Тематическое моделирование в BigARTM: новые возможности

Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ)

Содержание

Тематическое моделирование, ARTM, BigARTM

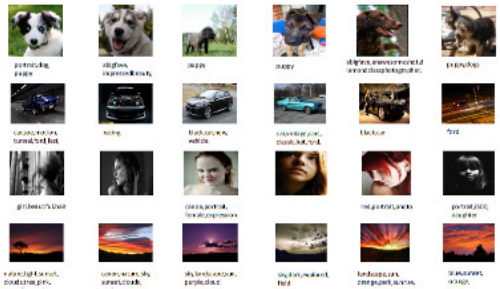
- *Тематическое моделирование* (Topic Modeling):
  выявление латентной тематической структуры текстов,
  на входе — большая текстовая коллекция,
  на выходе — частоты тем в документах и слов в темах
- *ARTM* — теория комбинирования тематических моделей для построения моделей с заданными свойствами
- *BigARTM* — проект с открытым кодом,
  «ЛЕГО-конструктор» тематических моделей
- *Основное отличие* от байесовских моделей — простой переход от постановки задачи к алгоритму

Приложения тематического моделирования

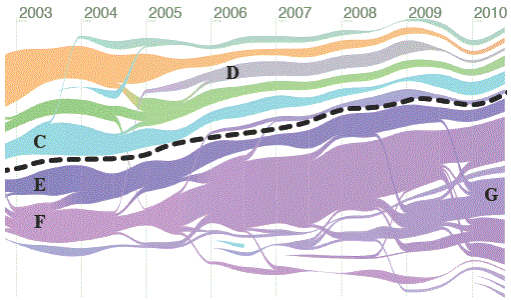
разведочный поиск в электронных библиотеках

персонализированный поиск в соцсетях

мультимодальный поиск текстов и изображений

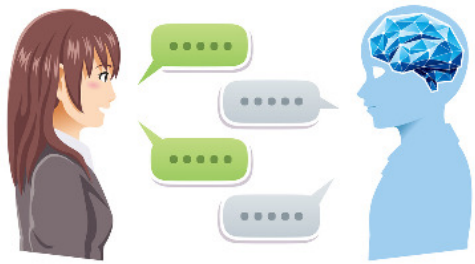


детектирование и трекинг новостных сюжетов



навигация по большим текстовым коллекциям

управлением диалогом в разговорном интеллекте



Постановка задачи тематического моделирования

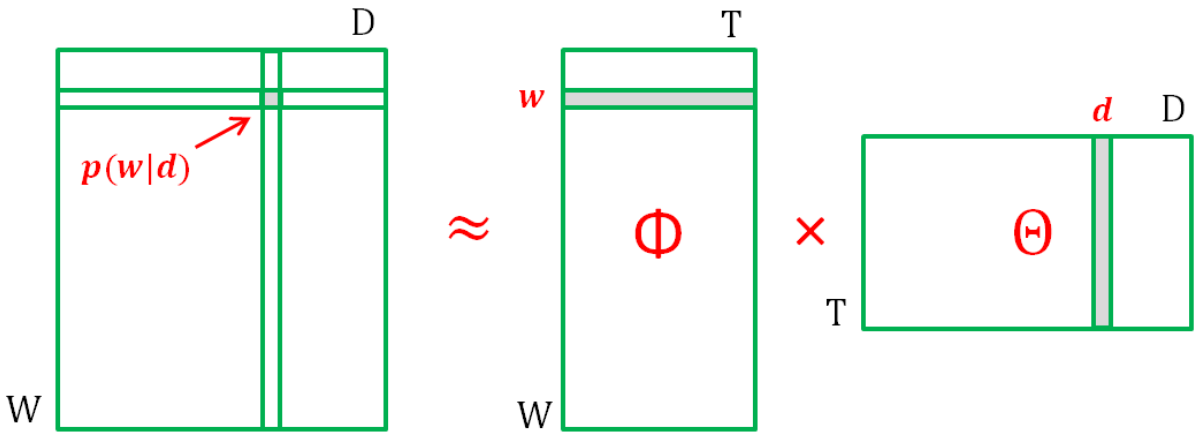
Дано: коллекция текстовых документов

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

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

- $\phi_{wt} = p(w|t)$ — вероятности слов 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:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

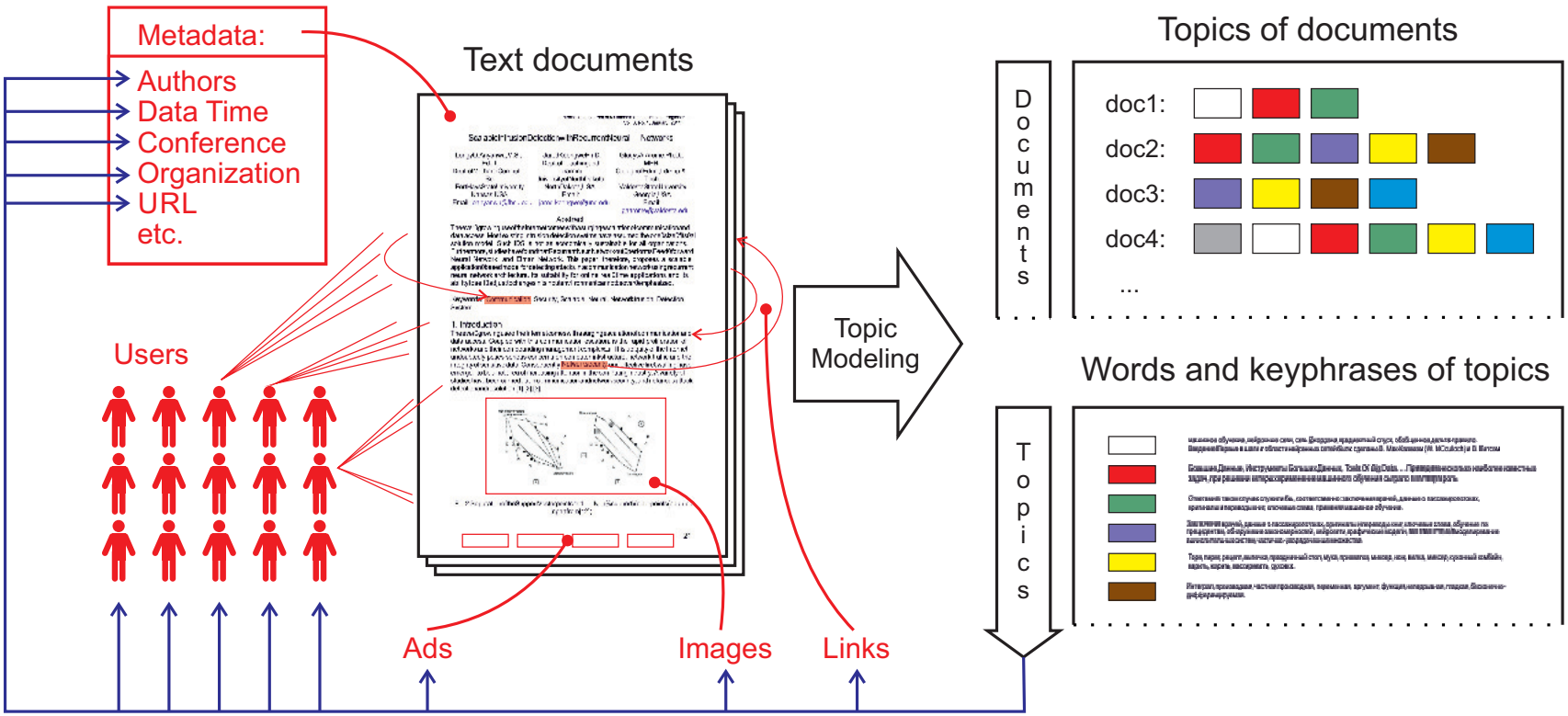
$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{d \in D} n_{dw} p(t|d,w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} n_{dw} p(t|d,w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}}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

где $\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s \in T} \max\{x_s,0\}}$ — операция нормировки вектора.

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

Задачи мультимодального тематического моделирования

Темы определяют распределения не только терминов $p(w|t)$, но и других *модальностей*: $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|\text{t}), \ldots$



Мультимодальная ARTM

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

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

K.Vorontsov, O.Frei, M.Apishev et al. Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

- 3.7M статей Википедии, 100K слов

	проц.	$T = 50$		$T = 200$	
		минут	перплексия	минут	перплексия
BigARTM	1	42	5117	83	3347
BigARTM async	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM async	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM async	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

TransARTM: тематизируем транзакционные данные

- *Тексты:* документы $\leftrightarrow$ слова
- *Рекомендательные системы:* субъекты $\leftrightarrow$ объекты
- *Транзакции:* взаимодействие трёх и более сущностей
- Обобщение легко строится в ARTM!
- И уже реализовано в BigARTM!

Илья Кариков. Гиперграфовые тематические модели транзакционных данных. (DataFest сегодня)

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки элементов разных модальностей.

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

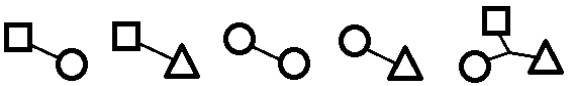
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

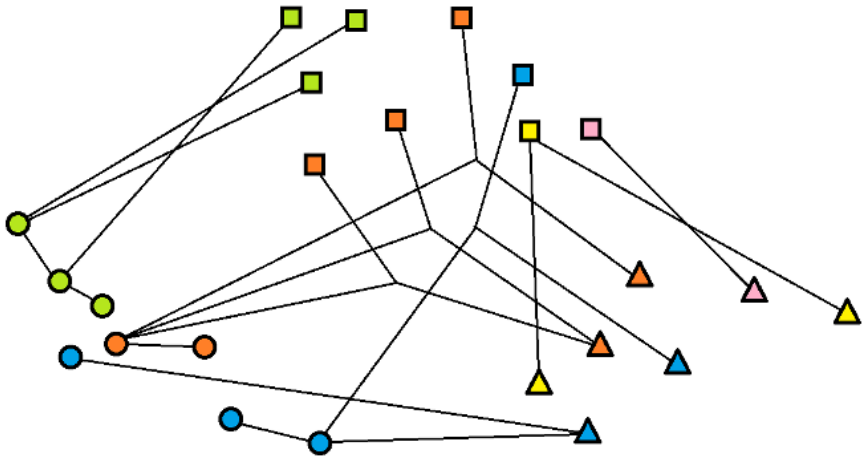
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p_k(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

$$p_k(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt},$$

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{kvt} = p_k(v|t)$ — для модальности v в теме t на рёбрах типа k

Задача максимизации log правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{kvt} \to \max_{\Phi, \Theta},$$

$$\phi_{kvt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{kvt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K} \tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \phi_{kvt} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{ktdx} = p_k(t|d,x)$:

E-шаг:
$$p_{ktdx} = \operatorname{norm}_{t\in T}\Big(\theta_{td} \prod_{v\in x} \phi_{kvt}\Big)$$

M-шаг:
$$\begin{cases} \phi_{kvt} = \operatorname{norm}_{v\in V^m}\Big(\sum_{(d,x)} [v\in x]\tau_k n_{dx} p_{ktdx} + \phi_{kvt}\frac{\partial R}{\partial \phi_{kvt}}\Big) \\ \theta_{td} = \operatorname{norm}_{t\in T}\Big(\sum_{k\in K}\sum_{(d,x)} \tau_k n_{dx} p_{ktdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Пост-обработка E-шага: обходим гипотезу мешка слов

- *Гипотеза «мешка слов»* — самое критикуемое предположение тематического моделирования
- Кажется, что оно заложено в самой конструкции разложения матрицы $p(w|d) = \frac{n_{dw}}{n_d}$
- Тем не менее, это ограничение можно обойти с помощью регуляризатора E-шага, учитывающего позиции слов
- *Лайфхак:* делать пост-обработку матриц $p(t|d,w)$, остальное оставить как есть
- Уже описано в ARTM и реализовано в BigARTM!

Николай Скачков. Тематическая сегментация текста. (DataFest сегодня)

Сегментная структура текста и пост-обработка E-шага

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

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

Гипотеза: пост-обработка E-шага — это неявная регуляризация

Между E- и M-шагом добавляется обработка матрицы $p_{tdw} = p(t|d,w)$ тематики слов документа:

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

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

Гипотеза

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

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

hARTM: выясняем отношения тем с их родителями

- Строить иерархии тем в BigARTM — это очень просто!
- 1) построить плоскую модель верхнего уровня
- 2) для каждого уровня: задать число тем,
  сделать из распределений $p(w|t)$ родительских тем
  псевдо-документы и добавить их в коллекцию
- Это эквивалентно регуляризации в ARTM!
- Класс hARTM уже реализован поверх BigARTM!

Послойное построение уровней тематической иерархии

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:

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

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

Родительская $\Phi^p \approx \Phi\tilde{\Psi}$, отсюда регуляризатор матрицы Φ:

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

Родительские темы t — *псевдо-документы* с частотами слов n_{wt}.

Двухуровневая модель коллекции postnauka.ru

20 тем на верхнем уровне, 58 тем на нижнем уровне

nGrammer: отбираем термины по тематичности

- Учёт n-грамм сильно улучшает интерпретируемость тем
- Как отсеять мусорные n-граммы и оставить *термины*?
- 1) статистически: например, алгоритм TopMine
- 2) синтаксически: например, с помощью SyntaxNet
- 3) тематически: по пиковости распределений $p(t|w)$
- Оказывается, 2) не нужно, если делать 1) и 3)
- Достаточно PLSA (без регуляризации) на 30–50 тем
- Реализовано и совместимо с BigARTM!

Три метода отбора терминов: 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

VisARTM: красим темы во все цвета радуги

- Web-приложение для визуализации ARTM моделей
- Открытый код: https://github.com/bigartm/visartm
- Автоматическое перестроение моделей через BigARTM
- Текстовые интерактивные визуализации документов, тем, терминов, модальностей
- Графическая визуализация иерархических моделей
- Графическая визуализация темпоральных моделей
- *Тематический спектр:* темы можно ранжировать так, чтобы семантически близкие темы стояли рядом
- Сбор асессорских оценок

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

VisARTM: Визуализация документа

Химические коммуникации планктона

Эколог Егор Задереев о типах химических сигналов, миграциях зоопланктона и образовании покоящихся яиц

Text | Bag of words

Что исследователи знают о химической коммуникации планктона в воде? Какими сигналами обменивается зоопланктон? Как размножается зоопланктон? Об этом рассказывает кандидат биологических наук Егор Задереев.

Планктон — это организмы, местоположение которых в водной толще в основном определяется течениями. То есть это что-то маленькое, то, что переносится течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рачки. То, как водные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В наземных экосистемах, мы знаем, есть феромоны, различные сигнальные системы, которые хорошо исследованы. Мы используем их для создания ловушек, например, для вредителей — феромонные ловушки. Вода — это среда, которая благоприятна для химической коммуникации.

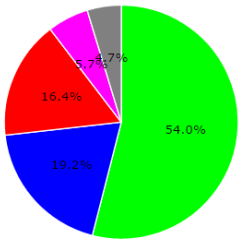
[post id="33793"]

Химические сигналы от хищников заставляют зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые ежесуточно происходят в океанах, морях и озерах. Зоопланктон ночью поднимается к поверхности, а днем уходит на глубину. Днем свет сверху помогает хищникам ловить животных, и животные уходят на глубину, а ночью поднимаются к поверхности, чтобы есть. Было показано, что эти вертикальные миграции регулируются двумя факторами. Первый — это освещенность. Очевидно, что, если не будет света, не будет сигнала. А второй — это химия, которую выделяют хищники.

В 2006 и 2009 годах выходили хорошие обзоры по химическим коммуникациям. То есть а) это очень маленькие молекулы, и б) они работают в очень низких концентрациях. Это до сих пор удивляет и поражает, потому что сообщества зоопланктона и вообще планктона в водных экосистемах — это сотни видов водорослей, рачков, которые живут в озерах, в морях, взаимодействуют между собой. А между ними есть очень сложная, судя по тому, что мы получаем в лаборатории, и разветвленная сеть химических сигналов и коммуникаций, которые влияют на разные поведенческие, физиологические и продуктивные функции. И эта сложная цепь, сеть взаимодействий до сих пор слабо исследована.

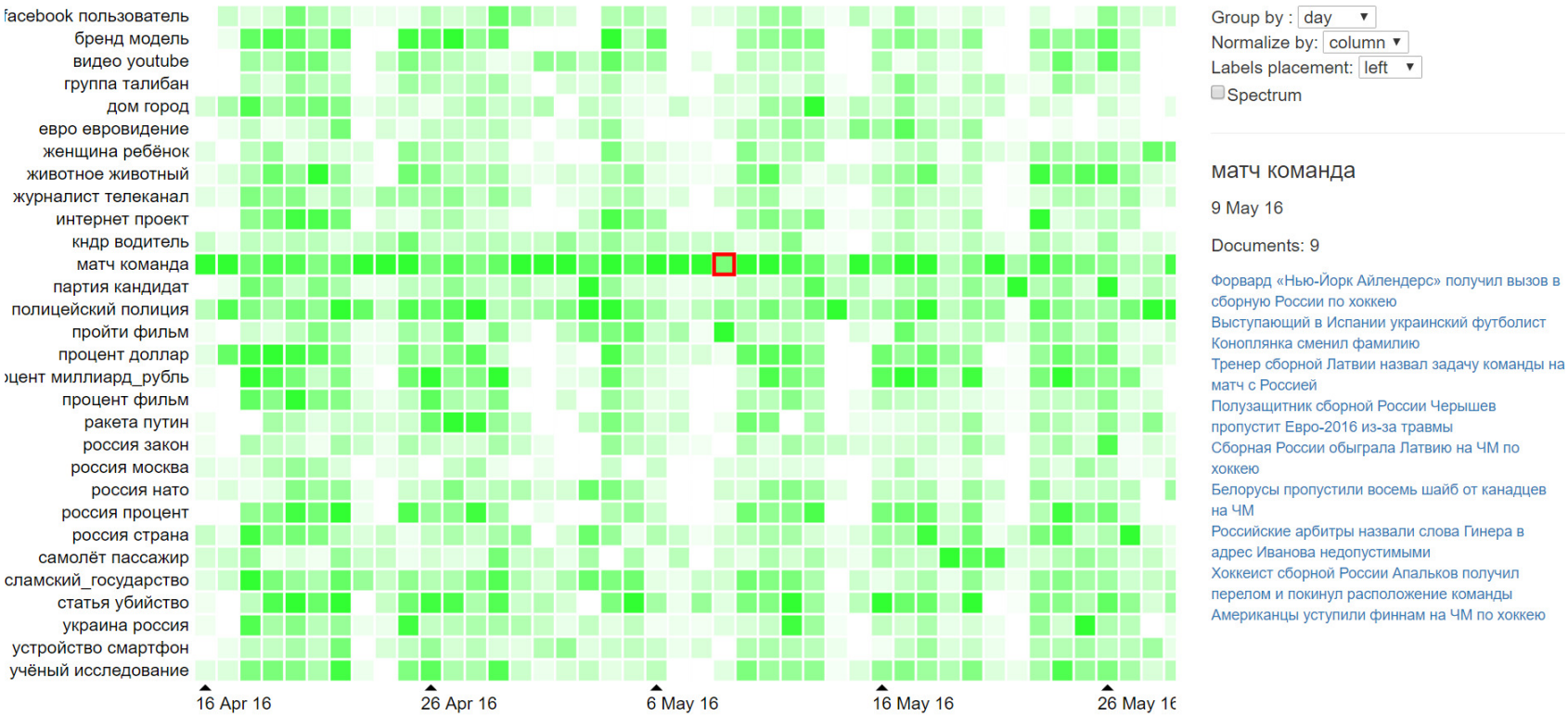
Dataset: postnauka
Time: Dec. 14, 2014, 3 p.m.
View original
index_id: 1866
text_id: 36719.txt
Terms count: 0
Unique terms count: 0
Model: flat-20
Highlighting: Words

Topic distribution

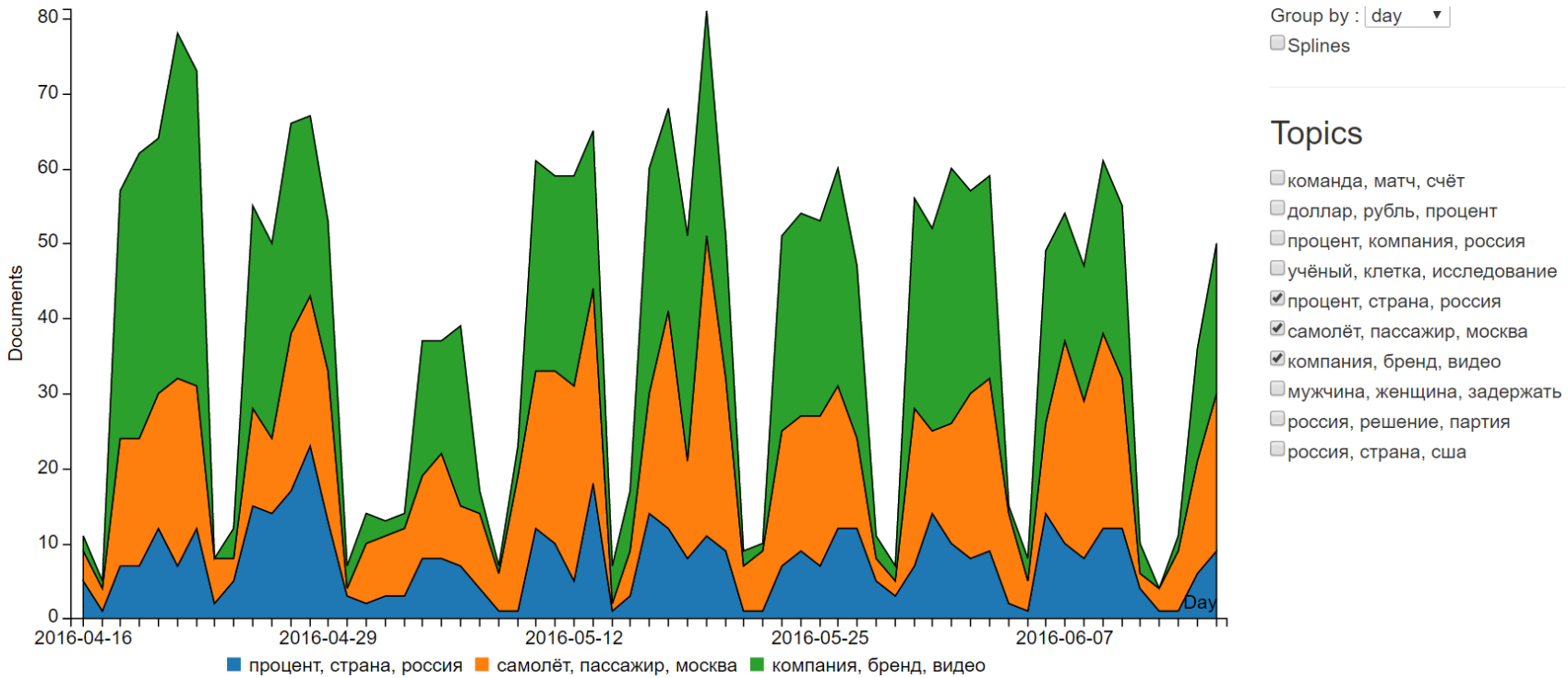


- земля, микроорганизм, вид
- вид, эволюция, ген
- материал, квантовый, структура
- город, социальный, пространство
- Other

VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация иерархической модели

Что такое «спектр тем» и зачем он нужен

Визуализация иерархии тем во времени (концепт):

- Интерпретируемые оси «время–темы»
- Близкие темы должны находиться рядом
- *Спектр тем* — одномерная линейная проекция (например, науки: гуманитарные $\rightarrow$ естественные $\rightarrow$ точные)

Построение спектра тем. Постановка задачи

Тематический спектр — такая перестановка тем $t_1, \ldots, t_{|T|}$, что сумма расстояний между соседними темами минимальна:

$$\sum_{i=2}^{|T|} \rho(t_i, t_{i-1}) \to \min$$

Функция расстояния $\rho(t, t')$ между темами, примеры:

- Манхэттенское: $\rho(t, t') = \sum_{w \in W} \left| \phi_{wt} - \phi_{wt'} \right|$
- Хеллингера: $\rho^2(t, t') = \frac{1}{2} \sum_{w \in W} \left(\sqrt{\phi_{wt}} - \sqrt{\phi_{wt'}} \right)^2$
- Жаккара: $\rho(t, t') = 1 - \frac{|W_t \cap W_{t'}|}{|W_t \cup W_{t'}|}$, $W_t = \left\{ w : \phi_{wt} > \frac{1}{|W|} \right\}$

Построение спектра тем — это задача коммивояжёра

Задача TSP (traveling salesman problem)

Найти путь минимальной суммарной стоимости, соединяющий T городов так, чтобы в каждом городе побывать один раз.

Алгоритм Лина–Кернигана в реализации Хельсгауна — лучший для решения задачи TSP, по данным *Encyclopedia of operations research* на 2013 год.
Вычислительная сложность $T^{2.2}$.

Другие алгоритмы оказались не только медленнее, но и хуже по качеству тематических спектров.

Keld Helsgaun. An effective implementation of the Lin–Kernighan traveling salesman heuristic. EJOR, 2000.

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

Пример спектра (коллекция postnauka.ru)

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

Пример спектра (коллекция postnauka.ru)

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

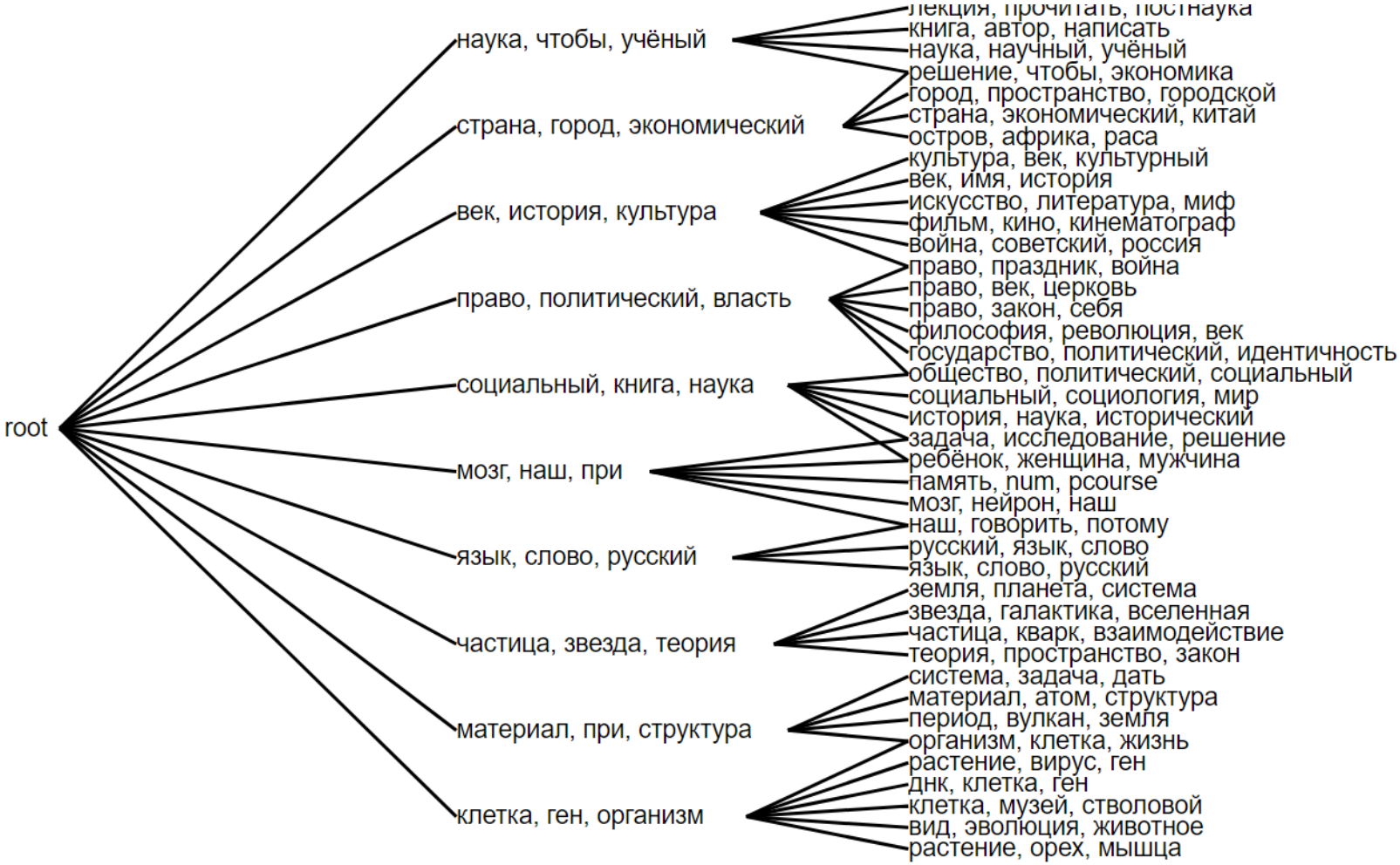
Пример спектра (коллекция lenta.ru)

1. спортсмен, допинг, олимпиада, рио, де, россия, проба, жанейро, wada, олимпийский_игра, соревнование
2. команда, матч, счёт, клуб, победа, чемпионат, турнир, минута, футболист, встреча, летний, футбол
3. евро, евровидение, страна, россия, конкурс, франция, болельщик, англия, украина, футбол, певец
4. пройти, мероприятие, россия, акция, фестиваль, москва, фильм, участник, картина, театр, музей
5. фильм, сериал, продукт, актёр, компания, продукция, процент, россия, книга, товар, картина, сезон
6. россия, москва, турист, процент, россиянин, страна, отель, рейс, путешественник, город, тысяча
7. процент, доллар, рубль, нефть, цена, россия, баррель, страна, уровень, вырасти, рынок, рост
8. компания, миллиард_рубль, процент, миллиард_доллар, россия, сумма, миллион_доллар, банк, банка
9. закон, законопроект, документ, реклама, использование, деятельность, поправка, внести, организация
10. россия, страна, керченский_пролив, российский, боинг, работа, чайка, ряд, гражданин, аэропорт
11. партия, кандидат, журналист, праймериза, выбор, единый_россия, госдума, выборы
12. россия, украина, крым, решение, киев, депутат, вопрос, отношение, страна, мнение, право, москва
13. россия, страна, турция, сша, ес, евросоюз, москва, санкция, отношение, украина, вопрос, государство
14. россия, сирия, исламский_государство, сша, нато, иго, запретить, террорист, страна, боевик
15. ракета, путин, россия, запуск, глава_государство, союз, спутник, президент
16. учёный, клетка, исследование, исследователь, ген, университет, оказаться, процент, помощь, организм
17. земля, животное, учёный, животный, тысяча, звезда, планета, обнаружить, кошка, территория, жизнь
18. самолёт, километр, машина, борт, пассажир, вертолёт, погибнуть, лайнер, пилот, час, район, яхта
19. полицейский, полиция, мужчина, задержать, автомобиль, улица, москва, пострадать, life
20. статья, убийство, задержать, суд, отношение, ук_рф, подозревать, следствие, обвинять, трамп, часть
21. ребёнок, женщина, мужчина, летний, дом, сын, семья, мальчик, жена, полиция, дочь, школа, врач
22. видео, youtube, ролик, фото, фотография, канал, снимка, auto, instagram, девушка, страница, группа
23. facebook, пользователь, интернет, страница, twitter, пост, написать, соцсеть, вконтакте, аккаунт
24. устройство, смартфон, компания, мотоциклист, игра, байкер, видео, миллион_доллар, робот, молодая
25. бренд, модель, компания, обувь, основать, одежда, релиз, коллекция, редакция, часы, поступить

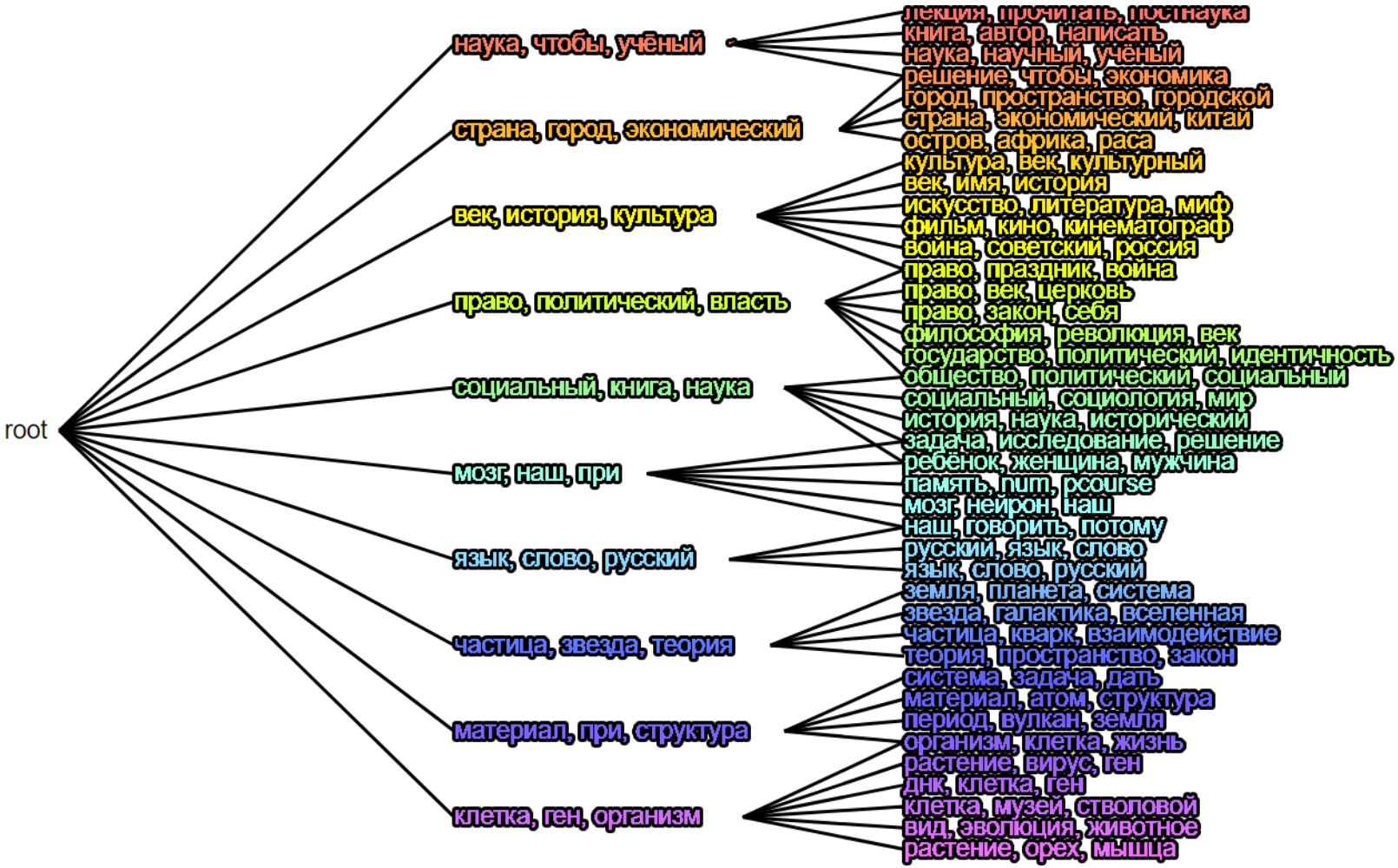
Пример спектра (коллекция lenta.ru)

1. спортсмен, допинг, олимпиада, рио, де, россия, проба, жанейро, wada, олимпийский_игра, соревнование
2. команда, матч, счёт, клуб, победа, чемпионат, турнир, минута, футболист, встреча, летний, футбол
3. евро, евровидение, страна, россия, конкурс, франция, болельщик, англия, украина, футбол, певец
4. пройти, мероприятие, россия, акция, фестиваль, москва, фильм, участник, картина, театр, музей
5. фильм, сериал, продукт, актёр, компания, продукция, процент, россия, книга, товар, картина, сезон
6. россия, москва, турист, процент, россиянин, страна, отель, рейс, путешественник, город, тысяча
7. процент, доллар, рубль, нефть, цена, россия, баррель, страна, уровень, вырасти, рынок, рост
8. компания, миллиард_рубль, процент, миллиард_доллар, россия, сумма, миллион_доллар, банк, банка
9. закон, законопроект, документ, реклама, использование, деятельность, поправка, внести, организация
10. россия, страна, керченский_пролив, российский, боинг, работа, чайка, ряд, гражданин, аэропорт
11. партия, кандидат, журналист, праймериза, выбор, единый_россия, госдума, выборы
12. россия, украина, крым, решение, киев, депутат, вопрос, отношение, страна, мнение, право, москва
13. россия, страна, турция, сша, ес, евросоюз, москва, санкция, отношение, украина, вопрос, государство
14. россия, сирия, исламский_государство, сша, нато, иго, запретить, террорист, страна, боевик
15. ракета, путин, россия, запуск, глава_государство, союз, спутник, президент
16. учёный, клетка, исследование, исследователь, ген, университет, оказаться, процент, помощь, организм
17. земля, животное, учёный, животный, тысяча, звезда, планета, обнаружить, кошка, территория, жизнь
18. самолёт, километр, машина, борт, пассажир, вертолёт, погибнуть, лайнер, пилот, час, район, яхта
19. полицейский, полиция, мужчина, задержать, автомобиль, улица, москва, пострадать, life
20. статья, убийство, задержать, суд, отношение, ук_рф, подозревать, следствие, обвинять, трамп, часть
21. ребёнок, женщина, мужчина, летний, дом, сын, семья, мальчик, жена, полиция, дочь, школа, врач
22. видео, youtube, ролик, фото, фотография, канал, снимка, auto, instagram, девушка, страница, группа
23. facebook, пользователь, интернет, страница, twitter, пост, написать, соцсеть, вконтакте, аккаунт
24. устройство, смартфон, компания, мотоциклист, игра, байкер, видео, миллион_доллар, робот, молодая
25. бренд, модель, компания, обувь, основать, одежда, релиз, коллекция, редакция, часы, поступить

Иерархический спектр (коллекция postnauka.ru)

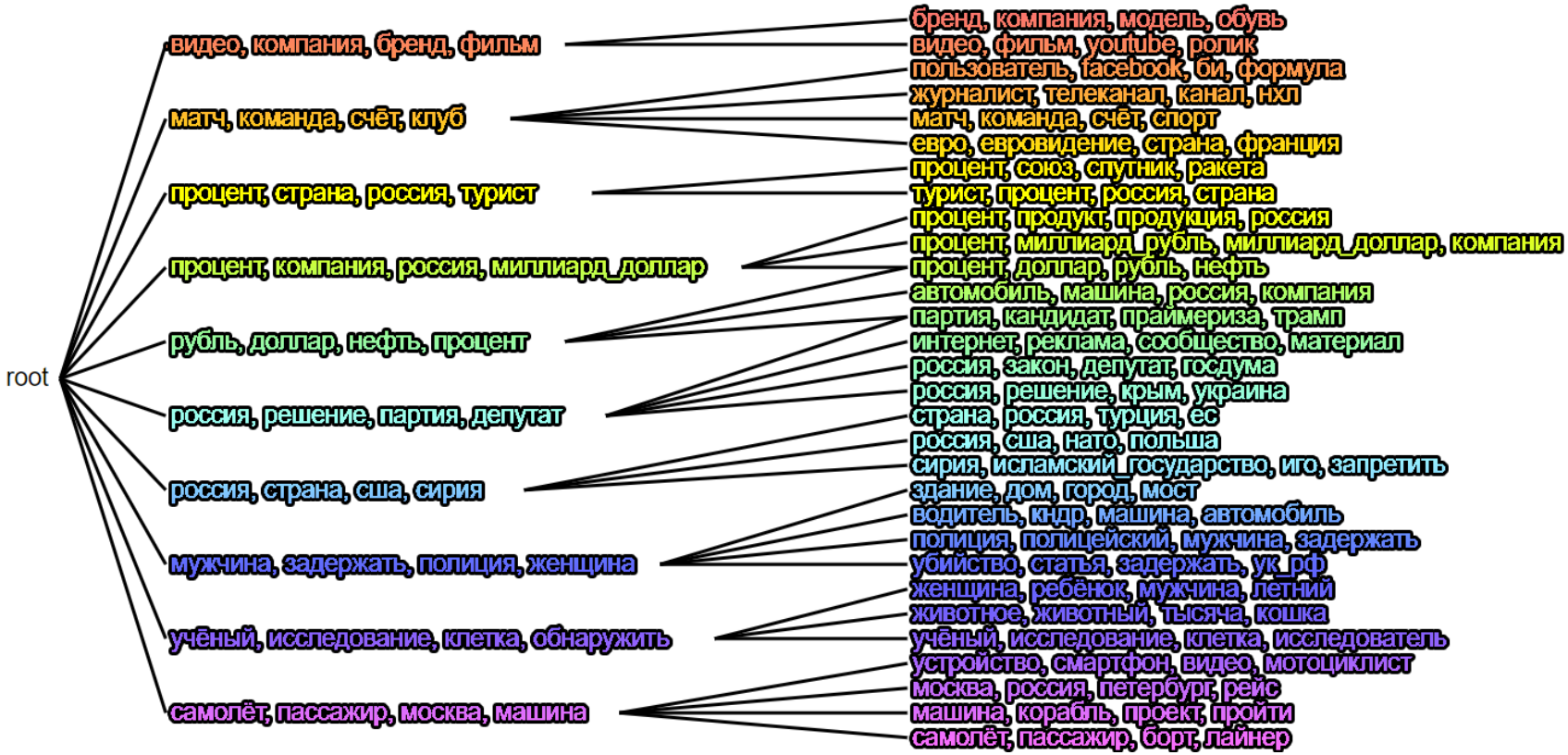


Иерархический спектр (коллекция postnauka.ru)



Иерархический спектр (коллекция lenta.ru)

Иерархический спектр (коллекция lenta.ru)



Тематические эмбедингги: обгоняем word2vec

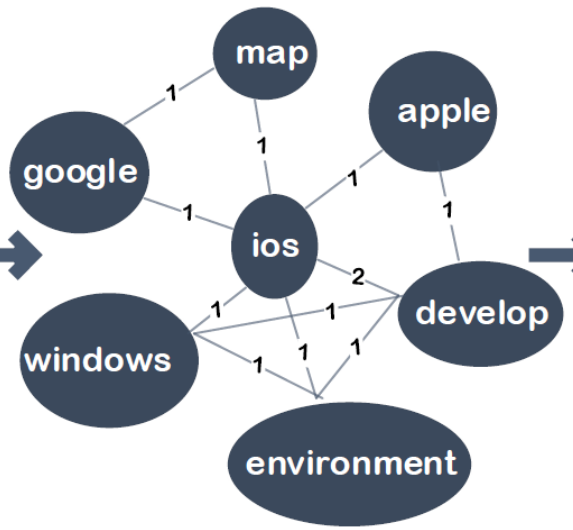
- $p(t|w)$ — тематические векторные представления слов
- Модели сети слов WTM (2009) и WNTM (2014) строятся не по документам, а по совстречаемости слов
- Тематические векторы — разреженные, интерпретируемые
- Задачи семантической близости — почти как word2vec
- Уже реализовано с использованием BigARTM!

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

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 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} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

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.

word2vec и ARTM на задачах 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

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

word2vec и ARTM на задачах семантической близости слов

Дамп Википедии 2016-01-13, $|W| = 100$K, разреженность 93%.

Конкуренты: LDA, SVD-PPMI, SGNS (word2vec).
Варианты ARTM: offline, online, online-with-sparsing.

	WordSim similarity	WordSim relatedness	WordSim joint	Bruni et al. MEN	Radinsky m.turk
LDA	0.530	0.455	0.474	0.583	0.483
SVD-PPMI	0.711	0.648	0.672	0.236	0.616
SGNS	**0.752**	0.632	0.666	**0.745**	**0.661**
ARTM off	0.701	0.615	0.647	0.707	0.613
ARTM on	0.718	**0.673**	**0.685**	0.669	0.639
ARTM on-sp	0.728	0.672	0.680	0.675	0.635

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

Сравнение word2vec и 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

ARTM — есть интерпретируемость:

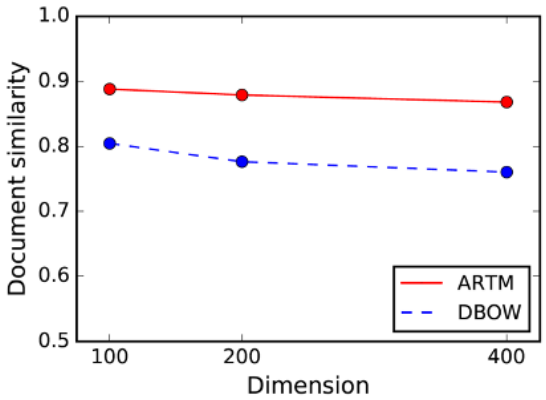
- scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus
- game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash 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.

word2vec и ARTM в задаче семантической близости документов

ArXiv triplets dataset: 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]

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.

Заводские настройки: выкатываем всепогодный ARTM

- **Проблема:** коэффициенты регуляризации приходится подбирать вручную, для каждой задачи заново
- *Относительный коэффициент регуляризации* $\tilde{\tau}_i$ (в %) — это «сила» воздействия регуляризатора на модель
- *Степень индивидуализации* γ_i (от 0 до 1) — чем больше, тем равномернее воздействие на столбцы Φ или Θ
- $\tilde{\tau}_i$ и γ подбираются универсально по многим задачам

Относительные коэффициенты регуляризации

Формула M-шага со взвешенной суммой регуляризаторов R_i:

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \sum_{i=1}^{k} \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right).$$

Суммарное воздействие r_{it} регуляризатора R_i на тему t и *суммарное воздействие* r_i регуляризатора R_i на все темы:

$$r_{it} = \sum_{w \in W} \left| \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right|, \qquad r_i = \sum_{t \in T} r_{it}.$$

Относительный коэффициент регуляризации $\tilde{\tau}_i$:

$$\tau_i = \tilde{\tau}_i \left(\gamma_i \frac{n_t}{r_{it}} + (1 - \gamma_i) \frac{n}{r_i} \right),$$

где γ_i — *индивидуализация* воздействия R_i на темы.

Модель ARTM «из коробки» с универсальными настройками

Рекомендуемый набор регуляризаторов и коэффициентов

	регуляризатор	$\tilde{\tau}_i$
Сглаживание ϕ_t фоновых тем	SmoothBcgPhi	2
Декоррелирование	DecorrelatorPhi	0.05
Сглаживание ϕ	SmoothAllPhi	0
Разреживание Θ	SparseTheta	-0.1
Сглаживание Θ фоновых тем	SmoothBcgTheta	0.1

Рекомендуемое значение *индивидуализации* $\gamma_i = 0.6$

Агрегация: доливаем море документов в модель

- Построили аккуратную модель по небольшой коллекции
- Хотим добавить большую коллекцию,
  1) не разрушив имеющиеся темы,
  2) отбросив нерелевантные документы,
  3) создав новые релевантные темы
- **Пример задачи:** агрегация научно-популярного контента
- Качество связей иерархии — когерентность (совстречаемость) топ-слов дочерней и родительской темы

М.Селезнёва, А.Белый, А.Шолохов. Quality evaluation and improvement for hierarchical topic modeling. Диалог 2018.

Приглашаем к сотрудничеству по проектам

- Создание веб-сервисов тематического поиска
- Разработка иерархических моделей больших коллекций
- Тематизация новостных потоков
- Анализ банковских транзакционных данных
- Тематическая сегментация диалогов
- Доработка VisARTM
- Создание датасетов, демо-моделей, тьюториалов

http://bigartm.org
k.v.vorontsov@phystech.edu

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2017. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016.
- *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.