НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Internet Studies Lab, Department of Applied Mathematics and Business Informatics

INTRODUCTION TO TOPIC MODELING

Анализ баз данных в публичном управлении

Кольцов С.Н.

Saint Petersburg, 05.10.2018

Введение в историю возникновения Topic modeling

Тематическое моделирование — это одно из современных направлений статистического анализа текстов (и не только), активно развивающееся с конца 90-х годов. Вероятностная тематическая модель (probabilistic topic model) коллекции текстовых документов предполагает, что документы и слова в коллекции можно представить в виде комбинации распределений по темам.

Тематические модели применяются:

1. для выявления трендов в новостных потоках или научных публикациях.
2. для анализа текстовых данных социальных сетей.
3. для классификации и категоризации документов.
4. для анализ изображений и видеопотоков.
5. для анализа нуклеотидных и аминокислотных последовательностей, а также в задачах популяционной генетики.
6. Для разных целей в физике.

Эволюция моделей.

1. Латентно-семантический анализ (**Latent Semantic Analysis**)
2. **PLSA**
3. **LDA** (классический вариант)
4. Регуляризационные модели **LDA**

Латентно-семантический анализ (Latent Semantic Analysis)

LSA используется для выявления латентных (скрытых) ассоциативно-семантических связей между термами (словами, н-граммами) путем сокращения факторного пространства термы-на-документы. Термами могут выступать как слова, так и их комбинации, т.наз. н-граммы.

Основная идея латентно-семантического анализа состоит в следующем: если в исходном вероятностном пространстве, состоящим из векторов слов (вектор = предложение, абзац, документ и т.п.), между двумя любыми словами из двух разных векторов может не наблюдаться никакой зависимости, то после **некоторого алгебраического преобразования данного векторного пространства** эта зависимость может появиться, причем величина этой зависимости будет определять силу ассоциативно-семантической связи между этими двумя словами.

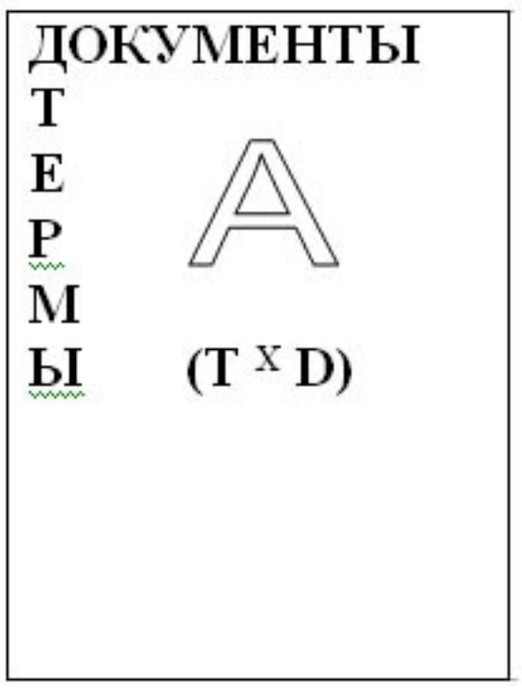
Как это работает:

В качестве исходной информации LSA использует матрицу термы-на-документы (термы – слова, словосочетания или н-граммы; документы – тексты, классифицированные либо по какому-либо критерию, либо разделенные произвольным образом – это зависит от решаемой задачи), описывающую набор данных, используемый для обучения системы. Элементы этой матрицы содержат, как правило, веса, учитывающие частоты использования каждого терма в каждом документе.

Некоторые понятия

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов. Терминами могут быть как отдельные слова, так и словосочетания. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $w_1, \ldots, w_n$ из словаря W .

Предполагается, что существует конечное множество факторов K , и каждое вхождение термина w в документ d связано с некоторым фактором. Термины w и документы d являются наблюдаемыми переменными, фактор $k \in K$ является латентной (скрытой) переменной.



Исходные данные (слова и документы можно представить в виде матрицы), где документы это колонки, первая колонка это список всех уникальных слов в заданной коллекции документов.

Таким образом, каждый документ это вектор в пространстве слов.

Латентно-семантический анализ (LSA)

LSA основан на использовании разложения вещественной матрицы по сингулярным значениям или SVD-разложения (SVD – Singular Value Decomposition). С помощью него любую матрицу можно разложить в виде произведения ортогональных матриц, комбинация которых является достаточно точным приближением к исходной матрице.
Согласно теореме о сингулярном разложении в самом простом случае матрица может быть разложена на произведение трех матриц:

$$A = U\,S\,V^T$$

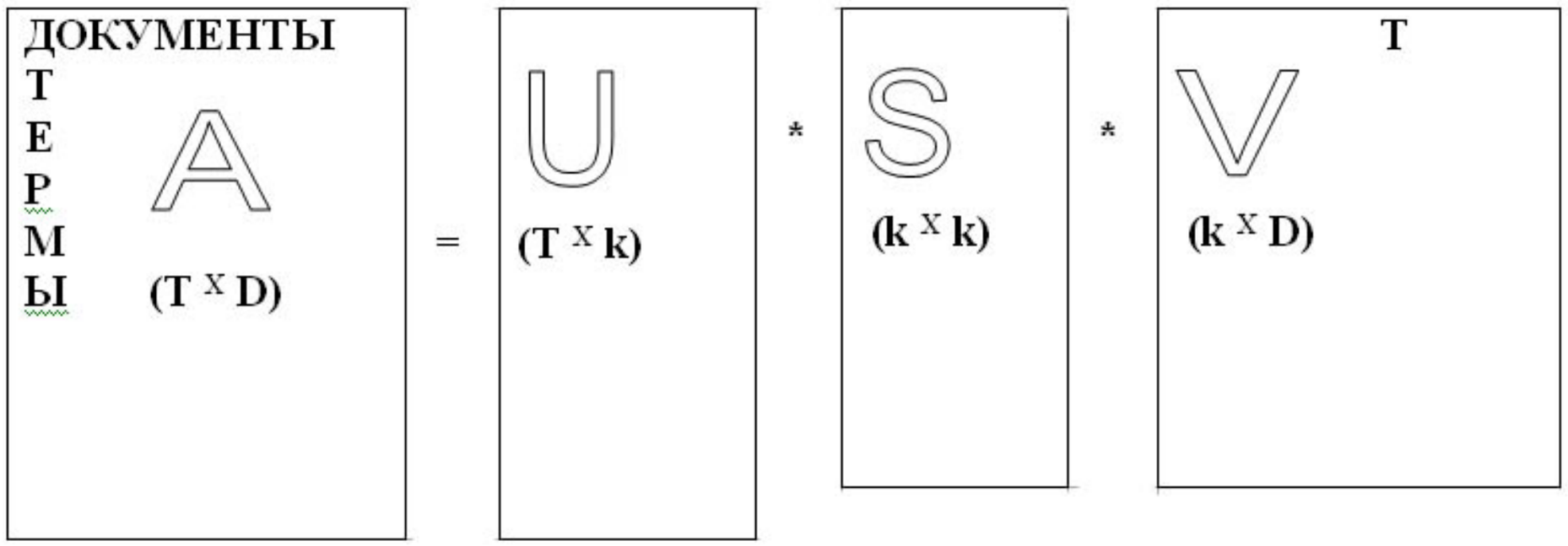
где матрицы U и V – ортогональные, а S – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы A. Символ T в обозначении матрицы означает транспонирование матрицы.

Транспонированная матрица — **матрица** , полученная из исходной **матрицы** заменой строк на столбцы.
Например:

$$\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}1 & 3\\2 & 4\end{bmatrix} \quad \text{и} \quad \begin{bmatrix}1 & 2\\3 & 4\\5 & 6\end{bmatrix}^{\mathrm{T}} = \begin{bmatrix}1 & 3 & 5\\2 & 4 & 6\end{bmatrix}$$

https://www.youtube.com/watch?v=JEYLfIVvR9I

Латентно-семантический анализ (LSA)



SVD - разложение матрицы A размерности (T * D) на матрицу термов U размерности (T * k), матрицу документов V размерности (k * D) и диагональную матрицу S размерности (k * k), где k – количество сингулярных значении диагональной матрицы S.

Выбор k зависит от поставленной задачи и подбирается эмпирически. Он зависит от количества исходных документов. Если документов не много, например сотня, то k можно брать 5-10% от общего числа диагональных значений; если документов сотни тысяч, то берут 0,1-2%.

Вероятностная модель коллекции документов (PLSA)

В отличии от LSA, PLSA работает с вероятностями появления слов и документов в темах. Для расчета этих вероятностей используется совершенно другой математический аппарат, который основан на правиле Байеса.

Вероятностная модель коллекции документов.
Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, \ldots, w_{nd})$ из словаря W . Термин может повторяться в документе много раз.

Вероятностное пространство и гипотеза независимости. Предполагается, что существует конечное множество тем T , и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. Коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения p(d, w, t), заданного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ является латентной (скрытой) переменной.

Вероятностная модель коллекции документов (PLSA)

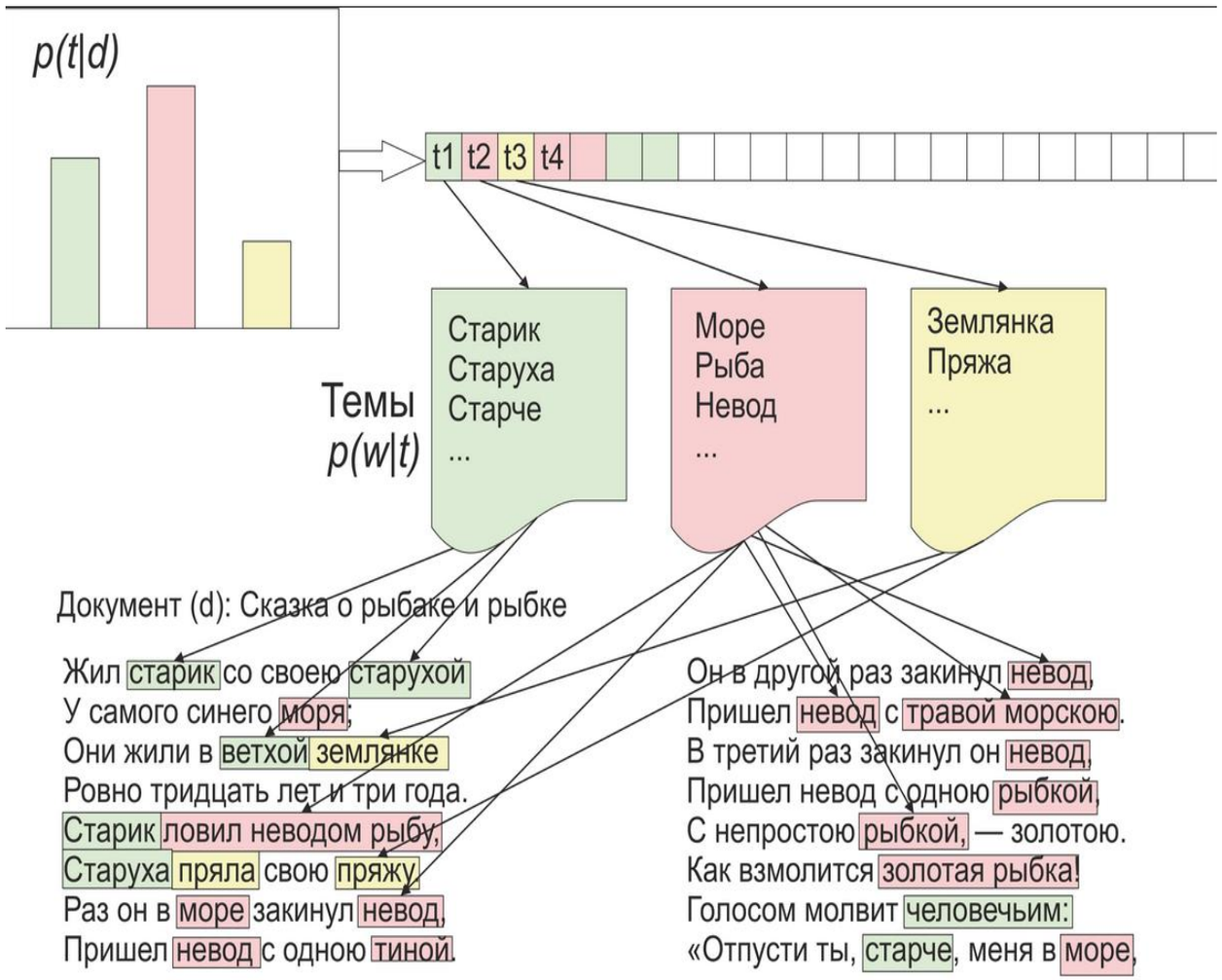
Bag of words.
Гипотеза о независимости элементов выборки эквивалентна предположению,
что порядок терминов в документах не важен для выявления тематики, то есть тематику документа можно узнать даже после произвольной перестановки терминов, хотя для человека такой текст теряет смысл. Это предположение называют гипотезой «мешка слов» (bag of words). Порядок документов в коллекции также не имеет значения; это предположение называют гипотезой «мешка документов».

Постановка задачи тематического моделирования.
Построить тематическую модель коллекции документов D — значит найти множество тем T , распределения $p(w \mid t)$ для всех тем $t \in T$ и распределения
$p(t \mid d)$ для всех документов $d \in D$. Можно также говорить о задаче совместной «мягкой» кластеризации множества документов и множества слов по множеству кластеров-тем. Мягкая кластеризация означает, что каждый документ или термин не жёстко приписывается какой-то одной теме, а распределяется по нескольким темам. Найденные распределения используются затем для решения прикладных задач. Распределение $p(t| d)$ является удобным признаковым описанием документа в задачах информационного поиска, классификации и категоризации документов.

Вероятностная модель коллекции документов (PLSA)

Вероятность появления слова w в документе d, то есть p(w | d) описывается следующей формулой:

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



p(w | d) – наблюдаемые величины.
p(t | d) – вероятность принадлежности документа к теме, нам не известно.
p(w | t) – вероятность принадлежности слова к теме, нам не известно. Таким образом, задачей тематического моделирования является задача восстановления распределений **p(t | d)** и **p(w | t)** на основе **p(w | d)** .

Вероятностная модель коллекции документов (PLSA)

Модель мешка слов – текст представлен в виде слов, расположение слов **не важно.**

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ – дискретное вероятностное пространство
- коллекция D – выборка троек $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$
- d_i, w_i – наблюдаемые, темы t_i – скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документа d:

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

Дано $\hat{p}(w|d) \equiv n_{dw}/n_d$, **найти:**

- $\phi_{wt} \equiv p(w|t)$ – распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ – распределение тем в документах $d \in D$.

Принцип максимума правдоподобия

Принцип максимума правдоподобия. Для оценивания параметров Φ, Θ тематической модели по коллекции документов D будем максимизировать правдоподобие (плотность распределения) выборки:

$$p(D;\Phi,\Theta) = C\prod_{d\in D}\prod_{w\in d} p(d,w)^{n_{dw}} = \prod_{d\in D}\prod_{w\in d} p(w\,|\,d)^{n_{dw}}\underbrace{Cp(d)^{n_{dw}}}_{\text{const}} \to \max_{\Phi,\Theta},$$

Прологарифмируем p(D; Φ, Θ), чтобы превратить произведения в суммы. Получим задачу максимизации логарифма правдоподобия (log-likelihood) при ограничениях неотрицательности и нормированности столбцов. Нормированность означает, что сумма вероятностей принадлежности документа ко всем темам равна 1, и сумма вероятностей всех слов по одной теме также равна 1.

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

$$\sum_{w\in W}\varphi_{wt} = 1; \qquad \varphi_{wt}\geqslant 0;$$

$$\sum_{t\in T}\theta_{td} = 1; \qquad \theta_{td}\geqslant 0.$$

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

Для решения задачи в PLSA применяется итерационный процесс, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization) Перед первой итерацией выбирается начальное приближение параметров ϕ_{wt} , θ_{td}.

На **E-шаге** по текущим значениям параметров ϕ_{wt}, θ_{td} с помощью формулы Байеса вычисляются условные вероятности p(t | d, w) всех тем t ∈ T для каждого термина w ∈ d в каждом документе d:

$$p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s \in T} \varphi_{ws}\theta_{sd}}.$$

На **M-шаге**, наоборот, по условным вероятностям тем p(t | d, w) вычисляется новое приближение параметров ϕ_{wt}, θ_{td}.

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt},$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt},$$

Таким образом гоняя E и M шаги можно рассчитать p(t | d) – вероятность принадлежности документа к теме, p(w | t) – вероятность принадлежности слова к теме.

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

Алгоритм PLSA-EM: рациональный EM-алгоритм для модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ, Φ;
Выход: распределения Θ и Φ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ и Φ не сойдутся;

Мультиномиальное распределение и Распределение Дирихле

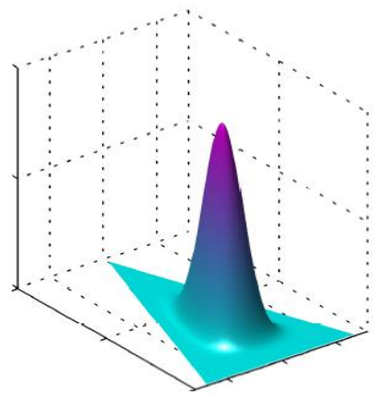
Мультиномиальное распределение имеет следующий вид (совместное распределение вероятностей случайных величин):

$$P(\mathrm{x} \mid \mathrm{p}) = \frac{n!}{\Pi_{i=1}^{K} x_i !} \Pi_{i=1}^{K} p_i^{x_i}$$

где x_i независимые одинаково распределенные случайные величины, p_i - функции распределений случайных величин.

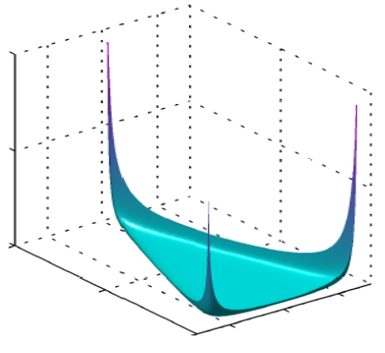
Распределение Дирихле имеет следующий вид:

$$P(\mathrm{p};\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$



Это означает, что если априорное распределение P является распределением Дирихле Dir(p;α), и x сгенерировано мультиномиальным распределением, то апостериорное распределение p(p x, α) также является распределением Дирихле:

$$P(\mathrm{p} \mid \mathrm{x}, \alpha) = Dir(\mathrm{p} \mid \mathrm{x} + \alpha) = \frac{1}{B(\mathrm{x} + \alpha)} \prod_{i=1}^{\alpha} p_i^{x_i + \alpha_i - 1}$$



Латентное размещение Дирихле (LDA)

Термины: **D** – пространство документов, **W** – пространство слов, **Z** – пространство тем. Темы являются скрытыми параметрами, которые должны быть найдены. Причем оценка основана на двух вещах: 1. Оценка производится как математическое ожидание. 2. В качестве функций используются мультиномиальные функции и функции Дирихле.

$$p(w \mid z, \beta) = \int p(w \mid z, \Phi)\, p(\Phi \mid \beta)\, d\Phi \qquad p(z \mid \alpha) = \int p(z \mid \Theta)\, p(\Theta \mid \alpha)\, d\Theta$$

Θ, Φ : матрица слова – темы и документы – темы.
Эти матрицы могут быть найдены двумя способами.

Модель LDA (вариационный подход)

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

Семплирование по Гиббсу

$$P(z_i = j \mid w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta} \cdot \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + \alpha T}$$

Регуляризация достигается за счет введения в уравнения функции Дирихле

Латентное размещение Дирихле (LDA) – Gibbs sampling

$$p(\mathrm{z},\mathrm{w}\mid\alpha,\beta)=\mathrm{p}(\mathrm{w}\mid z,\beta)\cdot\mathrm{p}(\mathrm{z}\mid\alpha)\qquad P(\mathrm{z}_i=j\mid w_i=m,\mathrm{z}_{-i},w_{-i})\approx\frac{C^{WT}_{m,j}+\beta}{\sum_{m'}C^{WT}_{m',j}+V\beta}\cdot\frac{C^{DT}_{d,j}+\alpha}{C^{DT}_{d,j'}+\alpha T}$$

$C^{WT}_{m,j}$ - Матрица; в каждой ячейке находится число сколько раз слово **w** было связанно с темой **t**,

$C^{DT}_{d,j}$ - Матрица; в каждой ячейке находится число сколько раз слово **w** в документе **d** связанно с темой **t**.

$\sum_{m'}C^{WT}_{m',j}=n_t$ - Вектор; в каждой ячейке находится общее число слов связанно с темой **t**,

$C^{DT}_{d,j'}=n_d$ - Дина документа **d** словах.

Результат моделирования:

1. Распределение слов по темам.
2. Распределение документов по темам.

$$\theta_{dj}=\frac{C^{DT}_{d,j}+\alpha}{C^{DT}_{d,j'}+T\alpha}$$

$$\phi_{m,j}=\frac{C^{WT}_{m,j}+\beta}{\sum_{m'}C^{WT}_{m',j}+V\beta}$$

Латентное размещение Дирихле (LDA)

Corpus

Documents D

d_1 d_2 d_3 d_4 d_5 d_n

Terms $w_{d,i}$

Lorem ipsum food lore ipsum safety lore GMO ipsum safe lore organic ipsum food regulation lore ipsum health lore genetically pesticides ipsum consumer lore ipsum farmer lore

Algorithm

LDA

ϕ

θ

Results

V	Topic_1	Topic_2	…	Topic_K
food	0.09	0.07	…	0.02
health	0.10	0.05	…	0.01
GMO	0.04	0.01	…	0.20
safety	0.03	0.03	…	0.05
farmer	0.01	0.10	…	0.07
…	…	…	…	…

D	Topic_1	Topic_2	…	Topic_K
d_1	0.10	0.04	…	0.02
d_2	0.05	0.03	…	0.02
d_3	0.02	0.03	…	0.09
d_4	0.07	0.05	…	0.01
…	…	…	…	…
d_D	0.08	0.01	…	0.10

Алгоритм сэмплирования

На входе: коллекция документов D, число тем |T|, Число итераций;
Initialization: φ(w,t), θ(t,d) для всех документов и слов d ∈ D, w ∈ W , t ∈ T;

Внешний цикл по документам (i). Длинна цикла=числу документов

Внутренний цикл. Длина цикла = количество слов в текущем документе.

1. Берем документ i.
2. Выбираем слово **k** из документа **i**.
3. Вычисляем номер темы **t** для слова **k.**

$$P(z_i = j \mid w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta} \cdot \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + \alpha T}$$

3.1. Вычисляем величину P(z)

для текущего слова и для каждой темы P(1…T), то есть присваиваем текущему слову определенный номер темы.

Конец внутреннего цикла.

Апдейтинг всех счетчиков

Конец внешнего цикла.

Расчет матриц φ(w,t), θ(t,d) но основании счетчиков.

$$\theta_{dj} = \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + T\alpha} \qquad \phi_{m,j} = \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta}$$

Результат: распределение слов по темам

Words with high probability

	1	2	3	4	5	6	7	8
1	ислам: 0,014347	образование: 0,003430	социальный: 0,006741	фотография: 0,018203	известный: 0,007305	власть: 0,019949	медведь: 0,007718	на: 0,039698
2	принимать: 0,013508	выступление: 0,003430	ru: 0,004531	сделать: 0,009500	мир: 0,006276	россия: 0,018902	алгоритм: 0,003370	reuters: 0,025063
3	ар: 0,007635	черт: 0,002323	гибельный: 0,002321	фотограф: 0,008775	премия: 0,005247	путин: 0,011571	лес: 0,003370	модель: 0,024453
4	photo: 0,007635	захват: 0,002323	белая: 0,002321	эта: 0,007325	кандидат: 0,005247	оппозиция: 0,011571	король: 0,002283	от: 0,018355
5	имя: 0,006796	жаловаться: 0,002323	обязательство: 0,002321	парк: 0,006599	нобелевский: 0,005247	сша: 0,011048	пикник: 0,002283	показ: 0,017135
6	год: 0,005957	дестабилизация: 0,002323	ваш: 0,002321	национальный: 0,005874	список: 0,004219	страна: 0,009477	какой-либо: 0,002283	коллекция: 0,0146
7	ты: 0,005118	nstarikov: 0,002323	счастие: 0,002321	заповедник: 0,005149	включать: 0,004219	демократия: 0,007906	слово: 0,002283	упасть: 0,011647
8	взять: 0,004279	физиологический: 0,002323	бессмысленный: 0,002321	африка: 0,005149	столица: 0,003190	под: 0,007383	отдыхать: 0,002283	инвалид: 0,011037
9	группа: 0,003440	родин: 0,002323	обама: 0,001216	рассказывать: 0,004424	номинантов: 0,003190	революция: 0,007383	несколько: 0,002283	мода: 0,010427
10	арт: 0,003440	развиваться: 0,002323	директива: 0,001216	отдыхать: 0,003699	лонг: 0,003190	война: 0,007383	книга: 0,002283	в: 0,009818
11	дейв: 0,003440	пусть: 0,002323	приближаться: 0,001216	южный: 0,003699	лист: 0,003190	голосовать: 0,006859	начало: 0,002283	февраль: 0,009818
12	блэйки: 0,003440	временить: 0,002323	привлекательный: 0,0012	индий: 0,003699	мэннинг: 0,002161	против: 0,006859	спокойно: 0,002283	полицейский: 0,009
13	стоун: 0,003440	что: 0,002323	неудобно: 0,001216	тупик: 0,003699	брэдли: 0,002161	этап: 0,006859	семейный: 0,002283	неделя: 0,009818
14	африка: 0,003440	лексический: 0,001217	зацеплять: 0,001216	сова: 0,002973	юлий: 0,002161	видео: 0,006335	вернуться: 0,002283	ла-пас: 0,008598
15	член: 0,003440	ржавый: 0,001217	маммограмму: 0,001216	рак: 0,002973	скрывать: 0,002161	народ: 0,005812	голодный: 0,002283	фотография: 0,007
16	q: 0,002601	баглан: 0,001217	лента: 0,001216	national: 0,002973	русский: 0,002161	оранжевый: 0,005812	на: 0,002283	боливия: 0,007378
17	tip: 0,002601	алсиндор: 0,001217	полуостров: 0,001216	рысь: 0,002973	тимошенко: 0,002161	общество: 0,005812	кровавый: 0,002283	во: 0,006769
18	нация: 0,002601	жать: 0,001217	дамба: 0,001216	малыш: 0,002973	великий: 0,002161	чего: 0,004241	ролик: 0,002283	время: 0,006769
19	али: 0,002601	камал: 0,001217	джами: 0,001216	побережье: 0,002973	отмечать: 0,002161	новый: 0,004241	мы: 0,002283	david: 0,006769
20	включать: 0,002601	зодиакальный: 0,001217	julie: 0,001216	род: 0,002973	его: 0,002161	манипуляция: 0,003717	изначально: 0,001196	подиум: 0,006159
21	мохаммед: 0,002601	контракт: 0,001217	жилье: 0,001216	птица: 0,002973	себе: 0,002161	снайпер: 0,003717	abbyy: 0,001196	mercado: 0,006159
22	x: 0,002601	демонстрантка: 0,001217	багор: 0,001216	park: 0,002973	падение: 0,002161	интернет: 0,003717	comas: 0,001196	путь: 0,004939
23	сейчас: 0,002601	нестабильный: 0,001217	бессрочный: 0,001216	запечатлевать: 0,002973	помещение: 0,002161	гражданский: 0,003717	дополнять: 0,001196	творение: 0,004939
24	коран: 0,002601	бесплодие: 0,001217	зерно: 0,001216	серенгети: 0,002973	комитет: 0,002161	выбор: 0,003717	гаммакурта: 0,001196	набрасываться: 0,0
25	верить: 0,002601	халена: 0,001217	невинный: 0,001216	под: 0,002973	буда: 0,002161	интерес: 0,003194	фашистский: 0,001196	падение: 0,004330

Каждая колонка это распределение слов. Соответственно просматривая эти колонки можно выбрать нужны темы для анализа.

Результат: распределение документов по темам

Documents with high probability

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	48: 0,421482	84: 0,146154	34: 0,106557	57: 0,596774	31: 0,452247	76: 0,435268	27: 0,236979	15: 0,906034	7: 0,391587	50: 0,563670	45: 0,140244	81: 0,077778	34: 0,139344	4: 0,350147
2	68: 0,064103	47: 0,042763	69: 0,080508	29: 0,145320	71: 0,087209	97: 0,323529	34: 0,090164	21: 0,748333	6: 0,339714	88: 0,218333	95: 0,083929	84: 0,069231	39: 0,054245	34: 0,057377
3	67: 0,054124	81: 0,033333	46: 0,060345	37: 0,077320	41: 0,069588	92: 0,306250	56: 0,087963	86: 0,100000	5: 0,325306	2: 0,206537	96: 0,062319	99: 0,058442	46: 0,043103	46: 0,043103
4	86: 0,047368	45: 0,030488	86: 0,047368	42: 0,066832	87: 0,059677	91: 0,292526	81: 0,055556	26: 0,087838	3: 0,051165	82: 0,048387	72: 0,053867	87: 0,043548	71: 0,040698	56: 0,041667
5	92: 0,043750	83: 0,030201	78: 0,041985	34: 0,040984	77: 0,055556	64: 0,284483	45: 0,042683	56: 0,078704	39: 0,049528	77: 0,043210	55: 0,053030	56: 0,041667	41: 0,038660	60: 0,034545
6	34: 0,040984	46: 0,025862	98: 0,041667	86: 0,036842	99: 0,045455	85: 0,253817	29: 0,036946	33: 0,049043	82: 0,048387	45: 0,042683	37: 0,046392	67: 0,038660	26: 0,033784	80: 0,028846
7	79: 0,033333	56: 0,023148	55: 0,037879	81: 0,033333	66: 0,034247	96: 0,242029	95: 0,030357	34: 0,040984	25: 0,044479	83: 0,030201	93: 0,039474	72: 0,031768	99: 0,032468	27: 0,028646
8	47: 0,029605	97: 0,022876	37: 0,036082	27: 0,023438	81: 0,033333	60: 0,241818	35: 0,026814	47: 0,036184	9: 0,043534	71: 0,029070	86: 0,036842	79: 0,023810	84: 0,023077	47: 0,023026
9	11: 0,026786	80: 0,022436	40: 0,035865	55: 0,022727	74: 0,023504	61: 0,240809	89: 0,026699	81: 0,033333	77: 0,043210	11: 0,026786	26: 0,033784	52: 0,022124	3: 0,022879	39: 0,021226
10	46: 0,025862	39: 0,021226	56: 0,023148	45: 0,018293	47: 0,023026	63: 0,224138	53: 0,019481	92: 0,031250	42: 0,042079	47: 0,023026	67: 0,028351	50: 0,020599	80: 0,022436	11: 0,020833
11	26: 0,020270	53: 0,019481	89: 0,021845	71: 0,017442	52: 0,022124	93: 0,223684	30: 0,017949	35: 0,023659	66: 0,029680	74: 0,019231	90: 0,027778	64: 0,020115	53: 0,019481	50: 0,020599
12	99: 0,019481	71: 0,017442	77: 0,018519	63: 0,017241	58: 0,022013	83: 0,218121	63: 0,017241	69: 0,021186	4: 0,029087	36: 0,018519	32: 0,027273	96: 0,018841	63: 0,017241	43: 0,018719
13	63: 0,017241	63: 0,017241	45: 0,018293	90: 0,016667	35: 0,020505	95: 0,205357	82: 0,016129	45: 0,018293	78: 0,026718	27: 0,018229	63: 0,017241	92: 0,018750	49: 0,016432	63: 0,017241
14	82: 0,016129	90: 0,016667	71: 0,017442	82: 0,016129	63: 0,017241	79: 0,195238	74: 0,014957	41: 0,018041	46: 0,025862	30: 0,017949	82: 0,016129	71: 0,017442	82: 0,016129	82: 0,016129
15	58: 0,015723	60: 0,016364	63: 0,017241	98: 0,013889	50: 0,016854	86: 0,194737	98: 0,013889	71: 0,017442	84: 0,023077	42: 0,017327	76: 0,015625	63: 0,017241	95: 0,016071	87: 0,014516
16	98: 0,013889	82: 0,016129	60: 0,016364	56: 0,013889	82: 0,016129	94: 0,183019	93: 0,013158	63: 0,017241	99: 0,019481	63: 0,017241	64: 0,014368	29: 0,017241	98: 0,013889	64: 0,014368
17	93: 0,013158	9: 0,015948	82: 0,016129	52: 0,013274	28: 0,015823	75: 0,159898	68: 0,012821	61: 0,016544	53: 0,019481	65: 0,016883	79: 0,014286	48: 0,016960	56: 0,013889	98: 0,013889
18	89: 0,012136	86: 0,015789	8: 0,013980	20: 0,013235	98: 0,013889	69: 0,148305	42: 0,012376	100: 0,016484	92: 0,018750	40: 0,014768	98: 0,013889	82: 0,016129	62: 0,013393	93: 0,013158
19	91: 0,011598	11: 0,014881	25: 0,013804	93: 0,013158	62: 0,013393	67: 0,146907	66: 0,011416	82: 0,016129	45: 0,018293	79: 0,014286	56: 0,013889	38: 0,015487	52: 0,013274	68: 0,012821
20	76: 0,011161	64: 0,014368	20: 0,013235	68: 0,012821	93: 0,013158	66: 0,139269	76: 0,011161	11: 0,014881	63: 0,017241	98: 0,013889	62: 0,013393	91: 0,014175	93: 0,013158	72: 0,012431
21	81: 0,011111	98: 0,013889	93: 0,013158	69: 0,012712	27: 0,013021	72: 0,122928	40: 0,010549	64: 0,014368	43: 0,016223	93: 0,013158	52: 0,013274	31: 0,014045	27: 0,013021	78: 0,011450
22	74: 0,010684	52: 0,013274	68: 0,012821	10: 0,012521	68: 0,012821	78: 0,110687	96: 0,010145	79: 0,014286	33: 0,015550	68: 0,012821	41: 0,012887	98: 0,013889	68: 0,012821	81: 0,011111

Каждая колонка это распределение документов. Соответственно просматривая эти колонки можно выбрать нужны темы для анализа

Проблема стабильности ТМ. Неоднозначность матричного разложения

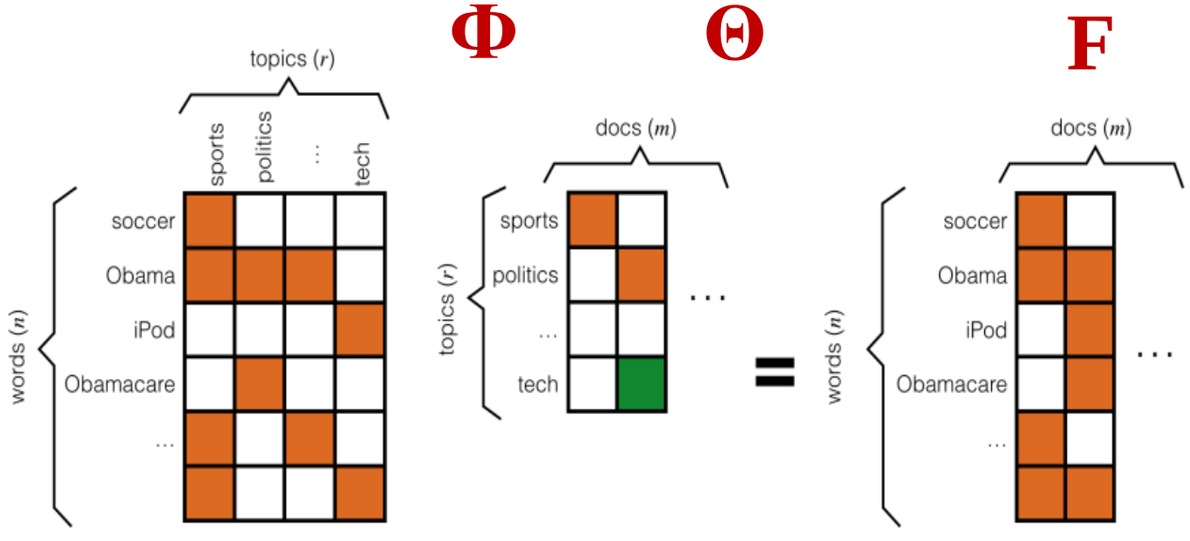
$$F[documents \times words] = \Theta[documents \times topics] \cdot \Phi[topics \times words]$$

Матрица F датасет (строки – документы, колонки список уникальных слов). Большой датасет может быть представлен в виде произведения двух относительно небольших матриц Φ and Θ. **Однако**:

$$F = \Theta \cdot \Phi = (\Theta \cdot R) \cdot (R^{-1}\Phi) = \Theta^{'} \cdot \Phi^{'}$$

Матрица может быть представлена в виде комбинации различных матриц, но такого же размера.

Это значит, что исходному набору слов и документов могут быть сопоставлены разные тематические решения, но имеющие одинаковое количество тем. Содержимое матриц и могут отличатся при разных матрицах преобразования.

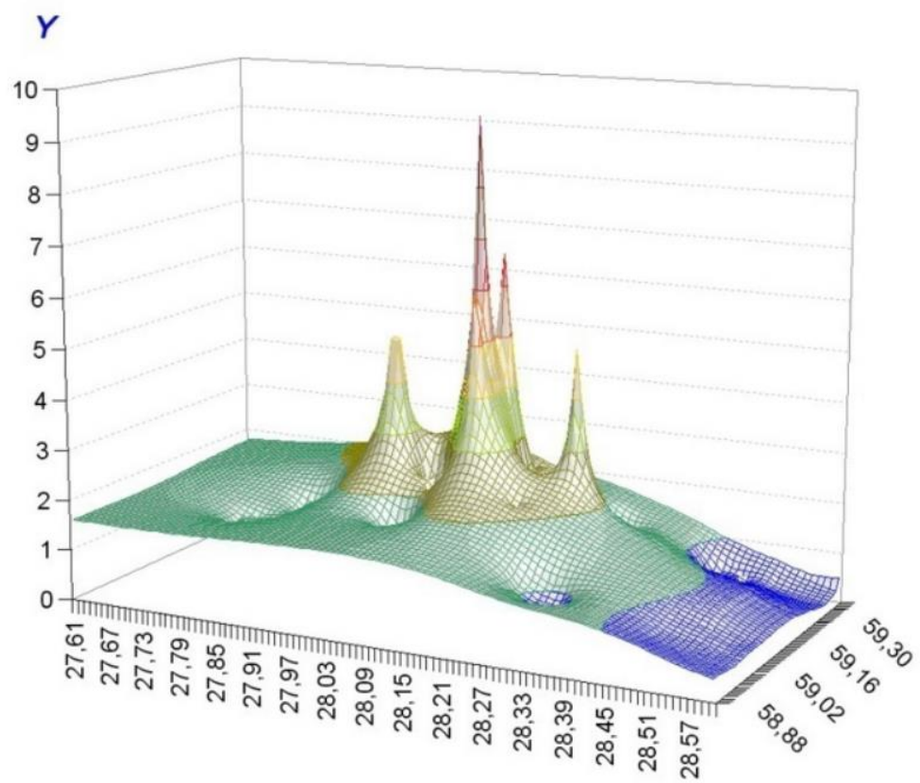


Проблема стабильности ТМ. Множество локальных минимумов и максимумов

$$\theta_{\mathrm{m,k}} = \int p(\mathrm{t}\,|\,\mathrm{d})\,\mathrm{p}(\theta_d, \alpha) = \frac{n(k;m) + \alpha_k}{(\sum_{k=1}^{K} \Omega(d;\mathrm{k}) + \alpha_k)}$$

$$\phi_{\mathrm{k},t} = \frac{n(\mathrm{t};\mathrm{k}) + \beta_t}{(\sum_{t'=1}^{V} n(\mathrm{t}';\mathrm{k}) + 1 + \beta_{t'})}$$

Под интегралом стоит произведение функций Дирихле. Итоговое подинтегральное выражение имеет множество локальным максимумов и минимумов.



Так как расчет интегралов основан на методе Монте - Карло, и сводится к подсчету счетчиков, то в итоге значения счетчиков могут существенно отличатся при разных запусках процедуры сэмплирования. Это связанно, с тем что в ходе сэмплирования не все минимумы могут быть обойдены, и процесс сэмплирования может приводит к тому, что итоговые матрицы распределений слов и документов по темам будут отражать какой то один (или несколько) минимум.

Результаты сравнения стабильности некоторых моделей

Topic model	Topic quality metrics		Topic stability metrics	
	coherence	tf-idf coherence	stable topics	Jaccard
pLSA	-237.38	-126.08	54	0.47
pLSA + Φ sparsity reg., $\alpha = 0.5$	-230.90	-126.38	9	0.44
PLSA + Θ sparsity reg., $\beta = 0.2$	-240.80	-124.09	87	0.47
LDA, Gibbs sampling	-207.27	-116.14	77	0.56
LDA, variational Bayes	-254.40	-106.53	111	0.53
SLDA	-208.45	-120.08	84	0.62
GLDA, $l = 1$	-183.96	-125.94	195	0.64
GLDA, $l = 2$	-169.36	-122.21	195	0.71
GLDA, $l = 3$	-163.05	-121.37	197	0.73
GLDA, $l = 4$	-161.78	-119.64	200	0.73

RESULT: (1) регуляризация может существенно влиять на результаты тематического моделирования. Регуляризация может как улучшать так и убивать стабильность тематического моделирования.

(модель LDA является регуляризованной версией pLSA, где регуляризация заключается в факте добавления информации о Dirichlet функциях).

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru