

Тематические векторные представления текстов, графов и транзакционных данных

Воронцов Константин Вячеславович
(МФТИ, ФИЦ ИУ РАН)



Интеллектуализация Обработки Информации
Италия, Гаэта • 8–12 октября 2018

1 Теория

- Задача тематического моделирования
- Аддитивная регуляризация и дальнейшие обобщения
- Проект с открытым кодом BigARTM

2 Тематизация текстов и графов

- Дистрибутивная семантика
- Тематические векторные представления слов
- Тематическая модель гиперграфа

3 Тематизация транзакционных данных

- Транзакции физических лиц
- Транзакции юридических лиц

Пусть

- W — конечное множество слов (терминов, токенов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- каждое слово w в документе d связано с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- 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: $p(w|d, t) = p(w|t)$

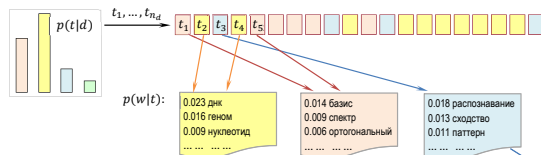
Тематическая модель, по формуле полной вероятности:

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

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

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t :

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



w_1, \dots, w_{n_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 — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

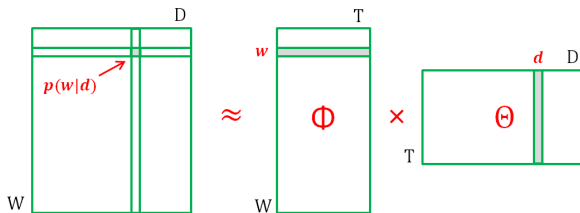
Дано: коллекция текстовых документов

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

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

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

Это задача стохастического матричного разложения:



Принцип 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

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

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

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

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Задача максимизации логарифма правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

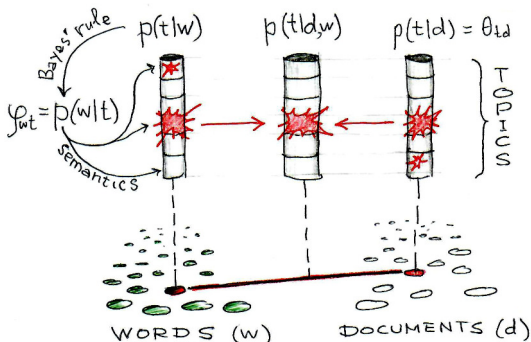
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$\begin{array}{l} \text{E-шаг:} \\ \text{M-шаг:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right) \end{array} \right.$$

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

Тематические векторные представления слов и документов

- Коллекция текстов — двудольный граф с рёбрами (d, w)
- Слово w встречается в d , когда у них есть общие темы
- Интерпретируемость тем возникает благодаря $p(w|t)$



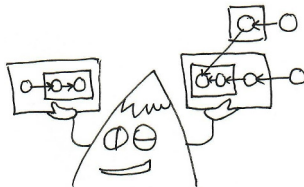
Обобщение №1

Проблема

В PLSA слишком много параметров, возможно переобучение.
Надо наложить ограничения на столбцы матриц Φ и Θ .
Желательно так, чтобы они стали более разреженными.

Решение

Модель латентного размещения Дирихле (2003).



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

Регуляризованный EM-алгоритм: модель LDA

Задача максимизации апостериорной вероятности:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{ln правдоподобия } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \beta_w - 1 \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \alpha_t - 1 \right) \end{cases} \end{cases}$$

Обобщение №2

Проблема

LDA — слишком простой и слабый регуляризатор.

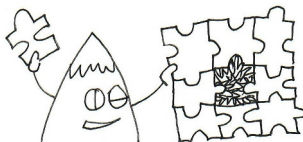
LDA не позволяет комбинировать разные регуляризаторы.

Решение

Ввести произвольный регуляризатор $R(\Phi, \Theta)$

или сумму регуляризаторов $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$.

Аддитивность → модульный подход к моделированию.



ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация \log правдоподобия с регуляризатором R :

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

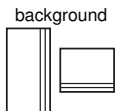
EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{aligned} \text{E-шаг:} & \left\{ \begin{aligned} p_{tdw} &= \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} &= \operatorname{norm}_{w \in W} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} &= \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in D} n_{dw} p_{tdw} \end{aligned} \right. \end{aligned}$$

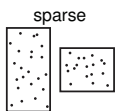
где $\operatorname{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

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

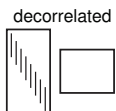
Регуляризаторы для улучшения интерпретируемости тем

Сглаживание фоновых тем $B \subset T$:

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

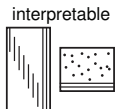
Разреживание предметных тем $S = T \setminus B$:

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



Декоррелирование для повышения различности тем:

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

Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

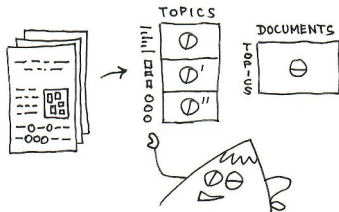
Обобщение №3

Проблема

Есть много задач, в которых документы содержат не только слова, но и элементы других модальностей

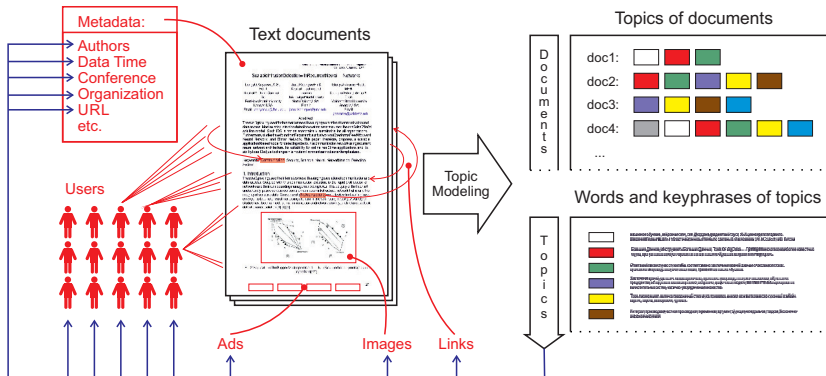
Решение

Ввести для каждой модальности свою матрицу Φ и максимизировать сумму критериев лог-правдоподобия



Задачи мультимодального тематического моделирования

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



Мультимодальная ARTM

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

Максимизация суммы \log правдоподобий с регуляризацией:

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

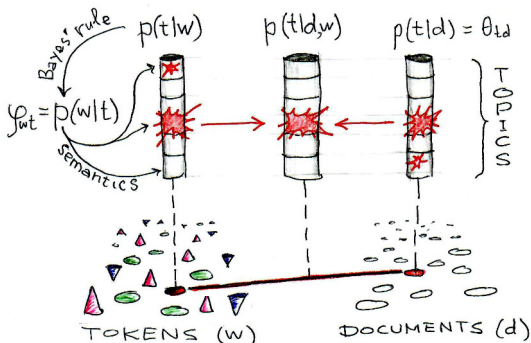
EM-алгоритм: метод простой итерации для системы уравнений

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

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

Мультимодальные тематические векторные представления

- Документы содержат слова и токены других *модальностей*
- Примеры модальностей: авторы, время, теги, пользователи, ...
- Через темы смыслы слов передаются другим модальностям



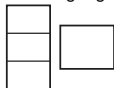
Регуляризаторы для мультимодальных тематических моделей

supervised



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

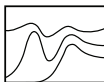
multilanguage



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

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

temporal



Темпоральные модели с модальностью времени i :

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

geospatial



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

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

Пример. Модальность n -грамм улучшает качество тем

Коллекция 1000 статей конференций MMPO, ИОИ на русском

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

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

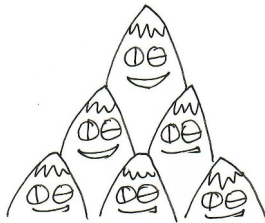
Обобщение №4

Проблема

Невозможно определить оптимальное число тем.
Хотелось бы разделять темы на подтемы иерархически.
Придумано много иерархических моделей, но они либо ограниченные, либо медленные, либо слишком сложные.

Решение

Нисходящая послойная иерархическая модель



Послойное построение уровней тематической иерархии

Шаг 1. Строим модель с небольшим числом тем.

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

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t \in T} n_t \text{KL}_w \left(p(w|t) \parallel \sum_{s \in S} p(w|s)p(s|t) \right) \rightarrow \min_{\Phi, \Psi}$$

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

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

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

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

Обобщение №5

Проблема

Гипотеза «мешка слов» — самое часто критикуемое допущение тематического моделирования.

Как строить модели, учитывающие порядок слов?

Решение

Построив матрицу тем $p(t|d, w_i)$, выполнить некоторые преобразования над локальными окрестностями слов w_{i-1}, w_i, w_{i+1}



Сегментная структура текста и пост-обработка E-шага

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

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



Регуляризация E-шага

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

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

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

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

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

Гипотеза

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

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

Регуляризаторы для моделирования последовательного текста

n-gram



Модели с модальностями n -грамм, коллокаций, именованных сущностей (например, TopMine)

syntax



Модели, учитывающие результаты автоматического синтаксического разбора (например, SyntaxNet)

coherence



Модели дистрибутивной семантики на основе совстречаемости слов (битермы, когерентность)

sentence



Тематические модели, учитывающие границы предложений, абзацев и секций документов

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

BigARTM: библиотека тематического моделирования

Ключевые возможности:

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

Сообщество:

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

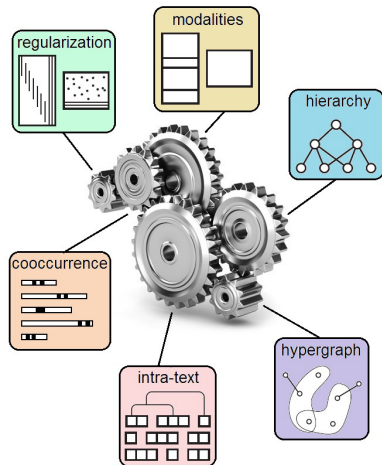


Лицензия и среда разработки:

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

Ключевые механизмы BigARTM

- 1 PLSA, LDA
- 2 регуляризация
- 3 модальности
- 4 иерархия тем
- 5 пост-обработка E-шага
- 6 встречаемость термов
- 7 гиперграфы транзакций



Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К слов

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

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

BigARTM упрощает разработку тематических моделей


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


Этапы моделирования

Bayesian TM

ARTM

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

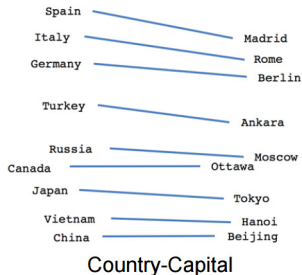
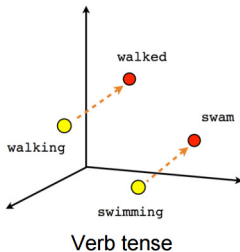
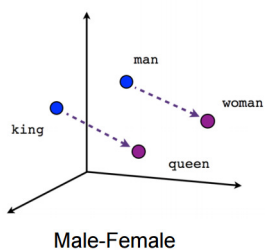
 -- нестандартизируемые этапы, уникальная разработка для каждой задачи

 -- стандартизуемые этапы

Задача семантического векторного представления слов

Найти для каждого слова w вектор $x_w \in \mathbb{R}^T$, чтобы близкие по смыслу слова имели близкие векторы.

Задача семантической аналогии слов:
по трём словам угадать четвёртое.



Дистрибутивная гипотеза

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

Синтагматическая близость слов:

со-встречаемость слов в одном контексте.



здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:

взаимозаменяемость слов в одном контексте.



здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.

J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.

P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR). 2010.

Формализация дистрибутивной гипотезы в программе word2vec

Дано: n_{uw} — совстречаемость слов u, w в окне $\pm h$ слов

Найти: семантические векторные представления слов x_w

Модель: вероятность слова w в контексте слова u :

$$p(w|u) = \text{SoftMax}_{w \in W} \langle x_w, x_u \rangle = \frac{\exp \langle x_w, x_u \rangle}{\sum_v \exp \langle x_v, x_u \rangle}$$

Критерий максимума log-правдоподобия и его аппроксимация:

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

$$\sum_{w, u \in W} n_{wu} \left(\ln \sigma \langle x_w, x_u \rangle + \sum_{i=1}^k \ln \sigma(-\langle x_{v_i}, x_u \rangle) \right) \rightarrow \max_{\{x_w\}}$$

где v_1, \dots, v_k — случайные k слов не из контекста u .

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

Модели векторных представлений для текстов и графов

word2vec: эмбединги слов

T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов

Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений

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

FastText: эмбединги символьных n -грамм

<https://github.com/facebookresearch/fastText>

node2vec: эмбединги вершин графа

A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах

A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research

L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

Недостаток: координаты векторов не интерпретируемы

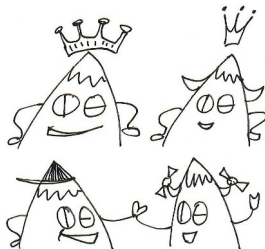
Обобщение №6

Проблема

Тематические векторные представления слов плохо решают задачи семантической близости, по сравнению с word2vec.

Решение

Использовать данные о совстречаемости слов (n_{uw}) вместо данных о встречаемости слов в документах (n_{dw}).



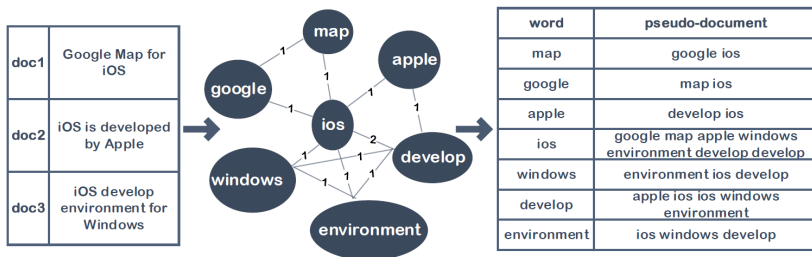
Модель сети слов WNTM для коротких текстов

Идея: моделировать не документы, а связи между словами.

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

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

Контекст — короткое сообщение / предложение / окно $\pm h$ слов.



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

Модели WNTM и WTM (Word Topic Model)

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

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

где d_u — псевдо-документ слова u .

Максимизация логарифма правдоподобия:

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

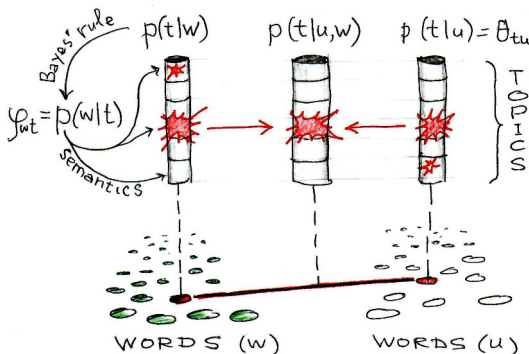
где n_{uw} — совстречаемость слов u, w (кстати, $n_{uw} = n_{wu}$).

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

Berlin Chen. **Word Topic Models** for spoken document retrieval and transcription. ACM Trans., 2009.

Интерпретируемые эмбединги совстречаемости слов

- Идея дистрибутивной семантики: “Words that occur in the same contexts tend to have similar meanings” [Harris, 1954].
- Слово индуцирует псевдо-документ всех его контекстов



word2vec и ARTM на задачах аналогии слов

Два подхода к синтезу векторных представлений слов:

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

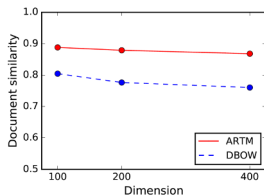
| Операция | Результат ARTM | Результат word2vec |
|-------------------------|--|--|
| king – boy + girl | <i>queen, princess, lord, prince</i> | <i>queen, princess, regnant, kings</i> |
| moscow – russia + spain | <i>madrid, barcelona, aires, buenos</i> | <i>madrid, barcelona, valladolid, malaga</i>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

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

word2vec и ARTM в задаче семантической близости документов

ArXiv triplets dataset: 20К троек статей:

⟨ статья А, схожая статья В, непохожая статья С ⟩



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

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

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

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

Обобщение №7

Проблема

Исходные данные могут быть сложнее, чем парные взаимодействия (транзакции) между объектами

Решение

Тематическая модель описывает транзакции, образуемые произвольными подмножествами объектов



Транзакционные данные

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

Примеры:

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

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

Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.

$V = V^1 \sqcup \dots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

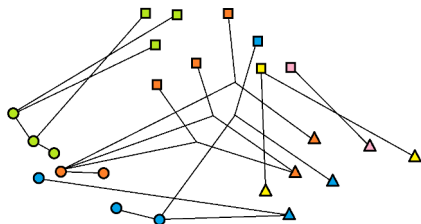
□ ○ △

K — множество типов рёбер:

□○ □△ ○○ ○△ ○△

T — множество тем:

● ● ● ● ●



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x) : вершина-контейнер $d \in V$ и вершины $x \subset V$,

n_{dx} — число вхождений ребра (d, x) в выборку X^k

$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k :

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

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k

$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации \log правдоподобия:

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

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

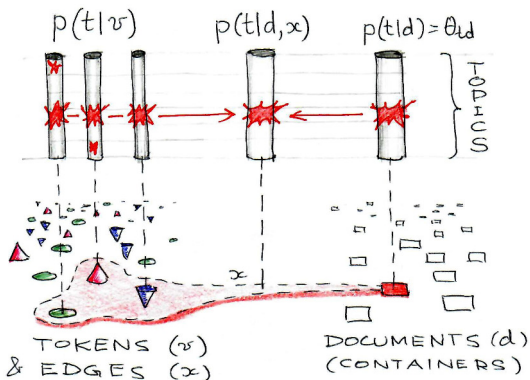
$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in X} \phi_{vt} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

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

$$\begin{cases} \text{E-шаг:} & p_{tdx} = \operatorname{norm}_{t \in T} \left(\theta_{td} \prod_{v \in X} \phi_{vt} \right) \\ \text{M-шаг:} & \begin{cases} \phi_{vt} = \operatorname{norm}_{v \in V^m} \left(\sum_{k \in K} \tau_k \sum_{(d,x)} [v \in X] n_{dx} p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{k \in K} \tau_k \sum_{(d,x)} n_{dx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

Интерпретируемые эмбединги транзакционных данных

- *Гиперграф* — множество подмножеств вершин-токенов
- Транзакция = подмножество токенов = ребро гиперграфа
- Транзакция происходит, когда токены имеют общие темы



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

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

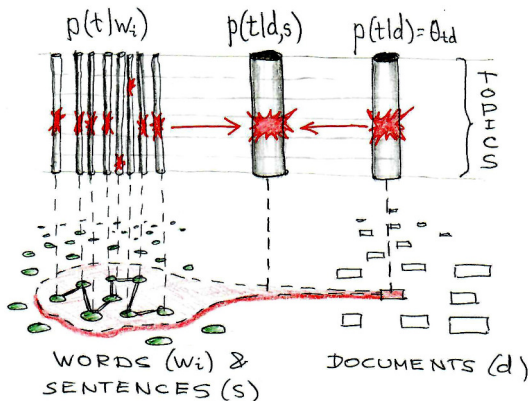
это частный случай гиперграфовой модели, в которой предложения являются «транзакциями» или гипер-рёбрами.

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.

Интерпретируемые эмбединги предложений

- Предложение — семантически однородная единица языка
- Предложение образуется из слов, имеющих общие темы
- Предложение = подмножество слов = ребро гиперграфа



Анализ данных о транзакциях 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
- обучение прогнозов объёма трат по метрике RMSLE_1

Интерпретация тематической модели

Прогноз объёма трат клиента d по категории w

$$n_{dw} = n_d \sum_{t \in T} p(w|t)p(t|d),$$

n_d — объём клиента d по всем категориям или его прогноз.

Проблема несбалансированности клиентов по объёмам:
приоритет получают «богатые» клиенты с наибольшими n_d

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

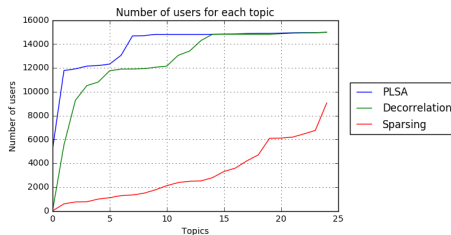
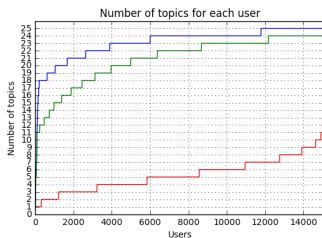
Два способа решения (первый сохраняет аддитивность объёмов):

$$\sum_{d \in D} \sum_{w \in d} \frac{1}{n_d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta};$$

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

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

- $\phi_{wt},\%$ МСС-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

- $\phi_{wt}, \%$ МСС-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

- $\phi_{wt}, \%$ МСС-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

- $\phi_{wt}, \%$ МСС-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

$\phi_{wt}, \%$ МСС-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

- $\phi_{wt}, \%$ МСС-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

- $\phi_{wt}, \%$ МСС-код (категория расходов)
- 28 Авиа линии, 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 Беспшлинные магазины Duty Free

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

Провинциальный малый бизнес

$\phi_{wt}, \%$ МСС-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 данных о транзакциях клиентов банка

Данные:

лесная отрасль, 2016 г., 10.7М транзакций, 1М компаний

Транзакция — ⟨покупатель, продавец, текст⟩

Документы — компании

Четыре модальности:

- компании-поставщики
- слова в текстах транзакций покупки
- компании-покупатели
- слова в текстах транзакций продажи

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 |

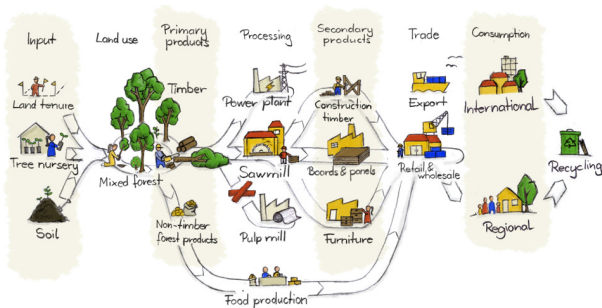
Примеры тем — видов деятельности компаний











| 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 | |

Конечные цели моделирования транзакционных данных

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



-  *K.B.Воронцов*. Обзор вероятностных тематических моделей. 2018. – **NEW!**
<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>
-  *K.B.Воронцов*. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
-  *K.Vorontsov, A.Potapenko*. Additive regularization of topic models. Machine Learning, 2015.
-  *O.Frei, M.Apishev*. Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
-  *N.Chirkova, K.Vorontsov*. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
-  *A.Ianina, K.Vorontsov*. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
-  *A.Potapenko, A.Popov, K.Vorontsov*. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
-  *V.Alekseev, V.Bulatov, K.Vorontsov*. Intra-Text Coherence as a Measure of Topic Models Interpretability. Dialogue, 2018.
-  *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov*. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
-  *N.Skachkov, K.Vorontsov*. Improving topic models with segmental structure of texts. Dialogue, 2018.