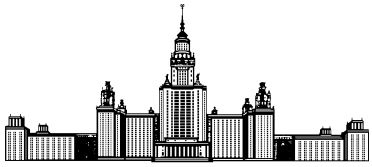


Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ШАДРИКОВ Андрей Алексеевич

Алгоритмы неотрицательных матричных разложений для тематического моделирования

ДИПЛОМНАЯ РАБОТА

Научный руководитель:
д.ф-м.н., доцент
К. В. Воронцов

Москва, 2015

Содержание

1 Введение

Задача обработки текстовой информации и естественных языков в целом непроста, но очень перспективна как задача машинного обучения. Она может иметь воздействия в областях анализа данных, где появляется необходимость в информационном поиске и интерпретации языковых особенностей.

Предполагается, что каждый написанный на естественном языке документ состоит не просто из набора слов, а из некоторых тем, которые раскрываются написанными словами. Поиск такой скрытой информации и является целью тематического моделирования. Представив набор текстов как матрицу частот встречаемости слов в документах, можно искать интересующие темы с помощью методов матричного разложения. Таким образом тематическое моделирование является более узкой задачей матричного разложения, где матрицы предполагаются неотрицательными.

Для решения обеих задач было предложено множество методов оптимизации, в частности тематическая модель вероятностного семантического анализа [14] или модифицированный метод градиентного спуска [16]. Но большинство методов из-за своей итерационной природы имеют схожие проблемы, как например остановка в локальных оптимумах и зависимость от начального приближения. Задача выявления скрытых тем предполагает участие человека для оценки качества модели по списку наиболее часто встречаемых слов в теме. На практике обученные модели часто сложны для интерпретации человеком из-за того, что некоторые темы непонятны, состоят из слишком большого числа общеупотребимых слов, имеют в себе слова из совершенно разных областей.

Одной из модификацией модификацией методов является поиск матриц с определённой структурой. Этого можно добиться, добавляя в задачу ограничения разреженности или декорреляции [7] или более жёстких требований однозначного соответствия между темами и небольшим набором слов [2], что модифицирует оптимизируемый функционал и позволяет избежать локальных оптимумов с нежелательными значениями параметров. С другой стороны, добиться лучшего решения можно с помощью выбора хорошего начального приближения, избежав попадания в неудачную область параметров.

Целью данной работы является анализ влияния инициализации методов неотрицательного матричного разложения и тематического моделирования на их сходимость и интерпретируемость получаемой модели.

Для инициализации рассматриваются способы, основанные на понятии якорного слова, что даёт инициализируемым параметрам структурность. Задача тематического моделирования невыпукла, и общие методы выдают лишь локальные оптимумы. Поиск хорошего начального приближения позволяет приблизиться к глобальному оптимуму.

Качество модели измеряется не только по минимизируемому функционалу (в большинстве задач неотрицательного матричного разложения по норме Фробениуса), но и по перплексии, которая используется в задаче тематического моделирования. Как показано в [18] экспертные оценки интерпретируемости коррелируют с точечной взаимной информацией. Кроме этого используются новые критерии интерпретируемости, показывающих различность найденных тем.

В разделе 2 приводится формальная постановка задач неотрицательного матричного разложения и тематического моделирования. В разеделе 3 приводится краткий обзор методов неотрицательного матричного

разложения и тематического моделирования. В разделе 4 рассматривается метод нахождения якорных слов и построения с их помощью тематической модели. В разделе 5 вводятся и исследуются в экспериментах способы инициализации, основанные на понятии якорного слова. В разделе 6 приводится новый способ визуализации обученной тематической модели с помощью проекции на плоскость.

Эксперименты на коллекциях англоязычных статей блога KOS и научной конференции NIPS [17] показали, что инициализация моделей не только позволяет достичь лучшего решения с точки зрения оптимизируемого функционала, но и повысить интерпретируемость модели.

2 Определения и обозначения

Исходную матрицу, содержащую частоты терминов, будем обозначать $V \in \mathbb{R}^{N \times M}$, где N — число уникальных терминов в коллекции, а M — число документов. Данная матрица неотрицательна $V_{ij} \geqslant 0$ и является стохастической $\sum_{p=1}^{N} V_{pj} = 1 \forall j$.

Искомые матрицы обозначим как $W \in \mathbb{R}^{N \times T}$ и $H \in \mathbb{R}^{T \times M}$. Эти матрицы также неотрицательные и стохастические, поскольку столбцы матриц задают распределения.

Введём меру сходства матриц $D(A, B)$, с помощью которой будем определять качество аппроксимации. Тогда можно записать проблему тематического моделирования как следующую задачу оптимизации:

$$\begin{aligned} \min_{W,H} \quad & D(V, WH) \\ \text{при условии} \quad & W_{ik} \geqslant 0 \; \forall i, k \\ & H_{kj} \geqslant 0 \; \forall k, j \\ & \textstyle\sum_i W_{ik} = 1 \; \forall k \\ & \textstyle\sum_k H_{kj} = 1 \; \forall j \end{aligned} \tag{1}$$

В качестве функции $D(A, B)$ потерь часто рассматриваются:

- норма Фробениуса:

$$D(A, B) = \|A - B\|_F^2 = \sum_{i,j} (A_{ij} - B_{ij})^2$$

- дивергенция Кульбака-Лейблера:

$$D(A, B) = KL(A||B) = \sum_{i,j} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right)$$

Задачу неотрицательного матричного разложения (НМР) можно записать с теми же обозначениями следующим образом:

$$\begin{aligned} \min_{W,H} \quad & D(V, WH) \\ \text{при условии} \quad & W_{ik} \geqslant 0 \; \forall i, k \\ & H_{kj} \geqslant 0 \; \forall k, j \end{aligned} \tag{2}$$

3 Существующие методы

Первые шаги в формировании тематического моделирования как отдельного класса задач обработки текстов были сделаны в работе [15], где был предложен простой метод решения задачи, основанный на индексировании числа встречаемых в текстах терминов. Предложенный метод

был модернизирован в ставший популярным метод вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA, [14]).

Последующие изменения методов можно грубо разделить на два класса: первые добавляли дополнительные априорные знания в модель (например, априорное распределение на слова, как в работе [4]), вторые добавляли регуляризацию на оптимизируемые параметры (например, декорреляция столбцов матрицы W [7]). Однако, как показано в [22], большинство изменений моделей, внесением априорных знаний, можно описать как некоторые регуляризаторы.

Большинство предложенных в литературе вариантов решения задачи тематического моделирования — итерационные методы блочно-координатной оптимизации исходной задачи. На каждой итерации фиксируется одна из матриц и решается задача оптимизации по другой матрице, затем наоборот. Так продолжается, пока не достигнется максимальное число итераций или не выполнится необходимый критерий останова. Общий способ решения приведён в алгоритме 1.

Алгоритм 1. Общий вид итерационного алгоритма решения задачи неотрицательного матричного разложения

Вход: матрица V, T, $\#$ итераций $iter_{\max}$;

Выход: матрицы W и H;

Инициализировать W_{ik}, H_{kj} $\forall i, k, j$;

для всех $iter = 1, \ldots, iter_{\max}$

  $H^{new} = G(V, W^{old}, H^{old})$;

  $W^{new} = F(V, W^{old}, H^{new})$;

Рассмотрим несколько примеров методов решения задачи матричного разложения.

3.1 Вероятностный латентный семантический анализ

Метод PLSA является примером решения задачи тематического моделирования 1, поэтому для вывода формул обновления матриц используется метод максимизации логарифма правдободобия:

$$\max_{W,H} \sum_{i,j} V_{ij} \ln \left(\sum_k W_{ik} H_{kj} \right)$$

$$\begin{aligned} W_{ik} &\geqslant 0,\ \forall i,k; \qquad \textstyle\sum_i W_{ik} = 1\ \forall k \\ H_{kj} &\geqslant 0,\ \forall k,j; \qquad \textstyle\sum_k H_{kj} = 1\ \forall j \end{aligned}$$

Решается данная оптимизационная задача с помощью EM-алгоритма, каждая итерация которого состоит из двух шагов. На E-шаге фиксируются матрицы W и H и высчитываются элементы трёхмерного массива:

$$n_{ikj} = V_{ij} \frac{W_{ik} H_{kj}}{\sum_{s=1}^{T} W_{is} H_{sj}} \tag{3}$$

На M-шаге посчитанные значения используются для обновления матриц W и H:

$$W_{ik} = \frac{n_{ik}}{n_k} \equiv \frac{\sum_j n_{ikj}}{\sum_{ij} n_{ikj}}, \qquad H_{kj} = \frac{n_{kj}}{n_j} \equiv \frac{\sum_i n_{ikj}}{\sum_{ik} n_{ikj}}, \tag{4}$$

Реализация такого метода проблематична из-за необходимости хранить значения n_{ikj}, которых вообще говоря $N \times T \times M$, что значительно больше не только искомых матриц W и H, но и исходной матрицы V.

Можно заметить, что в М-шаге для обновление матрицы H суммирование n_{ikj} по первому индексу приводит к формуле:

$$\sum_i n_{ikj} = \sum_i \frac{V_{ij} W_{ik} H_{kj}}{\sum_{s=1}^{T} W_{is} H_{sj}} = H_{kj} \sum_i W_{ik} \frac{V_{ij}}{\sum_{s=1}^{T} W_{is} H_{sj}} \tag{5}$$

Если ввести матрицу $F \in \mathbb{R}^{N \times M}$, $F_{ij} = \frac{V_{ij}}{\sum_{s=1}^{T} W_{is} H_{sj}}$, то итоговый вид формулы (5) будет следующий:

$$\sum_i n_{ikj} = H_{kj} \left(W^{\mathsf{T}} F \right)_{kj} \tag{6}$$

Подобным образом преобразуется и суммирование по последнему индексу в формуле обновления матрицы W:

$$\sum_j n_{ikj} = W_{ik} \left(F H^{\mathsf{T}} \right)_{ik} \tag{7}$$

Выведенные формулы позволяют отказаться от необходимости хранить $N \times T \times M$ значений одновременно, и делают шаги обновления матриц подходящими для общего алгоритма решения задачи тематического моделирования.

3.2 Градиентный спуск с мультипликативным шагом

В работе [16] был предложен специальный вид градиентного спуска, где шаг градиента не подбирается на каждой итерации путём решения дополнительной задачи оптимизации, а считается таким образом, чтобы обновление элементов матриц происходило не с помощью суммирования, а с помощью умножения.

В зависимости от выбранной метрики D для задачи неотрицательного матричного разложения формулы шагов получаются разными. Для нормы Фробениуса итерация будет следующей:

$$W_{ik} = W_{ik}\frac{(VH^\mathsf{T})_{ik}}{(WHH^\mathsf{T})_{ik}}, \qquad H_{kj} = H_{kj}\frac{(W^\mathsf{T}V)_{kj}}{(W^\mathsf{T}WH)_{kj}} \tag{8}$$

При этом выражения для дивергенции Кульбака-Лейблера принимают более громоздкую форму, и вычисление обновлённых матриц становится более ресурсозатратным:

$$W_{ik} = W_{ik}\frac{\sum_j H_{kj}\frac{V_{ij}}{(WH)_{ik}}}{\sum_m H_{km}}, \qquad H_{kj} = H_{kj}\frac{\sum_i W_{ik}\frac{V_{ij}}{(WH)_{ik}}}{\sum_n W_{nk}} \tag{9}$$

Можно заметить сильное сходство данных формул с шагом алгоритма PLSA (6), (7). Разница лишь в двух мелочах. Во-первых нормировка матриц в PLSA происходит таким образом, чтобы они оставались стохастическими. В формуле (9) нормировка происходит для поддержания стохастичности только произведения WH. Во-вторых PLSA обновляет обе матрицы сразу, а градиентный спуск по очереди.

3.3 Чередующиеся наименьшие квадраты

Поскольку общий алгоритм предполагает оптимизацию по каждой из матриц W, H независимо, то это почти сразу же приводит к варианту метода, когда в качетве нового приближения выдаётся решение линейной системы $V = WH$. Данный подход называется чередующиеся наименьшие квадраты (Alternating Least Squares, ALS) и используется не только в задачах матричного разложения. Однако первые применения метода к задаче НМР были предложены не так давно [19, 8]. Введя операцию об-

нуления отрицательных элементов и обозначив её $[\bullet]_+$, можно записать шаг обновления матриц следующим образом:

$$W^\mathsf{T} \leftarrow \left[\left(HH^\mathsf{T}\right)^{-1} HV^\mathsf{T}\right]_+$$
$$H \leftarrow \left[\left(W^\mathsf{T}W\right)^{-1} W^\mathsf{T}V\right]_+$$

3.4 Иерархические чередующиеся наименьшие квадраты

Задачу (2) не обязательно решать, оптимизируя по целым матрицам. Можно каждую итерацию оптимизировать по строке H или столбцу W. Тогда, взяв в качестве метрики норму Фробениуса, функционал удобнее переписать в следующем виде:

$$D(V\,,\,WH) = \|V - \sum_{k=1}^{T} w_k h_k\|_F^2 \tag{10}$$

За w_k обозначен столбец W, а за h_k — строка H. Данный метод исследовался независимо как иерархические чередующиеся квадраты (Hierarchical Alternating Least Squares, HALS [10, 9]) и одноранговая итерация невязки (Rank-one Residue Iteration, RRI [13]). Поскольку каждая задача оптимизации в отдельности является выпуклой, есть гарантия, что каждая найденная пара матриц W, H будет являться стационарной точкой для задачи НМР (2). При этом обновление строк H и столбцов W выглядят следующим образом:

$$h_k = \frac{\left[w_k^\mathsf{T}\left(V - \sum_{s=1,s\neq k} w_s h_s\right)\right]_+}{\|w_k\|_2^2} \tag{11}$$

$$w_k = \frac{\left[\left(V - \sum_{s=1, s\neq k} w_s h_s\right) h_k^{\mathsf{T}}\right]_+}{\|h_k\|_2^2} \tag{12}$$

Можно заметить, что в отличие от ALS, для обновления очередной строки матрицы H не нужно ждать обновления всей матрицы W и наоборот, что позволяет последовательно обновлять блоки тем в обеих матрицах сразу, что отображено в алгоритме 2.

Алгоритм 2. Шаг обновления матриц W и H с помощью метода HALS.

Вход: матрицы V, W, H, число тем T;

Выход: обновлённые матрицы W и H;

для всех $k = 1, \ldots, T$

    Обновить k-ю стоку матрицы H по формуле (11) ;

    Обновить k-й столбец матрицы W по формуле (12) ;

3.5 Применение методов NMP для тематического моделирования

Из-за схожести формулировок задачи неотрицательного матричного разложения и тематического моделирования следует ожидать, что методы NMP будут применимы в задаче тематического моделирования с минимальными изменениями. Для их использования в решении задачи тематического моделирования необходимо добавить шаг проекции на допустимое множество. В нашем случае этой проекцией будет простая нормировка матриц:

$$W_{ik} = \frac{W_{ik}}{\sum_{p=1}^{N} W_{pk}};\ H_{kj} = \frac{H_{kj}}{\sum_{q=1}^{T} H_{qj}} \tag{13}$$

4 Якорные слова и изменение задачи

Попробуем рассмотреть проблему нахождения тем с другой стороны. Допустим, в каждой теме должно быть такое слово, чтобы при его встрече в документе можно было сразу сказать о принадлежности темы к этому документу. Таким образом мы приходим к понятию якорных слов.

Определение 1. *Понятие якорного слова можно ввести несколькими эквивалентными способами:*

- *В матрице W для любого столбца k можно выделить якорного слово (строку) i такое, что $W_{ik} > 0$, $W_{is} = 0$, $\forall s \neq k$*
- *В матрице W можно выделить диагональную подматрицу размера $T \times T$, строки которой состоят из якорных слов.*
- *Для каждой темы существует якорное слово, которое исключительно выделяет данную тему.*

Введение нового понятия добавляет новые ограничения для задачи тематического моделирования. Теперь необходимо находить не просто неотрицательные стохастические матрицы. В матрице W должна выделяться диагональная подматрица.

Такая постановка задачи приводит к появлению у получаемых решений важных свойств. Главным из них является то, что при найденных якорных словах нет необходимости оптимизировать по матрицам W, H

в отдельности несколько итераций. Всего за одну итерацию метод тут же попадает в локальный оптимум [2]. Более того, в работах [12, 3] показывается, что найденный оптимум окажется глобальным при выполнении следующих условий::

- Строки матрицы V должны быть векторами из T-мерного симплекса.
- Каждая вершина T-мерного симплекса не должна лежать близко к линейной оболочке других.
- В «реальной» матрице W, с помощью которой по предположению была составлена матрица V, выделяется диагональная подматрица размера $T \times T$.
- Строки «реальной» матрицы H линейно независимы.

Под «реальными» матрицами в условиях подразумевается, что подаваемая на вход метода матрица V является произведением неизвестных «реальных» матриц W и H. К сожалению обычно на реальных данных не выполняется даже самое первое условие, потому что ранг матрицы V гораздо больше задаваемого числа тем T, что оставляет проблему тематического моделирования открытой.

4.1 Поиск якорных слов

Поиск якорных слов происходит быстрее работы итерационных алгоритмов матричного разложения [1]. Таким образом, основанные на поиске якорных слов методы должны выигрывать как минимум в скорости работы. Для нахождения необходимого набора терминов (строк матрицы

V), в работе [1] предлагается алгоритм 3. Иллюстрацию доказательства работы алгоритма можно увидеть на рис. 1.

Алгоритм 3. Алгоритм нахождения якорных слов

Вход: N точек $\{d_1, d_2, \ldots, d_N\}$, $\#$ якорных слов T;

Выход: якорные слова S;

$S \leftarrow \{d_i | d_i = \arg\max_i \|d_i\|\}$;

для всех $i = 1, \ldots, T-1$

    Берём d_i наиболее удалённое от $\text{span}(S)$;

    $S \leftarrow S \cup \{d_i\}$;

Переименуем вектора: $S = \{u_1, u_2, \ldots, u_T\}$;

для всех $i = 1, \ldots, T$

    Берём d_i наиболее удалённое от $\text{span}(\{u_1, \ldots, u_T\} \setminus u_i)$;

    Заменяем u_i на d_i;

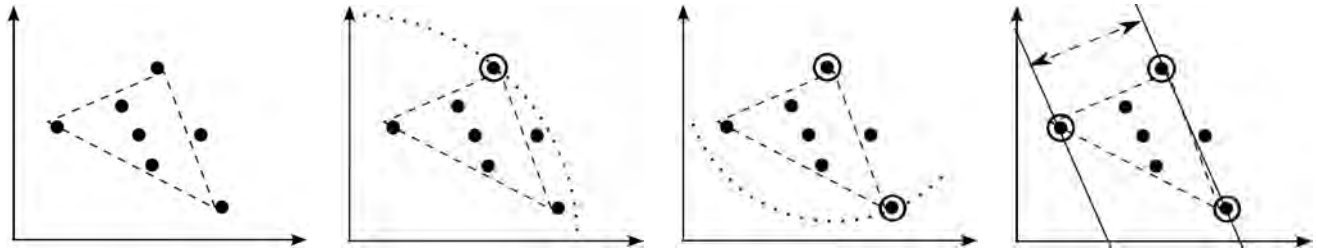


Рис. 1: (Взято из работы [1].) Иллюстрация работы алгоритма 3 нахождения якорных слов. Пример приводится для поиска приближения 3-мерного симплекса. Вначале выбирается вектор с наибольшой нормой. Затем наиболее удалённый от выбранного. И наконец последним берётся наиболее удалённый от выпуклой оболочки первых выбранных векторов. Оставшиеся вне симплекса точки считаются погрешностью и в последующих итерациях проецируются внутрь симплекса.

4.2 Использование якорных слов для решения задачи тематического моделирования

Полученные якорные слова можно использовать для ускорения получения решения задачи тематического моделирования. Поскольку найденные вектора являются строчками в V, то найдя якорные слова, мы фактически восстановили матрицу H. Действительно, если были выполнены условия 4, то строки матрицы V представляют из себя линейную комбинацию строк матрицы H. Так как в матрице W выделяется диагональная подматрица размера $T \times T$ и строчки H линейно независимы, то часть строк V являются помноженными на некоторый ненулевой коэффициент строками H.

Поскольку строки матрицы V из T-мерного симплекса, а строки H линейно независимы, то получаем, что строки H и есть вершины того симплекса, из которого матрица V. В работе [1] также доказано, что алгоритм 3 выдаёт вершины T-мерного симплекса.

Остаётся лишь по полученной матрице H восстановить матрицу W. Это можно сделать с помощью решения оптимизационной задачи, например, описанной в алгоритме 4.

Алгоритм 4. Алгоритм восстановления матрицы W по найденным якорным словам

Вход: Матрица V, набор якорных слов S

Выход: матрица W

1 Нормируем матрицу V по строкам для формирования $\hat{V}$;

2 Сохраним нормировочные костанты в p_w ;

3 Обозначим за s_k индекс для k-го якорного слова.;

4 **для всех** $i = 1, \ldots, N$

5 Решить оптимизационную задачу $C_i = \arg\min_{C_i} D\left(\hat{V}_i, \sum_{k \in S} C_{i,k} \hat{V}_{s_k}\right)$, с ограничениями $C_{i,k} \geqslant 0$, $\sum_k C_{i,k} = 1$;

6 $W \leftarrow diag(p_w)C$;

5 Инициализация тематической модели

Модель якорного слова — мощный инструмент для быстрого нахождения решения задачи тематического моделирования. Но требования выделения диагональной подматрицы может оказаться достаточно жёстким. Действительно, для однозначного определения темы иногда одного слова может быть недостаточно, особенно, если эта тема не является узко специализированной с большим набором общих слов.

Однако полученные якорные слова можно использовать в качестве начального приближения для итерационных методов. Для иллюстрации восстановления матрицы W, отвечающую за распределение слов в те-

мах, удобно использовать модельные данные, где эта матрица имеет ярко выраженную структуру. Пусть при этом в ней не будет выделяться диагональная подматрица, чтобы итерации методов после нахождения якорных слов имели смысл. Сгенерированную матрицу можно видеть на рис. 2. Матрицу H важно сгенерировать с линейно независимыми строчками, чтобы итоговая матрица V имела ранг не меньше T.

Для имитации работы с зашумлёнными данными матрицу информации $\hat{V}$ будем получать не произведением W на H, а генерировать с помощью порождающей тематической модели, описанной в алгоритме 5.

Алгоритм 5. Порождающая тематическая модель

Вход: Матрицы W, H

Выход: матрица частот терминов $\hat{V}$

для всех $j = 1, \ldots, M$
  для всех $i = 1, \ldots, N$
    $z \leftarrow$ Выбрать тему из распределения $H_{:,j}$;
    $\hat{V}_{ij} \leftarrow$ Выбрать термин из распределения $W_{:,z}$;

Полученные данные будем подавать на вход рассмотренным в разделе 3 методам, чтобы определить какие из методов лучше восстанавливают реальную матрицу W. Поскольку столбцы W задают распределения, сравнивать их будем с помощью расстояния Хеллигера:

$$H(W^1, W^2) = \frac{1}{\sqrt{2}} \sum_{k=1}^{T} \sqrt{\sum_{i=1}^{N} \left(\sqrt{W^1_{ik}} - \sqrt{W^2_{ik}} \right)^2} \tag{14}$$

Итоговые результаты работы методов, усреднённые по трём запускам, можно увидеть на рис. 2. Видно, что лучше всего сработал метод чередующихся наименьших квадратов (ALS), восстановив основную структуру W. Метод градиентного спуска с мультипликативным прави-

лом (MU) вместе с вероятностным латентным семантическим анализом (PLSA) сумели восстановить матрицу очень грубо.

5.1 Инициализация якорными словами

Из эксперимента на модельных данных видно, что не все методы достаточно хорошо восстанавливают структуру реальной матрицы W. Это можно исправить, проинициализировав их матрицами, которые в себе уже содержат часть искомой структуры. Метод поиска якорных слов, описанный в алгоритме 4 позволяет находить структурированную матрицу W. Использование найденной матрицы в качестве начального приближения для итерационных методов может улучшить их сходимость и итоговый результат.

Из рис. 3 видно, что все методы, кроме ALS восстановили матрицу W ближе к реальной. Это позволяет предположить, что начальное приближение должно иметь некоторую структуру, которая определяет желаемое решение задачи тематического моделирования.

5.2 Инициализация якорными ядрами

Проблема поиска якорных слов заключается в установленных ограничениях в их определении. Не всегда одно слово может однозначно определять тему. Если, однако, считать, что любая тема может быть однозначно определена некоторым конечным набором слов, то матрица W, сформированная таким образом, будет иметь другую структуру. В ней будет выделяться не диагональная, а d-диагональная подматрица.

Определение 2. *Совокупность наборов слов* $\{W^k | k = 1, \ldots, T\}$*, удовлетворяющих следующим свойствам:*

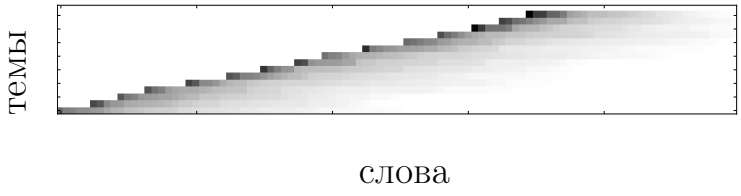
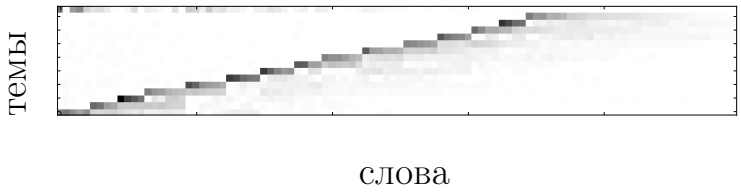
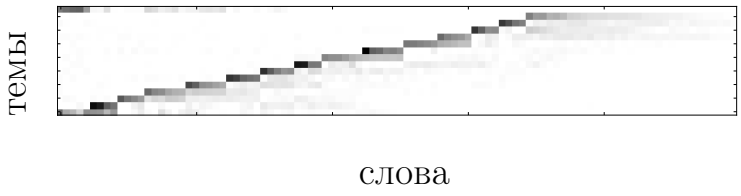
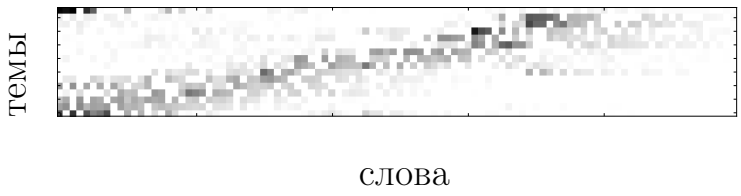
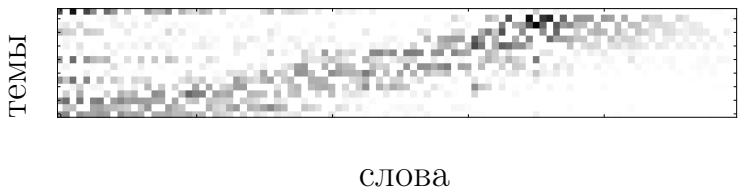
Метод	Изображение матрицы	Расстояние Хеллингера
Реальная матрица		0
ALS		0.1028
HALS		0.1555
MU		0.3126
PLSA		0.3179

Рис. 2: Сравнение восстановленных матриц W различными методами со случайной инициализацией.

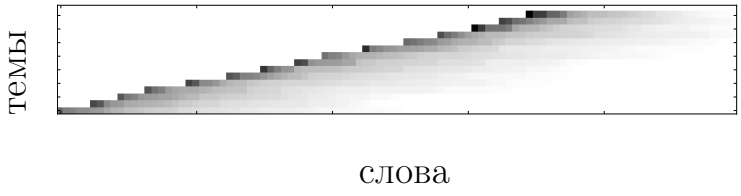
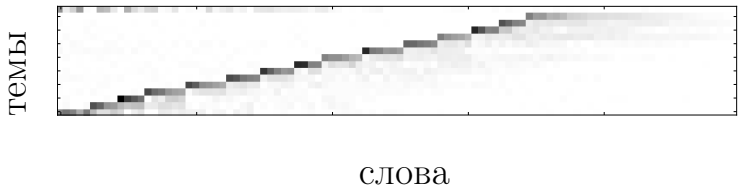
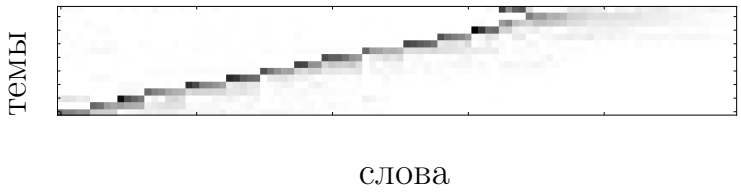
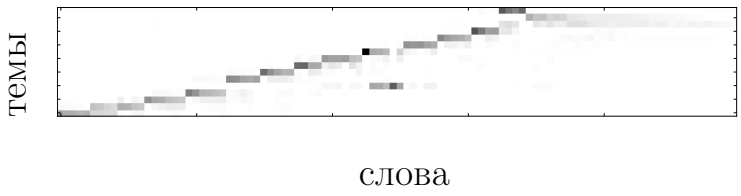
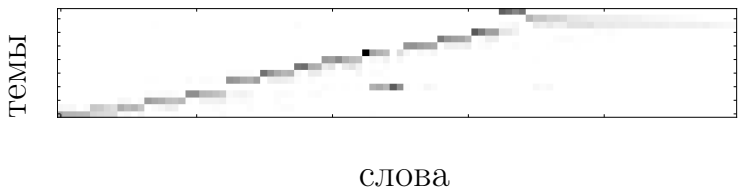
Метод	Изображение матрицы	Расстояние Хеллингера
Реальная матрица		0
ALS		0.1215
HALS		0.1472
MU		0.2787
PLSA		0.2986

Рис. 3: Сравнение восстановленных матриц W различными методами, проинициализиарованными якорными словами.

1. $W^k \not\subset W^s \; \forall s \neq k$;

2. $W_i^k > 0 \; \forall i$;

3. $\forall s \neq k \; \exists i : W_i^k > W_i^s$,

будем называть якорными ядрами тем.

Для поиска якорных ядер также можно использовать алгоритм 4, немного модифицировав матрицу V, которая подаётся на вход алгоритму поиска якорных слов 3. Все слова кластеризуются несколько раз на разное число кластеров не обязательно точным, но быстрым алгоритмом, например k-means. Затем из полученных центроидов всех проведённых кластеризаций формируется матрица $\hat{V}$, в которой ищатся якорные слова. Полученная после алгоритма 4 матрица $\hat{W}$ используется для получения матрицы H, решая следующую задачу оптимизации:

$$\begin{aligned} \min_H \quad & D(\hat{V}, \hat{W}H) \\ \text{при условии} \quad & H_{kj} \geqslant 0 \; \forall k, j \\ & \textstyle\sum_k H_{kj} = 1 \; \forall j \end{aligned}$$

Затем наоборот, зафиксировав H, ищем матрицу W:

$$\begin{aligned} \min_W \quad & D(V, WH) \\ \text{при условии} \quad & W_{ik} \geqslant 0 \; \forall i, k \\ & \textstyle\sum_i W_{ik} = 1 \; \forall k \end{aligned}$$

Полученная пара W, H и есть начальное приближения, основанное на якорных ядрах. Итоговый способ поиска ядер можно видеть в алгоритме 6. Хорошей эвристикой для множества M числа кластеров оказалось

использование числа кластеров от числа тем T до трети числа слов $\frac{N}{3}$, каждый раз умножая число тем в два раз.

Алгоритм 6. Алгоритм нахождения якорных ядер тем.

Вход: матрица V, число тем T, множество значений мощности кластеров M;

Выход: матрицы W и H;

C — матрица с нулевым числом строк ;

для всех $k \in M$

  Кластеризовать строки матрицы V быстрым k-means на k кластеров ;

  Центроиды добавить как строки в матрицу C ;

$\hat{W} \leftarrow$ якорные слова матрицы C ;

$H \leftarrow \underset{H \geqslant 0}{\arg\min} D(C, \hat{W}H)$;

$W \leftarrow \underset{W \geqslant 0}{\arg\min} D(V, WH)$;

Из рис. 4 видно, что подобный способ инициализации способствует улучшению работы всех методов по сравнению с инициализацией якорными словами.

5.3 Эксперименты

Вычислительные эксперименты производились на двух коллекциях данных, которые не подвергались дополнительной предварительной обработке:

- Коллекция KOS, состоящая из $M = 3430$ статей политического блога Daily Kos. Объём словаря $N = 6906$.

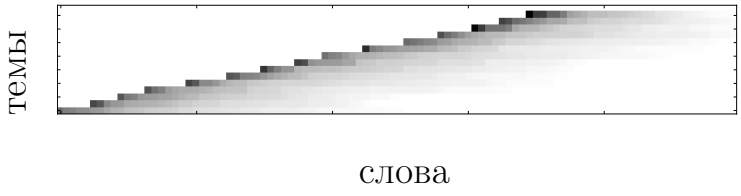
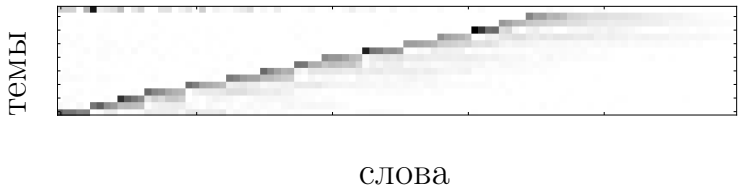
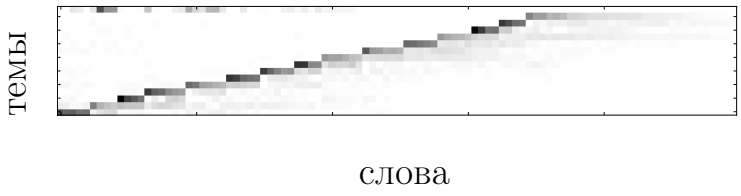
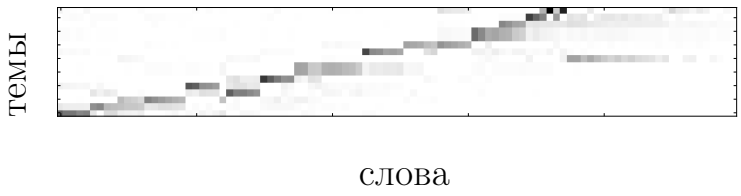
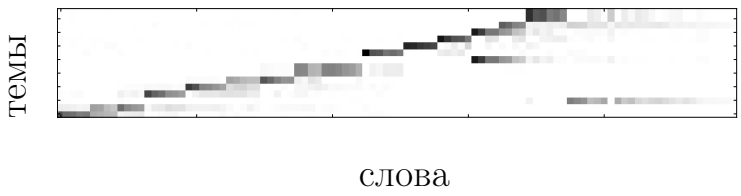
Метод	Изображение матрицы	Расстояние Хеллингера
Реальная матрица		0
ALS		0.1044
HALS		0.1343
MU		0.2415
PLSA		0.2762

Рис. 4: Сравнение восстановленных матриц W различными методами, проинициализиарованными якорными ядрами.

- Коллекция NIPS, состоящая из $M = 1500$ статей научной конференции Neural Information Processing Systems. Объём словаря $N = 12419$.

Для оценки качества работы методов кроме нормы Фробениуса вычислялась перплексия, по которой принято сравнивать методы тематического моделирования:

$$P(V, W, H) = \exp\left(-\frac{1}{n}\sum_{i,j} V_{ij} \ln \sum_{k} W_{ik} H_{kj}\right) \tag{15}$$

На рис. 5–8 изображены значения метрик качества в зависимости от номера итерации каждого из рассмотренных методов. Цветом обозначены методы, а стилем линии — способ инициализации. Пункцирной линией обозначена случаная инициализация. Число тем T везде было выбрано равным 25. Видны преимущества неслучайной инициализации: методы значительно быстрее сходятся, причём к лучшим оптимумам.

К сожалению метрики качества нельзя использовать для определения интерпретируемости тематической модели. Для её измерения в работе [18] использовалась точечная взаимная информация (PMI), показывающая насколько связаны друг с другом пары различных слов:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{16}$$

Для каждой темы выбирается некоторое число самых вероятных слов, между которыми считается PMI. Поскольку для каждой темы PMI получается различный, имеет смысл ввести два критерия:

- MPMI — среднее значение PMI по всем темам.
- IPMI — среднее значение максимальных PMI по всем темам.

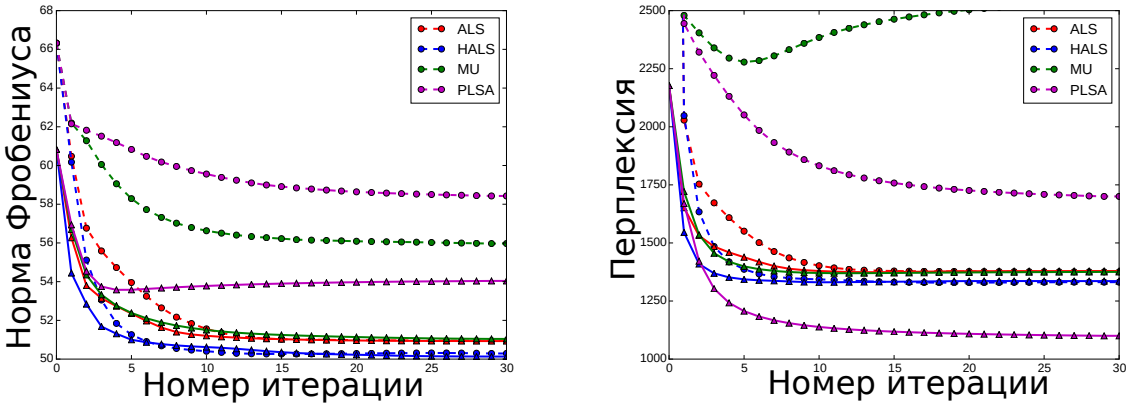


Рис. 5: Значения метрик во время обучения моделей на данных KOS, инициализированными якорными словами.

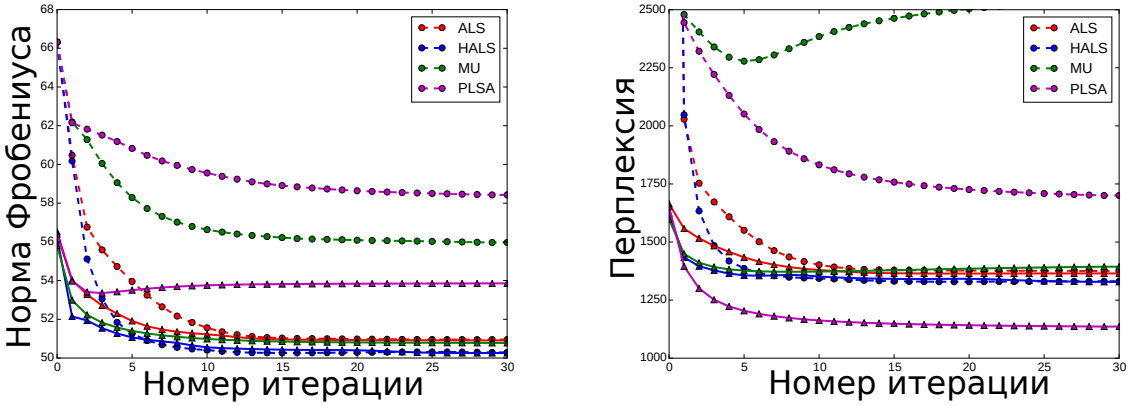


Рис. 6: Значения метрик во время обучения моделей на данных KOS, инициализированными якорными ядрами.

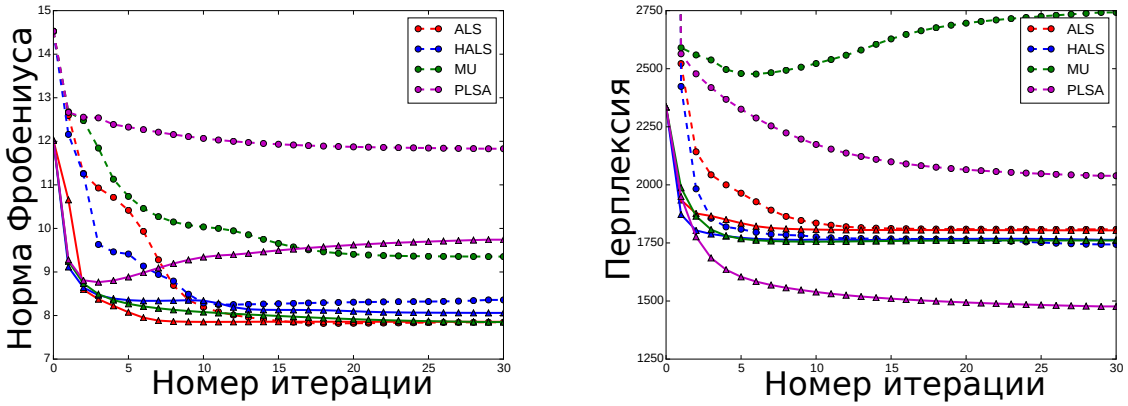


Рис. 7: Значения метрик во время обучения моделей на данных NIPS, инициализированными якорными словами.

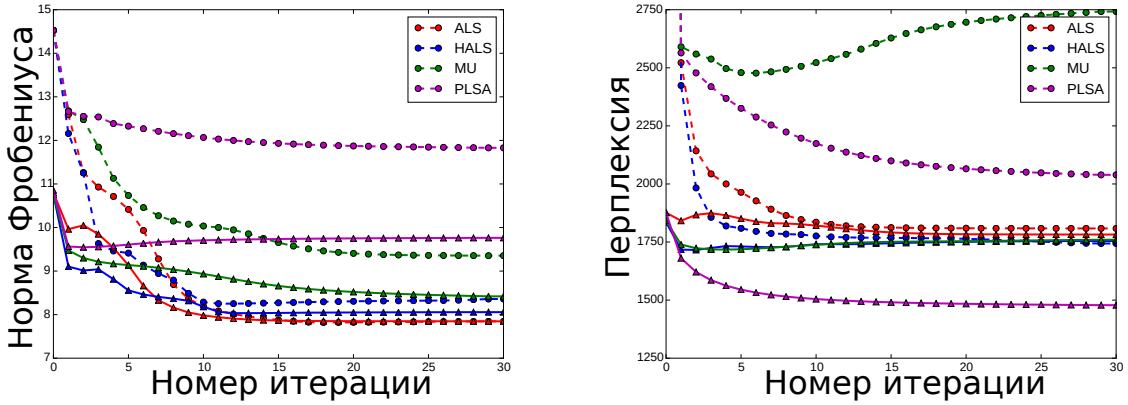


Рис. 8: Значения метрик во время обучения моделей на данных NIPS, инициализированными якорными ядрами.

Кроме точечной взаимной информации введём критерии, которые показывают различность получаемых тем. Будем считать расстояние Хеллингера (14) от каждой темы до ближайшей. В качестве метрики между темами возьмём расстояние.

- AverH — среднее расстояние между всеми темами.
- MeanH — среднее расстояние до ближайшей темы.
- MinH — минимальное расстояние до ближайшей темы.

В таблице 1 приведены значения метрик качества на коллекции NIPS. Можно заметить, что, несмотря на лучшие значения метрик, метод может быть хуже по другим критериям. Можно сделать следующие выводы:

- ALS лучший для уменьшения нормы Фробениуса, при этом почти не зависит от начального приближения.
- PLSA достигает самой низкой перплексии и самого высокого MPMI при использовании инициализации якорными ядрами.
- HALS показывает наивысший IPMI, при неслучайной инициализации, а также самые высокие AverH и MeanH по сравнению с остальными методами.
- MU хорошо минимизирует лишь норму Фробениуса и отличается от ALS меньшей перплексией и большим AverH.
- Инициализация якорными ядрами показывает лучше значения на критериях в 15 случаях из 20, чем инициализация якорными словами.

		Frob	Perp	MPMI	IPMI	AverH	MeanH	MinH
случайная	ALS	**7.849**	1807	0.761	3.871	0.794	0.718	0.657
	HALS	8.122	1742	1.024	3.878	0.816	0.716	0.612
	MU	9.346	2661	0.417	3.581	**0.833**	**0.754**	0.703
	PLSA	11.79	2013	0.491	1.901	0.785	**0.757**	**0.727**
я. слова	ALS	**7.856**	1803	0.801	3.904	0.792	0.718	0.643
	HALS	8.128	1761	1.004	**4.187**	**0.822**	0.727	0.562
	MU	**7.846**	1759	0.651	3.832	0.802	0.681	0.567
	PLSA	9.812	**1475**	1.065	3.025	0.734	0.619	0.469
я. ядра	ALS	**7.849**	1791	0.797	3.861	0.791	0.714	0.646
	HALS	8.089	1742	1.017	4.009	**0.824**	0.737	0.607
	MU	8.315	1791	0.741	3.931	0.816	0.706	0.604
	PLSA	9.639	**1456**	**1.091**	2.754	0.756	0.662	0.562

Таблица 1: Значения критериев на коллекции NIPS при $T = 25$ темах и пятидесяти итераций каждого метода, усреднённые по трём запускам.

6 Визуализация тематической модели

Результаты решения задачи тематического моделирования часто сравнивают не только по значениям метрик качества таких, как перплексия или норма Фробениуса. Для определения качества построенной модели также привлекаются эксперты, которые по самым вероятным словам в теме пытаются её охарактеризовать [6]. При этом очень важным является способ визуализации тематической модели. В литературе чаще всего это делается с помощью вывода столбцов матрицы W в отсортированном по уменьшению значений порядке. Схожий подход, но с интерактивным интерфейсом для просмотра связей между документами, использовался в работе [5].

Однако упорядоченные списки частот слов в темах не позволяют быстро понять общую картину тематической модели, а именно насколько похожи те или иные темы, особенно если в них встречаются одинаковые слова. Визуализированные на плоскости данные могут восприниматься существенно быстрее. Для проектирования многомерных данных на плоскость есть множество методов [11, 20] начиная с обычного сжатия данных с помощью PCA и до более продвинутых способов, использующих возможную структурную информацию данных. Но большинство подобных методов пытаются сохранить глобальную структуру данных вместе с локальной, что совершенно не нужно для визуализации тематической модели. Сжатием с сохранением именно локальной структуры занимается метод t-SNE (t-distributed stochastic neighbor embedding, [21]), который работает за счёт попытки сблизить распределения векторов в исходном пространстве с распределением векторов на плоскости.

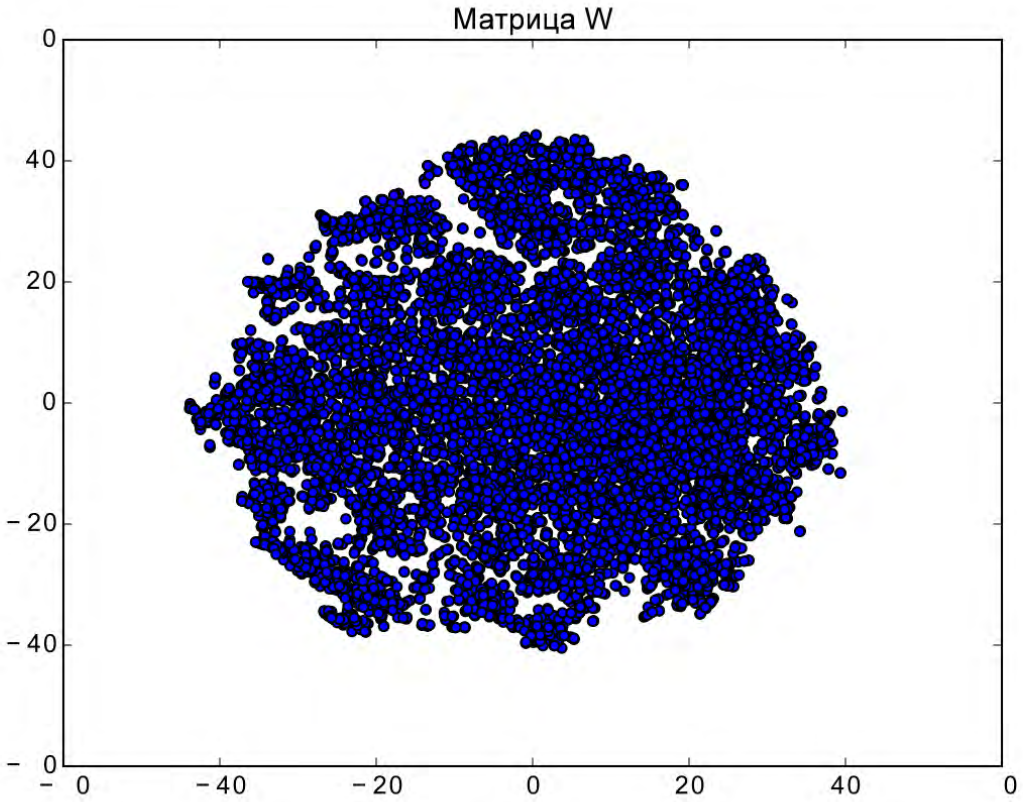


Рис. 9: Визуализация матрицы W после работы алгоритма ALS на данных NIPS. Каждая точка — проекция строки матрицы W на плоскость.

По рис. 9 можно убедиться в том, что изображение восстановленной матрицы само по себе не является визуализацией тематической модели, так как нет выделяемых тем и все слова выглядят одинаково. Необходимо визуализировать каждую тему с выделением порядка встречаемости слов в теме. Для этого сжатую с помощью t-SNE матрицу W будем изображать столько раз, сколько у неё столбцов (число тем), каждый раз выделяя интенсивностью цвета вероятность встретить слова в текущей теме. Пример изображения обученных тематических моделей можно видеть на рис. 10 – 11.

Важной особенностью t-SNE является то, что если данные разделимы в исходном пространстве, они будут разделимы после проекции на плоскость. Этот факт позволяет утверждать, что при получении изображения тематической модели, на которой точки разделяются на ярко выраженные кластеры, сами темы в построенной модели будут также сильно отличаться, что заметно повышает интерпретируемость модели для человека.

Интересно посмотреть на рис. 11, на котором модель определила две темы, которые состоят из имён и фамилий авторов статей, которые вошли в коллекцию NIPS. При этом эти темы явно выделяются среди остальных спроецированных тем.

К сожалению данный способ визуализации хорошо применим только к небольшому числу тем, поскольку наложение нескольких цветов в одной точке на изображении значительно уменьшают число цветов, которыми можно обозначить темы.

Тема 0		Тема 1		Тема 5		Тема 10	
kerry	0.134	november	0.063	administration	0.032	senate	0.076
bush	0.065	house	0.014	president	0.024	elections	0.026
general	0.017	poll	0.013	cheney	0.016	race	0.016
poll	0.014	senate	0.013	house	0.013	carson	0.013
edwards	0.013	republicans	0.013	white	0.010	coburn	0.010
polls	0.013	governor	0.012	national	0.006	oklahoma	0.010
voters	0.012	electoral	0.012	officials	0.006	results	0.010
john	0.012	account	0.012	commission	0.006	obama	0.009
results	0.010	polls	0.011	intelligence	0.005	illinois	0.009

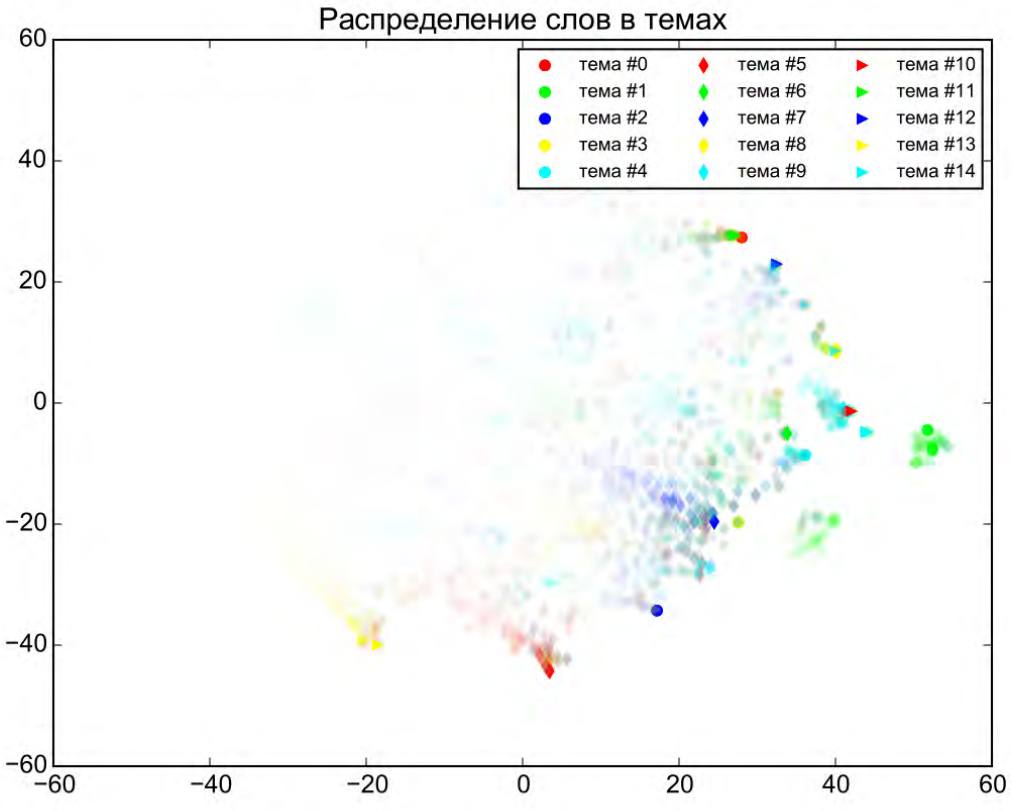


Рис. 10: Тематическая модель, обученная с помощью метода HALS, проинициализированного якорными ядрами, на данных KOS с 15-ю темами. Видно хорошо отделившиеся тимы 1, 2, 5. Также хорошо заметно, что темы 0 и 10 имеют мало слов с большой вероятностью встречаемости.

Тема 0		Тема 2		Тема 4		Тема 8	
amari	0.011	william	0.057	neuron	0.020	algorithm	0.023
dayan	0.010	objective	0.021	cell	0.014	learning	0.023
bishop	0.009	robert	0.020	input	0.012	function	0.012
giles	0.009	mark	0.019	model	0.011	problem	0.009
jaakkola	0.009	andrew	0.019	network	0.009	action	0.007
taylor	0.009	martin	0.018	system	0.007	system	0.006
atkeson	0.008	thomas	0.018	circuit	0.006	optimal	0.006
tresp	0.008	paul	0.017	neural	0.006	step	0.006
simard	0.008	eric	0.017	pattern	0.005	result	0.006

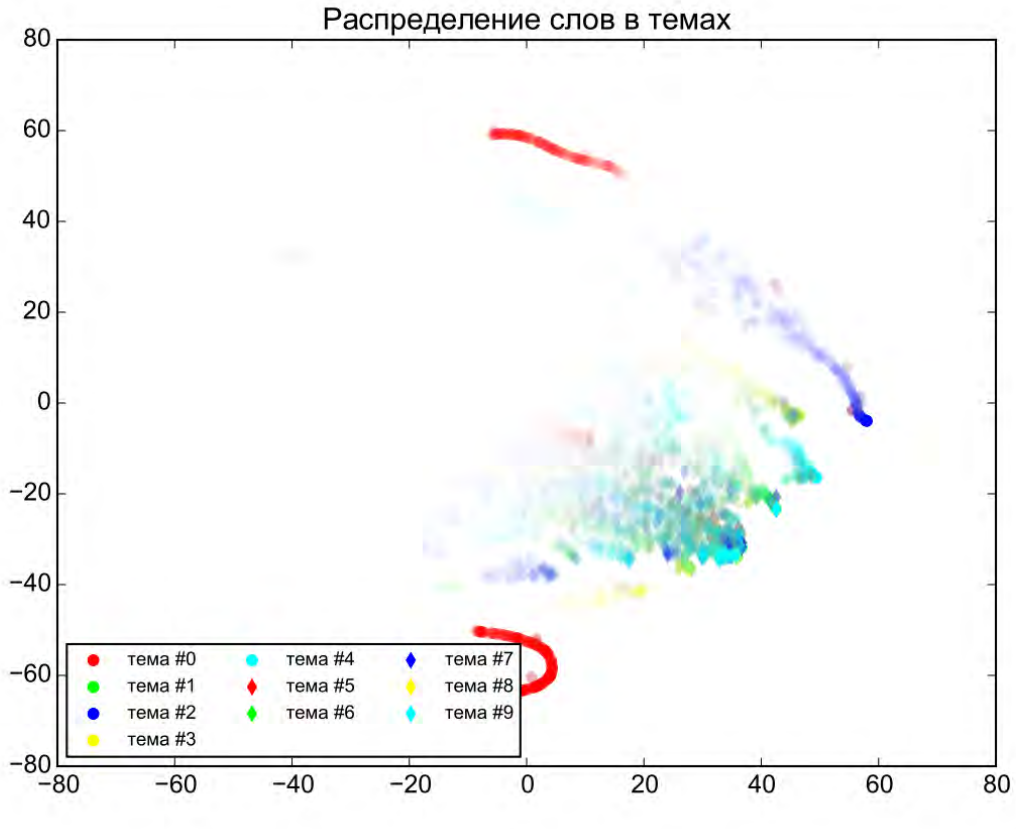


Рис. 11: Тематическая модель, обученная с помощью PLSA, проинициализированного якорными ядрами, на данных NIPS с 10-ю темами. Хорошо видна тема 0, которая являются набором фамилий авторов статей, а также тема 2, которая состоит из имён авторов статей.

7 Заключение

В работе исследуется проблема инициализации методов тематического моделирования, а также её влияние на интерпретируемость получаемой модели. Введённые в работе критерии исходят из предположения, что темы тем лучше интерпретируемы человеком, чем более они различны, а слова внутри них взаимосвязаны.

Хорошая инициализация позволяет избежать нежелательные области. Для инициализации предложен алгоритм, основанный на понятии якорных слов, которое позволяет быстро строить тематическую модель. Изменение ограничений, накладываемых якорными словами, позволяет строить тематические модели, имеющие лучшие метрики качества. Быстрая скорость построения таких моделей также делает их хорошей инициализацией более общих методов тематического моделирования, что показано экспериментально на коллекциях статей.

Список литературы

[1] *Arora S. et al.* A practical algorithm for topic modeling with provable guarantees // Proceedings of the 30th International Conference on Machine Learning (ICML-13). — 2013. — Pp. 280–288.

[2] *Arora S., Ge R., Kannan R., Moitra A.* Computing a nonnegative matrix factorization–provably // Proceedings of the forty-fourth annual ACM symposium on Theory of computing / ACM. — 2012. — Pp. 145–162.

[3] *Arora S., Ge R., Moitra A.* Learning topic models–going beyond svd // Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on / IEEE. — 2012. — Pp. 1–10.

[4] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // *the Journal of machine Learning research.* — 2003. — Vol. 3. — Pp. 993–1022.

[5] *Chaney A. J.-B., Blei D. M.* Visualizing topic models. // ICWSM. — 2012.

[6] *Chang J., Gerrish S., Wang C., Boyd-graber J. L., Blei D. M.* Reading tea leaves: How humans interpret topic models // Advances in neural information processing systems. — 2009. — Pp. 288–296.

[7] *Chen Z., Cichocki A.* Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints // *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep.* — 2005. — Vol. 68.

[8] *Chu M., Diele F., Plemmons R., Ragni S.* Optimality, computation, and interpretation of nonnegative matrix factorizations // SIAM Journal on Matrix Analysis / Citeseer. — 2004.

[9] *Cichocki A., Anh-Huy P.* Fast local algorithms for large scale nonnegative matrix and tensor factorizations // *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences.* — 2009. — Vol. 92, no. 3. — Pp. 708–721.

[10] *Cichocki A., Zdunek R., Amari S.-i.* Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization // Independent Component Analysis and Signal Separation. — Springer, 2007. — Pp. 169–176.

[11] *Dhillon I. S., Modha D. S., Spangler W. S.* Class visualization of high-dimensional data with applications // *Computational Statistics & Data Analysis.* — 2002. — Vol. 41, no. 1. — Pp. 59–90.

[12] *Donoho D., Stodden V.* When does non-negative matrix factorization give a correct decomposition into parts? // Advances in neural information processing systems. — 2003. — P. None.

[13] *Ho N.-D.* Nonnegative matrix factorization algorithms and applications: Ph.D. thesis / ÉCOLE POLYTECHNIQUE. — 2008.

[14] *Hofmann T.* Probabilistic latent semantic analysis // Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence / Morgan Kaufmann Publishers Inc. — 1999. — Pp. 289–296.

[15] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval / ACM. — 1999. — Pp. 50–57.

[16] *Lee D. D., Seung H. S.* Algorithms for non-negative matrix factorization // Advances in neural information processing systems. — 2001. — Pp. 556–562.

[17] *Lichman M.* UCI machine learning repository. — 2013.

[18] *Newman D., Karimi S., Cavedon L.* External evaluation of topic models // in Australasian Doc. Comp. Symp., 2009 / Citeseer. — 2009.

[19] *Paatero P., Tapper U.* Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values // *Environmetrics*. — 1994. — Vol. 5, no. 2. — Pp. 111–126.

[20] *Ultsch A.* Maps for the visualization of high-dimensional data spaces // Proc. Workshop on Self organizing Maps. — 2003. — Pp. 225–230.

[21] *Van der Maaten L., Hinton G.* Visualizing data using t-sne // *Journal of Machine Learning Research*. — 2008. — Vol. 9, no. 2579-2605. — P. 85.

[22] *Vorontsov K.* Additive regularization for topic models of text collections // Doklady Mathematics / Pleiades Publishing. — Vol. 89. — 2014. — Pp. 301–304.