Оптимизация на единичных симплексах для обучения тематических моделей и нейронных сетей

Воронцов Константин Вячеславович
k.vorontsov@iai.msu.ru

д.ф.-м.н., профессор РАН,
зав. лаб. машинного обучения и семантического анализа
Института искусственного интеллекта МГУ,
зав. каф. математических методов прогнозирования ВМК МГУ,
зав. каф. интеллектуальных систем МФТИ,
г.н.с. ФИЦ «Информатика и управление» РАН

«Data science: mathematical foundations and applications in medicine»
Научный семинар лаборатории «Вероятностные методы в анализе»
факультета математики и компьютерных наук, СПбГУ • 2025-03-20

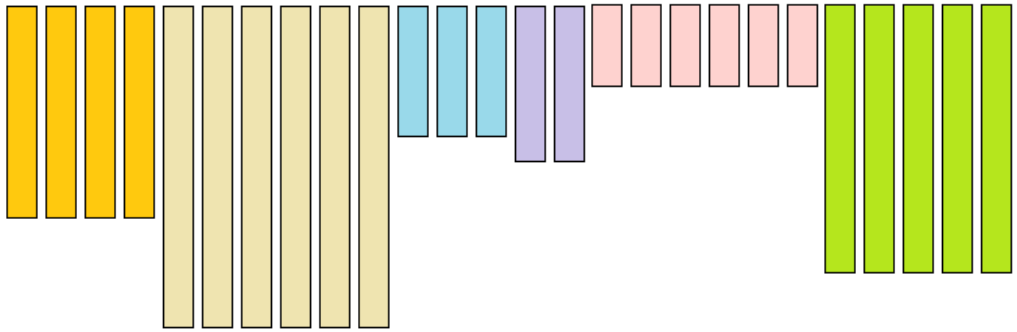
Содержание

1. **Оптимизация на единичных симплексах**
   - Задача максимизации на единичных симплексах
   - Основная лемма
   - Сходимость
2. **Вероятностное тематическое моделирование**
   - Постановка задачи
   - Аддитивная регуляризация тематических моделей
   - Тематические модели внимания
3. **Реализация и приложения**
   - Библиотека BigARTM
   - Примеры приложений
   - Приложения в области медицины и биоинформатики

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω.
Если ω_j — вектор локального экстремума нашей задачи
и $\exists i$: $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Доказательство леммы о максимизации на симплексах

Задача: $f(\Omega) \to \max\limits_{\Omega}$; $\sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

$$\mathscr{L}(\Omega; \mu, \lambda) = -f(\Omega) + \sum_{j \in J} \lambda_j \Big(\sum_{i \in I_j} \omega_{ij} - 1 \Big) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

Условия Каруша–Куна–Таккера для вектора ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij} \omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Умножим обе части первого равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

Согласно условию леммы $\exists i$: $A_{ij} > 0$. Значит, $\lambda_j > 0$.

Если $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Тогда $\omega_{ij} \lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname{norm}\limits_i (A_{ij})$.

■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)
- $\exists \lambda > 0 \;\; f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$ (монотонный рост f)

Тогда $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей. Труды Института математики и механики УрО РАН. 2020.

Открытая проблема: неудобное четвёртое условие

Определение. $H(\Omega^t)$ есть линейное приближение приращения функции f в окрестности точки Ω^t:

$$f(\Omega^{t+1}) - f(\Omega^t) = H(\Omega^t) + o(\Delta\Omega^t)$$

Лемма. Квадратичное представление функции $H(\Omega)$:

$$H(\Omega) = \frac{1}{2}\sum_{j\in J}\sum_{i,k\in I_j}\left(\frac{\partial f(\Omega)}{\partial \omega_{ij}} - \frac{\partial f(\Omega)}{\partial \omega_{kj}}\right)^2 \omega_{ij}\omega_{kj}$$

Следовательно, $H(\Omega^t) \geqslant 0$.

$f(\Omega^{t+1}) - f(\Omega^t) \approx H(\Omega^t)$ — согласно определению;

$f(\Omega^{t+1}) - f(\Omega^t) \geqslant \lambda H(\Omega^t)$, начиная с некоторой итерации t при некотором $\lambda > 0$ — хотелось бы получить это как результат, а не вводить как предположение. Доказать это пока не удалось.

A.M.Ostrowski. Solution of equations and systems of equations. New York, 1966.

Промежуточные итоги и направления исследований

- Метод похож на обычную градиентную оптимизацию, но не требует подбора градиентного шага η
- Ограничения неотрицательности и нормировки могут накладываться не на все векторы, а лишь на некоторые
- Операция norm может приводить к обнулению части координат, следовательно, к разреживанию векторов ω_j
- **Приложения:**
  — вероятностное тематическое моделирование
  — неотрицательные матричные разложения
  — нейронные сети с неотрицательными весами
- **Открытая проблема:** упростить четвёртое условие в теореме сходимости (оно представляется избыточным)
- **Открытая проблема:** оценить скорость сходимости

Пусть

- W — конечное множество *термов* (слов, терминов)
- D — конечное множество текстовых *документов*
- T — конечное множество *тем* (topics)
- каждый терм w в документе d связан с некоторой темой t
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен (bag of docs)
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- *гипотеза условной независимости*: $p(w|d,t) = p(w|t)$

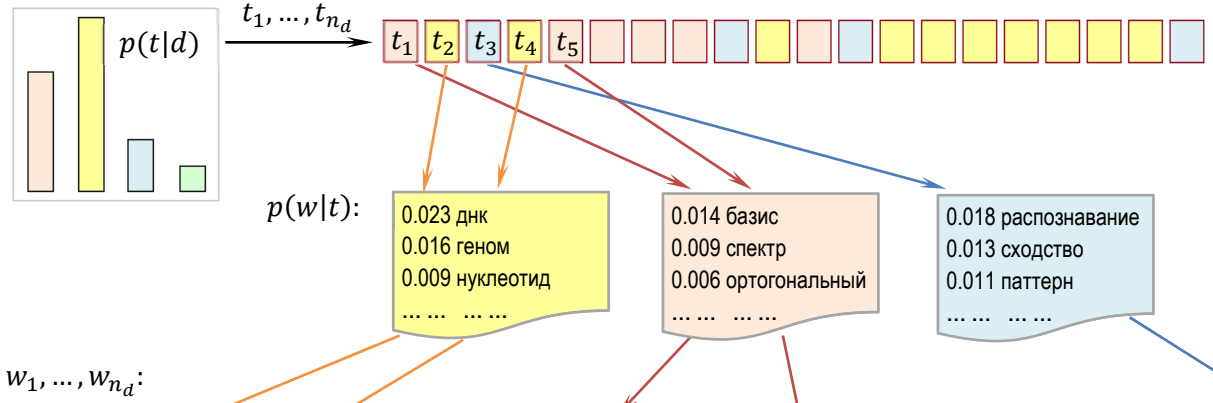
Тематическая модель, по формуле полной вероятности:

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d)$$

Прямая задача: порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление термов w по темам t в документах d:

$$p(w|d) = \sum_t p(w|t)\, p(t|d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача: «о чём все эти тексты?»

Дано: коллекция текстовых документов

- n_{dw} — частота слов (термов) $w \in W$ в документе $d \in D$
- $|T|$ — сколько тем хотим определить в коллекции D

Найти: тематическую языковую модель

- $p(w|d) = \sum_{t \in T} p(w|\cancel{d}, t)\, p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
- $p(w|t) = \phi_{wt}$ — из каких слов w состоит каждая тема $t \in T$
- $p(t|d) = \theta_{td}$ — из каких тем t состоит каждый документ d

Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

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

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Критерий максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \underbrace{p(d)}_{\text{const}} \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

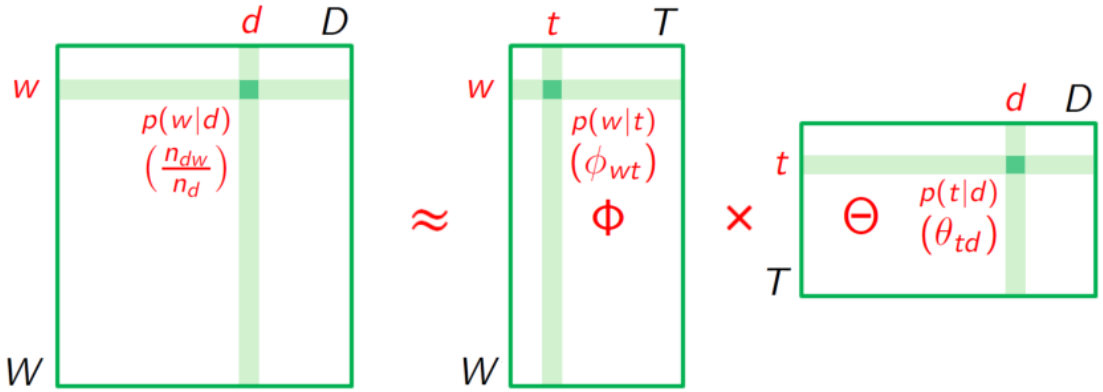
при ограничениях неотрицательности и нормировки столбцов (такие матрицы Φ, Θ называются *стохастическими*)

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Три трактовки задачи тематического моделирования

1. **Мягкая кластеризация** документов по кластерам-темам

2. Низкоранговое стохастическое **матричное разложение:**



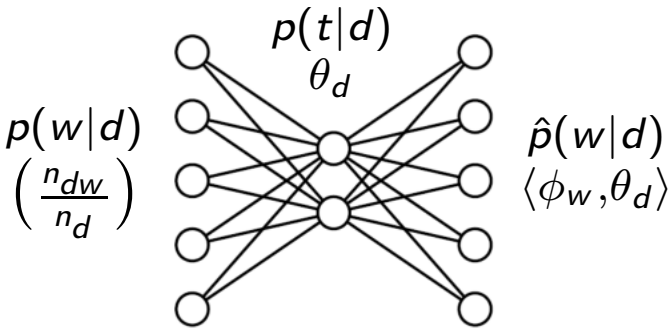
3. **Автокодировщик** документов в тематические эмбединги:

— кодировщик $f_\Phi : \frac{n_{dw}}{n_d} \to \theta_d$

— декодировщик $g_\Phi : \theta_d \to \Phi\theta_d$

задача реконструкции:

$$\sum_d n_d \,\mathrm{KL}\Big(\frac{n_{dw}}{n_d} \,\Big\|\, \langle \phi_w, \theta_d \rangle\Big) \to \min_{\Phi,\Theta}$$



Свойство интерпретируемости тематических моделей

Тематическая модель формирует тематические векторы:

- $p(t|d) = \theta_{td}$ для каждого документа d
- $p(t|w) = \dfrac{p(w|t)p(t)}{p(w)} = \phi_{wt}\dfrac{n_t}{n_w}$ для каждого терма w
- $p(t|d,w)$ для каждого локального контекста (d,w)

Интерпретируемость тематических векторов (эмбедингов):

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\left\{w\colon p(w|t) > \gamma p(w)\right\}$
- тема может «рассказать о себе» словами или фразами
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\left\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\right\}$

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их частоты $p(w|t)$ 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 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их частоты $p(w|t)$ 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.

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
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.

Цели и не-цели тематического моделирования

Цели:

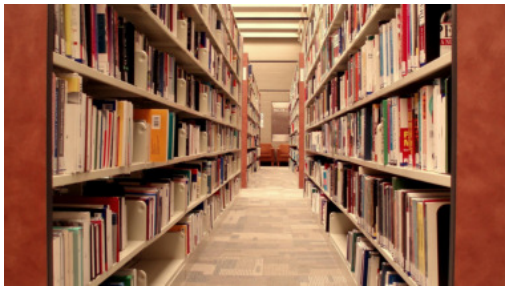
- Выявлять кластерную тематическую структуру текстовой коллекции, сколько в ней тем и о чём они
- Получать *интерпретируемые* тематические векторные представления (эмбединги) слов $p(t|w)$, $p(t|d,w)$, документов $p(t|d)$, фрагментов $p(t|s)$, объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать слова по контексту (TM слабы как модели языка)
- Понимать смысл текста
- Генерировать связный текст

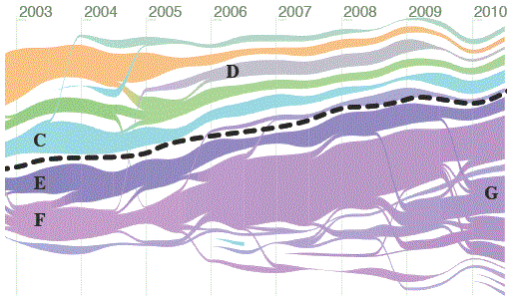
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

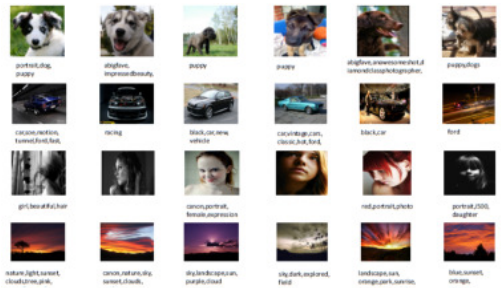


поиск тематических сообществ в соцсетях

выявление и отслеживание цепочек новостей

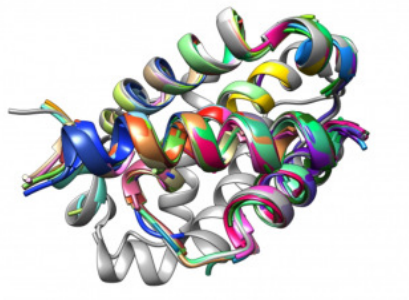


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

поиск паттернов в задачах биоинформатики



J.Boyd-Graber, Yuening Hu, D.Mimno. Applications of Topic Models. 2017.

H.Jelodar et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. 2019.

Модель PLSA (Probabilistic Latent Semantic Analysis)

Максимизация log-правдоподобия для стохастических матриц:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

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

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_d n_{dw} p_{tdw}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_w n_{dw} p_{tdw}\right) \end{cases}$$

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), \quad \text{rank} S = |T|$
- $L(\Phi', \Theta') \approx L(\Phi, \Theta)$

Регуляризация или стабилизация — доопределение решения добавлением второго оптимизационного критерия.

А.Н.Тихонов
(1906–1993)

Модель LDA (Latent Dirichlet Allocation)

Максимизация log-правдоподобия + байесовская регуляризация с априорными распределениями Дирихле на столбцы Φ, Θ:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td} \to \max_{\Phi,\Theta}$$

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

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

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

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. NIPS, 2001; JMLR, 2003.

Байесовская и классическая регуляризация

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (громоздкий, приближённый) ради точечной оценки Ω:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega} \big(\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma)\big)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Аддитивная Регуляризация Тематических Моделей (ARTM)

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 со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

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

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

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

Доказательство (по лемме о максимизации на ед.симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

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

Дифференцируя, выделим вспомогательную переменную p_{tdw}:

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

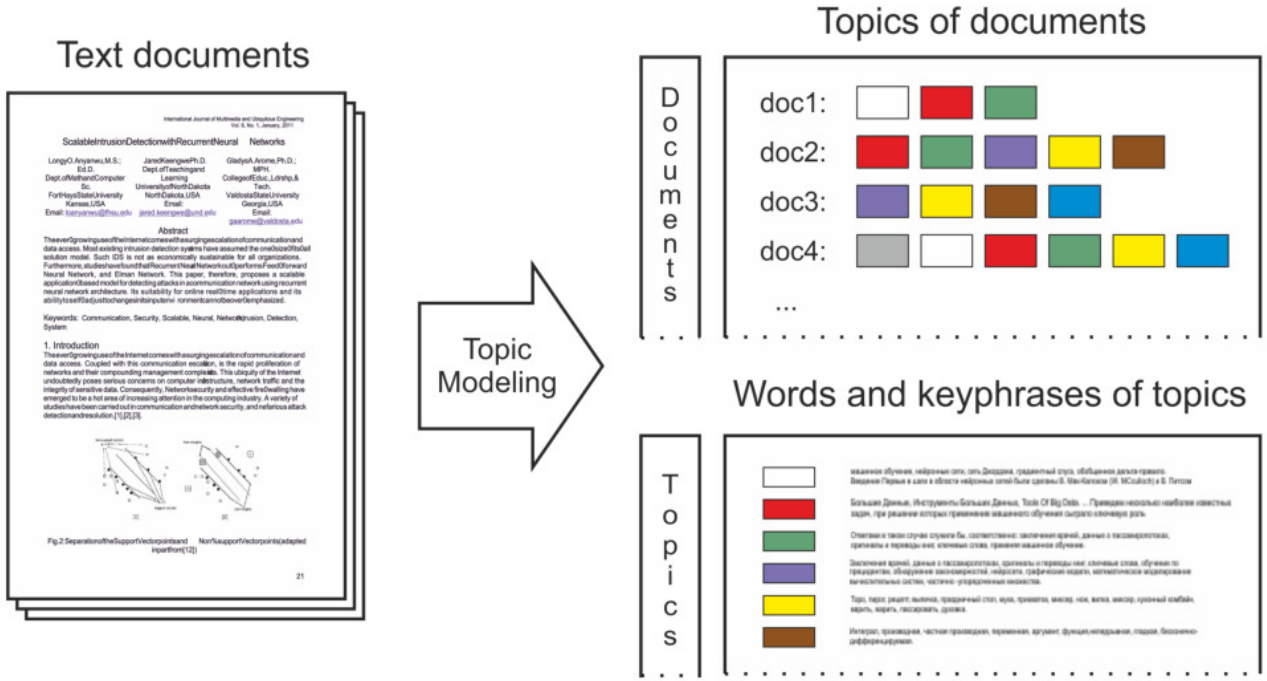
$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

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

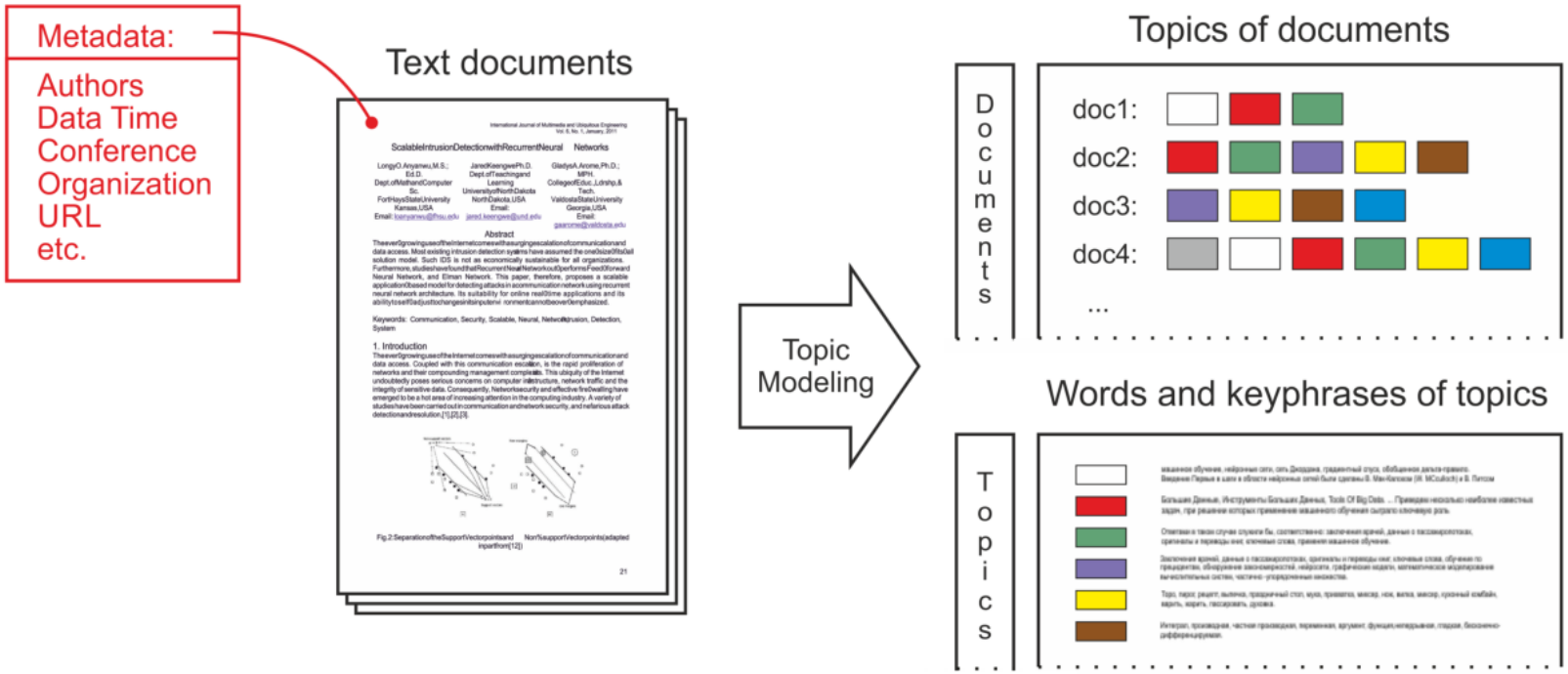
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$,



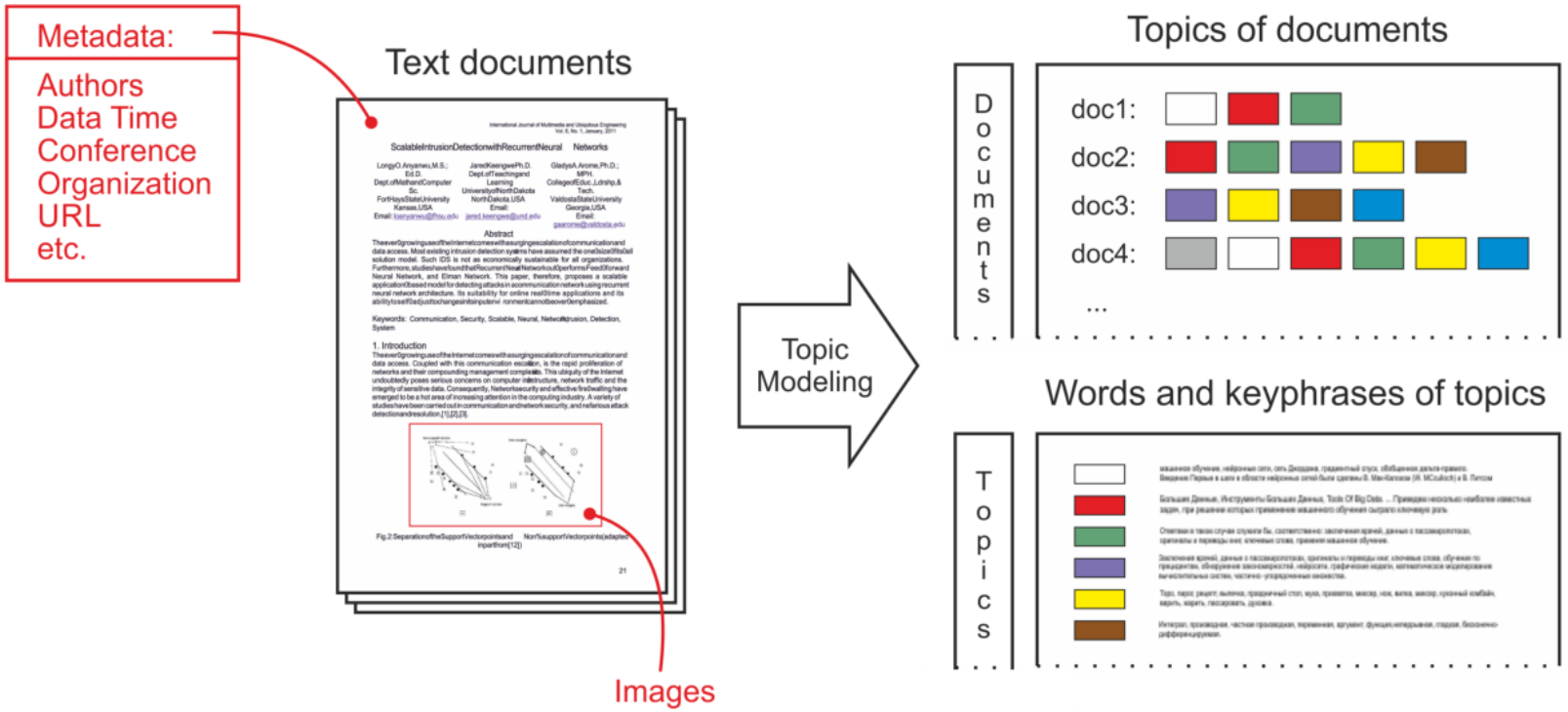
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,



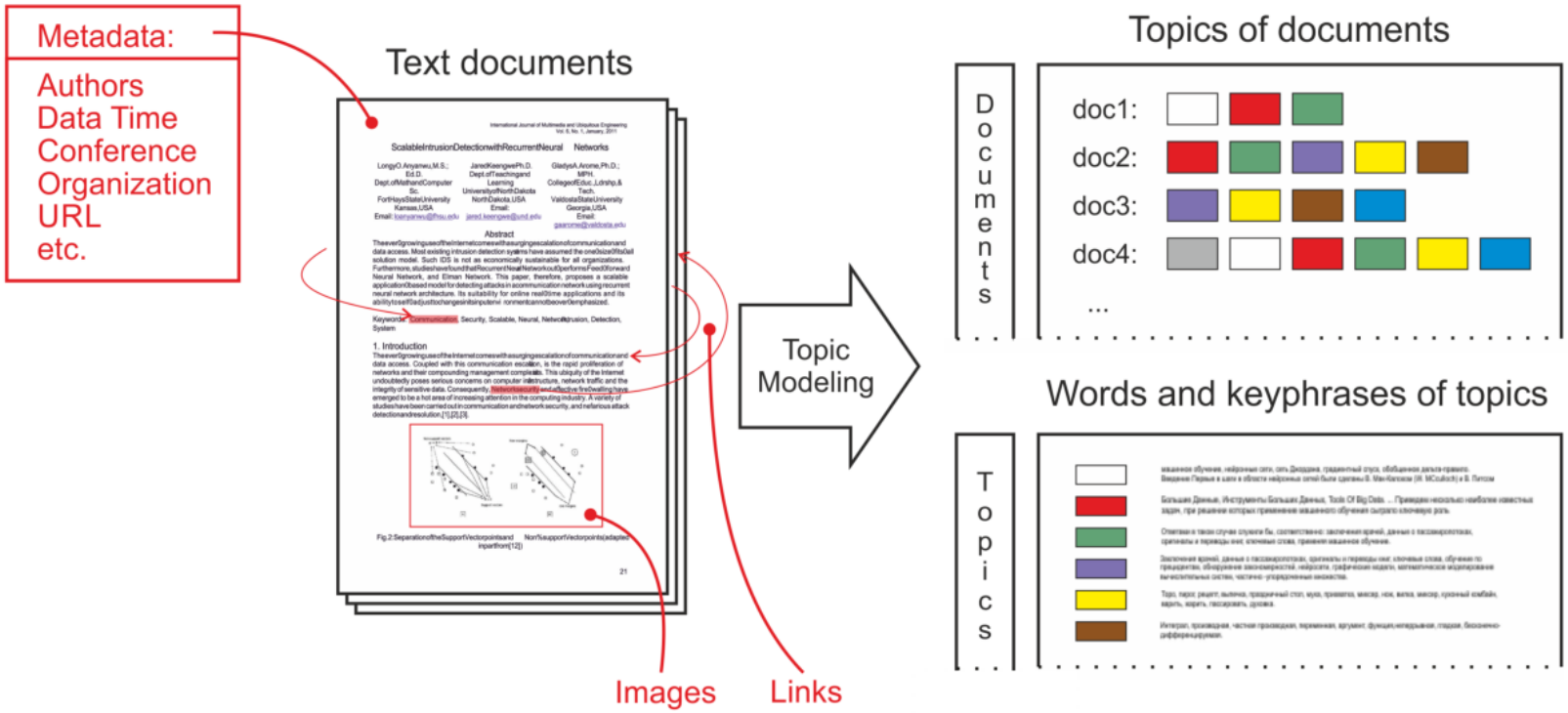
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$,



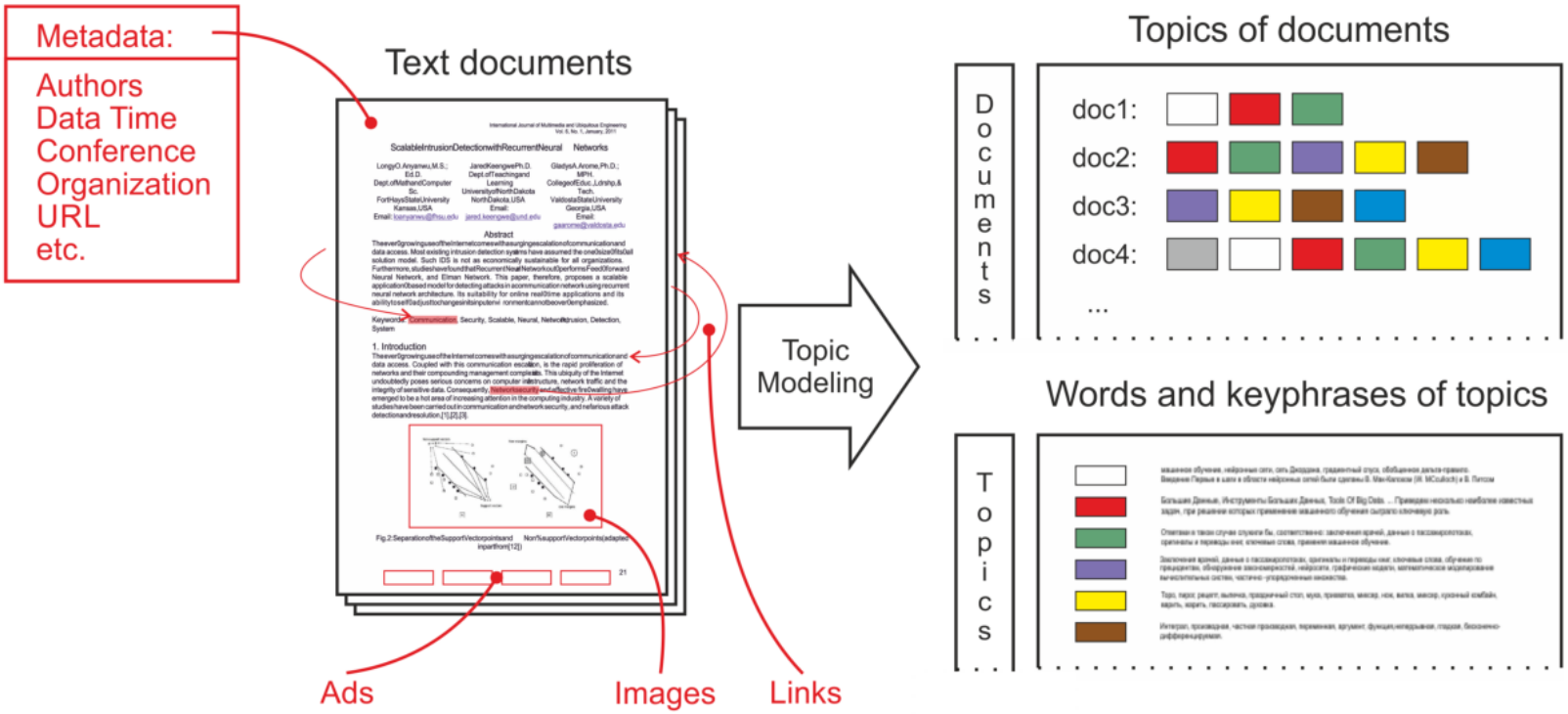
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$, $p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



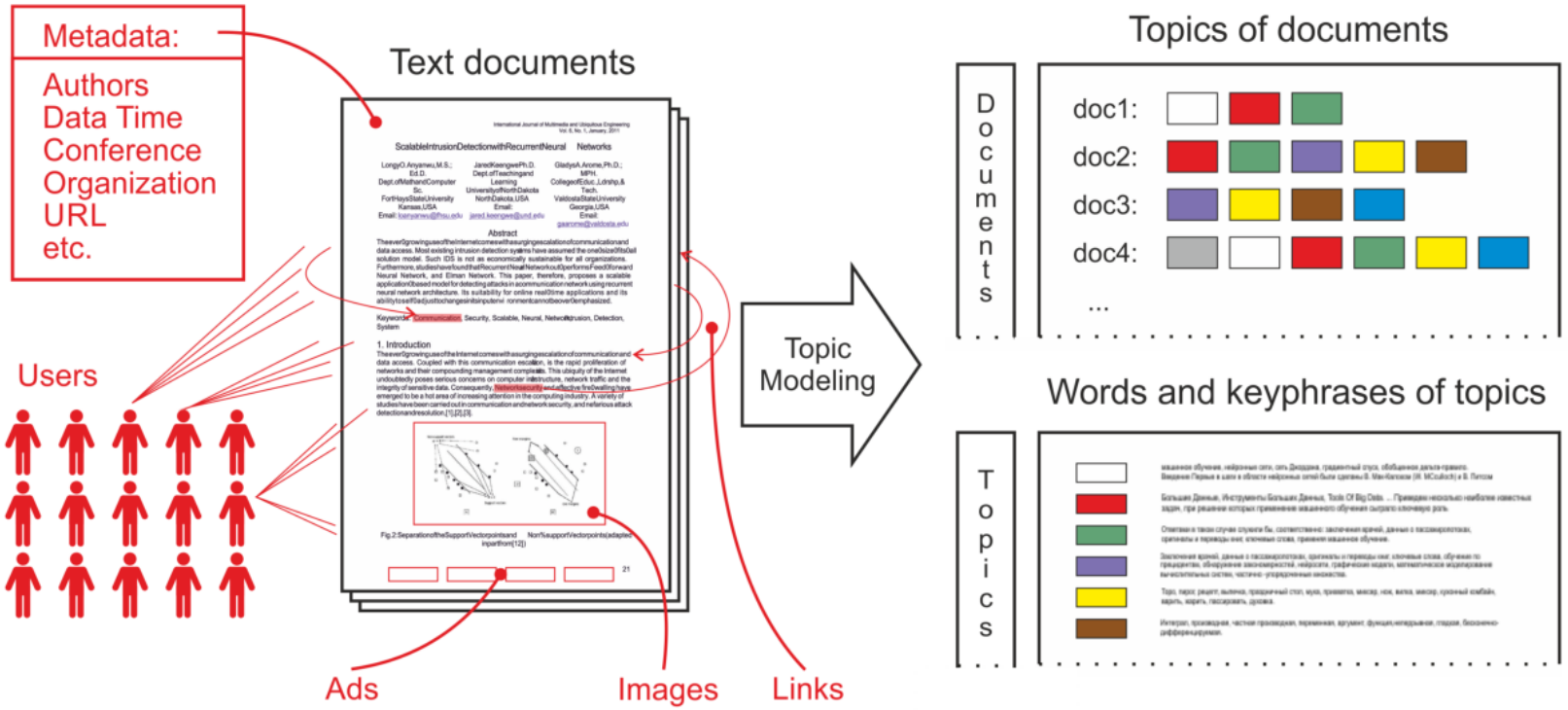
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



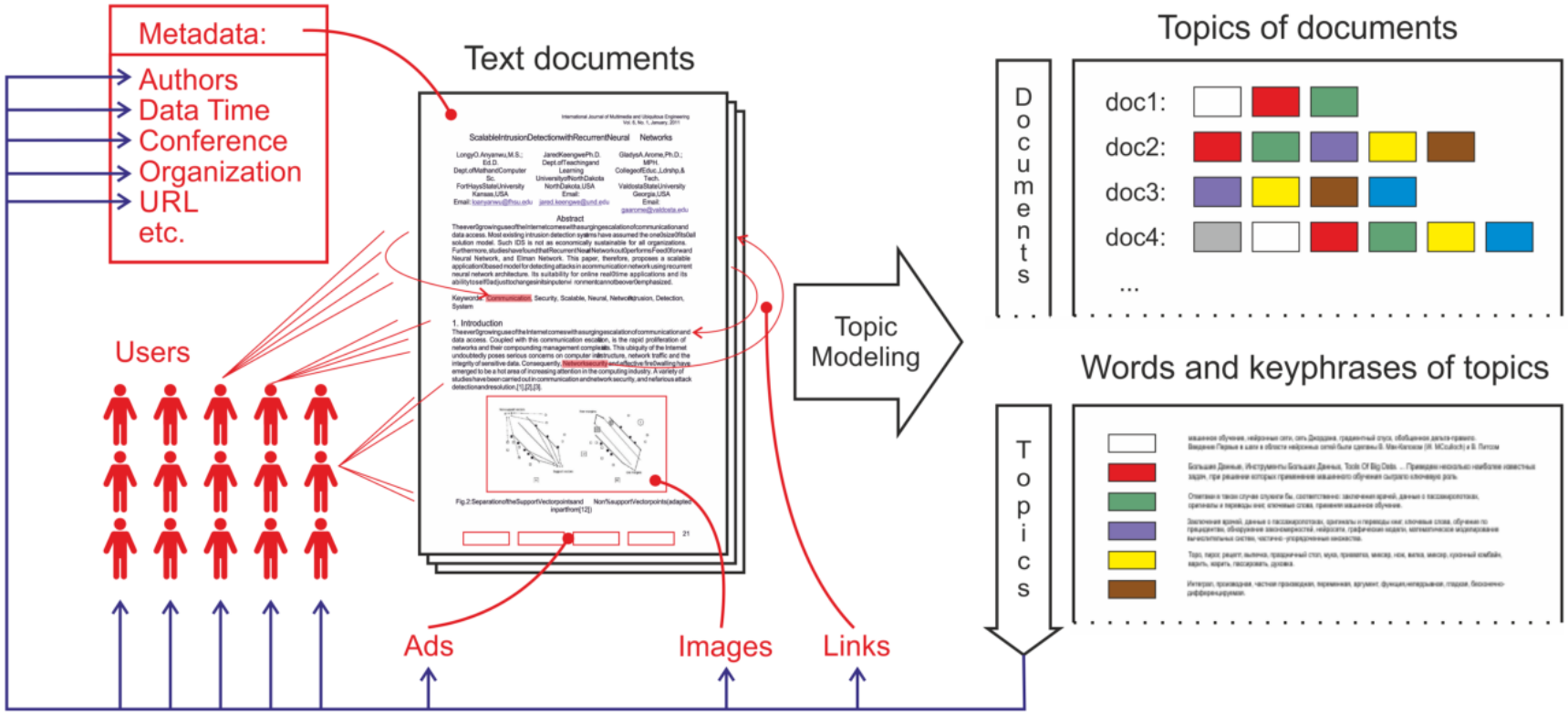
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная ARTM

W_m — словарь термов 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

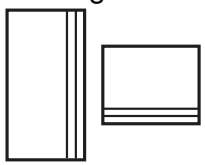
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}} \left(\phi_{wt} \theta_{td} \right)$$

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

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

Регуляризаторы для улучшения интерпретируемости тем

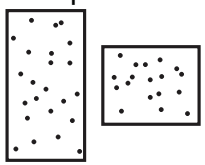
background



Сглаживание фоновых тем $B \subset T$:

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

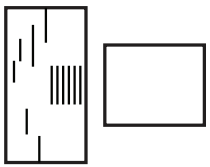
sparse



Разреживание предметных тем $S = T \setminus B$:

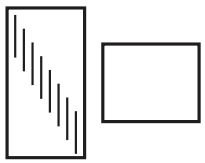
$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in S}\sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d}\sum_{t\in S} \alpha_t \ln \theta_{td}$$

seed words



Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов

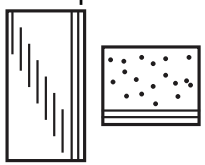
decorrelated



Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s}\sum_{w} \phi_{wt}\phi_{ws}$$

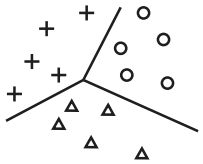
interpretable



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

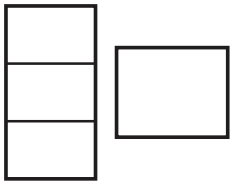
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов или категорий для задач классификации и категоризации текстов.

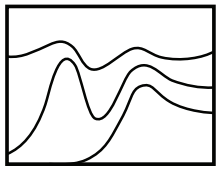
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

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

temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \left|\phi_{it} - \phi_{i-1,t}\right|.$$

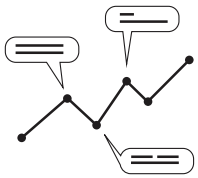
geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

Регуляризаторы для учёта дополнительной информации

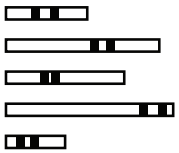
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

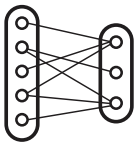
biterm



Связи сочетаемости слов (n_{uv} — частота битерма):

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

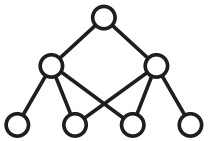
relational



Связи или ссылки между документами:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

hierarchy

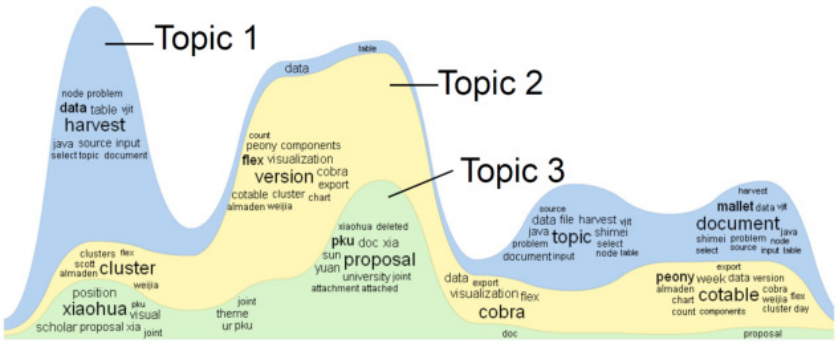


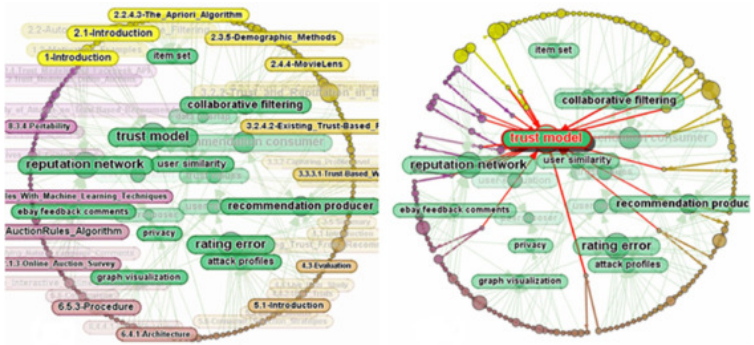
Связи родительских тем t с дочерними подтемами s:

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

Мотивации. Что хотим:

- вместо «мешка слов» — последовательность $w_1, \ldots, w_n$
- вместо документов — локальные контексты слов
- определять тематику любого фрагмента текста
- быстро находить фрагменты, относящиеся к данной теме
- в том числе фразы для суммаризации документа или темы
- разделять документ на тематически однородные сегменты
- визуализировать тематическую структуру документа





Идея тематизации текста за один проход

Дано: s — фрагмент текста d, Φ — тематическая модель
Найти: $p(t|s)$ — тематический вектор фрагмента текста

Проблемы:

- как не переобучить вектор $p(t|s)$, если текст короткий?
- как согласовать $p(t|s)$ с объемлющим контекстом $p(t|d)$?
- как согласовать $p(t|s)$ с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ термов $w \in s$?

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in W} n_{dw} p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}^0\big)$$

- формула полной вероятности + гипотеза усл. независ.:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\, p(t|w,\cancel{d}) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt} p_t\big)$$

EM-алгоритм для ARTM с явным выражением Θ через Φ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big); \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

$$p'_{tdw} = p_{tdw} + \frac{1}{n_{dw}} \sum_{s \in T} \frac{n_{sd}}{\theta_{sd}} \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}}$$

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

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по Лемме о максимизации на симплексах)

Оптимизационная задача M-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du}p_{sdu}\ln\big(\phi_{us}\theta_{sd}(\Phi)\big) + R\big(\Phi,\Theta(\Phi)\big) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,s,u} n_{du}p_{sdu}\frac{\phi_{wt}}{\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial\theta_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{\phi_{wt}}{\theta_{sd}}\underbrace{\left(\sum_{u\in d} n_{du}p_{sdu} + \theta_{sd}\frac{\partial R}{\partial\theta_{sd}}\right)}_{n_{sd}}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right) + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\underbrace{\left(p_{tdw} + \frac{1}{n_{dw}}\sum_{s\in T}\frac{n_{sd}}{\theta_{sd}}\phi_{wt}\frac{\partial\theta_{sd}}{\partial\phi_{wt}}\right)}_{p'_{tdw}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}.$$

■

EM-алгоритм для ARTM с линейной тематизацией документов

$$\theta_{td}(\Phi) = \sum_{w \in d} \frac{n_{dw}}{n_d} \underset{t \in T}{\operatorname{norm}}(\phi_{wt} p_t) \;\Rightarrow\; \phi_{wt} \frac{\partial \theta_{sd}}{\partial \phi_{wt}} = \frac{n_{dw}}{n_d} \phi'_{tw} \left(\delta_{st} - \phi'_{sw}\right)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w \in d} \frac{n_{dw}}{n_d} \phi'_{tw}$$

$$p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw}$$

$$n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

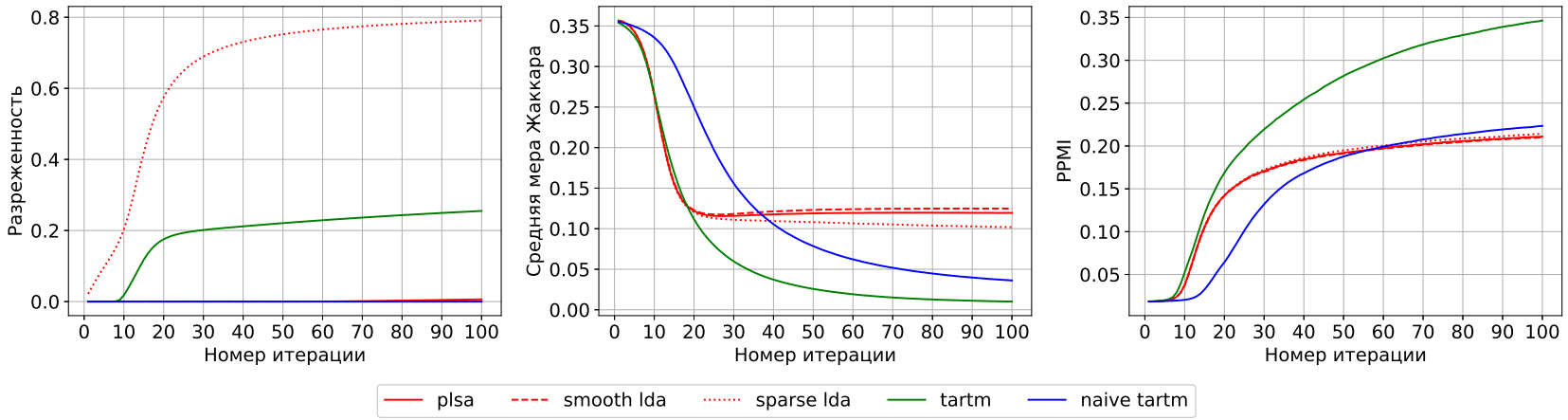
$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{s \in T} \phi'_{sw} \frac{n_{sd}}{\theta_{sd}} \right)$$

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

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма



- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

https://github.com/ilirhin/python_artm

Упрощение EM-алгоритма для линейной тематизации

- Нет регуляризации по Θ, следовательно, $\frac{\partial R}{\partial \theta_{td}} = 0$
- Значение отношения $\frac{n_{td}}{\theta_{td}} \approx n_d$ не зависит от t, подстановка в формулу M-шага приводит к упрощению: $p'_{tdw} = p_{tdw}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} n_t\right); \qquad \theta_{td} = \sum_{w \in d} \frac{n_{dw}}{n_d} \phi'_{tw};$$

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\left(\phi_{wt} \theta_{td}\right); \qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dw} p_{tdw};$$

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

Это обычный EM-алгоритм, только с однопроходным E-шагом!
ОГО! И ТАК МОЖНО БЫЛО?!

Линейная тематизация: от документа к локальным контекстам

Тематизация документа $d = (w_1, \ldots, w_{n_d})$ за один проход:

$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i) = \frac{1}{n_d} \sum_{i=1}^{n_d} \phi'_{tw_i}$$

Тематизация *локального контекста* $C_i = (\ldots, w_i, \ldots)$ терма w_i:

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \frac{1}{|C_i|} \sum_{u \in C_i} p(t|u) = \frac{1}{|C_i|} \sum_{u \in C_i} \phi'_{tu}$$

Тематизация локального контекста с распределением весов:

$$\theta_{ti}(\Phi) \equiv p(t|C_i) = \sum_{u \in C_i} \phi'_{tu} \alpha(u|i), \quad \sum_{u \in C_i} \alpha(u|i) = 1, \quad \alpha(u|i) \geqslant 0$$

Локализованная тематическая модель:

$$p(w|C_i) = \sum_{t \in T} p(w|t)\, p(t|C_i) = \sum_{t \in T} \phi_{wt} \sum_{u \in C_i} \phi'_{tu} \alpha(u|i)$$

EM-алгоритм с локализованным E-шагом

$w_1, \ldots, w_n$ — сквозная нумерация термов во всей коллекции
C_i — локальный контекст (окружение) терма w_i
$\alpha(u|i)$ — распределение важности термов $u \in C_i$ для терма w_i

- не нужна гипотеза «мешка слов»
- не нужно разбиение коллекции на документы

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} \equiv p(t|w) = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt} p_t\big); \qquad \theta_{ti} \equiv p(t|C_i) = \sum_{u\in C_i} \phi'_{tu}\alpha(u|i);$$

$$p_{ti} \equiv p(t|C_i, w_i) = \underset{t\in T}{\operatorname{norm}}\big(\phi_{w_i t}\theta_{ti}\big); \qquad p_t \equiv p(t) = \frac{1}{n}\sum_{i=1}^{n} p_{ti};$$

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{i=1}^{n}[w_i = w]p_{ti} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right).$$

Быстрое вычисление двунаправленных векторов контекста

Два прохода по тексту — «слева направо» и «справа налево» для вычисления *экспоненциальных скользящих средних* (ЭСС):

$$\vec{p}(t|i) = \vec{\gamma}_i\, p(t|w_i) + (1-\vec{\gamma}_i)\, \vec{p}(t|i-1),\ \ i = 1,\ldots,n,\ \ \vec{\gamma}_1 = 1$$
$$\overleftarrow{p}(t|i) = \overleftarrow{\gamma}_i\, p(t|w_i) + (1-\overleftarrow{\gamma}_i)\, \overleftarrow{p}(t|i+1),\ \ i = n,\ldots,1,\ \ \overleftarrow{\gamma}_n = 1$$

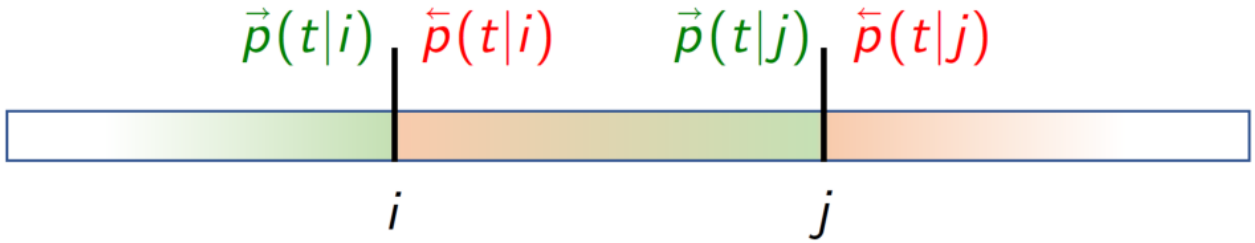
где $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i

Основное свойство: если $\gamma_i = \gamma$, то $\alpha(w_k|i) = \gamma(1-\gamma)^{|i-k|}$

Несколько соображений, как распоряжаться выбором $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм
- γ_i можно умножать на оценку важности терма

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{p}(t|i)$ — тематика левого контекста терма w_i
- $\overleftarrow{p}(t|i)$ — тематика правого контекста терма w_i
- $\frac{1}{2}\big(\vec{p}(t|i) + \overleftarrow{p}(t|i)\big)$ — тематика двустороннего контекста w_i
- $p(t|i\dots j) = \frac{1}{2}\big(\overleftarrow{p}(t|i) + \vec{p}(t|j)\big)$ — тематика сегмента $[i\dots j]$
- $\overleftarrow{p}(t|i) \approx \vec{p}(t|j)$ — однородность тематики сегмента $[i\dots j]$
- $\max_i \big\|\vec{p}(t|i) - \overleftarrow{p}(t|i)\big\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты

Гипотеза: есть аналогия с моделью внимания и трансформером

Онлайновый EM-алгоритм с локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры β, $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, α, δ;
Выход: матрица Φ, векторы термов документов p_{ti};

инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} := \text{random}$;
для всех документов $d \in D$

$p_{ti} := \text{norm}_t(\phi_{w_i t} n_t)$, $i = 1, \ldots, n_d$, $t \in T$;
$\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1 - \vec{\gamma}_i)\, \vec{\theta}_{t,i-1}$, $i = 1, \ldots, n_d$, $\vec{\gamma}_1 = 1$, $t \in T$;
$\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1 - \overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}$, $i = n_d, \ldots, 1$, $\overleftarrow{\gamma}_{n_d} = 1$, $t \in T$;
$p_{ti} := \text{norm}_t\big(\phi_{w_i t}\big(\beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\big)\big)$, $i = 1, \ldots, n_d$, $t \in T$;
$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}$; $n_t := n_t + p_{ti}$, $i = 1, \ldots, n_d$, $t \in T$;
если пора обновить матрицу Φ **то**

$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$;
$\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right)$;

Модель внимания Query–Key–Value

q — *вектор-запрос*, трансформируемый в контекстный вектор z.
Контекст задаётся последовательностью n пар ключ-значение:
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*,
$V = (v_1, \ldots, v_n)$ — *векторы-значения*.

Модель внимания — это выпуклая комбинация векторов v_i, взвешенных по сходству их ключей k_i с запросом q:

$$z = \text{Attn}(q, K, V) = \sum_{i=1}^{n} v_i \operatorname*{SoftMax}_i \langle k_i, q \rangle$$

Модель само-внимания (self-attention) трансформирует
$X = (x_1, \ldots, x_n)$ — входные *бесконтекстные векторы* в
$Z = (z_1, \ldots, z_n)$ — выходные *контекстные векторы*:

$$z_i = \text{Attn}(W_q x_i, W_k X, W_v X),$$

где W_q, W_k, W_v — обучаемые матрицы параметров.

Vaswani et al. Attention is all you need. 2017.

BERT — Bidirectional Encoder Representations from Transformers

Трансформер BERT — двунаправленный кодировщик текста, предобучаемый для решения различных задач NLP

Схема преобразования данных:

- $S = (w_1, \ldots, w_n)$ — токены входного текста
  $\downarrow$ обучение векторов (эмбедингов) токенов
- $X = (x_1, \ldots, x_n)$ — бесконтекстные векторы токенов
  $\downarrow$ многократная трансформация через само-внимание
- $Z = (z_1, \ldots, z_n)$ — контекстные векторы токенов
  $\downarrow$ дообучение на конкретную задачу
- Y — разметка текста / классификация и т.п.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language) BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Dichao Hu. An introductory survey on attention mechanisms in NLP problems. 2018.

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \ldots, h_n)$
$d = \dim x_i, p_i, h_i = 512$
$\dim H = 512 \times n$

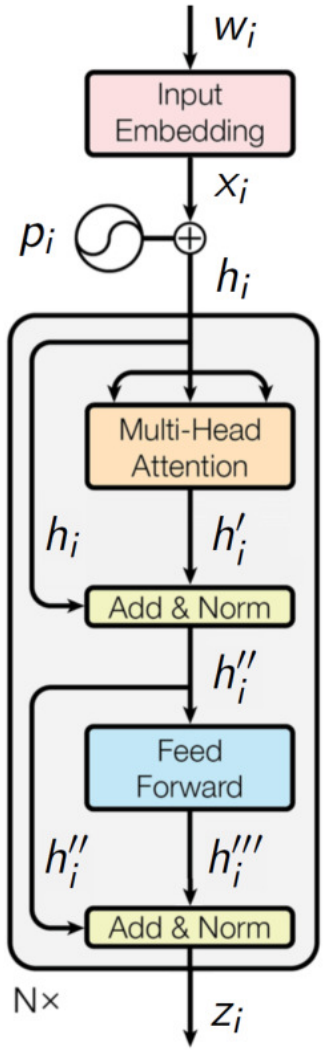
2. Многомерное само-внимание:
$h_i^j = \mathrm{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$
$j = 1, \ldots, J = 8$
$\dim h_i^j = 64$
$\dim W_q^j, W_k^j, W_v^j = 64 \times 512$

3. Конкатенация (multi-head attention):
$h'_i = \mathrm{MH}_J(h_i^j) \equiv \left[h_i^1 \cdots h_i^J\right]$
$\dim h'_i = 512$

4. Сквозная связь + нормировка уровня:
$h''_i = \mathrm{LN}(h'_i + h_i; \mu_1, \sigma_1)$
$\dim h''_i, \mu_1, \sigma_1 = 512$

5. Полносвязная 2х-слойная сеть FFN:
$h'''_i = W_2\,\mathrm{ReLU}(W_1 h''_i + b_1) + b_2$
$\dim W_1 = 2048 \times 512$
$\dim W_2 = 512 \times 2048$

6. Сквозная связь + нормировка уровня:
$z_i = \mathrm{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$
$\dim z_i, \mu_2, \sigma_2 = 512$



Критерий обучения MLM (Masked Language Modeling)

Критерий маскированного языкового моделирования MLM, строится автоматически по текстам (self-supervised learning):

$$\sum_S \sum_{i \in M(S)} \ln p(w_i | i, S, W) \to \max_W,$$

где $M(S)$ — подмножество (15%) маскированных токенов из S,

$$p(w|i, S, W) = \underset{w}{\text{SoftMax}}(W_z z_i(S, W_T) + b_z)$$

— языковая модель, предсказывающая i-й токен в тексте S;
$z_i(S, W_T)$ — контекстный вектор i-го токена текста S на выходе Трансформера с параметрами W_T;
$W = (W_T, W_z, b_z)$ — все параметры языковой модели

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Аналогия локализованного E-шага с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\big(\phi_{w_i t}\theta_{ti}\big) = \underset{t\in T}{\text{norm}}\Big(\sum_{u\in C_i}\phi'_{tu}\phi_{w_i t}\alpha(u|i)\Big)$$

Контекстный вектор на выходе модели само-внимания:

$$z_i = \sum_{u\in C_i} W_v x_u\,\alpha(u|i) = \sum_{u\in C_i} W_v x_u \underset{u\in C_i}{\text{SoftMax}}\langle W_q x_i, W_k x_u\rangle$$

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ'_u из контекста терма w_i,
- наиболее (семантически) схожих с вектором терма w_i.

Отличия:

- адамарово умножение вектора ϕ'_u на вектор-фильтр ϕ_{w_i};
- нет обучаемых матриц W_q, W_k, W_v как у модели внимания;
- проецирование итогового вектора на единичный симплекс.

Аналогия локализованного E-шага с моделью трансформера

Один проход документа аналогичен модели внимания:

— для каждого $d \in D$, для каждой позиции $i = 1, \ldots, n_d$ вычисляются 5 тематических векторов, связанных с термом w_i:

$\phi'_{tw_i} = \text{norm}_t(\phi_{w_i t} p_t)$ — бесконтекстный вектор терма $p(t|w_i)$

$\vec{p}(t|i) = \vec{\theta}_{ti}$, $\overleftarrow{p}(t|i) = \overleftarrow{\theta}_{ti}$ — векторы левого и правого контекста

$\theta_{ti} = \beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}$ — вектор двустороннего контекста

$p_{ti} = \text{norm}_t\left(\phi_{w_i t}\theta_{ti}\right)$ — контекстный вектор терма $p(t|C_i, w_i)$

Несколько таких проходов аналогичны трансформеру:

контекстный вектор терма $p_{ti} = p(t|C_i, w_i)$ с предыдущего прохода используется вместо его бесконтекстного вектора $\phi'_{tw_i} = p(t|w_i)$

L таких итераций аналогичны проходу L блоков внимания

Онлайновый EM с многопроходным локализованным E-шагом

Вход: коллекция, число тем $|T|$, параметры L, β, $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, α, δ;
Выход: матрица Φ, векторы термов документов p_{ti};
инициализация: $n_{wt} := 0$; $\tilde{n}_{wt} := 0$; $n_t := 1$; $\phi_{wt} :=$ random;
для всех документов $d \in D$

$p_{ti} := \text{norm}_t(\phi_{wt} n_t)$;
для всех $l = 1, \ldots, L$ (аналог L блоков внимания)

$\vec{\theta}_{ti} := \vec{\gamma}_i\, p_{ti} + (1-\vec{\gamma}_i)\, \vec{\theta}_{t,i-1}, \quad i = 1, \ldots, n_d, \quad \vec{\gamma}_1 = 1;$

$\overleftarrow{\theta}_{ti} := \overleftarrow{\gamma}_i\, p_{ti} + (1-\overleftarrow{\gamma}_i)\, \overleftarrow{\theta}_{t,i+1}, \quad i = n_d, \ldots, 1, \quad \overleftarrow{\gamma}_{n_d} = 1;$

$p_{ti} := \text{norm}_t\big(\big(\beta \vec{\theta}_{ti} + (1-\beta)\overleftarrow{\theta}_{ti}\big) p_{ti}/n_t\big)$;

$\tilde{n}_{w_i t} := \tilde{n}_{w_i t} + p_{ti}$; $n_t := n_t + p_{ti}$;
если пора обновить матрицу Φ **то**

$n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$; $\tilde{n}_{wt} := 0$;

$\phi_{wt} := \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\right)$;

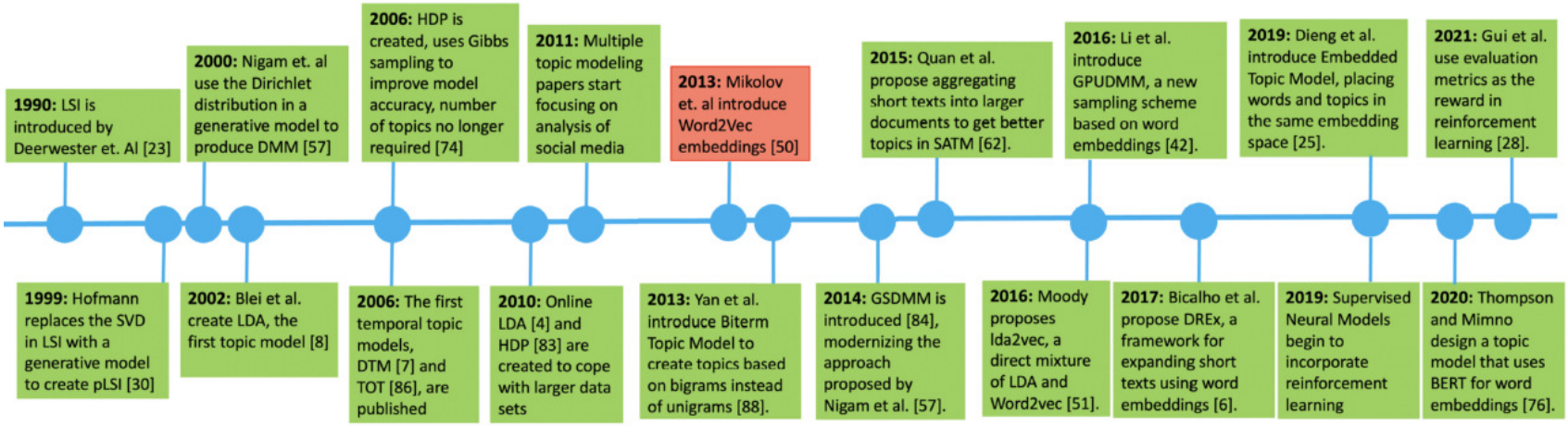
Открытые проблемы и постановки задач

- надо ли исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$?
- какие другие варианты $\alpha(u|i)$ кроме скользящих средних?
- как найти баланс β левого и правого контекста?
- правильно ли подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге?
- имеет ли смысл увеличивать число проходов L?

- как (и нужно ли) параметризовать модель внимания?
- как обучать её параметры, разные для разных проходов?
- как (и нужно ли) ввести аналог многих голов внимания?

- слишком много эвристических преобразований сделано... мы всё ещё решаем исходную оптимизационную задачу?
- действительно ли на E-шаге можно подвергать $p(t|d, w_i)$ всяким модификациям, почему и в каких пределах?

Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Модульный подход к синтезу моделей с заданными свойствами

Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	**Bayesian TM**	**ARTM**	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

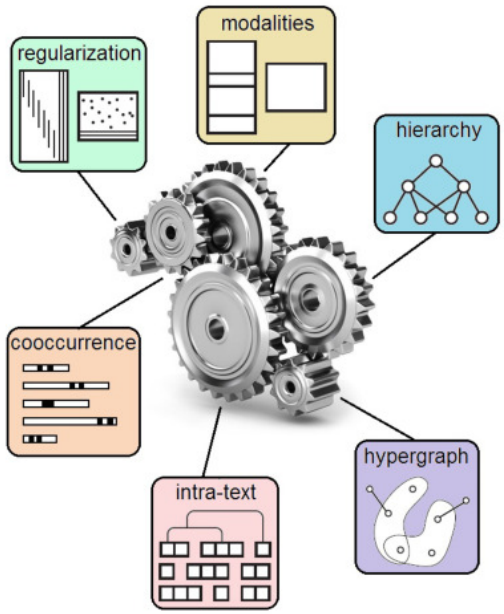
Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация результатов тематического моделирования

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

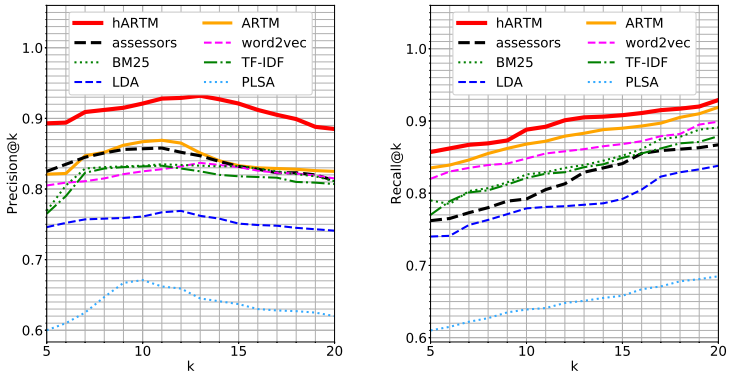
3.7M статей Википедии, 100K слов: время min (перплексия)

проц.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

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

Разведочный поиск в технологических блогах

Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175K документов),
— TechCrunch.com (760K док.).



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{hierarchy}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- Точность и полнота 93%, превосходит асессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов:
  200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (по словарю из 300 этнонимов).

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатако, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боления, фидель, глаза, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Регуляризаторы:

$$\mathscr{L}\left(\Phi, \Theta\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) + R\left(\text{sentiment}\right) \to \max$$

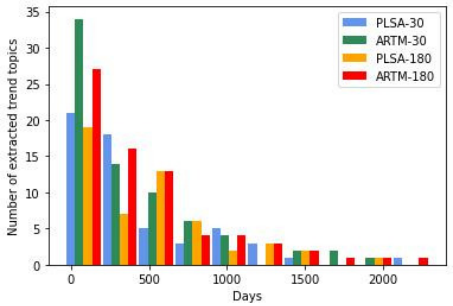
Результаты: число релевантных тем: 45 (LDA) $\to$ 83 (ARTM).

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.

Mining ethnic content online with additively regularized topic models. 2016.

Выявление трендов в коллекции научных публикаций

Цель: раннее обнаружение трендовых тем с начальным экспоненциальным ростом; проверка модели на трендах в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{dynamic}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов.
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Поиск и рубрикация научных публикаций на 100 языках

Цель: мультиязыковой поиск и классификация научных публикаций по рубрикаторам УДК, ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая TM	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\right) + R\left(\right) + R\left(\right) + R\left(\right) \to \max$$

PLSA · interpretable · multimodal · multilanguage · supervised

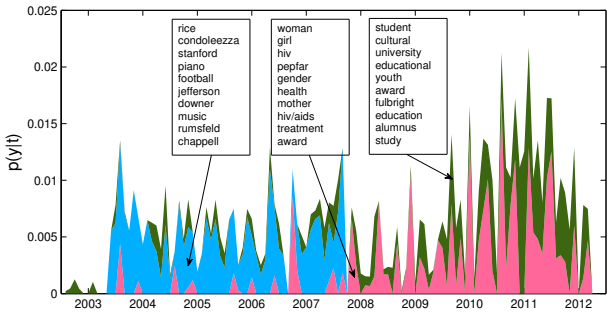
Результаты:

- точность мультиязычного поиска 94%
- сокращение модели 128 Гб $\to$ 4.8 Гб при редукции словарей (BPE-токенизация) до 11K токенов на каждый язык.

П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Н.Зиновкин, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{multimodal}\right)$$
$$+ R\left(\text{n-gram}\right) + R\left(\text{multilanguage}\right) \to \max$$

Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: 5.5 $\to$ 6.5

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

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым событийная тема разделяется на кластеры-мнения

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi}{\square}\,\underset{\Theta}{\square}\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) \to \max$$

PLSA · interpretable · multimodal · n-gram · syntax

Результаты:

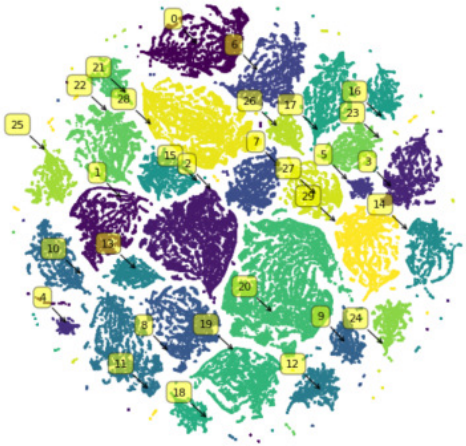
- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей: факты «субъект–предикат–объект», семантические роли слов по Филлмору, тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка, причём

- документы $\to$ клиенты,
- слова $\to$ MCC-коды продавцов.



Регуляризаторы:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{supervised}\right) \to \max$$

Результаты:

- темы — паттерны потребительского поведения
- предсказание пола, возраста, достатка клиентов

E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Поиск обсуждений болезней в социальных сетях

Цель: поиск сообщений в Twitter о болезнях, симптомах, способах лечения, побочных эффектах.

Регуляризаторы:

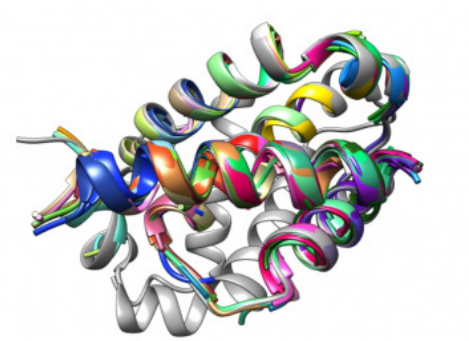
$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) +$$
$$+ R\left(\text{hierarchy}\right) + R\left(\text{temporal}\right) + R\left(\text{geospatial}\right) \to \max$$

Модель ATAM (Ailment Topic Aspect Model) похожа на поиск этно-релевантных тем и легко реализуема в BigARTM

M.J.Paul, M.Dredze. Discovering Health Topics in Social Media Using Topic Models, 2014

Обработка последовательностей нуклеотидов или аминокислот

Цель: поиск мотивов и предсказание функций по нуклеотидным или аминокислотным последовательностям.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{seed words}\right) + R\left(\text{interpretable}\right) +$$
$$+ R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) + R\left(\text{segmentation}\right) \to \max$$

Такая модель легко реализуема в BigARTM.

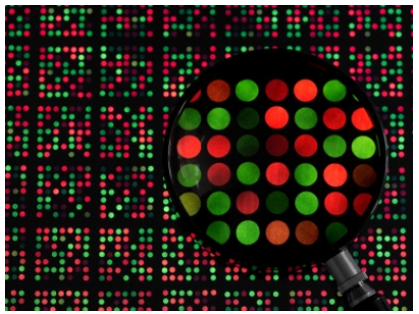
J.B.Gutierrez, K.Nakai. A study on the application of topic models to motif finding algorithms. 2016.

Lin Liu, Lin Tang, Libo He, Shaowen Yao, Wei Zhou. Predicting protein function via multi-label supervised topic model on gene ontology. 2017.

Lin Liu, Lin Tang, Xin Jin, Wei Zhou. A multi-label supervised topic model conditioned on arbitrary features for gene function prediction. 2019

Обработка данных ДНК-микрочипов

Цель:
кластеризация или классификация генов без предположения о функциональной независимости между генами.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{supervised}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) \to \max$$

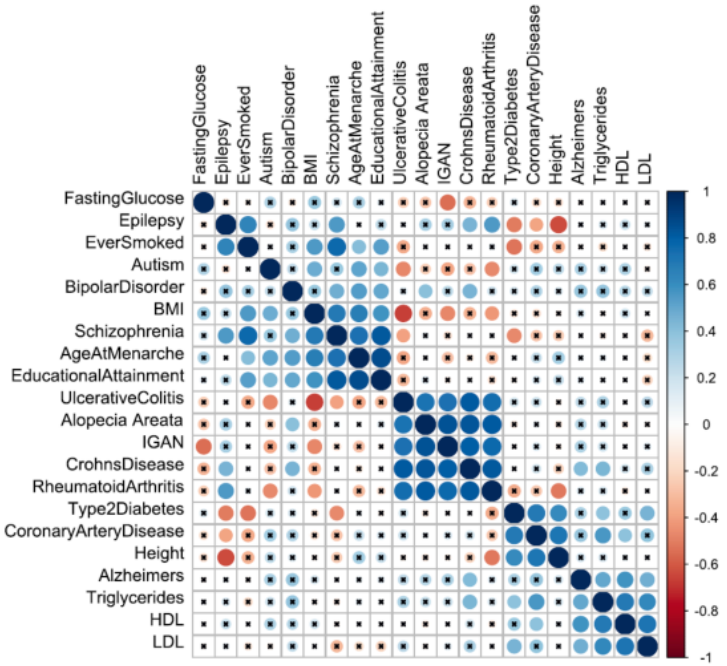
Такая модель легко реализуема в BigARTM.

M.Bicego, P.Lovato, et al. Investigating Topic Models' Capabilities in Expression Microarray Data Classification. 2012.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. 2016.

Функциональная аннотация генома человека

Цель: прогноз тканеспецифических функций некодирующих генетических вариантов для каждой позиции в геноме человека в 127 различных тканях и типах клеток.



Регуляризаторы (гипотеза):

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{supervised}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{hierarchy}\right) + R\left(\text{n-gram}\right) \to \max$$

Такая модель легко реализуема в BigARTM.

D.Backenroth et al. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. 2018

В сухом остатке

- В вероятностном тематическом моделировании (PTM) отказ от байесовской регуляризации в пользу обычной (классической, не-байесовской) сильно упрощает теорию
- Теперь PTM — это теория одной леммы
- BigARTM — скорость, масштабируемость, гибкость
- **Открытая проблема:** объединение PTM и DeepNN не поверхностное (на уровне моделей как чёрных ящиков) а концептуальное (на уровне итерационных процессов)

Vorontsov K. V. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization. 2023.

Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2025 (принято к публикации в издательстве УРСС)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022

He Zhao, Dinh Phung , Viet Huynh , Yuan Jin , Lan Du , Wray Buntine. Topic Modelling Meets Deep Neural Networks: A Survey. 2021