Вероятностные тематические модели
Лекция 9. Мультимодальные тематические модели

К. В. Воронцов
vokov@forecsys.ru

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

МФТИ – ФИЦ ИУ РАН • 6 ноября 2019

Содержание

Напоминание. Мультимодальная ARTM: постановка задачи

Дано: W^m — словарь термов m-й модальности, $m \in M$,
D — коллекция текстовых документов $d \subset W = \bigsqcup_m W^m$,
n_{dw} — сколько раз терм w встретился в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W^m \times T}{\Phi^m}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терма w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума регуляризованного правдоподобия:

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

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

Vorontsov K.V., Frei O.I., Apishev M.A., Romov P.A., Suvorova M.A., Yanina A.O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // CIKM. Workshop on Topic Models, 2015.

Напоминание. Мультимодальная ARTM: EM-алгоритм

Максимизация log правдоподобия с регуляризатором R:

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

где $\tilde{n}_{dw} = \tau_{m(w)} n_{dw}$, $m(w)$ — модальность терма w.

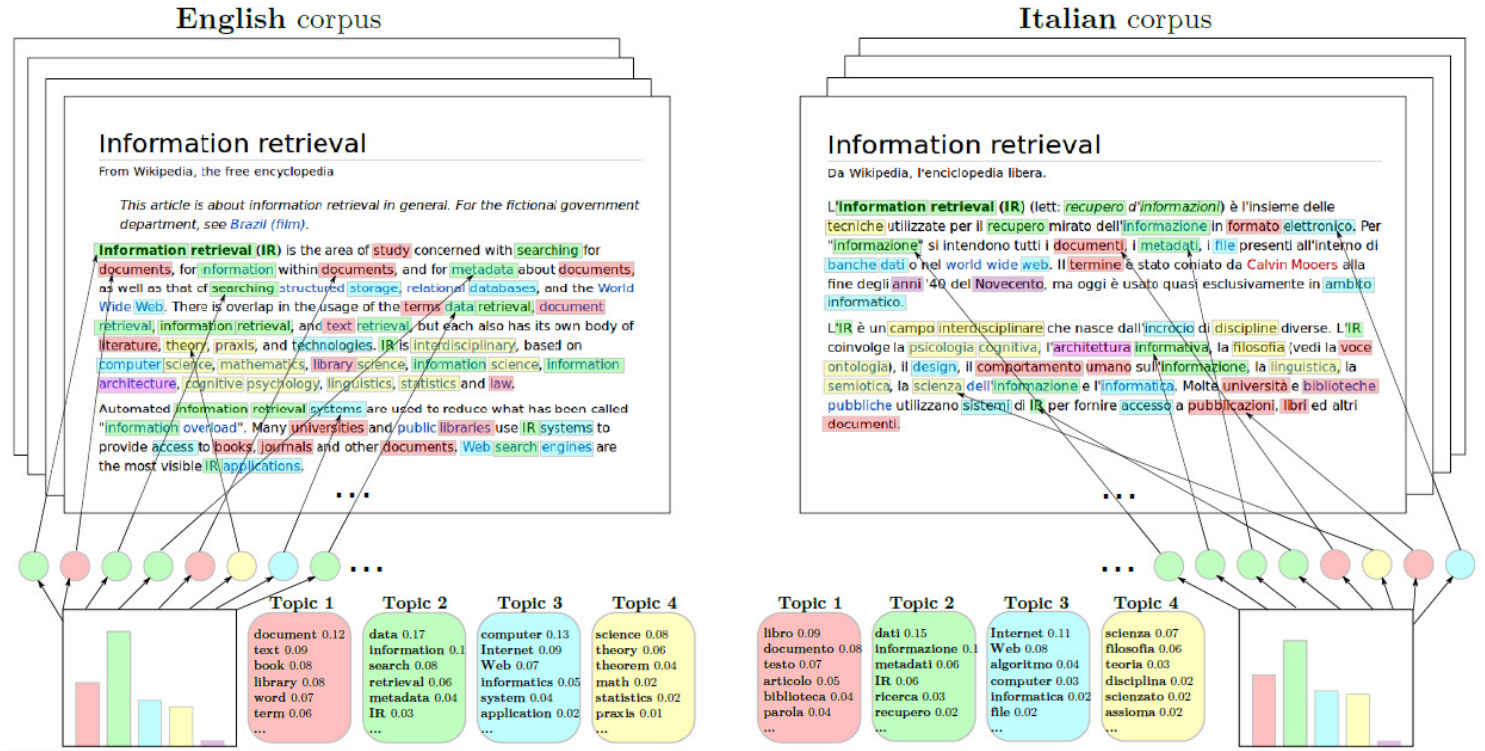
EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Многоязычные модели параллельных коллекций



Для построения мультиязычных тем достаточно иметь парные документы, без выравнивания, без двуязычных словарей!

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

Параллельные и сравнимые корпуса текстов

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

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

W^ℓ — словарь языка ℓ из множества языков L.

Модель ML-P (MultiLingual Parallel)

- каждый язык — отдельная модальность
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup\limits_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

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

---

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

Пример тем. Мультиязычная модель Википедии

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

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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример тем. Мультиязычная модель Википедии

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

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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Регуляризация по двуязычным словарям. Модель ML-TD

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

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

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

Модель ML-TD (MultiLingual Translation Dictionary)

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

Недостатки. Модель ML-TD не учитывает два обстоятельства:

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

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

Матрица вероятностей переводов. Модель ML-TDP

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

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

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

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

Модель ML-TDP (MultiLingual Translation Dictionary Probability)

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

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

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

ML-TD (MultiLingual Translation Dictionary):

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

ML-TDP (MultiLingual Translation Dictionary Probability):

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

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

Смысл регуляризации:

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

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

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

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

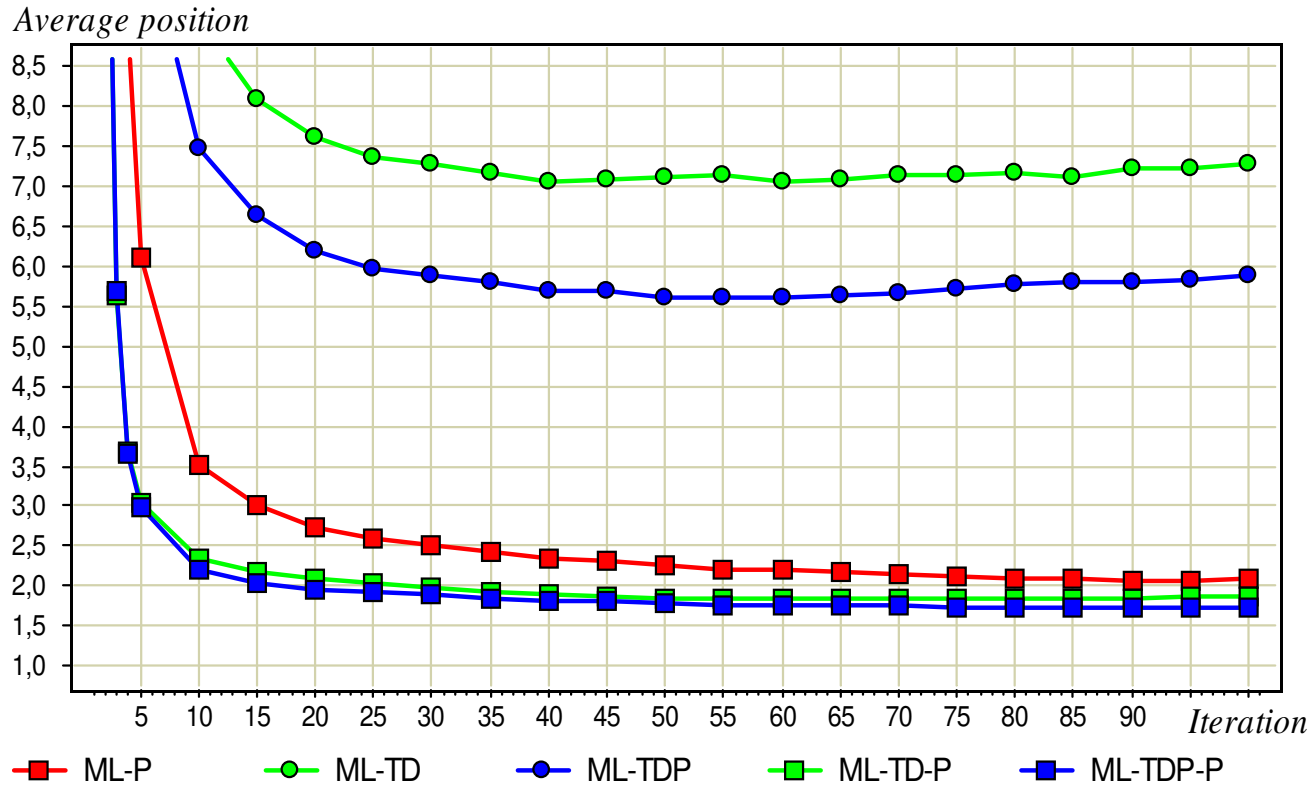
Темы, в которых $p(\text{«total»}|\text{«сумма»},t) > 0.9$

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

Кросс-язычный поиск: ищем документ по его переводу

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

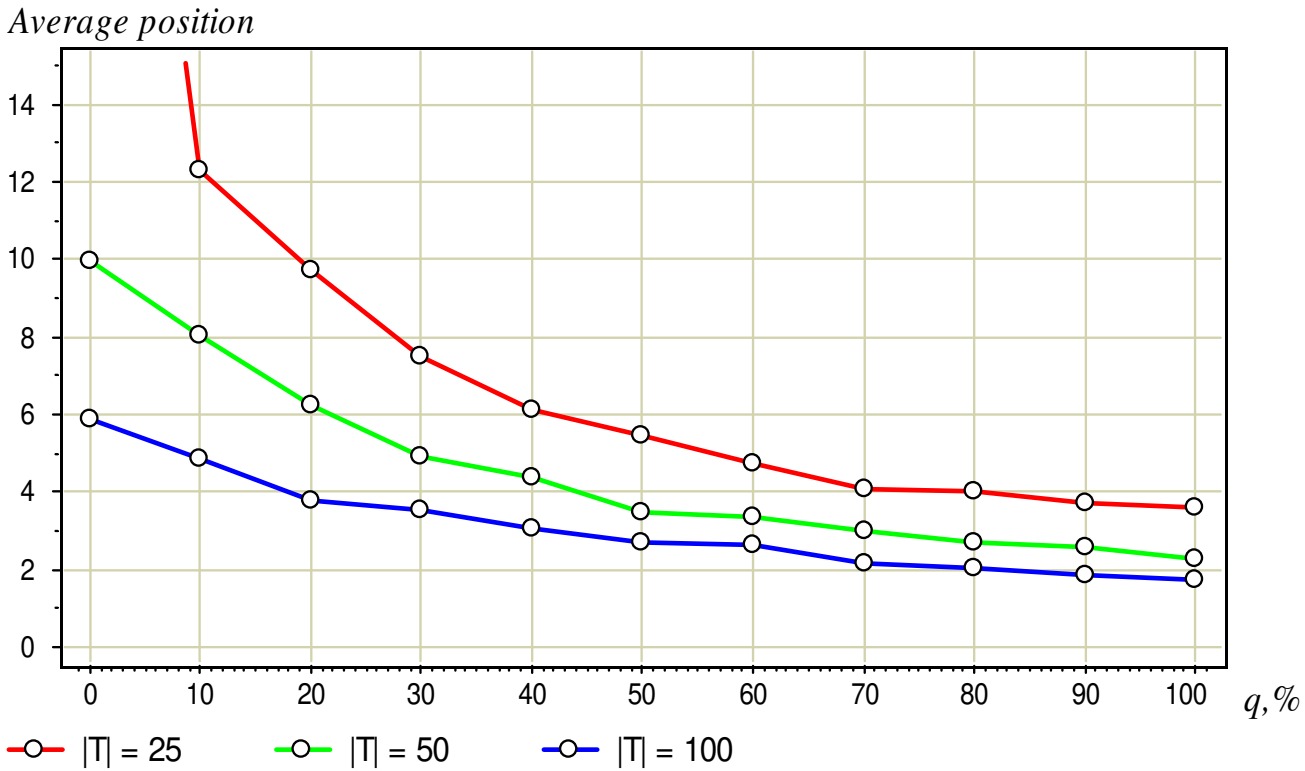
Качество поиска — средняя позиция перевода в выдаче:



Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 25, 50, 100$, $|W^{\text{рус}}| = 19\,305$, $|W^{\text{eng}}| = 23\,413$, переводов 82 642 пар.

Зависимость средней позиции перевода в выдаче
от числа тем $|T|$ и доли q параллельных текстов в коллекции:



Резюме по мультиязычным моделям

- Главное чудо: для построения мультиязычных тем достаточно иметь сравнимые корпуса
- Сравнимая коллекция является более сильным источником многоязычной информации, чем словарь переводов (!)
- Модель с вероятностями переводов — самая сильная
- Не обязательно, чтобы все документы имели параллельные
- Главное применение — по запросу на одном языке ищем:
  — тексты на другом языке — *кросс-язычный поиск*
  — тексты на всех языках — *мульти-язычный поиск*
- Применение в статистическом машинном переводе:
  выбор варианта перевода согласно тематике документа

Тематическая модель с порождающей модальностью

Основные предположения:

- Модальность C (категории, авторы) порождает темы
- $D \times W \times T \times C$ — дискретное вероятностное пространство
- коллекция — i.i.d. выборка $(d_i, w_i, t_i, c_i)_{i=1}^n \sim p(d, w, t, c)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- два предположения об условной независимости:
  $p(w|d,t) = p(w|t), \;\; p(t|c,d) = p(t|c)$

Вероятностная модель порождения документа d:

$$p(w|d) = \sum_{t \in T} p(w|t) \sum_{c \in C} p(t|c)\, p(c|d) = \sum_{t \in T} \phi_{wt} \sum_{c \in C} \psi_{tc} \pi_{cd}$$

- $\phi_{wt} \equiv p(w|t)$ — распределение термов в темах
- $\psi_{tc} \equiv p(t|c)$ — распределение тем в категориях
- $\pi_{cd} \equiv p(c|d)$ — распределение категорий в документах

ARTM для трёх-матричных разложений ФΨП

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C} \phi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \;\to\; \max_{\Phi,\Psi,\Pi};$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$p_{tcdw} \equiv p(t,c|d,w) = \underset{(t,c)\in T\times C}{\operatorname{norm}}\big(\phi_{wt}\psi_{tc}\pi_{cd}\big);$$

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right); & n_{wt} = \sum\limits_{d,c} n_{dw}p_{tcdw} \\ \psi_{tc} = \underset{t\in T}{\operatorname{norm}}\left(n_{tc} + \psi_{tc}\dfrac{\partial R}{\partial \psi_{tc}}\right); & n_{tc} = \sum\limits_{d,w} n_{dw}p_{tcdw} \\ \pi_{cd} = \underset{c\in C}{\operatorname{norm}}\left(n_{cd} + \pi_{cd}\dfrac{\partial R}{\partial \pi_{cd}}\right); & n_{cd} = \sum\limits_{w,t} n_{dw}p_{tcdw} \end{cases}$$

Автор-тематическая модель (Author-topic model)

$C_d \subset C$ — множество порождающих категорий документа d

- Если $\pi_{cd} = \frac{1}{|C_d|}\left[c \in C_d\right]$, вклады авторов равны,
  то матрица Π фиксирована, EM-алгоритм на Π отдыхает :)
- Если $\pi_{cd} = 0, \quad c \notin C_d$, вклады авторов определяет модель,
  фиксирована структура разреженности матрицы Π,
  EM-алгоритм определяет только ненулевые элементы.
- Если множество C_d задано неточно или частично:

$$R(\Pi) = \sum_{d \in D} \sum_{c \in C_d} \ln \pi_{cd} \to \max$$

- Если множества C_d неизвестны, но Π разрежена:

$$R(\Pi) = -\sum_{d \in D} \sum_{c \in C} \ln \pi_{cd} \to \max$$

M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth. The author-topic model for authors and documents. 2004.

Тематическая модель для анализа видеопотоков

- Документ d — 1-секундный видеоклип
- Категория c — *поведение* (behaviour), сочетание действий
- Тема t — *действие* (action), сочетание событий
- Терм w — элементарное *визуальное событие* (event)

Задача: выделить в клипе одно основное поведение.

T. Hospedales, Shaogang Gong, Tao Xiang. Video behaviour mining using a dynamic topic model. 2011.

Транзакционные данные

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

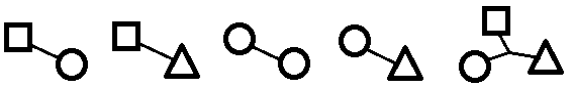
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

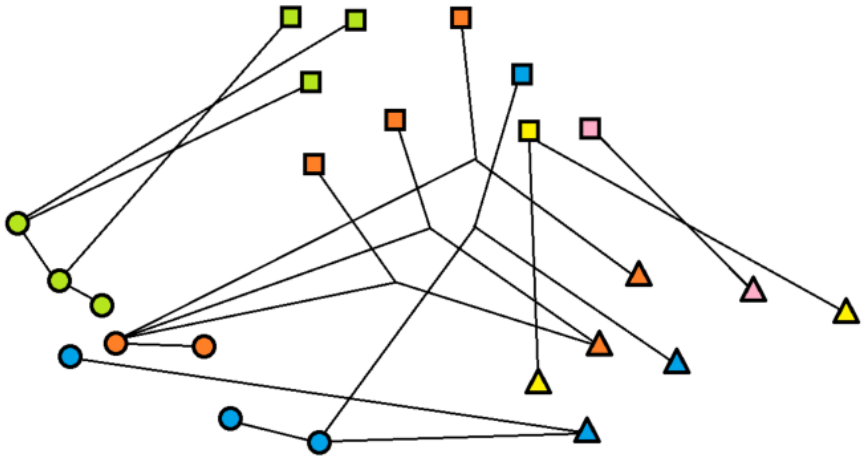
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



Исходные данные:

E_k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{kdx} — число вхождений ребра (d, x) в выборку E_k

Тематическая модель гиперграфа: основные предположения

- в ребре (d, x) подмножество $x \subset V$ может быть любым, независимо от типа ребра k
- первая *гипотеза условной независимости*: тематика контейнера $p(t|d)$ не зависит от типа ребра k
- вторая *гипотеза условной независимости*: распределение $p(v|t)$ термов v модальности V^m в теме t не зависит ни от контейнера d, ни от типа ребра k
- третья *гипотеза условной независимости*: термы $v \in x$ в ребре (d, x) независимы друг от друга
- *гипотеза «мешка транзакций»*: выборка транзакций типа k порождается случайно и независимо из

$$p_k(d, x) = p(d) \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t)$$

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k

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

Задача максимизации взвешенной суммы log-правдоподобий по всем типам рёбер:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \to \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки:

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

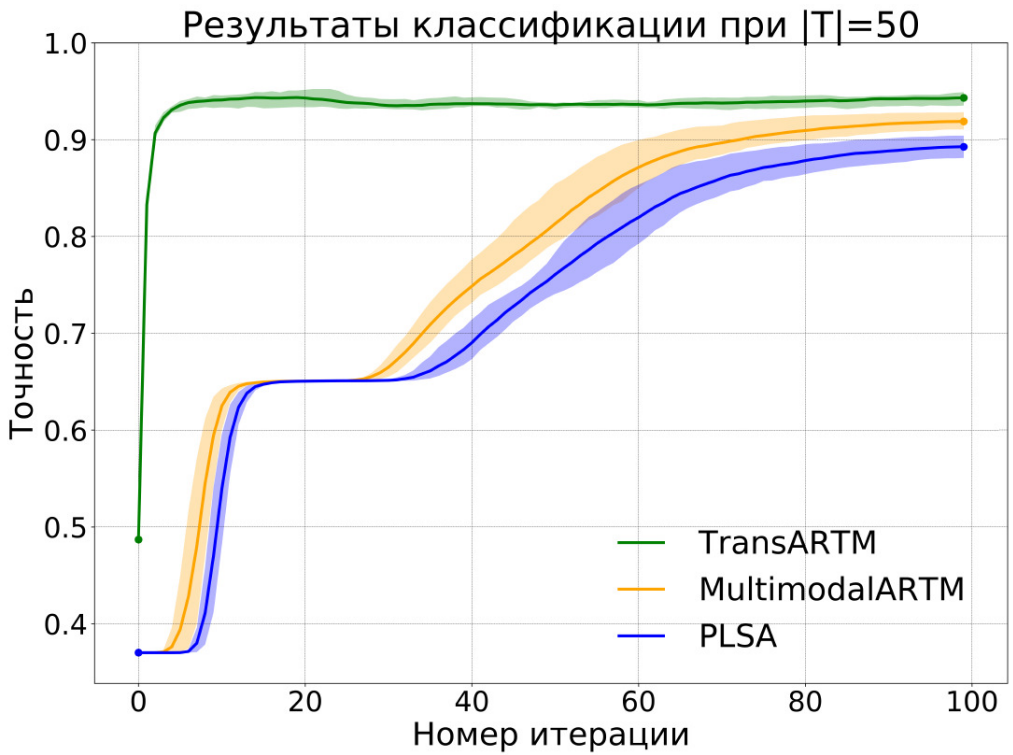
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

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

M-шаг:
$$\begin{cases} \phi_{vt} = \underset{v\in V^m}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} \big[v\in x\big] n_{kdx} p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}} \Big) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx} p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \Big) \end{cases}$$

Эксперименты на модельных данных

13M транзакций, 3 модальности, 5 классов, 9 типов рёбер



Вывод: обычные модели не могут восстановить гиперграф.

Илья Жариков. Гиперграфовые тематические модели транзакционных данных. Магистерская диссертация, МФТИ, 2018.

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t \in T} p(t|d) \prod_{w \in s} p(w|t)^{n_{sw}} = \sum_{t \in T} \theta_{td} \prod_{w \in s} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

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

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Гиперграфовые тематические модели языка

Что ещё может быть ребром гиперграфа?

Любое подмножество термов, связанных друг с другом по смыслу, и порождаемых одной общей темой.

- предложение
- синтагма, ветка синтаксического дерева
- именная группа
- факт «объект, субъект, действие»
- пары термов в соседних предложениях:
  два синонима, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст сообщения и его автор
- финансовая транзакция с текстом платёжного поручения

Анализ транзакций розничных клиентов банка

Дано (Sberbank Data Science Contest):
D — множество клиентов (15 000)
W — категории = MCC-коды (Merchant Category Code) (328)
n_{dw} — сумма транзакций клиента d по категории w

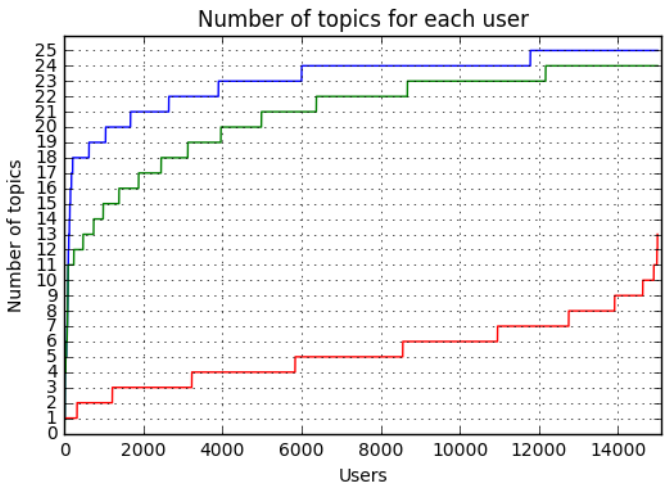
Найти: темы — типы экономического поведения (потребления)
$\phi_{wt} = p(w|t)$ — структура потребления для темы t
$\theta_{td} = p(t|d)$ — типы потребления клиента d

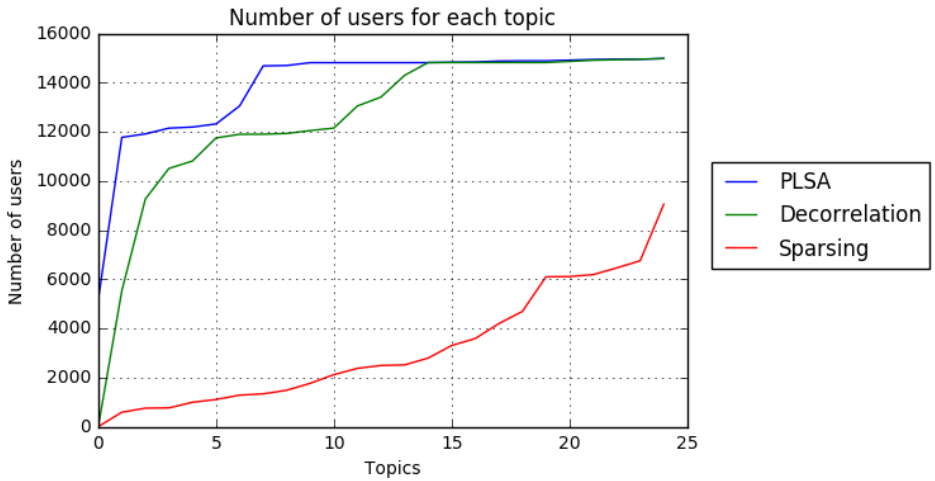
Регуляризаторы:

- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала

Построение модели ARTM, 25 тем

- 30 итераций PLSA — без регуляризаторов
- 10 итераций — повышение различности тем
- 10 итераций — разреживание $p(t|d)$





Декоррелирование Φ и разреживание Θ определяют минимальное число типов экономического поведения каждого клиента, достаточное для описания его расходов.

Пользуюсь картой только чтобы снять наличные

ϕ_{wt},%	MCC-код (категория расходов)
72	Финансовые институты — снятие наличности вручную
27	Финансовые институты — снятие наличности автоматически
0.23	Денежные переводы MasterCard MoneySend
0.1	Денежные переводы
0.012	Финансовые институты — снятие наличности вручную
0.0055	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
0.0027	Магазины игрушек

Наличные + авто, спорт, компьютеры

ϕ_{wt},%	MCC-код (категория расходов)
55	Финансовые институты — снятие наличности автоматически
44	Денежные переводы
0.111	Станции техобслуживания
0.105	Автозапчасти и аксессуары
0.094	Компьютерная сеть/информационные услуги
0.043	Спортивная одежда, одежда для верховой езды и езды на мотоцикле
0.024	Финансовые институты — снятие наличности вручную
0.020	СТО общего назначения
0.018	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
0.015	Магазины мужской и женской одежды
0.015	Финансовые институты — снятие наличности вручную
0.013	Магазины спорттоваров
0.012	Садовые принадлежности (в том числе для ухода за газонами) в розницу
0.011	Паркинги и гаражи
0.011	Бакалейные магазины, супермаркеты
0.010	Различные магазины одежды и аксессуаров

Цивилизованный потребитель: разные магазины, связь, авто

ϕ_{wt},%	MCC-код (категория расходов)
27	Станции техобслуживания
20	Различные продовольственные магазины, рынки, полуфабрикаты
15	Звонки с использованием телефонов, считывающих магнитную ленту
12	Финансовые институты — снятие наличности автоматически
4.7	Горючее топливо — уголь, нефть, разжиженный бензин, дрова
4.1	Универсальные магазины
3.4	Автозапчасти и аксессуары
1.4	Аптеки
1.2	Магазины с продажей спиртных напитков на вынос
1.1	Бакалейные магазины, супермаркеты
0.57	Автошины
0.37	Прямой маркетинг — торговля через каталог
0.35	Товары для дома
0.33	Универмаги
0.32	Плавательные бассейны — распродажа
0.21	Места общественного питания, рестораны

Всего 24 категории с $\phi_{wt} > 0.1\%$; 61 категория с $\phi_{wt} > 0.01\%$

Продвинутые мамки

ϕ_{wt},%	MCC-код (категория расходов)
56	Бакалейные магазины, супермаркеты
8.6	Финансовые институты — снятие наличности автоматически
5.4	Аптеки
4.0	Звонки с использованием телефонов, считывающих магнитную ленту
2.2	Рестораны, закусочные
1.8	Обувные магазины
1.5	Различные продовольственные магазины — рынки, полуфабрикаты
1.4	Магазины спорттоваров
1.4	Детская одежда, включая одежду для самых маленьких
1.3	Магазины игрушек
1.3	Места общественного питания, рестораны
1.1	Магазины мужской и женской одежды
1.1	Магазины с продажей спиртных напитков на вынос
1.1	Магазины косметики
1.0	Садовые принадлежности в розницу
0.73	Одежда для всей семьи

Всего 41 категория с $\phi_{wt} > 0.1\%$; 95 категорий с $\phi_{wt} > 0.01\%$

Бизнес-леди: забыла про наличку — всё по карте

ϕ_{wt},%	MCC-код (категория расходов)
12	Магазины мужской и женской одежды
7.3	Оборудование, мебель и бытовые принадлежности
7.0	Места общественного питания, рестораны
5.6	Магазины по продаже часов, ювелирных изделий и изделий из серебра
5.3	Обувные магазины
4.7	Магазины косметики
4.6	Одежда для всей семьи
3.8	Универмаги
3.2	Готовая женская одежда
2.8	Практикующие врачи, медицинские услуги
1.8	Прямой маркетинг — торговля через каталог
1.5	Салоны красоты и парикмахерские
1.3	Детская одежда, включая одежду для самых маленьких
1.3	Аптеки
1.0	Изготовление и продажа меховых изделий
1.0	Центры здоровья

Всего 70 категорий с $\phi_{wt} > 0.1\%$; 134 категории с $\phi_{wt} > 0.01\%$

Продвинутый активный потребитель всего, и по карте

ϕ_{wt},%	MCC-код (категория расходов)
20	Финансовые институты — снятие наличности вручную
15	Универсальные магазины
13	Туристические агентства и организаторы экскурсий
11	Автозапчасти и аксессуары
8.8	Коммунальные услуги — электричество, газ, санитария, вода
4.2	Веломагазины — продажа и обслуживание
3.7	СТО общего назначения
0.9	Услуги курьера — по воздуху и на земле, агентство по отправке грузов
0.8	Рекламные услуги
0.7	Компьютеры, периферия, программное обеспечение
0.5	Образовательные услуги
0.4	Бакалейные магазины, супермаркеты
0.4	Практикующие врачи, медицинские услуги
0.3	Продажа мотоциклов
0.3	Оборудование, мебель и бытовые принадлежности
0.2	Автошины

Всего 35 категорий с $\phi_{wt} > 0.1\%$; 93 категории с $\phi_{wt} > 0.01\%$

Бизнес-класс: авиа, отели, казино, рестораны, ценные бумаги

ϕ_{wt},%	MCC-код (категория расходов)
28	Авиалинии, авиакомпании
19	Финансовые институты — торговля и услуги
9.5	Отели, мотели, базы отдыха, сервисы бронирования
8.6	Транзакции по азартным играм (плюс)
5.2	Финансовые институты — торговля и услуги
3.2	Места общественного питания, рестораны
3.1	Не-финансовые институты: ин.валюта, переводы, дорожн.чеки, квази-кэш
2.2	Пассажирские железнодорожные перевозки
1.7	Бизнес-сервис
1.4	Жилье — отели, мотели, курорты
1.3	Галереи/учреждения видеоигр
1.3	Транзакции по азартным играм (минус)
0.6	Ценные бумаги: брокеры/дилеры
0.5	Туристические агентства и организаторы экскурсий
0.3	Лимузины и такси
0.3	Беспошлинные магазины Duty Free

Всего 50 категорий с $\phi_{wt} > 0.1\%$; 103 категории с $\phi_{wt} > 0.01\%$

Провинциальный малый бизнес

ϕ_{wt},%	MCC-код (категория расходов)
27	Финансовые институты — снятие наличности автоматически
8.5	Лесо- и строительный материал
8.4	Бытовое оборудование
6.6	Плавательные бассейны — распродажа
5.5	Продажа электронного оборудования
4.1	Бакалейные магазины, супермаркеты
3.3	Универсальные магазины
3.0	Садовые принадлежности в розницу
2.6	Телекоммуникационное оборудование, включая продажу телефонов
2.4	Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
2.2	Товары для дома
2.1	Пассажирские железнодорожные перевозки
1.5	Оборудование, мебель и бытовые принадлежности
1.3	Скобяные товары в розницу
1.2	Магазины спорттоваров
1.1	Аптеки

Всего 54 категории с $\phi_{wt} > 0.1\%$; 104 категории с $\phi_{wt} > 0.01\%$

Анализ транзакций корпоративных клиентов банка

Данные:
лесная отрасль, 2016 г., 10.7M транзакций, 1M компаний.
Транзакция — это тройка $\langle$покупатель, продавец, текст$\rangle$.

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Семь модальностей:

- компании: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели
- ОКВЭДы данной компании
- ОКВЭДы контрагентов: поставщики / покупатели

Примеры тем — видов деятельности компаний

покупка	продажа
0.11: услуга	0.12: лдсп
0.07: классик	0.08: дсп
0.05: дрова	0.03: мдф
0.05: пиловочник	0.03: поставка
0.05: материал	0.02: услуга
0.03: порода	0.02: охранный
0.03: лесоматериал	0.02: ламинировать
0.03: сертум	0.02: хдф
0.02: хвойный	0.02: материал
0.01: дерево	0.01: накл
0.01: транспортный	0.01: товар

покупка	продажа
0.19: право	0.16: арендный
0.17: сбис	0.10: часть
0.16: использование	0.08: плата
0.03: аккаунт	0.04: минимальный
0.02: электронный	0.04: участок
0.02: лицевой	0.04: использование
0.02: устный	0.02: земля
0.01: устройство	0.02: лесничество
0.01: генерация	0.02: земельный
0.01: хранение	0.01: фонд
0.01: ключевой	0.01: федеральный

Примеры тем — видов деятельности компаний

покупка	продажа
0.09: ткань	0.16: мебель
0.09: поставка	0.05: плёнка
0.02: мебельный	0.04: стул
0.02: деревянный	0.03: кресло
0.02: транспортный	0.03: изделие
0.02: фанера	0.02: краска
0.02: поролон	0.02: фанера
0.01: механизм	0.01: лкм
0.01: плата	0.01: лакокрасочный
0.01: частичный	0.01: лак
	0.01: материал
	0.01: клеить

покупка	продажа
0.06: лдсп	0.37: товар
0.05: фурнитура	0.15: мебель
0.02: плёнка	0.04: поставка
0.02: материал	0.04: накладный
0.02: мебельный	0.03: накл
0.02: стекло	0.03: рубль
0.02: мдф	
0.02: кромка	
0.01: транспортный	
0.01: клеить	
0.01: профиль	
0.01: пвх	

Примеры тем — видов деятельности компаний

покупка		продажа
0.52: гсм		0.14: вывоз
0.43: далее		0.09: тбо
		0.04: мусор
		0.03: отход
		0.02: утилизация
		0.01: тко

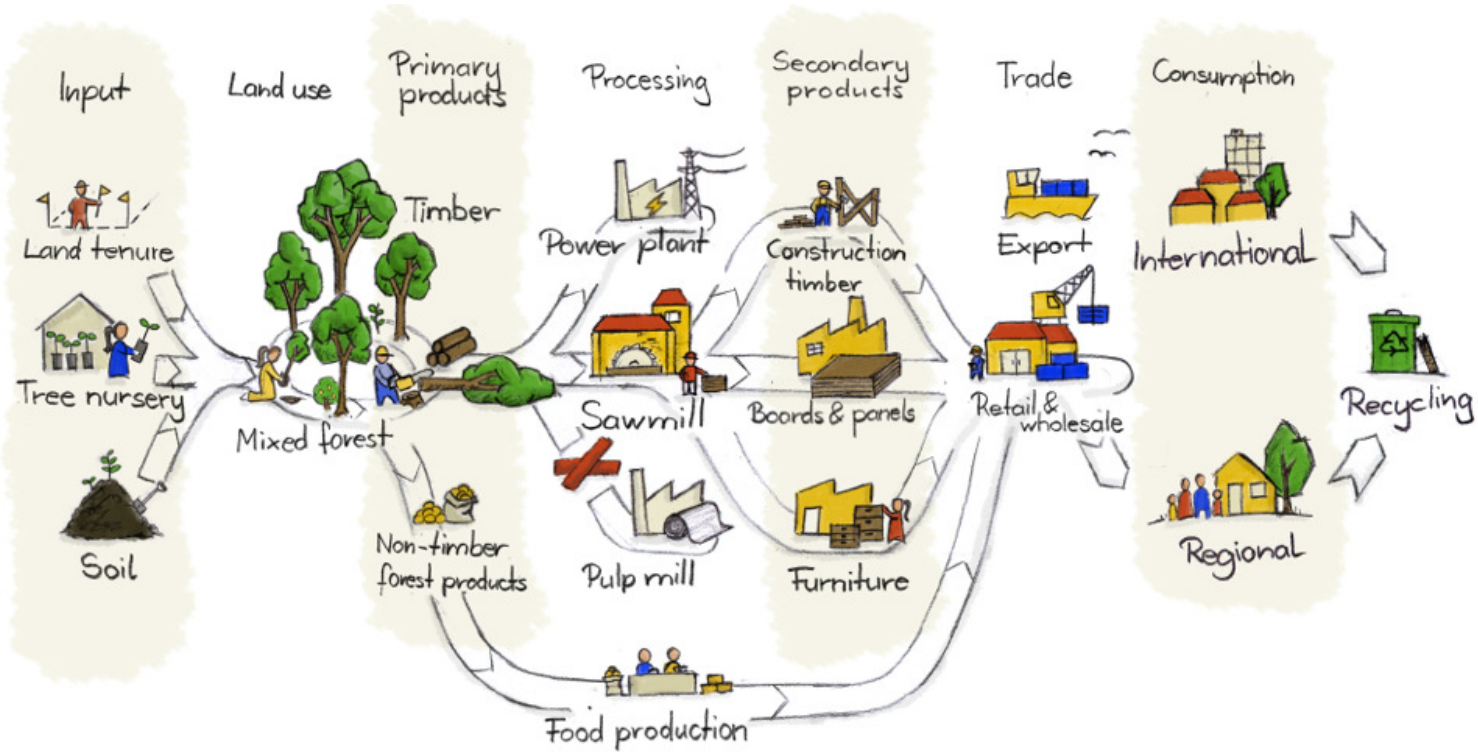
покупка		продажа
0.19: налог		0.11: бумага
0.06: услуга		0.08: гофроящик
0.04: макулатура		0.04: гофрокартон
0.03: поставка		0.03: гофрокороб
0.03: транспортный		0.03: поставка
0.02: лесопродукция		0.03: фактура
0.02: автоуслуга		0.02: гофропродукция
0.01: перевозка		0.02: гофротару
0.01: плата		0.02: гофрирование
		0.02: гофролоток
		0.02: товар
		0.01: лоток

Примеры тем — видов деятельности компаний

покупка	продажа	продажа	продажа
0.15: программа	0.13: фурнитура	0.14: рекламный	0.21: тмц
0.11: право	0.09: материал	0.13: размещение	0.06: накл
0.09: сертификат	0.08: лдсп	0.09: материал	0.04: инструмент
0.07: эвм	0.04: кромка	0.05: проект	0.03: пила
0.07: использование	0.04: мебельный	0.05: яндекс	0.02: заточка
0.07: лицензия	0.04: фрз	0.04: директ	0.02: нож
0.04: криптопро	0.04: мдф	0.04: реклама	0.02: материал
0.03: абонентский	0.03: клеить	0.02: рубль	0.02: фреза
0.02: обслужа	0.03: пвх	0.01: стек	0.02: клеить
0.02: пользование	0.02: тмц		0.01: товар
0.02: контур	0.02: комплект		0.01: перчатка
0.01: проверка	0.02: профиль		
	0.02: столешница		

Конечные цели моделирования транзакционных данных

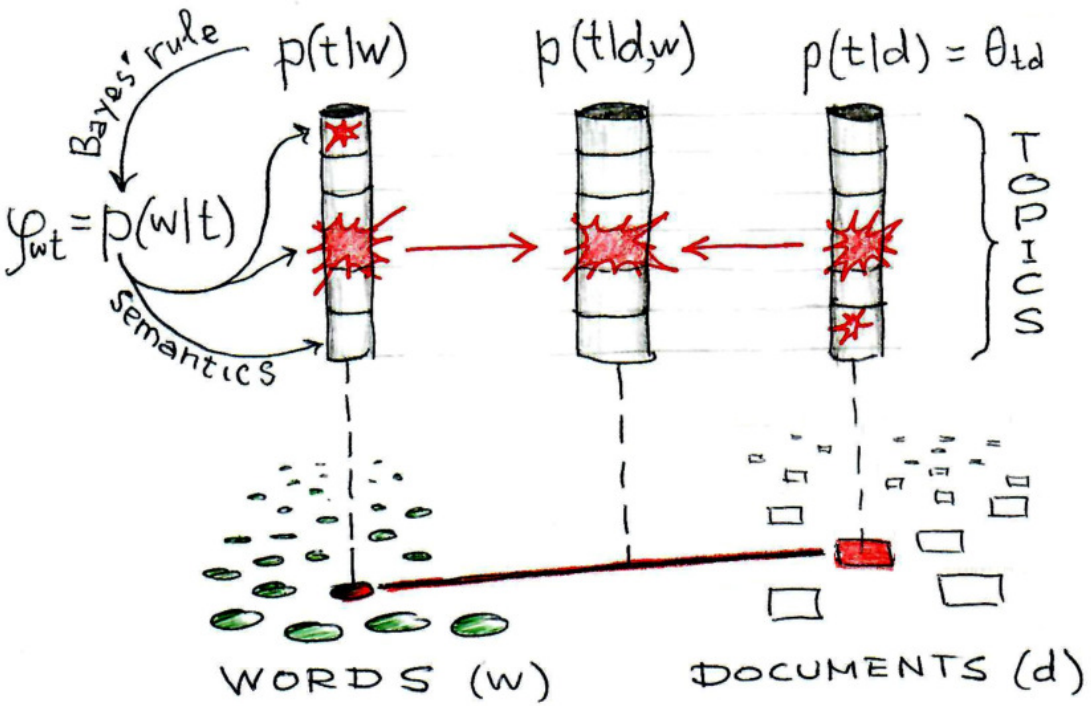
- Получение векторных представлений компаний
- Поиск схожих и конкурирующих компаний
- Восстановление структуры товарных потоков отрасли



Вместо резюме: общий взгляд на тематические модели

Интерпретируемые тематические эмбединги слов и документов

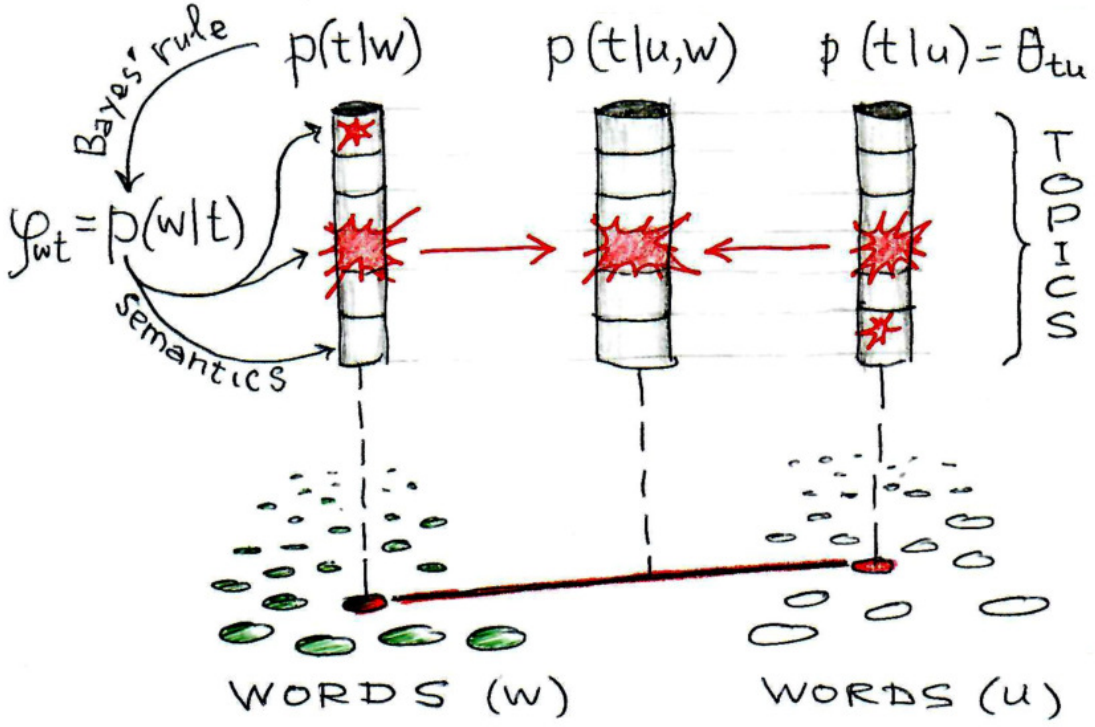
- Коллекция текстов — двудольный граф с рёбрами (d, w)
- Тематические эмбеддинги: $p(t|d)$, $p(t|w)$, $p(t|d,w)$
- Интерпретируемость тем через частоты термов $p(w|t)$
- Слово w встречается в d, когда у них есть общие темы



Вместо резюме: общий взгляд на тематические модели

Интерпретируемые эмбединги сочетаемости слов

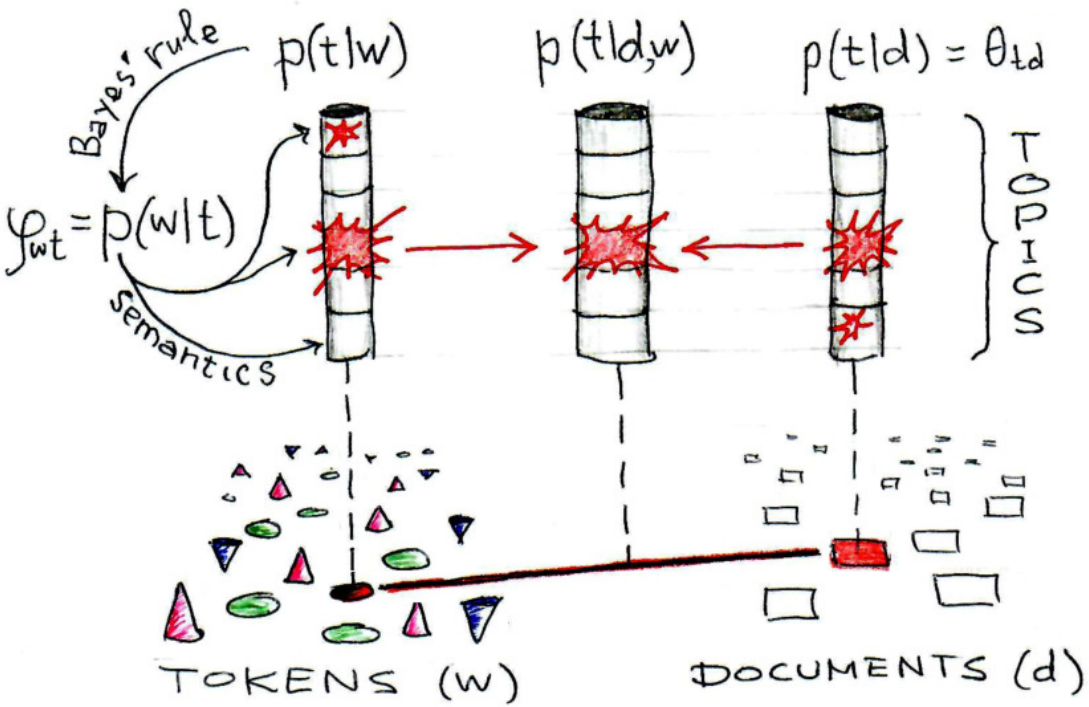
- Постулат *дистрибутивной семантики*: «смысл слова определяется распределением слов всех его контекстов»
- Слово u порождает псевдо-документ всех его контекстов
- Слова w, u сочетаются, когда у них есть общие темы



Вместо резюме: общий взгляд на тематические модели

Интерпретируемые эмбединги мультимодальных документов

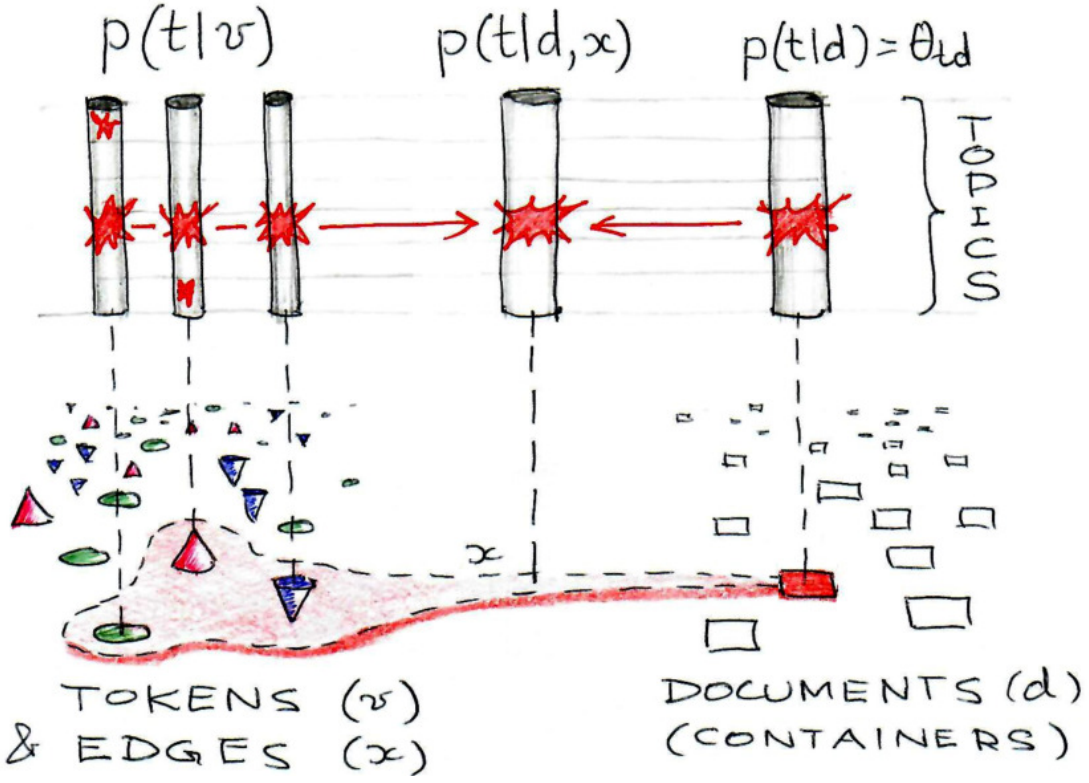
- Документы содержат термы различных *модальностей*
- Примеры: слова, n-граммы, теги, категории, авторы, ...
- Через темы смыслы слов передаются другим модальностям
- Терм w встречается в d, когда у них есть общие темы



Вместо резюме: общий взгляд на тематические модели

Интерпретируемые эмбединги транзакционных данных

- Транзакция — взаимодействие двух *или более* термов
- Примеры: $\langle$buyer, seller, item$\rangle$, $\langle$user, site, banner$\rangle$
- Транзакционные данные — это выборка рёбер гиперграфа
- Транзакция происходит, когда её термы имеют общие темы



Вместо резюме: общий взгляд на тематические модели

Интерпретируемые эмбединги предложений

- Ребро гиперграфа — семантически связанные слова
- Примеры: предложение, синтагма, лексическая цепочка, именная группа, факт «объект, субъект, действие»
- Слова связываются, когда они имеют общие темы