тематические модели предложений используются для сегментации и суммаризации текстов
- WindowDiff — мера качества сегментации