Вероятностные тематические модели
Лекция 8. Модели сочетаемости слов

Константин Вячеславович Воронцов
k.v.vorontsov@phystech.edu

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

МФТИ.ФПМИ.ИС.ИАД — ФИЦ ИУ РАН • ~~2025-10-30~~ 2025-11-06

Содержание

1. **Мультиграммные модели и выделение терминов**
   - Интерпретируемость n-граммных моделей
   - Автоматическое выделение терминов
   - Синтаксическое и тематическое выделение фраз
2. **Тематические модели дистрибутивной семантики**
   - Дистрибутивная гипотеза и word2vec
   - Тематическая модель битермов BitermTM
   - Тематическая модель сети слов WNTM
3. **Регуляризация E-шага**
   - Постобработка E-шага
   - Регуляризация E-шага
   - Примеры регуляризаторов E-шага

Несколько терминов из лингвистики

Единица языка, в зависимости от уровня членения текста — фонема, морфема, слово, словосочетание, фраза, предложение.

Сочетаемость (co-occurrence) — свойство языковых единиц сочетаться в речи, образуя единицы более высокого уровня.

Виды сочетаемости: *контактная* и *дистантная*.

n-грамма — последовательность из n единиц языка (n-грамма — это пример контактной сочетаемости).

Коллокация — *n-грамма* слов, встречающаяся в корпусе гораздо чаще, чем ожидается при их случайном соединении.

Словосочетание — *n-грамма* слов, связанных по смыслу и грамматически, служащая для обозначения единого понятия.

n-граммы радикально улучшают интерпретируемость тем

Коллекция 20Conf заголовков научных статей DBLP, тема «Information Retrieval»

Terms	*Phrases*
search	information retrieval
web	social networks
retrieval	web search
information	search engine
based	support vector machine
model	information extraction
document	web page
query	question answering
text	text classification
social	collaborative filtering
user	topic model

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.

Напоминание. Биграммы улучшают интерпретируемость тем

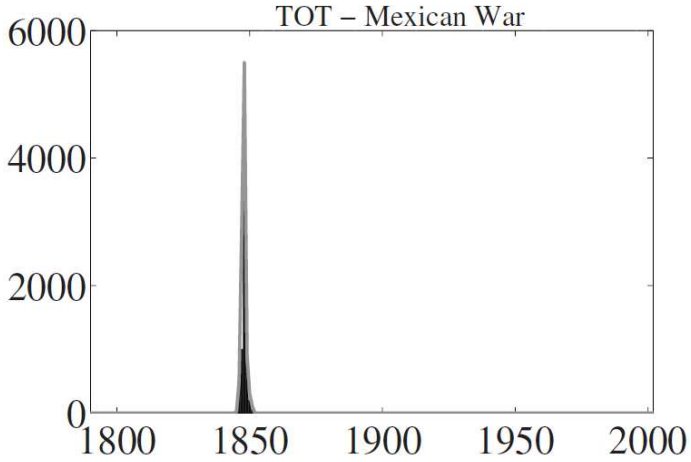
Коллекция 1000 статей конференций ММРО, ИОИ на русском

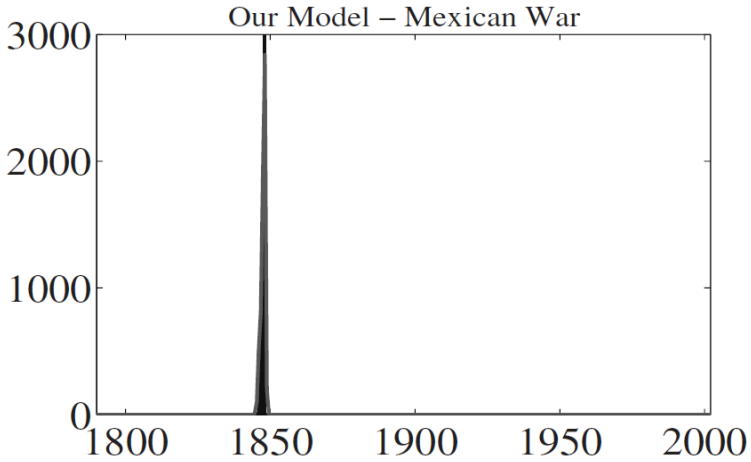
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Напоминание. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





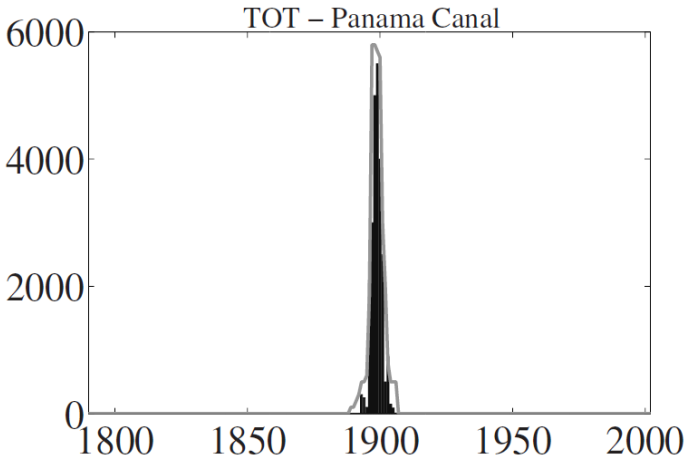
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

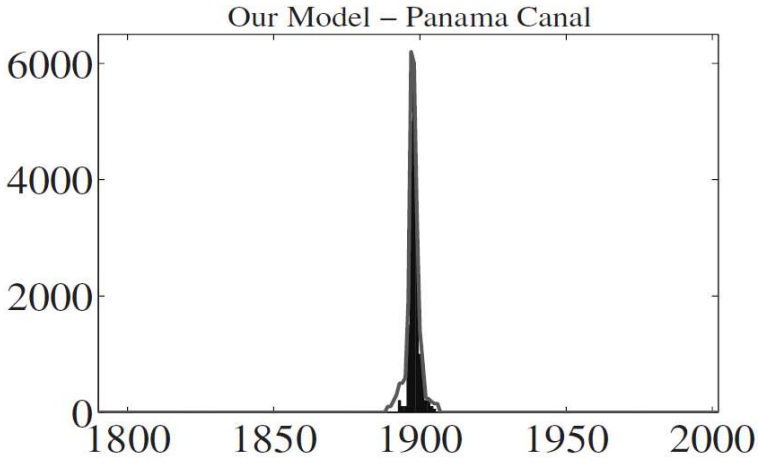
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Напоминание. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Задача автоматического выделения терминов

Термин — фраза (n-грамма) со следующим набором свойств:

1. *высокая частотность* (frequency):
   много раз встречается в коллекции;
2. *контактная сочетаемость слов* (collocation):
   состоит из слов, неслучайно часто встречающихся вместе;
3. *полнота* (completeness):
   является максимальной по включению цепочкой слов;
4. *синтаксическая связность* (syntactic connectedness):
   является грамматически корректным словосочетанием;
5. *тематичность* (topicality):
   часто встречается в узком подмножестве тем.

Сумма технологий для ATE (Authomatic Term Extraction):
TopMine 1 2 3 + UDPipe 4 + BigARTM 5

Алгоритм TopMine: определения и основные идеи

- $C(a_1, \ldots, a_k)$ — хэш-таблица частот k-грамм, $a_i \in W$, $C(w) = n_w$ для всех униграмм $w \in W$: $n_w \geqslant \varepsilon_1$
- ε_k — пороговое значение частоты частых k-грамм
- $A_{d,k}$ — множество позиций i в документе d, с которых начинаются все частые k-граммы:

$$C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k$$

- Свойство антимонотонности:

$$C(a_1, \ldots, a_k) \geqslant C(a_1, \ldots, a_k, a_{k+1})$$

- Основной шаг алгоритма: для всех $i = 1, \ldots, n_d$
  если $(i \in A_{d,k})$ **и** $(i+1 \in A_{d,k})$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015.

Алгоритм TopMine: быстрый поиск всех частых k-грамм

Вход: коллекция D, пороги ε_k;
Выход: хэш-таблица частот $C(a_1, \ldots, a_k)$, $k = 1, \ldots, k_{\max}$;

$C(w) := n_w$ для всех $w \in W$;
$A_{d,0} := \{1, \ldots, n_d\}$;
для $k := 1, \ldots, k_{\max}$
  для всех $d \in D$
    $A_{d,k} := \left\{ i \in A_{d,k-1} \;\middle|\; C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k \right\}$;
    для всех $i \in A_{d,k}$
      если $i + 1 \in A_{d,k}$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$;
  оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$;

Преимущество алгоритма: линейная память и скорость.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015.

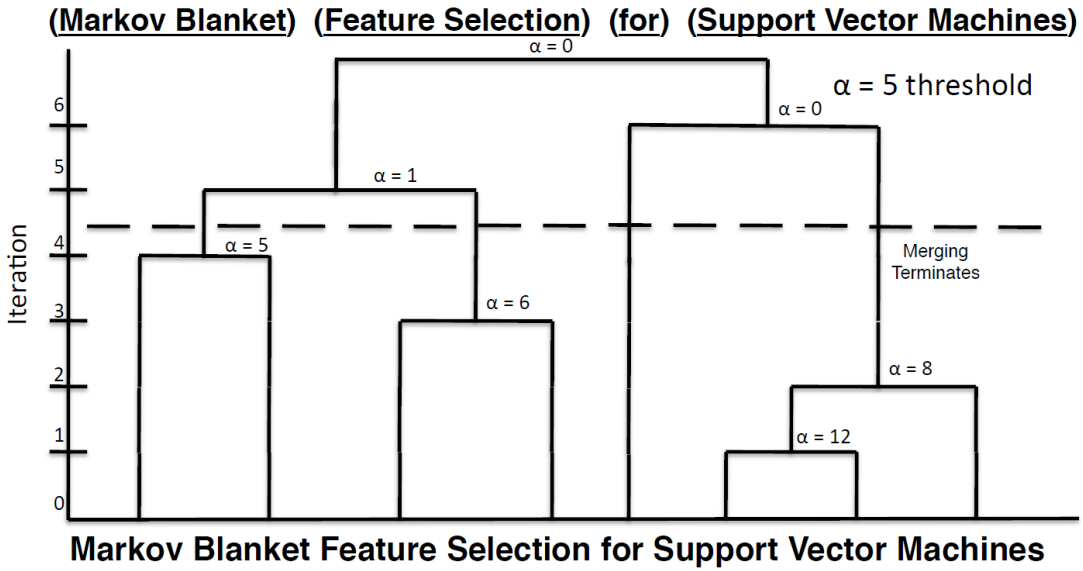
Алгоритм TopMine: отбор фраз по частоте и полноте

Итеративное слияние фраз с понижением значимости α.

p_u — оценка вероятности встретить фразу u

p_{uv} — оценка вероятности встретить фразу uv

Критерии: $\text{SignificanceScore} = \dfrac{p_{uv} - p_u p_v}{\sqrt{p_{uv}}}$ или $\text{PMI} = \log \dfrac{p_{uv}}{p_u p_v}$



Синтаксические анализаторы (UDPipe, SyntaxNet)

Вход: список предложений

Выход, для каждого слова в каждом предложении:

- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- нормальная форма
- часть речи: `NOUN`, `VERB`, `ADJ`, `ADV`, ...
- член предложения: `nsubj`, `dobj`, `conj`, `cc`, `nmod`, ...

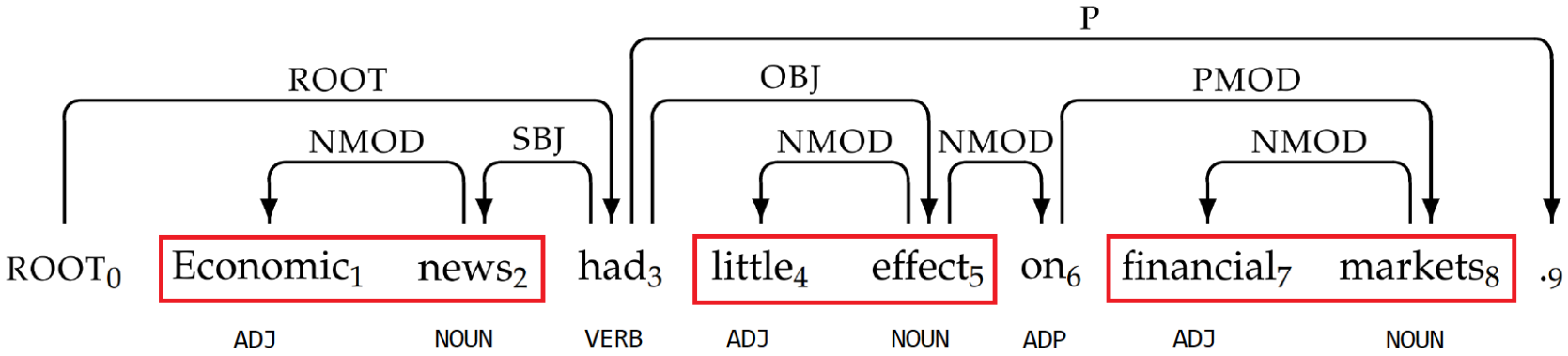
UDPipe (Universal Dependencies), 60 языков, включая русский

Google SyntaxNet — предобученная нейросеть поверх TensorFlow, поддерживает 40 языков, включая русский.

D.Andor, C.Alberti, D.dWeiss, A.Severyn, A.Presta, K.Ganchev, S.Petrov, M.Collins. Globally Normalized Transition-Based Neural Networks. 2016.

Использование дерева зависимостей для отбора терминов

Пример дерева зависимостей:



Варианты стратегий отбора терминов-кандидатов:

- брать все поддеревья
- брать все именные группы (корень — NOUN)
- не брать CONJ, SCONJ, DET, AUX, INTJ, PART, PUNCT, SYM

Announcing SyntaxNet: the world's most accurate parser goes open source.
https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

Денис Кирьянов. Изучаем синтаксические парсеры для русского языка. 2018.
https://habr.com/ru/company/sberbank/blog/418701

Критерии тематичности фраз

Насколько далеко $p(t|w) = \phi_{wt}\frac{n_t}{n_w}$ от равномерного $p_0(t) = \frac{1}{|T|}$.

Дивергенция Кульбака-Лейблера:

$$\mathrm{KL}(w) = \mathrm{KL}(p_0 \| p) = \sum_{t \in T} \frac{1}{|T|} \ln \frac{\frac{1}{|T|}}{p(t|w)} \to \max$$

Дивергенция Йенсена-Шеннона (метрика, не имеет проблем с нулевыми вероятностями), где $\bar{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|})$:

$$\mathrm{JS}(w) = \frac{1}{2}\mathrm{KL}(p_0 \| \bar{p}) + \frac{1}{2}\mathrm{KL}(p \| \bar{p}) \to \max$$

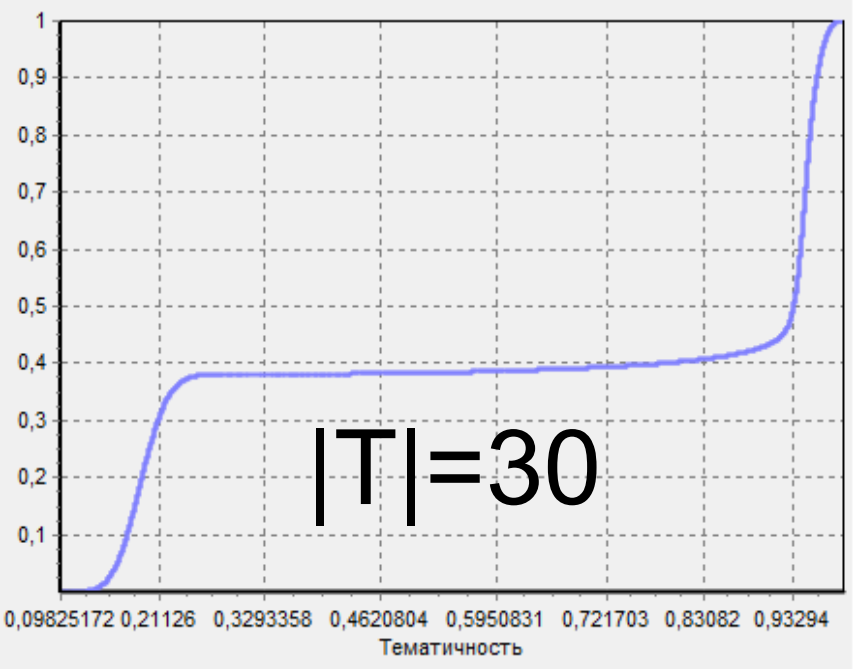
Нормированная сумма степенных функций, $\gamma > 1$:

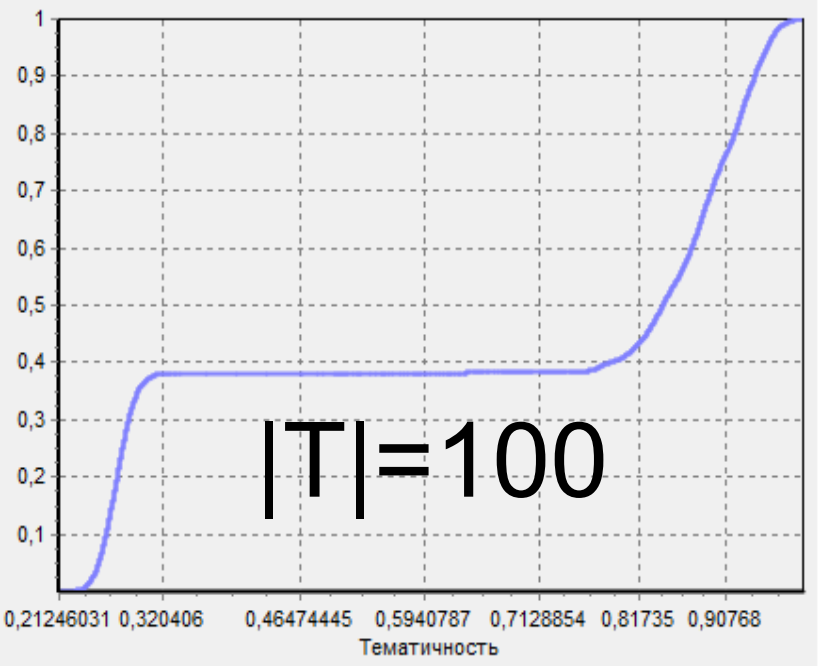
$$\text{Тематичность}(w) = |T|^{\gamma-1} \sum_{t \in T} p(t|w)^{\gamma} \to \max$$

Фразы чётко разделяются на тематичные и нетематичные

Коллекция SyntagRus $|D| = 600$, словарь $|W| = 46\,000$ фраз, тематические модели LDA с числом тем $T = 30,\ 100$.

Эмпирическое распределение фраз по тематичности:





Пограничный слой между тематичными и нетематичными фразами очень узкий $\approx \frac{200}{46\,000}$ и слабо зависит от числа тем.

Основной эксперимент ATE: SyntaxNet + TopMine + BigARTM

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого *случайного* подмножества (2000 n-грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
  логистическая регрессия, градиентный бустинг

Владимир Полушин. Тематические модели для ранжирования рекомендаций текстового контента. Бакалаврская диссертация, ВМК МГУ, 2017.

Сравнение методов автоматического отбора терминов

Найти *как можно больше терминов* — полнота важнее точности

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	**0.95**	**0.38**	**0.91**	**0.97**	**0.41**	**0.99**

Стат < Син < Син+Стат < Тем < Стат+Тем, Син+Тем < Стат+Син+Тем

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

Конкурс RuTermEval конференции «Диалог»

Идентификация и классификации терминов по 3 классам:

- **specific term:** термины, специфичные доменно (для конкретной предметной области) и лексически
- **common term:** термины, специфичные только доменно (могут быть известны и употребляться неспециалистами)
- **nomen:** номенклатурные наименования доменно специфичных объектов, материальных объектов данного домена (программы, базы данных, наборы данных, языки, корпуса, словари и т.д.)

Данные: 850 аннотаций по домену компьютерной лингвистики: 65K токенов, 18K разметок терминов (уникальных 6534), неразмеченные тексты того же домена и двух других доменов

---

https://dialogue-conf.org/evaluation/rutermeval-2024/

Проблема коротких текстов в тематическом моделировании

Короткие тексты (short text):

- Twitter и другие микроблоги
- социальные медиа
- заголовки статей и новостных сообщений

Тривиальные подходы (у каждого свой комплекс недостатков):

- считать каждое сообщение отдельным документом
- разреживать $p(t|d)$ вплоть до единственной темы
- объединить сообщения по автору/времени/региону/и т. п.
- объединить посты с комментариями
- дополнить коллекцию длинными текстами (Википедия и др.)

Более интересная идея:

- использовать сочетаемость пар слов в сообщениях

Дистрибутивная гипотеза и виды семантической близости слов

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
сочетаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

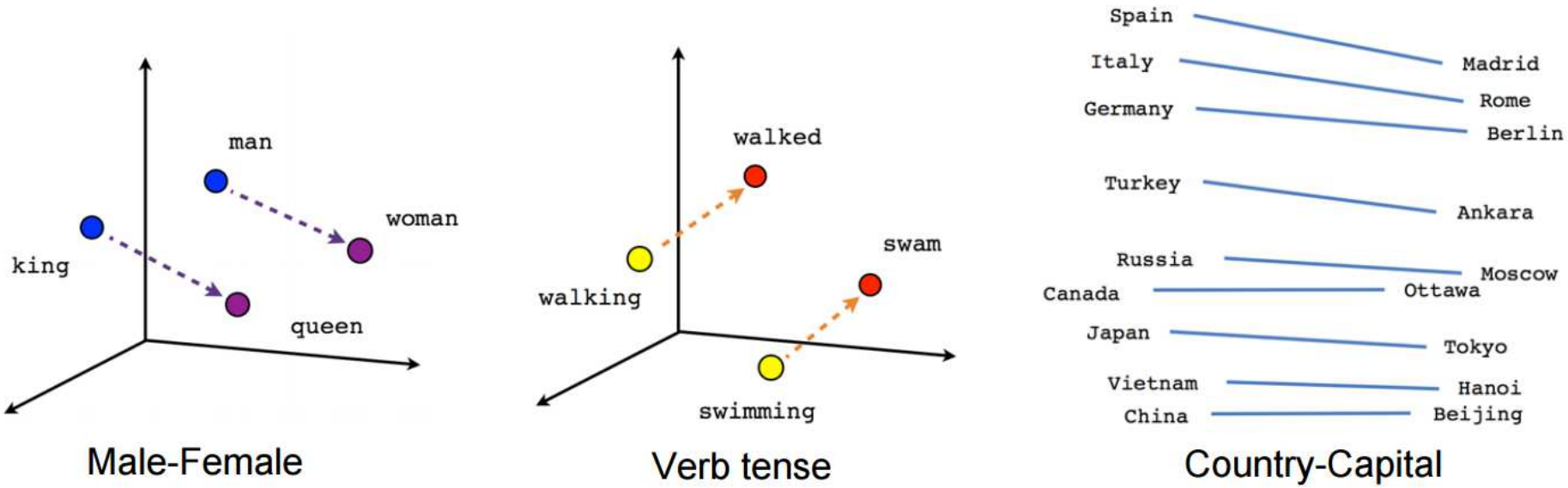
здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR). 2010.

Задача семантического векторного представления слов

Задача: по наблюдаемой синтагматической близости слов построить *векторные представления слов* (word embedding) $x_w \in \mathbb{R}^T$, $w \in W$, отражающие их парадигматическую близость, т.е. близкие по смыслу слова должны иметь близкие векторы.

Способ проверки — задача семантической аналогии слов: по трём словам угадать четвёртое.



Формализация дистрибутивной гипотезы в программе word2vec

Дано: n_{uw} — частота пары слов u, w в окне $\pm h$ слов
Найти: векторные представления слов x_w и контекстов y_u
в вероятностной языковой модели

$$p(w|u) = \underset{w \in W}{\mathrm{SoftMax}} \langle x_w, y_u \rangle = \underset{w \in W}{\mathrm{norm}} \big(\exp \langle x_w, y_u \rangle \big)$$

Критерий максимума правдоподобия для предсказания слов

$$\sum_{w,u \in W} n_{wu} \ln p(w|u) \;\to\; \max_{\{x_w, y_u\}}$$

или для классификации пар слов (Skip-Gram Negative Sampling):

$$\sum_{w,u \in W} n_{wu} \Big(\ln \sigma \langle x_w, y_u \rangle + \sum_{v \in V_k(u)} \ln \sigma \big(-\langle x_v, y_u \rangle \big) \Big) \;\to\; \max_{\{x_w, y_u\}}$$

где $V_k(u) \subset W$ — случайные k слов не из контекста u.

T.Mikolov, K.Chen, G.Corrado, J.Dean. Efficient estimation of word representations in vector space, 2013.

Связь word2vec с матричными разложениями

T — размерность векторов слов x_w и контекстов y_u
$X = (x_w)_{W \times T}$ — матрица векторов слов
$Y = (y_u)_{W \times T}$ — матрица векторов контекстов

SGNS строит матричное разложение $P \approx XY^{\mathsf{T}}$ матрицы Shifted PMI (Point-wise Mutual Information):

$$P_{wu} = \ln \frac{n_{wu} n}{n_w n_u} - \ln k,$$

n_{wu} — частота пары слов w, u в окне $\pm h$ слов,
n_w, n_u — число пар с участием слова w и u соответственно,
n — число всех пар слов в коллекции.

В качестве эвристики используют также Shifted Positive PMI:

$$P^{+}_{wu} = \left(\ln \frac{n_{wu} n}{n_w n_u} - \ln k \right)_{+}.$$

O.Levy, Y.Goldberg. Neural word embedding as implicit matrix factorization, 2014.

Модели векторных представлений для текстов и графов

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных n-грамм
https://github.com/facebookresearch/fastText

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: эмбединги фраз и предложений от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

Недостаток: координаты векторов не интерпретируемы

Преимущества и недостатки нетематических эмбедингов

- ⊕ Удивительно высокое качество на задачах семантической аналогии и близости слов.
- ⊕ Возможность нейросетевой реализации методом SG.
- ⊕ Имеются готовые реализации от Google, Facebook и др.
- ⊕ Имеются готовые векторы слов, предобученные по большим текстовым коллекциям на разных языках
- ⊖ Неинтерпретируемые компоненты векторов
- ⊖ Не ясно, почему XY^{T}, а не XX^{T} (обычно Y игнорируют)

Тематические модели BitermTM, WordNetworkTM, WordTM обучаются по частотам сочетания слов, аналогично word2vec.

Битермы: модель сочетаемости слов в коротких текстах

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

Тематическая модель битермов (Biterm Topic Model):

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

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

Критерий максимума логарифма правдоподобия:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{vt} \geqslant 0; \quad \textstyle\sum_v \phi_{vt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

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:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

где n_{uv} — частота битерма (u,v) в документах коллекции.

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned}
&\text{Е-шаг:} && p_{tuv} \equiv p(t|u,v) = \operatorname*{norm}_{t\in T}\big(\phi_{ut}\phi_{vt}\pi_t\big) \\
&\text{М-шаг:} && \phi_{vt} = \operatorname*{norm}_{v\in W}\Big(n_{vt} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}}\Big), \quad n_{vt} = \sum_{u\in W} n_{uv}p_{tuv} \\
& && \pi_t = \operatorname*{norm}_{t\in T}\Big(n_t + \pi_t\frac{\partial R}{\partial \pi_t}\Big), \quad n_t = \sum_{u,v\in W} n_{uv}p_{tuv}
\end{aligned}$$

И данные, и модель симметричны: $n_{uv} = n_{vu}$, $p(u,v) = p(v,u)$.

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

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

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \operatorname*{norm}_{w} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big); \quad p_{tuw} = \operatorname*{norm}_{t \in T} \big(n_t \phi_{wt} \phi_{ut} \big).$$

Это эквивалентно обработке *псевдо-документов* d_u, где каждый d_u объединяет все контексты слова u, причём $\theta_{tu} \propto n_t \phi_{ut}$; n_{uw} — число вхождений слова w в псевдо-документ d_u.

2. Регуляризатор разреживания Θ для коротких текстов:

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

Модель сети слов WNTM для коротких текстов

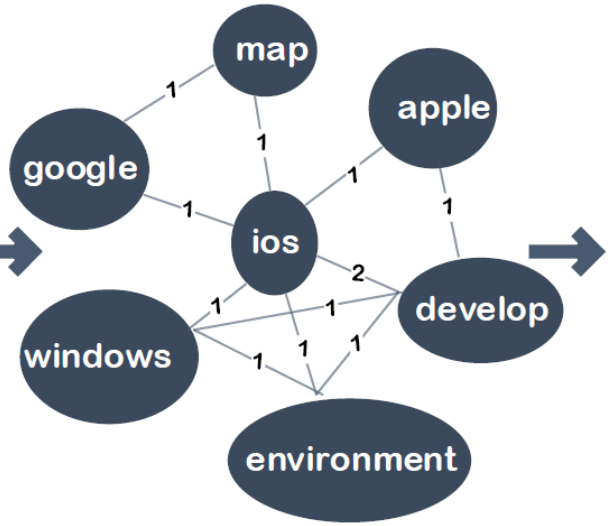
Идея: моделировать не документы, а связи между словами.

d_u — псевдо-документ, объединение всех контекстов слова u.

n_{uw} — число вхождений слова w в псевдо-документ d_u.

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 (Word Network) и WTM (Word Topic Model)

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

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

где n_{uw} — частота сочетания пары слов (w, u).

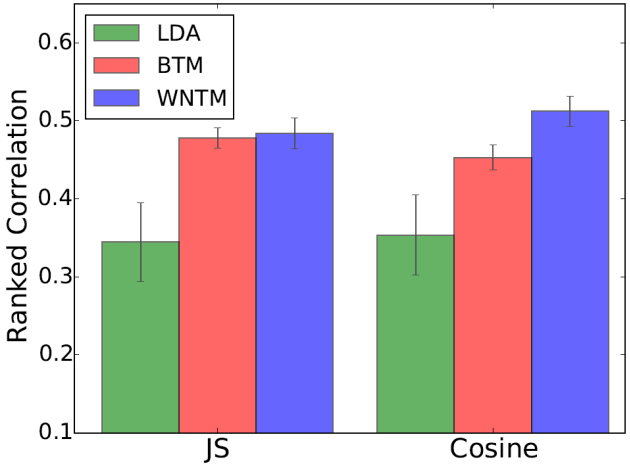
Отличие модели битермов: там $\Theta = \operatorname{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathsf{T}}$, здесь данные и модель НЕ симметричны: $n_{uw} \neq n_{wu}$, $p(u|d_w) \neq p(w|d_u)$

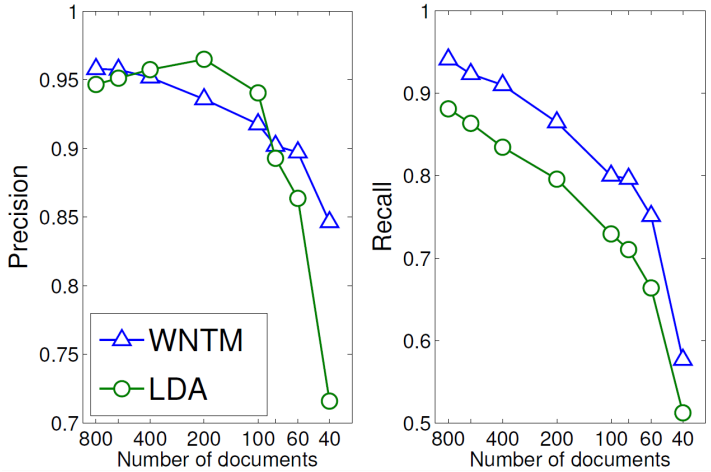
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

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 BitermTM; 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.

WN-ARTM на задачах семантической аналогии слов

Два подхода к синтезу векторных представлений слов:

- **WN-ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат WN-ARTM	Результат word2vec
king − boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow − russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india − russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars − car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Сравнение word2vec и WN-ARTM по интерпретируемости тем

SGNS (word2vec) — нет интерпретируемости:

- avg hearth soc protector decomposition whip stochastic sewer splinter accessory howie thief thermodynamic boltzmann equilibrium kingship unconscious
- rainy miocene snowy horner cfb triassic eleventh amadeus dams tenth mesozoic fourteenth thirteenth ninth diaries bight demographics seventh almanac eocene
- gnis usda bloomberg usgs regulator nhk gerd magnetism capacitor fed classifies capacitance stadt bipolar multilateral trpod kunst reciprocal smiths potassium

WN-ARTM — есть интерпретируемость:

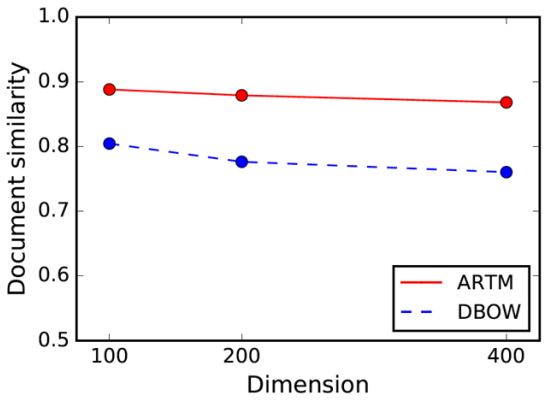
- scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus
- game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash zelda
- election party elected elections parliament assembly seats members minister legislative electoral liberal council representatives parliamentary democratic

---

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

WN-ARTM на задачах семантической близости документов

ArXiv triplets dataset [Dai et. al, 2015]: 20K троек статей:
⟨статья A, схожая статья B, непохожая статья C⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент: DBOW paragraph2vec [Dai et. al, 2015]

WN-ARTM превосходит модель DBOW (distributed bag-of-words)

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Сегментная структура текста и пост-обработка Е-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 Е-шага

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

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

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

ЕМ-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{Е-шаг:} & p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ & \tilde{p}_{tdw} = p_{tdw}\Big(1 + \frac{1}{n_{dw}}\Big(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\Big)\Big) \qquad (*) \\ \text{М-шаг:} & \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ & \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

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 М-шага:

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

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

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

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

$$\theta_{td}\frac{\partial\tilde{R}}{\partial\theta_{td}} = \theta_{td}\frac{\partial R}{\partial\theta_{td}} + \sum_w p_{tdw}Q_{tdw}; \quad Q_{tdw} = \frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}$$

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

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

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

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 М-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}:

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

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

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

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

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

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует пост-обработка $p_{tdw} \to \tilde{p}_{tdw}$.

Оказывается, верно и обратное:

Теорема. Если на k-й итерации ЕМ-алгоритма для каждого (d, w): $n_{dw} > 0$ в формулах М-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?!

Доказательство

В системе $(*)$ дифф. уравнений относительно R введём переменные x_{tdw}:

$$\underbrace{p_{tdw}^k \frac{\partial R}{\partial p_{tdw}}}_{x_{tdw}} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k) + p_{tdw}^k \sum_{z \in T} \underbrace{p_{zdw}^k \frac{\partial R}{\partial p_{zdw}}}_{x_{zdw}}, \quad t \in T.$$

Для любой пары (d, w) такой, что $n_{dw} > 0$, это система $|T|$ линейных уравнений относительно $|T|$ переменных x_{tdw}, $t \in T$. Подстановкой убеждаемся, что $x_{tdw} = n_{dw}(\tilde{p}_{tdw}^k - p_{tdw}^k)$ — решение системы. Взяв это решение, получим систему дифф. уравнений относительно R:

$$\frac{\partial R}{\partial p_{tdw}} = \frac{x_{tdw}}{p_{tdw}}, \quad d \in D, \ w \in d, \ t \in T.$$

Система декомпозируется по переменным p_{tdw}: каждой тройке (d, w, t) соответствует частное решение $R(\Pi) = x_{tdw} \ln p_{tdw} + C$. Общее решение:

$$R(\Pi) = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} x_{tdw} \ln p_{tdw} + C.$$

Подставляя сюда найденное решение x_{tdw}, получаем требуемое. ■

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 Е-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: Соседние векторы разреживаются независимо.

Пример 2. Тематическая модель сегментированного текста

S_d — множество сегментов (предложений) 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 $p(t|d,s)$:

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

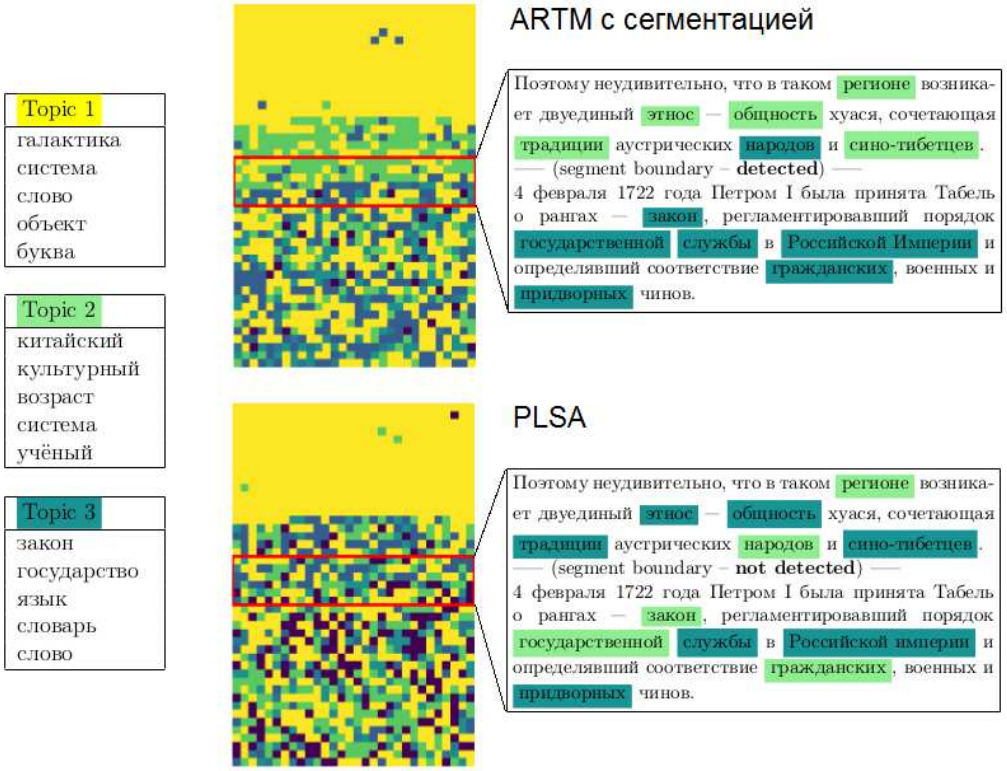
Формула регуляризованного Е-шага:

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

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

Теперь мы знаем пять механизмов тематического моделирования связного текста (от сильного к слабому):

1. *Линейная тематизация с локальным E-шагом* (лекция 2): вместо тематики документа $p(t|d)$ вычисляется тематика локального контекста $p(t|i)$ и $p(t|d, w_i)$, $i = 1, \ldots, n_d$.
2. *Пост-обработка E-шага* (сегодня): тематические векторы $p(t|d, w_i)$ подвергаются эвристическим преобразованиям, что эквивалентно регуляризации E-шага.
3. *Гиперграфовая тематическая модель* (лекция 7): термы, порождаемые общей темой (предложения, фразы, факты, и т. д.), объединяются в «транзакции».
4. *Тематическая модель сети слов или битермов* (сегодня): по каждому слову формируется псевдодокумент путём объединения (в «мешок») всех его локальных контекстов.
5. *Тематическая модель n-грамм* (сегодня): словари n-грамм строятся заранее и используются как модальности, для сокращения словарей оценивается тематичность термов.

Задания по курсу

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	$5X$
обзор по последним NeuralTM	$5X$
интеграция ARTM в pyTorch	$5X$
участие в одном из проектов	$10X$
работа над открытой проблемой	$10X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\big(10, \lfloor \text{score}/5 \rfloor\big)$ по 10-балльной шкале.

Теоретическое задание к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$
Найти параметры модели ξ_{dw}.

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
Найти параметры модели ξ_w.

Подсказка: применить условия ККТ или основную лемму.

3. Творческое задание (возможны разные решения)
Предложите модель, определяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Теоретическое задание к лекции 2

4. Запишите критерий логарифма правдоподобия с регуляризацией для тематической модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя исходные данные $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}. Выведете из него EM-алгоритм, докажите его эквивалентность обычному EM-алгоритму для ARTM.

5. Запишите критерий логарифма правдоподобия для локализованной тематической модели $p(w|C_i) = \sum_t \phi_{wt} p(t|C_i)$. Выведете из него EM-алгоритм с локализованным E-шагом.

Какие приближения пришлось сделать в процессе вывода? Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw}?

6. Творческое задание (возможны разные решения) Предложите «какую-нибудь разумную» параметризацию для тематической модели внимания. Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательское задание к лекции 2

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайнового EM-алгоритма
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$»

Теоретическое задание к лекции 3

7. Выведете формулы EM-алгоритма в случае, когда логарифм в функции потерь заменяется гладкой монотонно возрастающей функцией ℓ:

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

Подумайте, какие замены логарифма полезны, и почему.

8. Замените ln гладкой монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W}\beta_w\mu(\phi_{wt}) + \sum_{d\in D}\sum_{t\in T}\alpha_t\mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

9. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_{w}\bigl(n_{wt}[n_{wt} > \gamma n_t]\bigr)$$

Теоретическое задание к лекции 3

Аналитик построил тематическую модель Φ^0, Θ^0
и отметил среди столбцов матрицы Φ^0 темы двух типов:
удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

10. Предложите регуляризаторы для этого.

11. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

12. Предложите способ инициализации Φ для новой модели.

Исследовательские задания к лекции 4

- Проблема несбалансированности тем
  - генераторы синтетических несбалансированных коллекций
  - модели локального контекста лишены этой проблемы?
  - регуляризаторы декоррелирования + семантической однородности
- Семейство средневзвешенных статистик
  - генераторы синтетических коллекций, удовлетворяющих гипотезе условной независимости
  - как (и нужно ли) определять пороги для построения статистических тестов условной независимости?
  - как ослабить проверку гипотезы условной независимости в модели локального контекста?
  - как перестраивать несогласованные темы?
- Критерий внутритекстовой когерентности
  - найти лучший вариант критерия с помощью калибровки по размеченным тематическим цепочкам
  - вычисление критерия должно естественным образом встраиваться в модель локального контекста

Теоретическое задание к лекции 5

13. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, гарантирующий, что
1) у каждой родительской темы будет хотя бы одна дочерняя;
2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу М-шага для матрицы Ψ.

14. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

15. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum\limits_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum\limits_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 5

Участие в проекте «Мастерская знаний»

Дано:

- подборки, сгенерированные SciRus по одной статье
- асессорская разметка статей подборки по релевантности
- несколько вариантов токенизации
  — в том числе с автоматическим выделением терминов

Найти:

- тематическую модель
- модель ранжирования подборки по релевантности
- оптимальные: токенизацию, число тем, регуляризаторы
- распределение терминов по тематичности

Критерий:

- качество ранжирования
- (визуально) интерпретируемость тем
  — в том числе автоматического именования тем

Примеры датасетов для практических заданий по курсу

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

Проекты

- «Мастерская знаний» для научного поиска
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus.
  — задача: показать пользователю тематику подборки
  — понадобится автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем
  — конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
  — пользователь задаёт грубый фильтр текстового потока
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме
  — конечная цель: q&q аналитика проблемной среды

Открытые проблемы тематического моделирования

1. Проблема несбалансированности тем в коллекции
2. Обеспечение 100%-й интерпретируемости тем
3. Тематические модели внимания последовательного текста
4. Обнаружение новых тем или трендов в потоке текстов
5. Автоматическое именование и аннотирование тем
6. Обзор подходов в нейросетевых тематических моделях
7. Обеспечение полноты и устойчивости множества тем
8. Автоматический подбор гиперпараметров, AutoML
9. Оптимизация гиперпараметров в потоковом режиме
10. Проблема несбалансированности текстов по длине
11. Бережное слияние моделей нескольких коллекций
12. Гиперграфовые тематические модели в RecSys