Оптимизация и регуляризация вероятностных тематических моделей (лекция №2)

Воронцов Константин Вячеславович

(ФИЦ ИУ РАН • МФТИ • МГУ • ВШЭ • Яндекс • FORECSYS • Aithea)

• Традиционная молодёжная летняя школа •

14–20 июня 2017

Содержание

Задача тематического моделирования

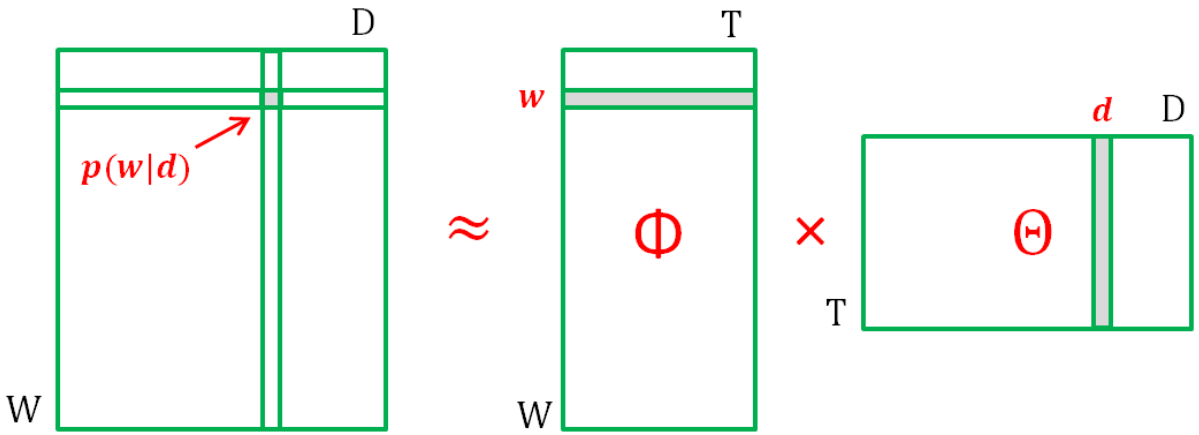
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $p(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_{i} \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

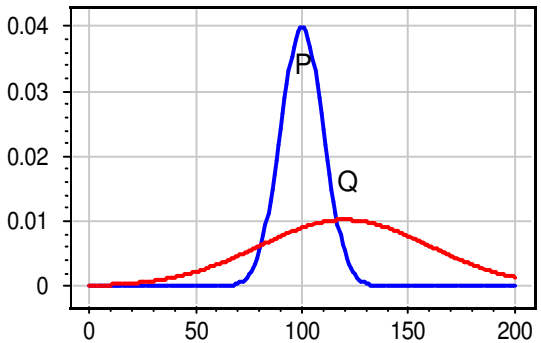
Дивергенция Кульбака–Лейблера

1. $\mathrm{KL}(P\|Q) \geqslant 0; \qquad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q;$

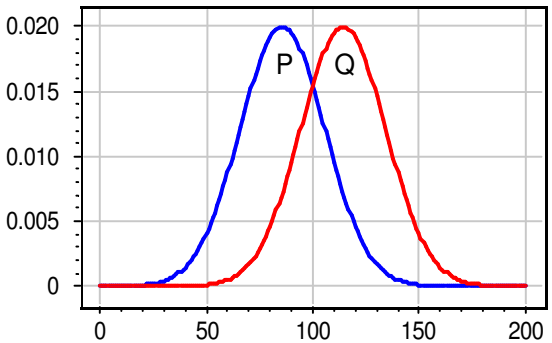
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Leftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha}$$

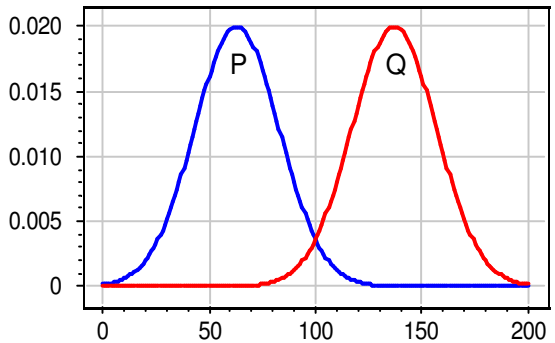
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P вложено в Q:



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 2.97$



$\mathrm{KL}(P\|Q) = 0.44$
$\mathrm{KL}(Q\|P) = 0.44$



$\mathrm{KL}(P\|Q) = 2.97$
$\mathrm{KL}(Q\|P) = 2.97$

Примеры регуляризаторов на основе KL-дивергенции

1. разреживание предметных тем $S \subset T$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \to \max$$

2. сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \to \max$$

3. частичное обучение по подмножествам $W_t \subset W$, $D_t \subset D$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W_t} \ln \phi_{wt} + \alpha_0 \sum_{t \in T} \sum_{d \in D_t} \ln \theta_{td} \to \max$$

4. удаление неинформативных тем:

$$R(\Theta) = -\tau \sum_{t \in S} \ln p(t) \to \max, \qquad p(t) = \sum_{d \in D} \theta_{td} p(d)$$

Параллельные и сравнимые корпуса текстов

Parallel — точный перевод (с выравниванием предложений), пример: EuroParl, протоколы европарламента, 21 язык.

Comparable — не перевод, а пересказ на другом языке, пример: Википедия.

Мультиязычные модели (ML-LDA, PLTM, BiLDA)

- каждый язык — отдельная модальность, W^ℓ — словарь языка ℓ из множества языков L.
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup\limits_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

- $\Pi_k(w) \subset W^k$ — все переводы слова $w \in W^\ell$ в языке k
- выравнивание документов по предложениям

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Регуляризация по двуязычным словарям. Модель ML-TD

Гипотеза. Если $u \in \Pi_k(w)$, то тематика слов w и u близка:

$$\mathrm{KL}\big(\hat{p}(t|u) \,\big\|\, p(t|w)\big) \to \min,$$

где $\hat{p}(t|u) = \frac{n_{ut}}{n_u}$, $\quad p(t|w) = p(w|t)\frac{p(t)}{p(w)} = \phi_{wt}\frac{n_t}{n_w}$.

Модель ML-TD (MultiLingual Translation Dictionary)

$$R(\Phi) = \tau \sum_{\ell,k \in L} \sum_{w \in W^\ell} \sum_{u \in \Pi_k(w)} \sum_{t \in T} n_{ut} \ln \phi_{wt} \;\to\; \max_{\Phi}.$$

Недостатки. Модель ML-TD не учитывает два обстоятельства:

- тематику омонимов сближать не нужно,
- слово может иметь разные переводы в разных темах.

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Матрица вероятностей переводов. Модель ML-TDP

Гипотеза. Переводы слов зависят от тем: $\pi_{uwt}^{k\ell} = p(u|w,t)$, темы согласуются в разных языках через переводы слов:

$$\mathrm{KL}\big(\hat{p}(u|t) \,\big\|\, p(u|t)\big) \to \min;$$

$\hat{p}(u|t) = \frac{n_{ut}}{n_t}$ — частотная оценка по модальности (языку) k,
$p(u|t)$ — модель темы t в языке k по языку ℓ:

$$p(u|t) = \sum_{w \in \Pi_\ell(u)} p(u|w,t)p(w|t) = \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt}.$$

Модель ML-TDP (MultiLingual Translation Dictionary Probability)

$$R(\Phi,\Pi) = \tau \sum_{\ell,k \in L} \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{k\ell}\phi_{wt} \to \max_{\Phi,\Pi}.$$

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Формулы M-шага для моделей ML-TD и ML-TDP

ML-TD (MultiLingual Translation Dictionary):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} n_{ut} \right)$$

ML-TDP (MultiLingual Translation Dictionary Probability):

$$\phi_{wt} = \underset{w \in W^\ell}{\text{norm}} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} \pi^{\ell k}_{wut} n_{ut} \right)$$

$$\pi^{k\ell}_{uwt} = \underset{u \in W^k}{\text{norm}} \left(\pi^{\ell k}_{wut} n_{ut} \right)$$

Смысл регуляризации:

условные вероятности $\phi_{wt} = p(w|t)$ согласуются с их частотными оценками по словам других языков

Тематические переводы слов $\pi_{uwt}^{k\ell} = p(u|w,t)$

Темы, в которых $p(\text{«sum»}|\text{«сумма»},t) > 0.9$

Тема №6		Тема №12		Тема №20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer

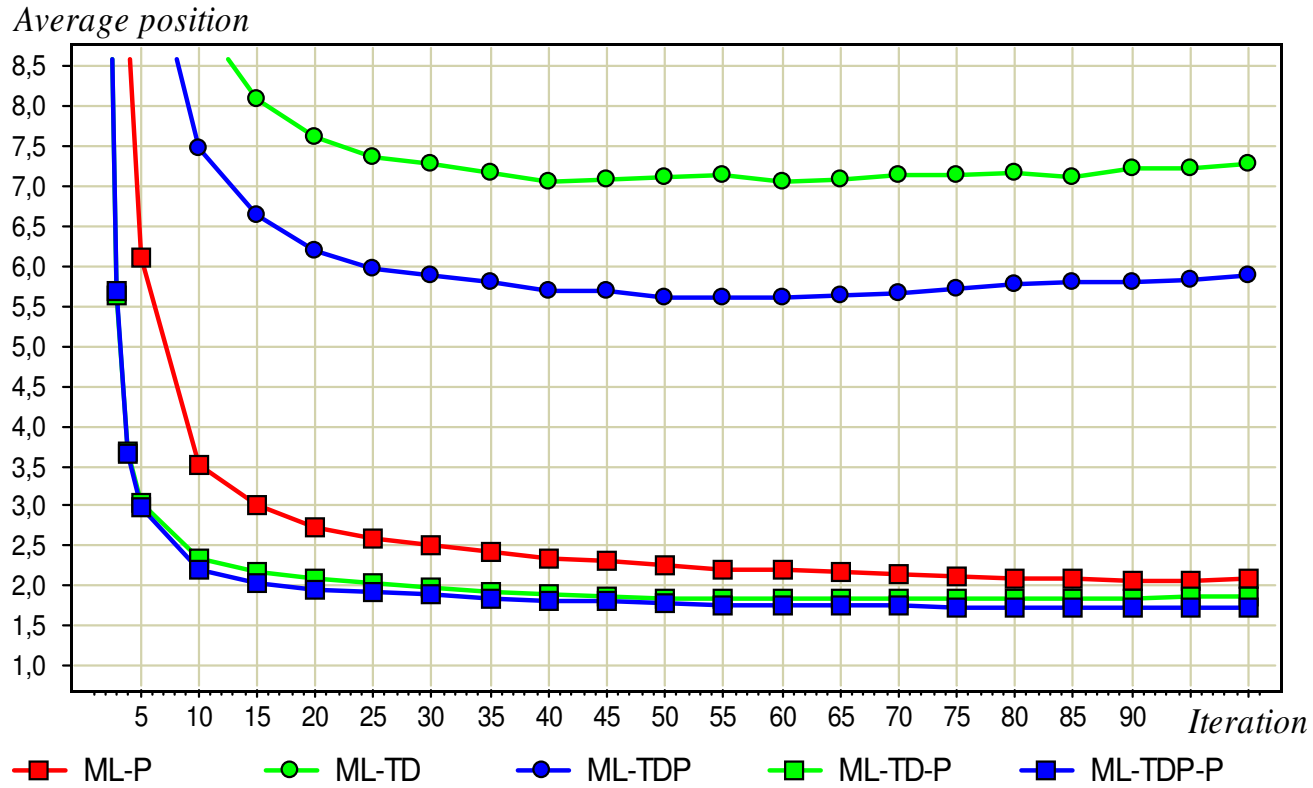
Темы, в которых $p(\text{«total»}|\text{«сумма»},t) > 0.9$

Тема №5		Тема №19		Тема №22	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 100$,
$|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

Качество поиска — средняя позиция перевода в выдаче:



Резюме по мультиязычным моделям

- Главное чудо: для построения мультиязычных тем достаточно иметь сравнимые корпуса.
- Сравнимая коллекция является более сильным источником многоязычной информации, чем словарь переводов (!)
- Модель с вероятностями переводов — самая сильная

Возможные применения:

- *Кросс-язычный поиск:* ищем тексты на другом языке
- *Мульти-язычный поиск:* ищем тексты на всех языках
- *Статистический машинный перевод:* выбираем вариант перевода, наиболее подходящий тематике документа.

Иерархические тематические модели

Стратегии построения тематических иерархий:

- структура иерархии: дерево / многодольный граф
- направление: снизу вверх / сверху вниз / одновременно
- наращивание: повершинное / послойное

Открытые проблемы:

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of hierarchical PTMs is also an open issue."

Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. 2011.

Регуляризатор родительских тем (реализован в BigARTM)

Шаг 1. Строим модель с небольшим числом тем.

Шаг k. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\Psi},$$

где $\Psi = (\psi_{st})_{S\times T}$ — *матрица связей*, $\psi_{st} = p(s|t)$.

$\Phi^p \approx \Phi\Psi$, отсюда регуляризатор матрицы Φ:

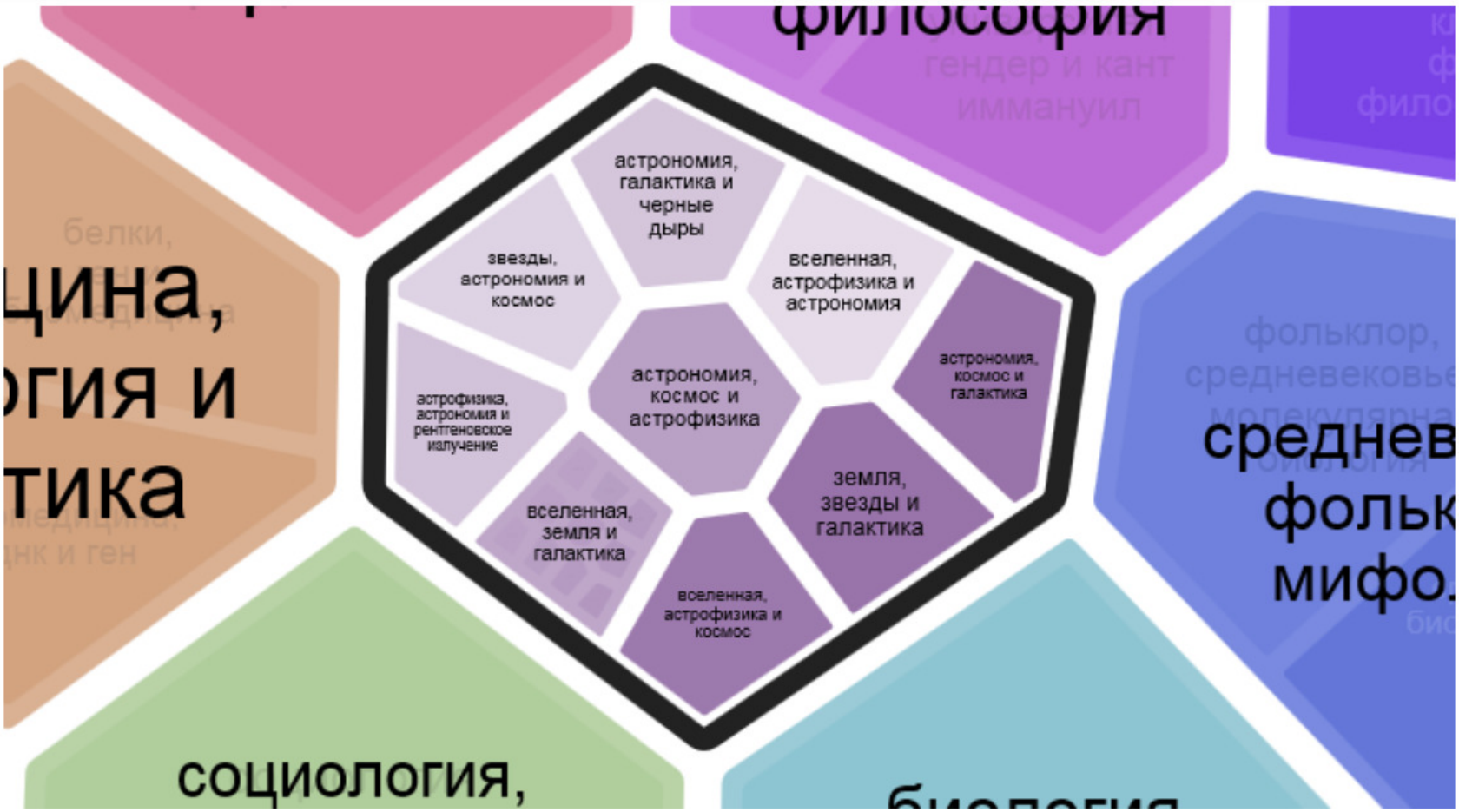
$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} \;\to\; \max.$$

Родительские темы t — «документы» с частотами слов n_{wt}.

Визуализация древовидных иерархий (FoamTree)

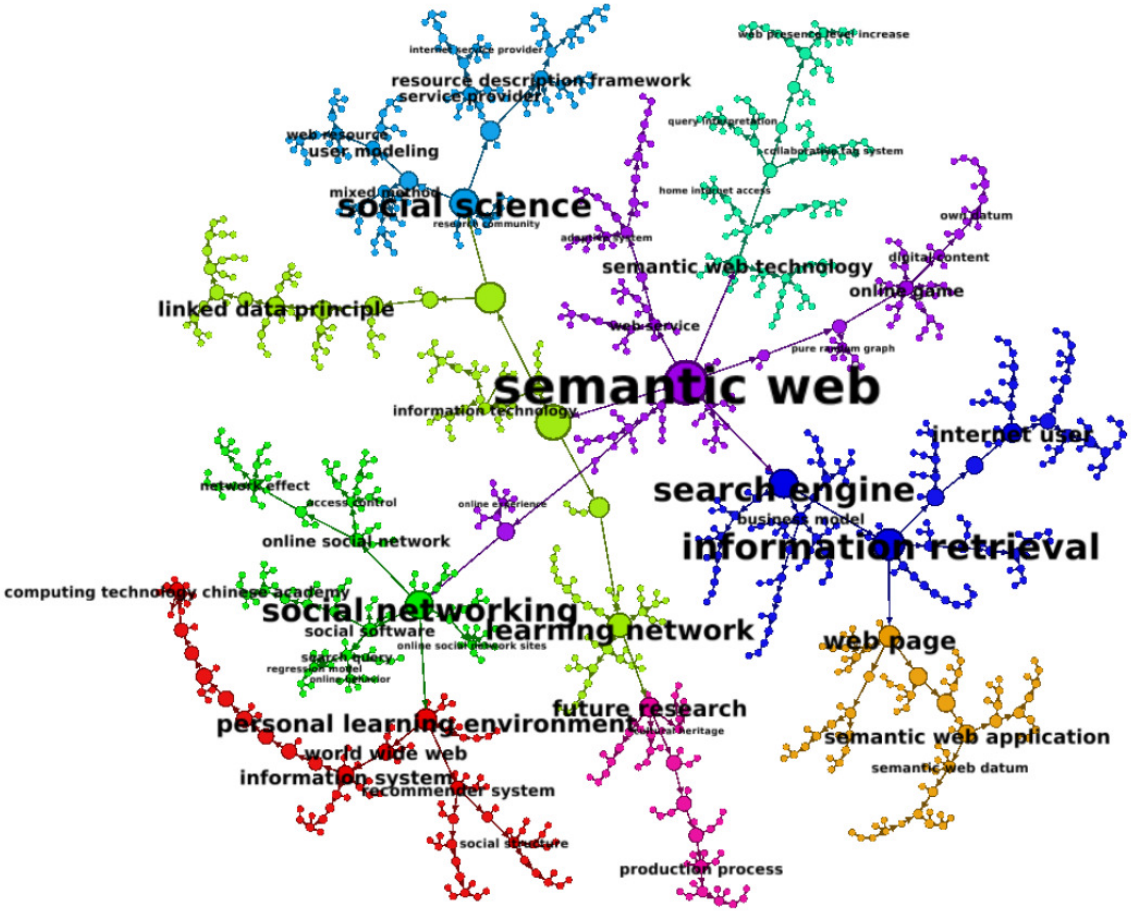
https://carrotsearch.com/foamtree-overview

Визуализация древовидных иерархий (FoamTree)



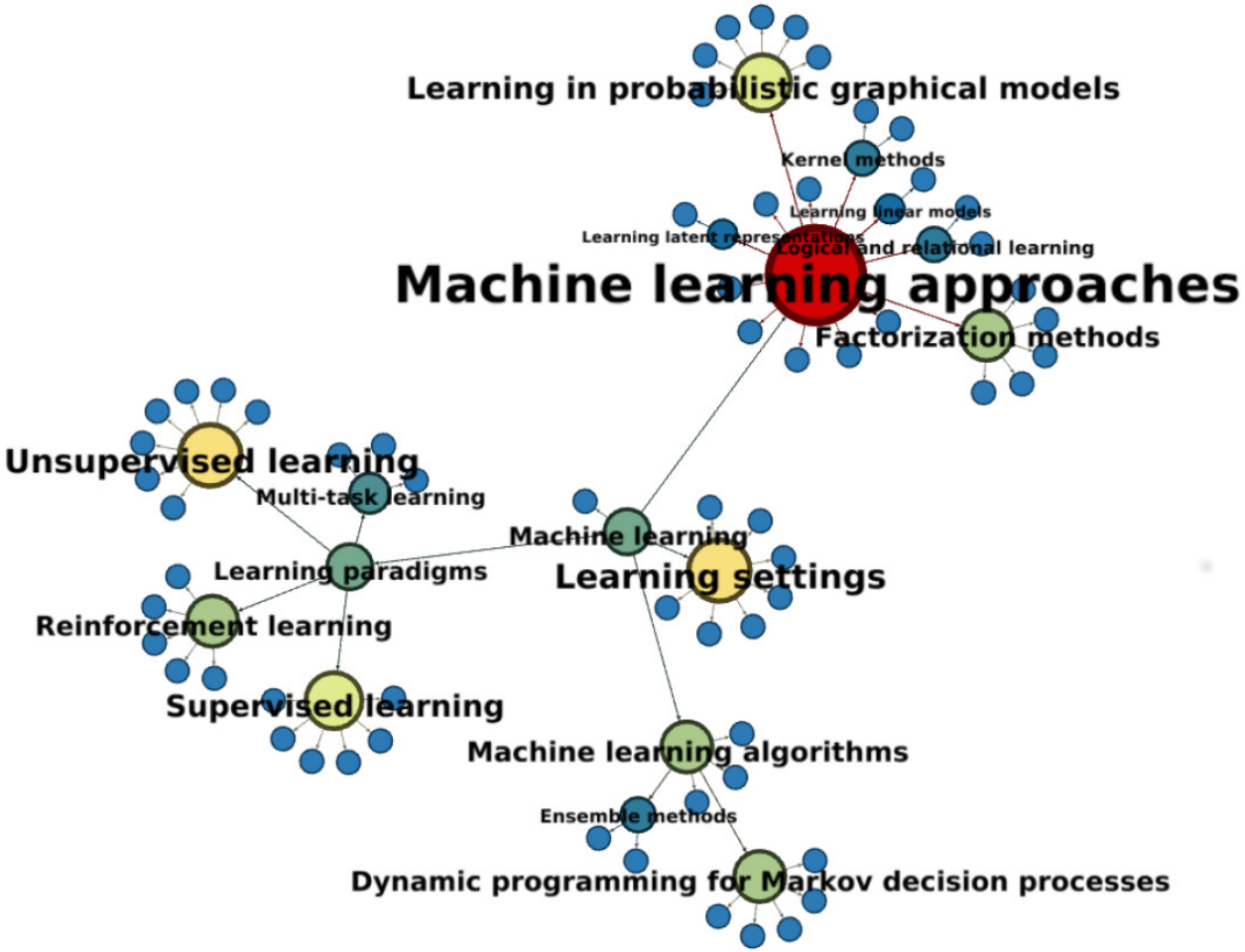
https://carrotsearch.com/foamtree-overview

Визуализация древовидных иерархий



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Визуализация древовидных иерархий



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Биграммная тематическая модель

n_{dvw} — частота пары слов «vw» в документе d
$\phi^v_{wt} = p(w|v,t)$ — распределение слов после слова v в теме t

Модель BTM (Bigram Topic Model):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \phi^v_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Это мультимодальная модель:
$M = W$, каждому слову v соответствует отдельная модальность,
$W^v = W$ — все слова, которые могут следовать за v.

Недостатки биграммной модели BTM:

- все пары соседних слов образуют биграммы;
- модель не описывает отдельные слова (униграммы);
- общее число токенов $O(|W|^2)$.

Hanna Wallach. Topic modeling: beyond bag-of-words // ICML 2006

Объединение униграмм и биграмм в одной модели

Модель TNG (Topical n-grams):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \underbrace{\big(\xi_{vt}\phi^{v}_{wt} + (1-\xi_{vt})\phi_{wt}\big)}_{p(w|v,t)} \theta_{td} \to \max_{\Phi,\Theta,\Xi},$$

где $\xi_{vt} = \mathrm{P}($слово v начинает биграмму в теме $t)$.

Мультимодальная модель ARTM:

W^n — словари n-грамм, отфильтрованные по трём критериям:

1) наличие подчинительных синтаксических связей,
2) статистически значимая совстречаемость (коллокация),
3) высокая тематичность $\mathrm{KL}\big(\frac{1}{|T|} \,\big\|\, p(t|vw)\big) \to \max$.

Xuerui Wang, Andrew McCallum, Xing Wei. Topical *N*-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. 2007.

Биграммы радикально 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.

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

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

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

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:

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

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

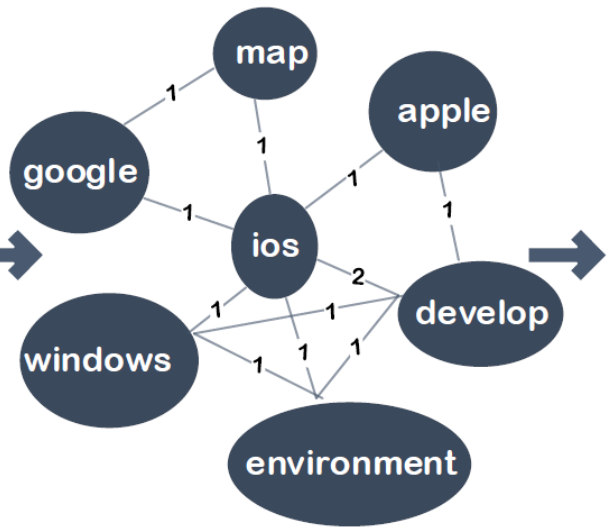
2. Регуляризатор разреживания для матрицы Θ:

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

Модель 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 $\Theta = \mathrm{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathsf{T}}$.

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.

Примеры векторных операций в задаче аналогии слов

Два подхода к синтезу векторных представлений слов:

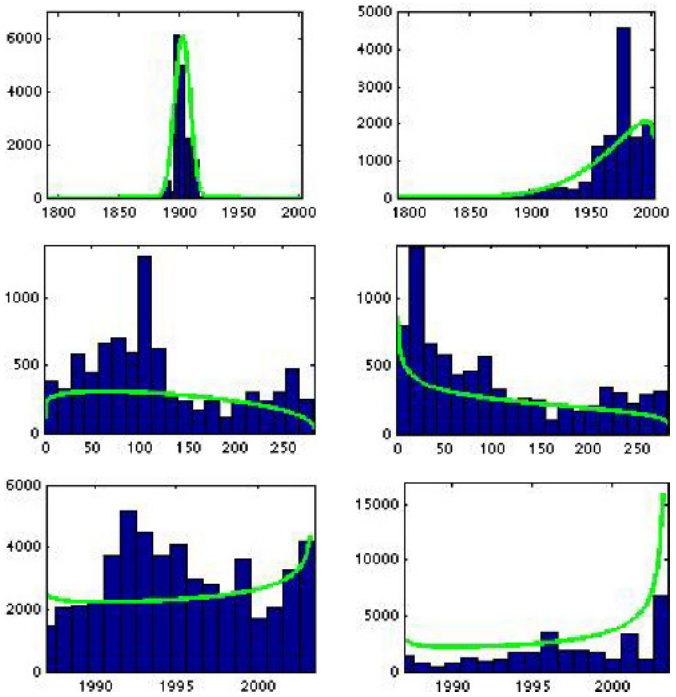
- **ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

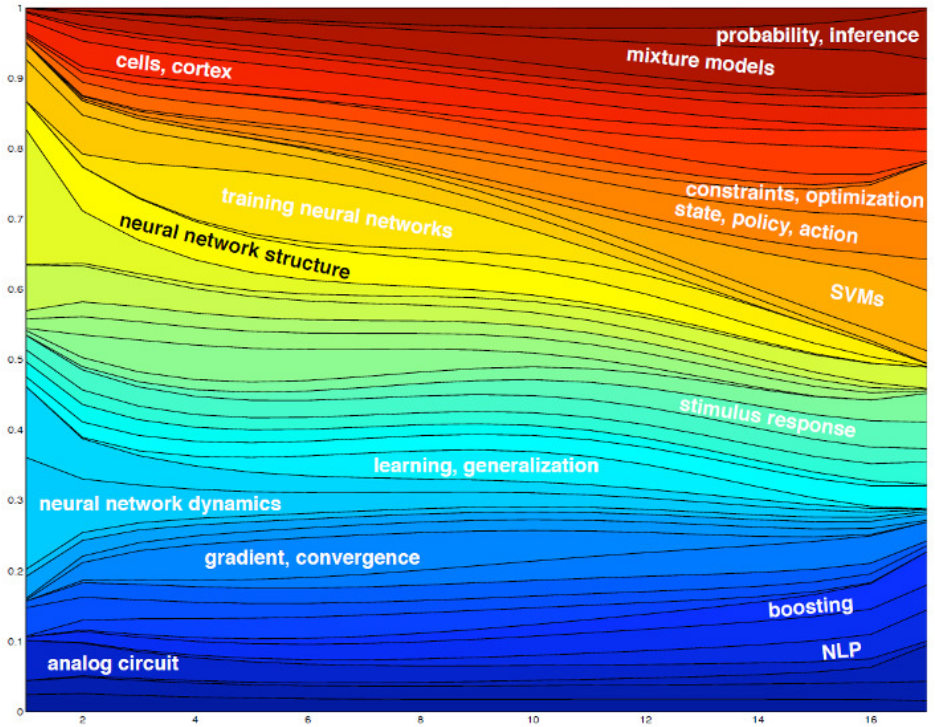
Операция	Результат ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah,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

Артём Попов. Регуляризация тематических моделей для векторных представлений слов. 2017. ВМК МГУ.

Модель TOT (Topics over Time)

1. Каждая тема имеет непрерывное β-распределение во времени
2. Каждое слово имеет метку времени





Xuerui Wang, *Andrew McCallum*. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends // ACM SIGKDD-2006

Темпоральные тематические модели

Неадекватность TOT очевидна даже по картинкам из статьи!

Наши предположения:

- Время дискретно, $i \in I$ — интервалы времени
- Как и в TOT, темы $p(w|t)$ не меняются во времени
- *Перманентные* темы имеют медленно меняющиеся $p(i|t)$
- *Событийные* темы имеют $p(i|t) = 0$ почти всё время
- Метки времени приписываются документам, а не словам
- Параметрические модели не используются

Цели моделирования:

- Выделить событийные и перманентные темы.
- Проследить развитие тем во времени.
- Выделить тренды (в новостях, в научных публикациях).

Регуляризаторы Θ для темпоральных тематических моделей

I — интервалы времени (например, годы публикаций),
$D_i \subset D$ — все документы, относящиеся к интервалу $i \in I$.
$n_i = \sum_{d \in D_i} n_d$ — доля коллекции, относящаяся к интервалу i.

1. Разреживание $p(t|i) = \sum_{d \in D_i} \theta_{td} \frac{n_d}{n_i}$ в каждом интервале i:

$$R_1(\Theta) = \tau_1 \sum_{i \in I} \mathrm{KL}\Big(\tfrac{1}{|T|} \| p(t|i)\Big) \to \max.$$

2. Сглаживание $p(i|t) = \sum_{d \in D_i} \theta_{td} \frac{n_d}{n_t}$ в соседних интервалах i, $i-1$:

$$R_2(\Theta) = -\tau_2 \sum_{i \in I} \sum_{t \in T} \big|p(i|t) - p(i-1|t)\big| \to \max.$$

Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Время как модальность. Регуляризатор Φ

Проблема регуляризатора Θ в онлайновом EM-алгоритме: соседние по времени документы могут попасть в разные пакеты.

Документы содержат слова $w \in W^1$ и время $i \in W^2 = I$
W^2 — модальность интервалов времени (time stamps)

1. Разреживание $p(t|i)$ эквивалентно разреживанию $p(i|t) = \phi_{it}$:

$$R_1(\Phi^2) = -\tau_1 \sum_{i \in I} \sum_{t \in T} \ln \phi_{it} \to \max$$

2. Сглаживание $p(i|t) = \phi_{it}$ в соседних интервалах i, $i-1$:

$$R_2(\Phi^2) = -\tau_2 \sum_{i \in I} \sum_{t \in T} \left|\phi_{it} - \phi_{i-1,t}\right| \to \max$$

Задача анализа потока пресс-релизов

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.

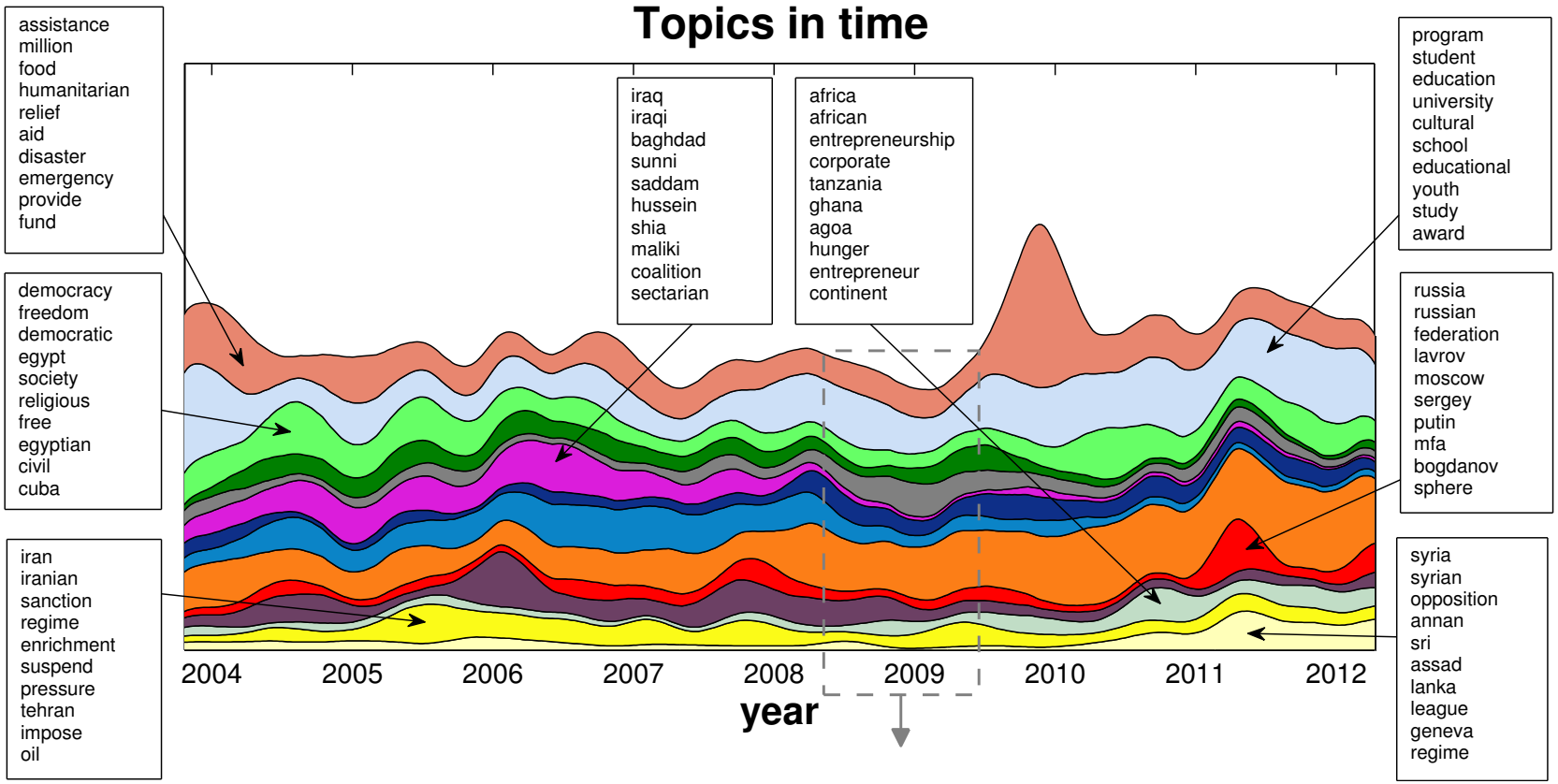
Цели исследования:

- какие темы общие, какие специфичны для источников?
- какие темы событийные, какие перманентные?
- какие темы и когда коррелируют с заданной темой?

Модальности и регуляризаторы:

- две модальности: источники, интервалы времени
- разреживание, сглаживание, декоррелирование
- сглаживание тем во времени

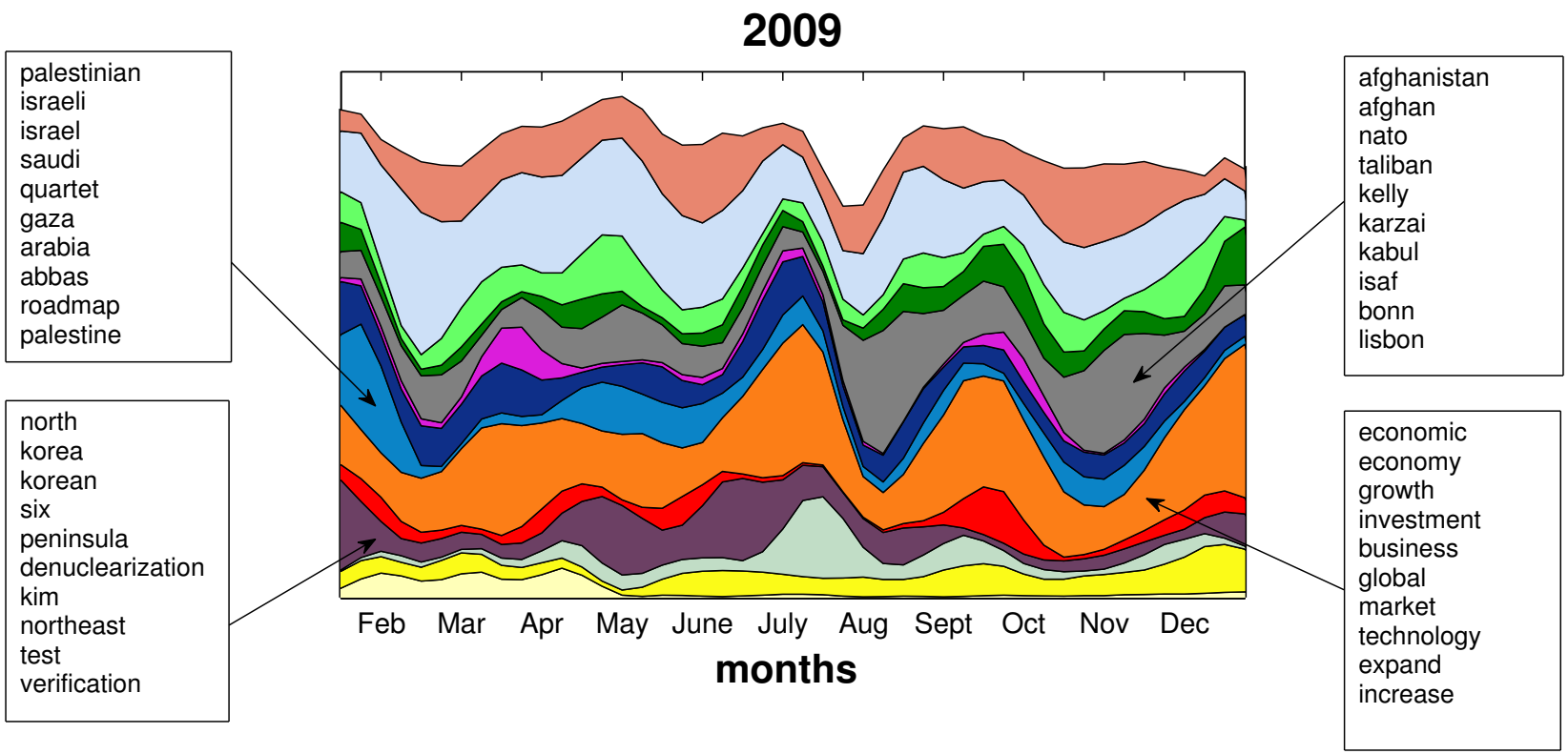
Динамика тем во времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

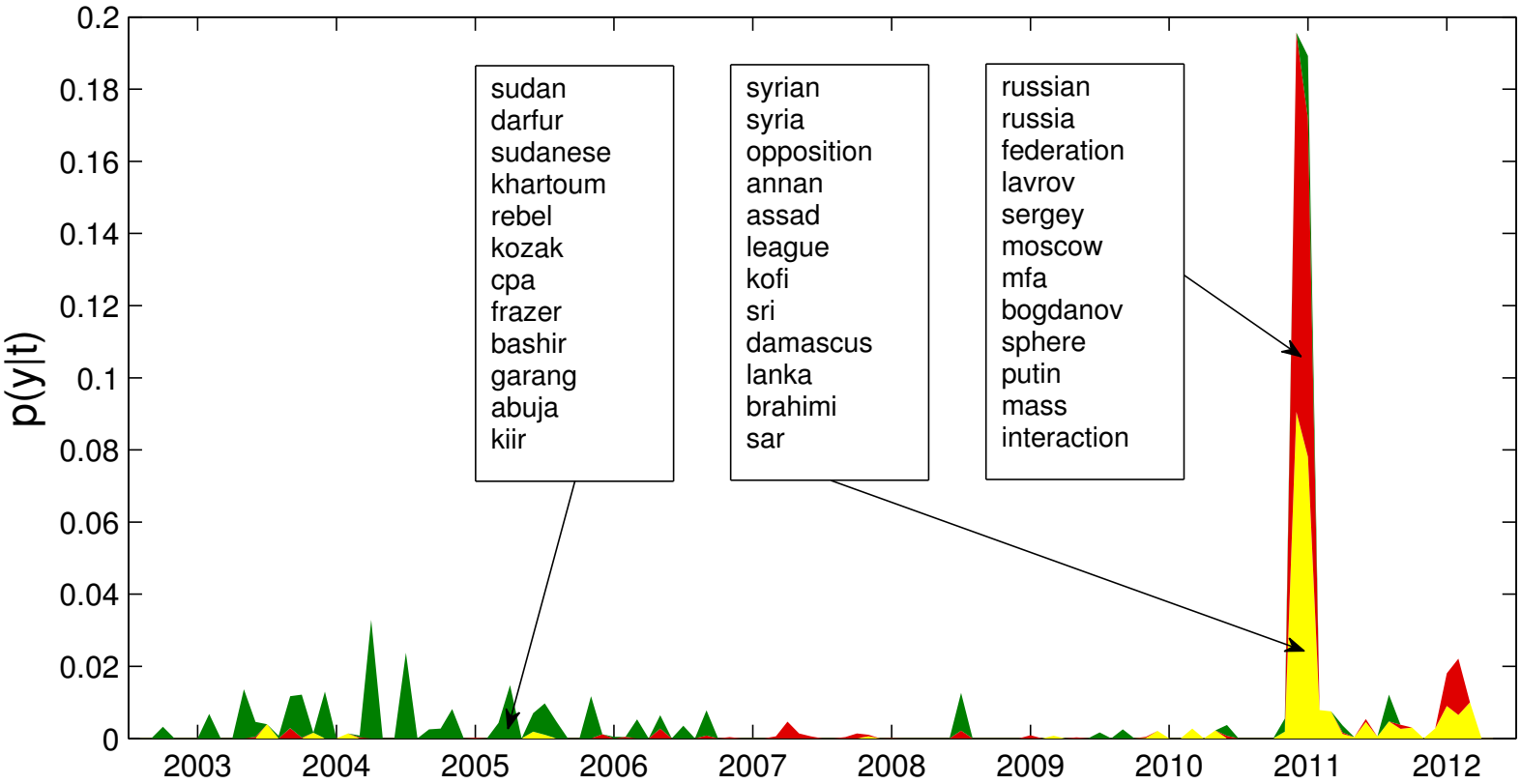
Укрупнение масштаба времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

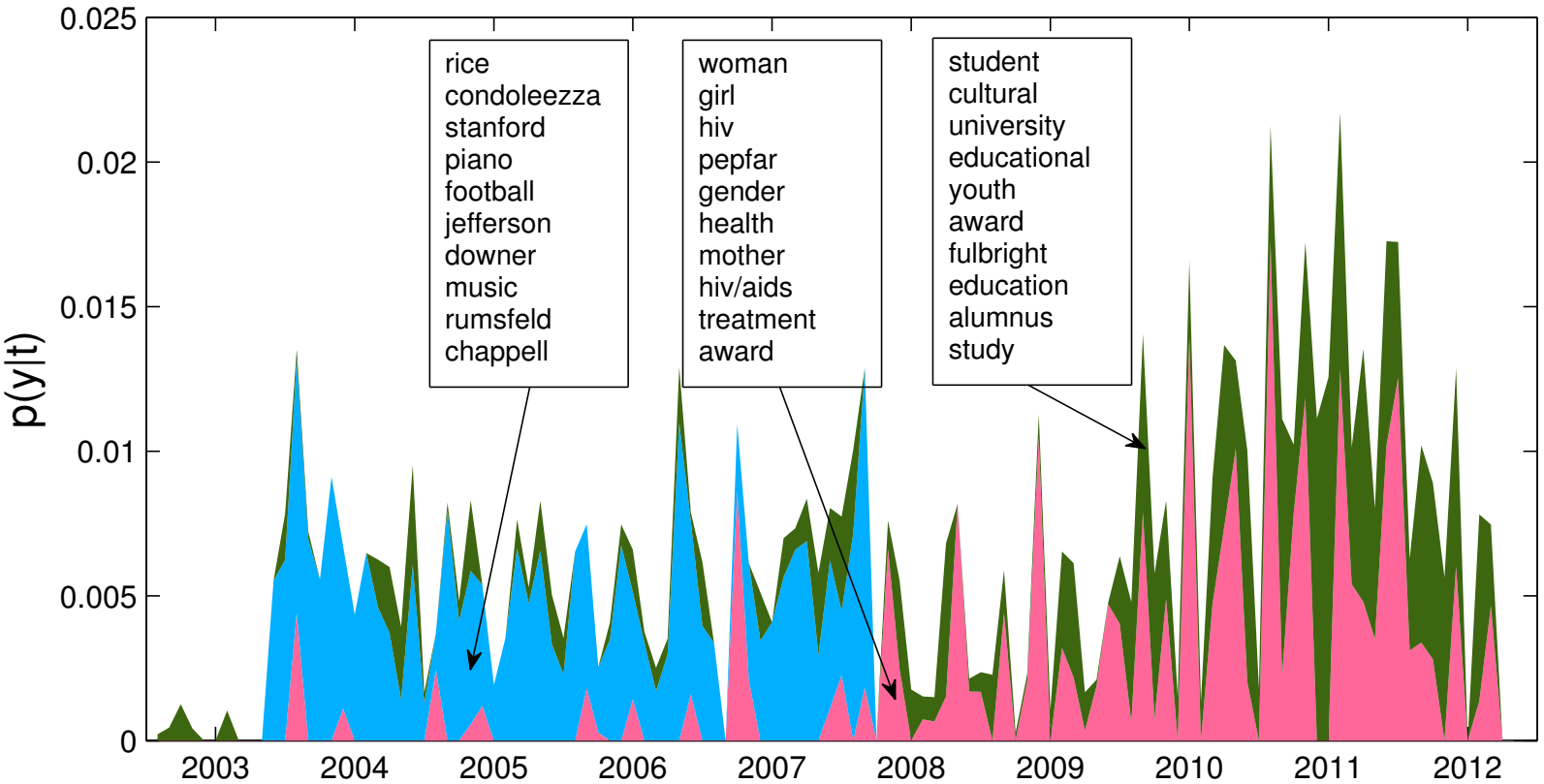
Пример: событийные темы и момент их совместного всплеска



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

Примеры перманентных тем (сглаживание отключено)



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Метод тематической сегментации TopicTiling

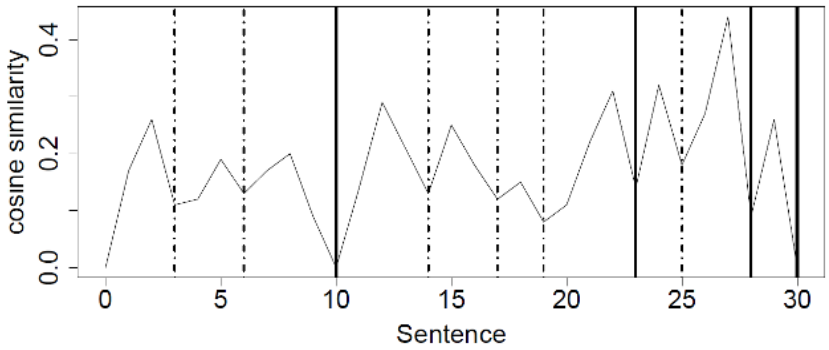
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

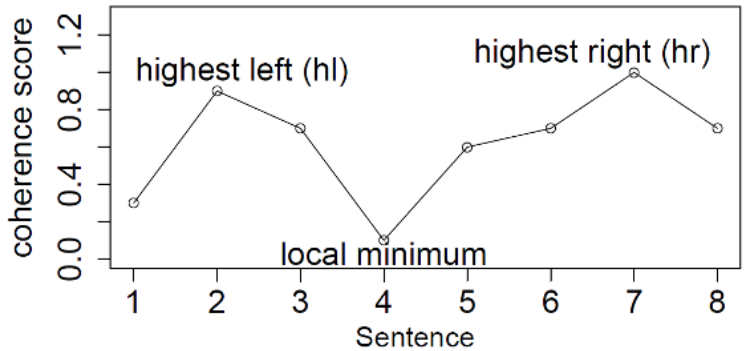
$p(t|d,s) = \frac{1}{|s|}\sum_{w\in s} p(t|d,w)$ — тематика предложения s

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

Задача тематической сегментации документов

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

Матрица *тематики слов в документах* $p(t|d, w_i)$ размера $T \times n_d$:

Регуляризация E-шага

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

Максимизация log правдоподобия с регуляризаторами R и $\tilde{R}$:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

Доказательство

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 функцию от вспомогательных переменных Π:

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

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

$$\phi_{wt}\frac{\partial R(\Pi)}{\partial \phi_{wt}} = \sum_{d\in D} p_{tdw}Q_{tdw}(\Pi); \qquad \theta_{td}\frac{\partial R(\Pi)}{\partial \theta_{td}} = \sum_{w\in d} p_{tdw}Q_{tdw}(\Pi).$$

Лемма 3. Формулы M-шага:

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

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

Тематическая модель сегментированного текста

S_d — множество сегментов, на которые разбит документ d
n_{sw} — число вхождений слова w в сегмент s длины n_s
Тематика сегмента $s \in S_d$ — средняя тематика его слов:

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

Регуляризатор разреживания $\mathrm{KL}\left(\frac{1}{|T|} \,\middle\|\, p(t|d,s)\right) \to \max$:

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

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.

Что такое «тема» в записи разговора контакт-центра

Типы тем в холодных продажах:

- Представление
- Продукт
- Свойство продукта
- Тип возражения клиента
- Аргумент оператора
- Дальнейшее действие клиента
- Прощание

Тематическое моделирование коллекции разговоров:

- **Вход:** каждый документ — последовательность слов.
- **Выход:** каждая тема t — частотный словарь слов $p(w|t)$; каждый разговор d — распределения тем $p(t|d)$, $p(t|d, w_i)$.

Пример темы: «Перевод баланса»

имеет смысл перевести потому что три месяца вы без процентов гасите это уже как согласитесь выигрыш но далее уже процент

с нашей помощью мы вы могли бы погасить ваши кредиты в других банках и беспроцентный период в этом случае увеличится на срок до девяноста дней

воспользоваться услугой перевод баланса услуга позволит вам погасить долг в другом банке оплачивая его вы не будить платить проценты в течении месяцев

предоставляет возможность воспользоваться услугой перевод баланса

беспроцентный период в этом случае увеличится на срок до девяноста дней

услуга бесплатная с помощью этой услуги вы можете частично или полностью закрыть действующие кредиты в других банках закрыть

Пример темы: «Оформление заявки»

для этого нужно сначала оформить заявку

мы в телефонном режиме оформляем заявку буквально десять минут вашего времени

если заявку на сайте сделать равно вам необходима помощь операторов в том моменте что там бывают такие нюансы что ну правильно заполнение то есть без оператора

предлагаю только составить заявку

в заявке указываете в кредит который вы желаете да по своим возможностям по своим по своим потребностям

готовы сейчас оформить заявку

оформить заявку на получение кредитной карты

для получения нашей кредитной карты предлагаю сейчас заполнить заявку для этого потребуется пять семь времени

вы можете оставить заявку

Пример темы: «Льготный период»

льготный период до пятидесяти пяти дней позволяет людям отдохнуть за границей при этом вернуться и пользуясь льготным периодом восполнить средства по карте то есть не потерять на процентах

льготный период до пятидесяти пяти дней у вас практически два месяца на беспроцентное погашение

беспроцентный льготный период до пятидесяти пяти дней вы совершаете покупки в обычном режиме в магазине на заправке в аптеке и при этом не платить проценты

льготный период когда вы можете пользоваться деньгами по карте и не платить за это проценты

льготный период беспроцентный когда вы можете погашать задолженность абсолютно при этом ничего не переплачивая

Пример. Сравнение асессорской и модельной разметки

- цветом выделяются темы
- подчёркиванием выделяется асессорская разметка

Оформление заявки | Индивидуальный подход | Решение банка | Доставка
Бонусная программа | Бесплатная доставка/оформление

вот на данный момент я звоню предлагаю только составить заявку чтобы банк изучил вашу кредитную историю и подобрал под вас индивидуальный тарифный план после чего на ваш мобильный поступит уведомление в котором будет указано каким образом в случае положительного ответа будут доставлены бумаги у нас есть два способа доставки это либо курьерская доставка либо заказным письмом почтой России

ну вот бонусы значит на все абсолютно покупки один процент а если вы совершаете покупки у банка будет полный перечень магазинов у вас в личном кабинете до тридцати процентов бонусов можете то есть вот две тысячи что то купили а ориентировочно шестьсот вернулось вам на это уже плюсов согласитесь что это довольно таки это одна покупка вот так и хочу сказать что вы абсолютно ничего не теряетесь соглашаясь оформить заявку ничего за что не платите потому что вам карту выпускают доставляют абсолютно бесплатно вам либо представитель банка привозит либо по почте она приходит

Анализ ошибок выделения сегментов по темам

- Part% — доля правильных сегментов в теме
- Error Part% — вклад темы в суммарную ошибку

Topic name	success	Total	Part %	Error Part %	бонусная программа	879	1009	87.12	1.650
Приветствие	224	232	96.55	0.102	использование карты	50	495	10.10	5.646
Цель звонка	75	425	17.65	4.441	карту не кредит	119	133	89.47	0.178
Удобно разговаривать	142	186	76.34	0.558	с кем разговариваю	23	142	16.20	1.510
Кредиты в других банках	118	223	52.91	1.332	дистанционность	116	168	69.05	0.660
престиж карты	37	56	66.07	0.241	общий вариант отказа	63	88	71.59	0.317
кредитный лимит	103	178	57.87	0.952	звонили ранее	14	27	51.85	0.165
беспроцентный период	275	581	47.33	3.883	данные абонента	58	89	65.17	0.393
доставка карты	88	113	77.88	0.317	откуда номер	24	51	47.06	0.343
доставка карты	65	77	84.42	0.152	точно сотрудник	46	46	100.00	0.000
индивидуальный подход	138	220	62.73	1.040	желаемая сумма	69	91	75.82	0.279
решение о карте	404	792	51.01	4.923	снятие/партнеры	13	22	59.09	0.114
перевод баланса	308	370	83.24	0.787	решение банка	30	93	32.26	0.799
общие вопросы клиенту	161	400	40.25	3.033	большой процент	4	9	44.44	0.063
процентная ставка	235	283	83.04	0.609	есть карта другого банка	20	48	41.67	0.355
оформление заявки	69	182	37.91	1.434	интернет банк	25	52	48.08	0.343
наличие паспорта	125	158	79.11	0.419	мобильное приложение	38	46	82.61	0.102
сравнить условия	58	96	60.42	0.482	годовое обслуживание	56	62	90.32	0.076
связь с банком	81	196	41.33	1.459	подумаете	22	22	100.00	0.000
перезвонить	229	292	78.42	0.799	слышите?	6	7	85.71	0.013
до свидания	5	8	62.50	0.038	нет полномочий	11	11	100.00	0.000

Резюме

- Модальности
  - классы, категории
  - языки
  - биграммы
- Псевдо-документы
  - родительские темы в иерархиях
  - контексты слов (модель WNTM)
- Регуляризаторы
  - разреживание, сглаживание, частичное обучение
  - декоррелирование тем
  - битермы (для коротких текстов)
- Метрики качества
  - внутренние (перплексия, когерентность, различность тем)
  - внешние (качество классификации, поиска, рекомендаций)