Вероятностные тематические модели
Лекция 8.
Короткие тексты и 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)»

ВМК МГУ • весна 2016

Содержание

Проблема коротких текстов

Короткие тексты (short text): Twitter и другие микроблоги, социальные медиа, заголовки статей и новостных сообщений.

Основные проблемы коротких сообщений:

- огромный объём ($\sim 10^9$ твитов в день)
- опечатки и намеренное искажение слов языка
- концентрация распределения $p(t|d)$ в одной теме
- выделение редких тем на фоне основных тем микроблогов (личная переписка, life style, репосты новостей)
- раннее обнаружение новых тем

Тривиальные подходы и их недостатки

- Считать каждое сообщение отдельным документом
  - для коротких сообщений $p(t|d)$ оценивается не надёжно
- Разреживать $p(t|d)$ вплоть до единственной темы; добавить модальности авторов, времени, регионов и т.п.
  - решение в духе APTM, пока не попробовали...
- Объединить сообщения по автору (времени, региону и т.п.)
  - появится дисбаланс документов по длине
  - появятся тематически неоднородные документы
- Объединить посты с комментариями
  - комментарии могут отсутствовать у большинства постов
- Дополнить коллекцию длинными текстами (Википедия и др.)
  - часть тем может не покрываться внешней коллекций
  - лексикон социальной сети может существенно отличаться

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 документов, $p(b|d) = \pi$.

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

Аналог Twitter-LDA с точностью до переобозначений:

1. Каждый документ d состоит из предложений $s \in S_d$.
2. Каждое предложение относится к одной теме $p(t|s) \in \{0, 1\}$.
3. Наблюдаемая выборка образуется тройками $(d_i, s_i, w_i)_{i=1}^n$.
4. Гипотеза условной независимости: $p(s, w|t) = p(s|t)p(w|t)$.

Тематическая модель сегментированного текста:

$$p(w, s|d) = \sum_{t \in T} p(w|t)p(s|t)p(t|d) = \sum_{t \in T} \phi_{wt}\psi_{st}\theta_{td}$$

Критерий максимума регуляризованного правдоподобия:

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

где n_{dsw} — частота термина w в предложении $s \in S_d$.

EM-алгоритм для модели сегментированного текста

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); & n_{wt} = \sum\limits_{d,s} n_{dsw}p_{tdsw} \\ \psi_{st} = \underset{s\in S_d}{\text{norm}}\left(n_{st} + \psi_{st}\frac{\partial R}{\partial \psi_{st}}\right); & n_{st} = \sum\limits_{w} n_{dsw}p_{tdsw} \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); & n_{td} = \sum\limits_{s,w} n_{dsw}p_{tdsw} \end{cases}$$

Битермы: модель совстречаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm topic model):

$$p(u,w) = \sum_{t\in T} p(w|t)p(u|t)p(t) = \sum_{t\in T} \phi_{wt}\phi_{ut}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,w} n_{uw} \ln \sum_t \phi_{wt}\phi_{ut}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts // WWW 2013.

Необходимые условия точки максимума правдоподобия

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{u,w} n_{uw} \ln \sum_{t} \phi_{wt}\phi_{ut}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

n_{uw} — частота битерма (u,w) в документах коллекции.

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tuw} \equiv p(t|u,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\phi_{ut}\pi_t\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum\limits_{u\in W} n_{uw} p_{tuw} \\ \pi_t = \underset{t\in T}{\text{norm}}\left(n_t + \pi_t \dfrac{\partial R}{\partial \pi_t} \right), \quad n_t = \sum\limits_{u,w\in W} n_{uw} p_{tuw} \end{cases}$$

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \to \max$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big);$$

$$p_{tuw} \equiv p(t|u,w) = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

2. Регуляризатор разреживания для матрицы Θ:

$$R(\Theta) = -\tau' \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Модель всплесков в темпоральной тематической модели

Гипотеза. В каждом временном интервале $i = 0, 1, \ldots, n$ каждое слово $w \in W$ либо связано с событием, либо фоновое.

Событийность (bursty probability) слова w в интервале i:

$$\eta_{iw} = \frac{(n_{iw} - \bar{n}_{iw} - k\sigma_{iw})_+}{n_{iw}},$$

$\bar{n}_{iw} = \frac{1}{i}\sum_{j<i} n_{jw}$ — среднее, $\sigma^2_{iw} = \frac{1}{i}\sum_{j<i}(n_{jw} - \bar{n}_{iw})^2$ — дисперсия.

Модель смеси событийных и фоновых слов в документе $d \in D_i$.

$$p(w|d) = \eta_{iw}\sum_{t\in S}\phi_{wt}\theta_{td} + (1 - \eta_{iw})\sum_{t\in B}\phi_{wt}\theta_{td},$$

В статье предлагался частный случай: $k = 0$, $|B| = 1$.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, Xueqi Cheng. A Probabilistic Model for Bursty Topic Discovery in Microblogs. 2015.

Модель всплесков BBTM (Bursty Biterm Topic Model)

Всё то же самое, только на битермах:

$$p(u,w) = \eta_{iw} \sum_{t \in S} \phi_{ut}\phi_{wt}\pi_t + (1-\eta_{iw}) \sum_{t \in B} \phi_{ut}\phi_{wt}\pi_t$$

Идея упрощения задачи: оценим снизу log-правдоподобие:

$$\sum_{i,u,w} n_{iuw} \ln\bigl(\eta_{iw} p_{uw} + (1-\eta_{iw}) q_{uw}\bigr) \geqslant$$

$$\sum_{i,u,w} n_{iuw}\eta_{iw} \ln p_{uw} + \sum_{u,w} n_{uw}\cancel{(1-\eta_{iw})} \ln q_{uw} \to \max_{\Phi}.$$

Можно ли взять только событийные слова, $n_{iuw} := n_{iuw}\eta_{iw}$, и вообще не моделировать фоновые слова?

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, Xueqi Cheng. A Probabilistic Model for Bursty Topic Discovery in Microblogs. 2015.

Интерпретируемости и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- Экспертные оценки:
  — интерпретируемость темы по балльной шкале;
  — каждую тему оценивают несколько экспертов.
- Метод интрузий (intrusion):
  — в список топовых слов внедряется лишнее слово;
  — измеряется доля ошибок экспертов его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

$$\mathrm{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j)$$

где w_i — i-й термин в порядке убывания ϕ_{wt}.

$\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\text{norm}}\left(n_{wt} + \tau \sum_{u\in W} C_{uw} n_{ut}\right).$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Альтернативный регуляризатор когерентности

Квадратичный регуляризатор Quad-Reg:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max,$$

$C_{uv} = N_{uv} \big[\mathrm{PMI}(u,v) > 0 \big]$,
N_{uv} — число документов, в которых u, v хотя бы раз встречаются на расстоянии не более 10 слов,
N_u — число документов, в которых u встречается хотя бы раз,
$\mathrm{PMI}(u,v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация.

В литературе пока не выработан окончательный вариант регуляризатора когерентности.

Newman D., Bonilla E. V., Buntine W. L. Improving topic coherence with regularized topic models. 2011.

Модель сети слов WNTM для коротких текстов

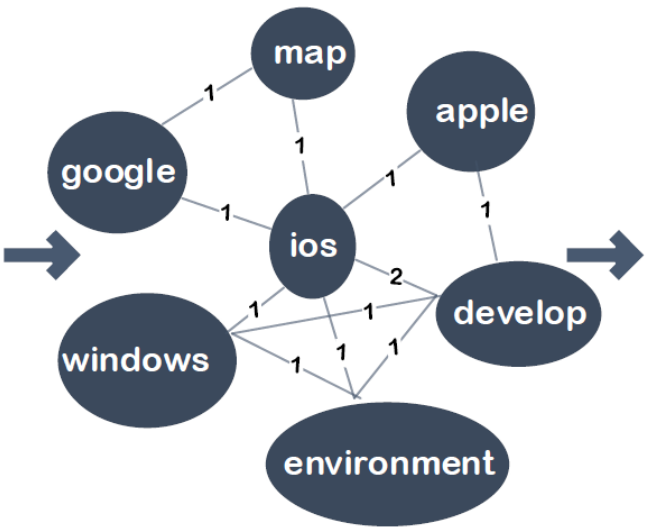
Идея: моделировать не документы, а связи между словами.

d_w — псевдо-документ, объединение всех контекстов слова w.

n_{wu} — число вхождений слова u в псевдо-документ d_w.

Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(u|d_w) = \sum_{t \in T} p(u|t)p(t|d_w) = \sum_{t \in T} \phi_{ut}\theta_{tw},$$

где d_w — псевдо-документ слова w.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{wu} \log \sum_{t \in T} \phi_{ut}\theta_{tw} \to \max_{\Phi,\Theta},$$

где n_{wu} — совстречаемость слов w, u.

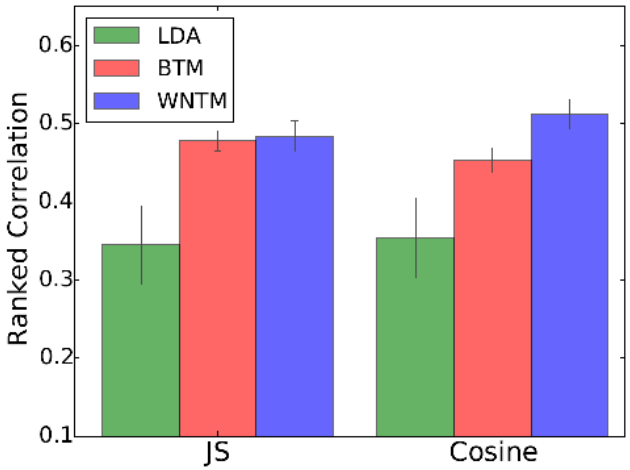
Отличие от модели битермов: там $\Theta = \mathrm{diag}(p_1, \ldots, p_t)\Phi^{\mathrm{T}}$.

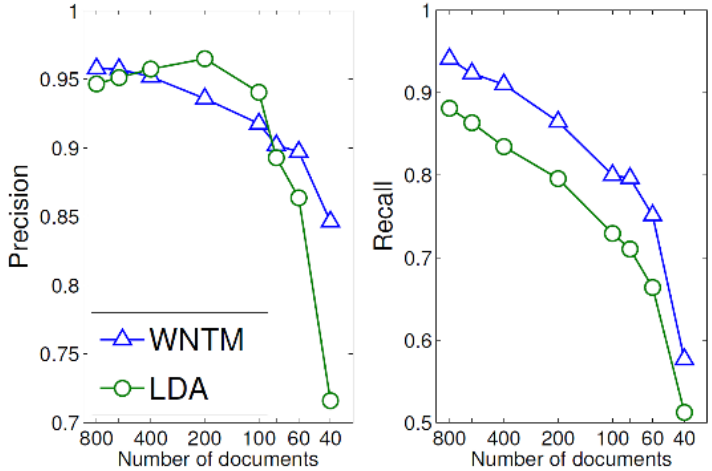
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription // ACM Trans., 2009.

Результаты оценивания модели WNTM

- Когерентность на коротких текстах лучше, чем у LDA и Biterm TM; на длинных текстах преимуществ нет.
- *Слева:* оценивание семантической близости слов по $p(t|w)$, корреляция с 10-балльными экспертными оценками.
- *Справа:* полнота и точность распознавания новой темы в зависимости от числа документов.



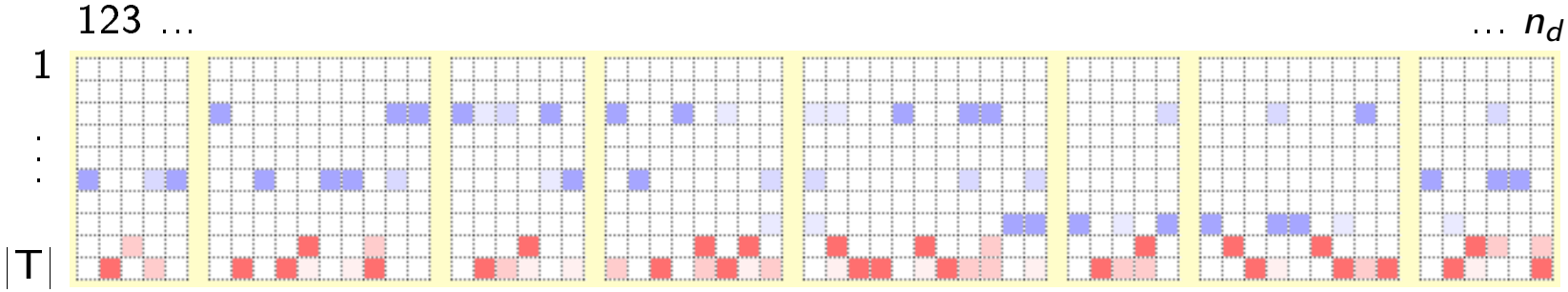


Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Тематическое моделирование последовательного текста

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

Матрица *тематических профилей слов* $p(t|d, w_i)$ размера $T \times n_d$:



Предположения разреженности и непрерывности тематики:

- каждое предложение относится к 1–2 предметным темам
- соседние предложения часто имеют одинаковые темы
- слова общей лексики не влияют на тематику предложений
- между абзацами вероятность смены темы выше
- между секциями она ещё выше

Позиционный регуляризатор Е-шага

Позиционный регуляризатор R_{di} зависит от позиции слова i в документе d и от параметров Φ, Θ через $p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$\mathscr{L}(\Phi,\Theta) + \sum_{d\in D}\sum_{i=1}^{n_d} R_{di}(p_{1dw_i},\ldots,p_{Tdw_i}) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\tilde{p}_{tdw} = p_{tdw}\frac{1}{n_{dw}}\sum_{\substack{i=1\\ w_i=w}}^{n_d}\left(1 + \frac{\partial R_{di}}{\partial p_{tdw}} - \sum_{s\in T} p_{sdw}\frac{\partial R_{di}}{\partial p_{sdw}}\right);$$

$$\phi_{wt} = \operatorname*{norm}_{w}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Регуляризатор разреживания распределений $p(t|d,w)$

Гипотеза разреженности распределений $p_{tdw} = p(t|d,w)$:
в документе слово может относиться только к одной теме.

Максимизируем KL-дивергенции между $\hat{p}(t) = \frac{1}{|T|}$ и $p(t|d,w)$:

$$R(\Phi,\Theta) = -\tau \sum_{d\in D} \sum_{w\in d} n_{dw} \frac{1}{|T|} \sum_{t\in T} \ln p_{tdw}.$$

Подставляем, получаем формулу модифицированного Е-шага:

$$\tilde{p}_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T} \phi_{ws}\theta_{sd}}(1+\tau) - \frac{\tau}{|T|}.$$

Эффект: разреживание p_{tdw} ведёт к разреживанию Φ и Θ:

если $p(t|w) < \frac{1}{|T|}$, то ϕ_{wt} уменьшается;

если $p(t|d) < \frac{1}{|T|}$, то θ_{td} уменьшается.

Регуляризатор сглаживания распределений $p(t|d,w)$ по контексту

Контекст слова w_i — множество слов w_j недалеко от слова w_i
$\hat{p}_{tdi}$ — эмпирическая оценка $p_{tdw_i} = p(t|d,w_i)$ по контексту,

$$\hat{p}_{tdi} = \frac{\sum_j K_{ij} p_{tdw_j}}{\sum_j K_{ij}},$$

где K_{ij} — оценка важности слова w_j в контексте w_i, возможно, с учётом границ предложений, абзацев, секций документа.

Минимизируем KL-дивергенции между $\hat{p}_{tdi}$ и p_{tdw_i}:

$$R(\Phi,\Theta) = \tau \sum_{d\in D} \sum_{i=1}^{n_d} \sum_{t\in T} \hat{p}_{tdi} \ln p_{tdw_i}.$$

Подставляем, получаем формулу модифицированного Е-шага:

$$\tilde{p}_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}(1-\tau) + \frac{\tau}{n_{dw}} \sum_{i=1}^{n_d} \big[w_i = w\big] \hat{p}_{tdi}.$$

Резюме

- Лучшие способы тематизации коротких текстов — битермы (BTM) и сети слов (WNTM)
- Сети слов (WNTM) легко моделировать в BigARTM, сформировав коллекцию локальных контекстов слов
- Регуляризация E-шага — пока не реализована в BigARTM, потенциальные применения — сегментация и аннотирование

Открытые проблемы:

- как разреживать модели сегментированного текста?
- стоит ли совмещать модель битермов с обычной?
- можно ли выкинуть весь фон в модели всплесков?
- эквивалентна ли модель сегментированного текста какой-то позиционной регуляризации E-шага?