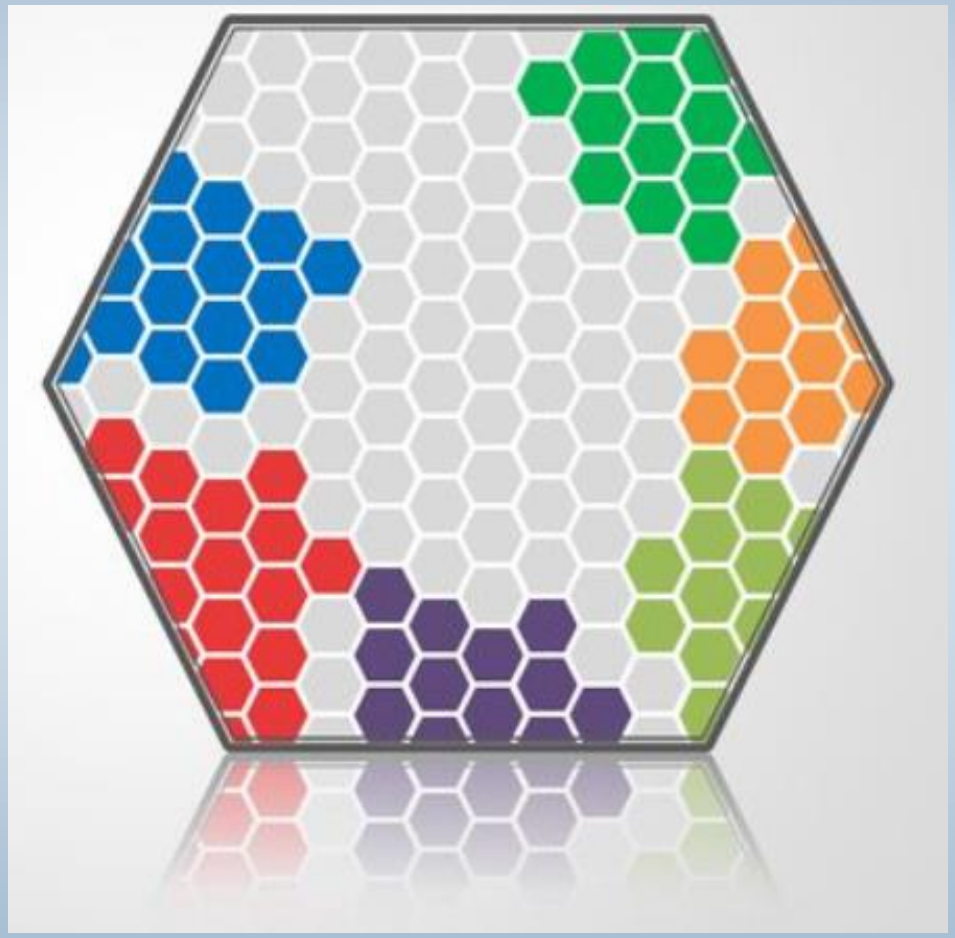
Topic Modeling

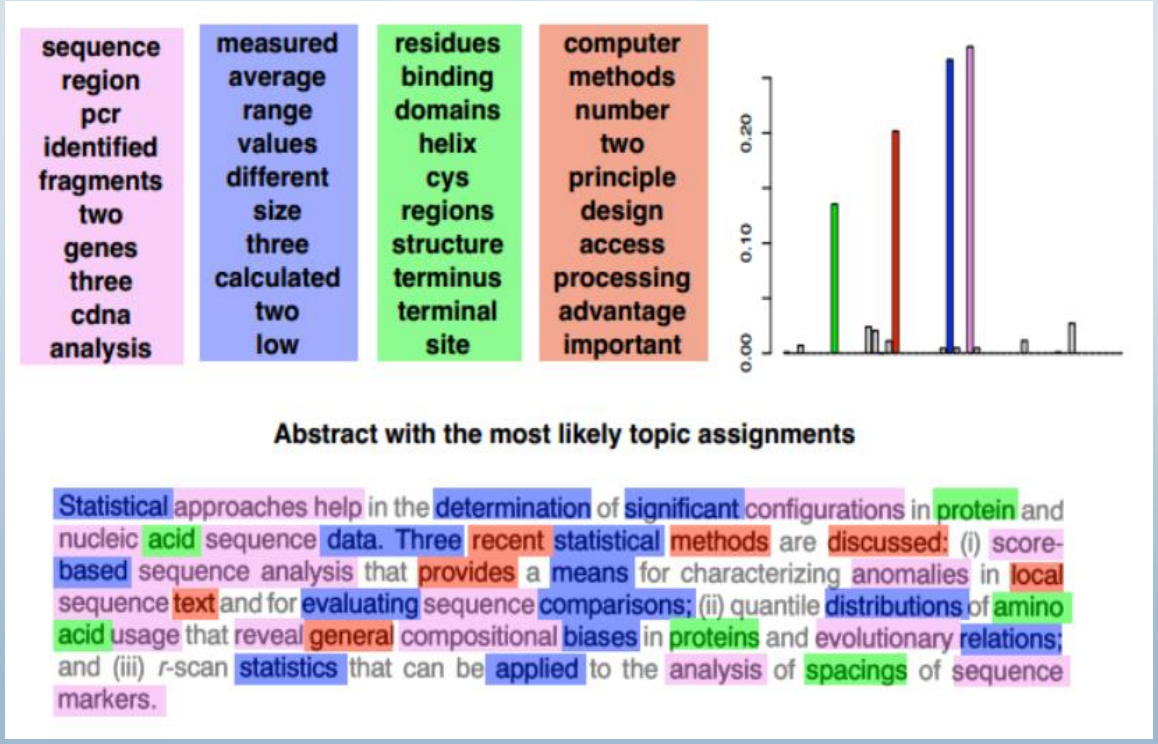


Кольцов С.Н.

Введение в тематическое моделирование

Тематическое моделирование в это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. **Тематическая модель (topic model)** коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (**термины**) образуют каждую тему.

Вероятностные тематические модели применяются в следующих направлениях:
1. Информационный поиск
2. Выявление трендов в научных публикациях и новостных потоках.
3. Классификация и категоризация документов, изображений, аудио и видеосигналов.
4. Определение тематики сообществ в социальных сетях

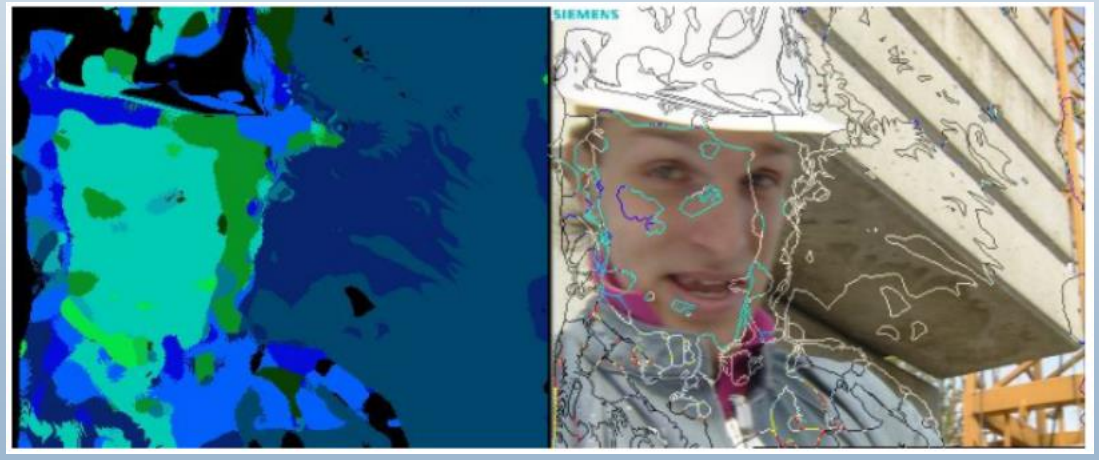


Введение в тематическое моделирование

Применение тематических моделей позволяет получить ответ на целый ряд нетривиальных вопросов.

1. Как выявлять смысл или тематику документов по их содержимому?
2. Как осуществлять классификацию документов на основе этих скрытых тематических закономерностей?
3. Как выявлять тенденций в развитии научных направлений?
4. Как выявлять роли людей в социальных сетях?
5. Как осуществлять индексацию и автоматическое аннотирование документов?
6. Как осуществлять распознавание образов.

Тематические модели основаны на теории вероятности, в частности на правиле Байеса.



Элементы теории вероятности

Различия в подходах к теории вероятностей

Случайная величина — это величина, которая принимает в результате опыта одно из множества значений, причём появление того или иного значения этой величины до её измерения нельзя точно предсказать.

1. В частотном подходе (классический подход) предполагается, что случайность есть объективная неопределенность . Вероятность рассчитывается из серии экспериментов и является мерой случайности как эмпирической данности. Исторически частотный подход возник из практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл.

2. В байесовском подходе предполагается, что случайность характеризует наше незнания. Например, случайность при бросании кости связана с незнанием динамических характеристик игральной кости, сопротивления воздуха и так далее. Многие задачи частотным методом решить невозможно (точнее, вероятность искомого события строго равна нулю). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ.

Элементы теории вероятности

Вероятность события — Вероятностью события А называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов.

Например. Вероятность того, что на кубике выпадет четное число, равна следующему отношению P=3/6=1/2.

Понятие условной вероятности

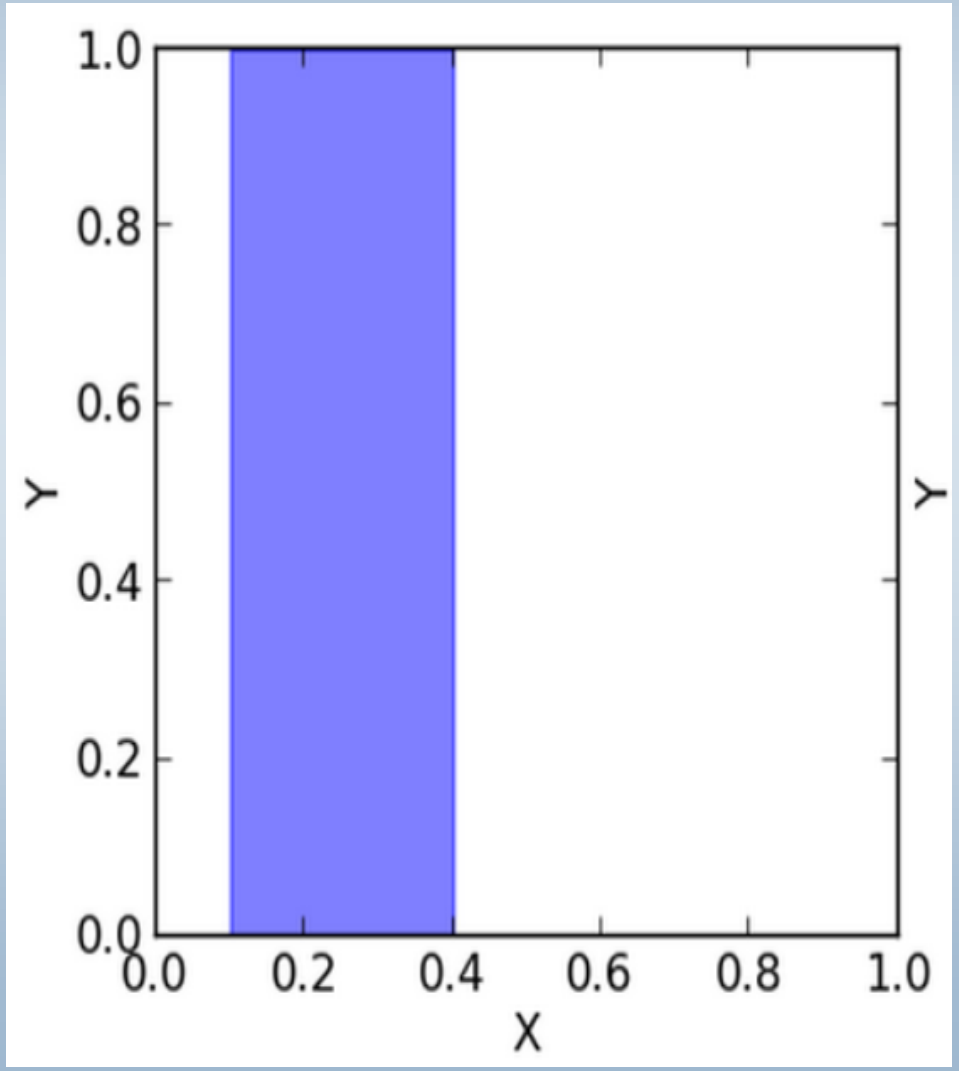
Условной вероятностью события **А** при условии, что произошло событие **В**, называется число **P(A|B)=P(B, A)/ P(B)**,
P(B, A) – произведение вероятностей, **P(B)** – вероятность события **В.**

Например. В урне 3 белых и 3 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (событие В), если при первом испытании был извлечен черный шар (событие А).
Вероятность события А=3/6=1/2
Произведение вероятностей P(B, A) =(3/6)*(3/5)=9/30

Итоговый результат: (9/30)/(1/2)=3/5

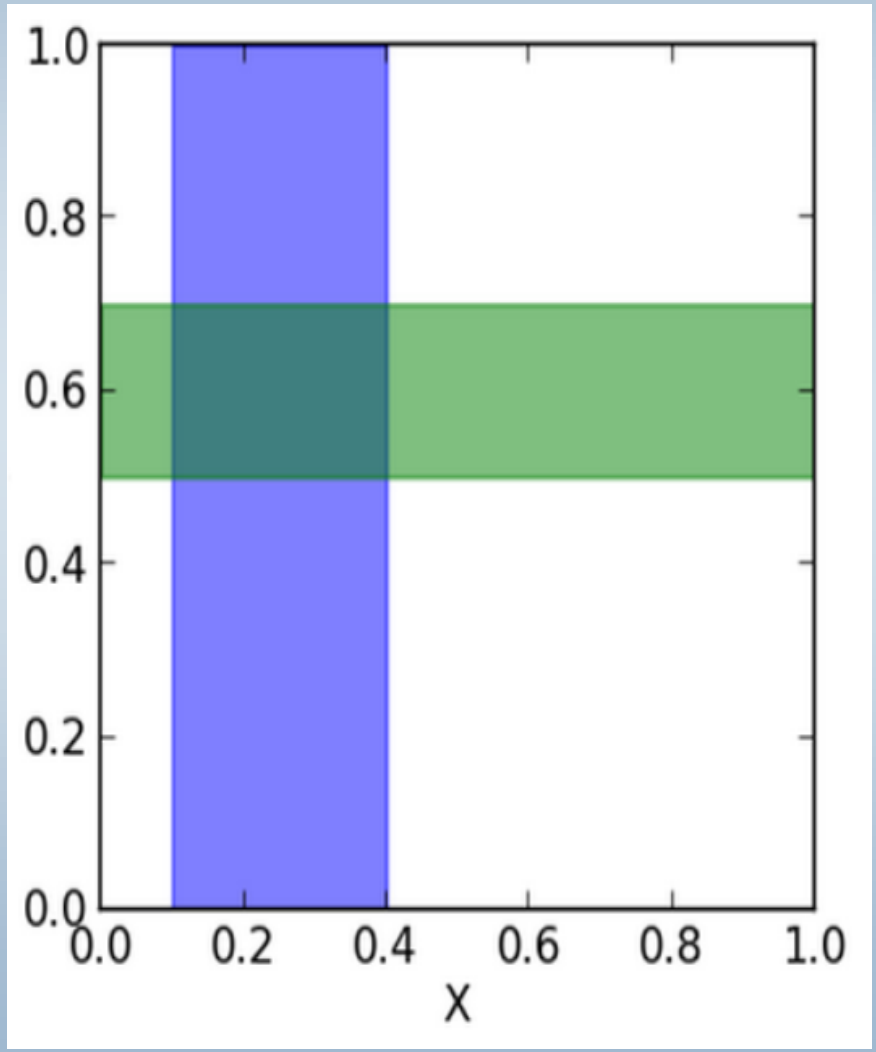
Элементы теории вероятности

Геометрическая интерпретация вероятности:
Рассмотрим следующий эксперимент: мы называем любое число из отрезка [0, 1] и смотрим за тем, что это число будет между, например, 0.1 и 0.4. Как нетрудно догадаться, вероятность этого события будет равна отношению длины отрезка [0.1, 0.4] к общей длине отрезка [0, 1] (другими словами, отношение «количества» возможных равновероятных значений к общему «количеству» значений), то есть (0.4 — 0.1) / (1 — 0) = 0.3, то есть вероятность попадания в отрезок [0.1, 0.4] равна 30%.



Элементы теории вероятности

Геометрическая интерпретация вероятности теоремы : вероятность того, что y находится внутри отрезка [0.5, 0.7] равна отношению площади зеленой области к площади всего квадрата p(0.5 <= y <= 0.7) = 0.2, или для краткости **p(Y)** = 0.2. **А теперь допустим мы хотим знать какова вероятность того, что y находится в интервале [0.5, 0.7], если x уже находится в интервале [0.1, 0.4].** Мы можем записать эту вероятность как **p(Y|X).** Очевидно, что эта вероятность равна отношению площади темной области (пересечение зеленой и синей областей - **p(X, Y)**) к площади синей области.



Элементы теории вероятности

Формула Байеса

Байесовская вероятность — это интерпретация понятия вероятности, используемое в байесовской теории. Вероятность определяется как степень уверенности в истинности суждения**.**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

P(A) — априорная вероятность гипотезы *A (заранее известная вероятность)*;
P (A|B) — вероятность гипотезы *A* при наступлении события *B* **(апостериорная вероятность**);
P(B|A) — вероятность наступления события *B* при истинности гипотезы *A*;
P(B) — полная вероятность наступления события *B*.
P(A|B)) — вероятность наступления события *A* при истинности гипотезы *B*;

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Таким образом, формула Байеса может быть использована для разработки алгоритмов классификации.

Пример применении теории вероятности

Пример: случайному пациенту сделали тест на наличие СПИД, и получили положительный результат. Пусть точность теста 99.8% (т.е. он дает положительный результат у 0.2% здоровых людей). Какова вероятность, что у этого пациента СПИД?

Априорная вероятность: P(больной) – доля больных в стране (пусть 0.3%)

$$P(\text{больной} \mid \text{тест}+) = \frac{P(\text{тест}+ \mid \text{больной}) \cdot P(\text{больной})}{P(\text{тест}+ \mid \text{больной})P(\text{больной}) + P(\text{тест}+ \mid \text{здоровый})P(\text{здоровый})} =$$

$$= \frac{1 \cdot 0.003}{1 \cdot 0.003 + 1 \cdot 0.002} = 60\%$$

Априорные и апостериорные суждения в теории вероятности

1. Предположим, мы хотим узнать значение некоторой неизвестной величины.
2. У нас имеются некоторые знания, полученные до (a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания.
3. В процессе наблюдений эти знания подвергаются постепенному уточнению. После (a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении.
4. Будем считать, что мы пытаемся оценить неизвестное значение величины P(A|B) посредством наблюдений некоторых ее косвенных характеристик (гипотез).

Формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений.

Вероятностная постановка задачи классификации

Пусть имеется множество объектов X и конечное множество классов Y . Требуется построить алгоритм способный классифицировать произвольный объект X в рамках заданного множества Y.
Апостериорная вероятность принадлежности объекта X классу Y по формуле Байеса:

$$P(A|B) = \frac{p(A,B)}{P(A)} = \frac{p(A)P(B|A)}{P(A)}$$

$P(A|B)$ - Апостериорная вероятность

$p(A,B)$ - Априорная вероятность

Задача классификации заключается в расчете (оценке) апостериорной информации на основании априорной информации. Такая оценка может быть реализована при помощи формулы Байеса. Однако существует проблема оценивания априорной величины p(A,B)

Задача восстановления априорного распределения

p(x,y)

Оценка функции ***p(x,y)*** может быть реализован при помощи трех методов.

1. Непараметрическое восстановление плотности основано на локальной аппроксимации плотности p(x) в окрестности классифицируемого объекта x ∈ X. Пример, Алгоритм Парзена-Розенблатта (метод парзеновского окна).

2. Параметрическое восстановление плотности основано на предположении, что плотность распределения известна с точностью до параметра, ***p(x,y)*** = ϕ(x; θ), где ϕ фиксированная функция. Пример. Нормальный дискриминантный анализ. LSA – в основе лежит метод SVD разложения.

3. Восстановление смеси плотностей. Если функцию плотности ***p(x,y)*** не удаётся смоделировать параметрическим распределением, можно попытаться описать её смесью нескольких распределений:

Собственно именно третий метод является основой тематического моделирования

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x; \theta_j), \qquad \sum_{j=1}^{k} w_j = 1,$$

Тематическое моделирование

Тема

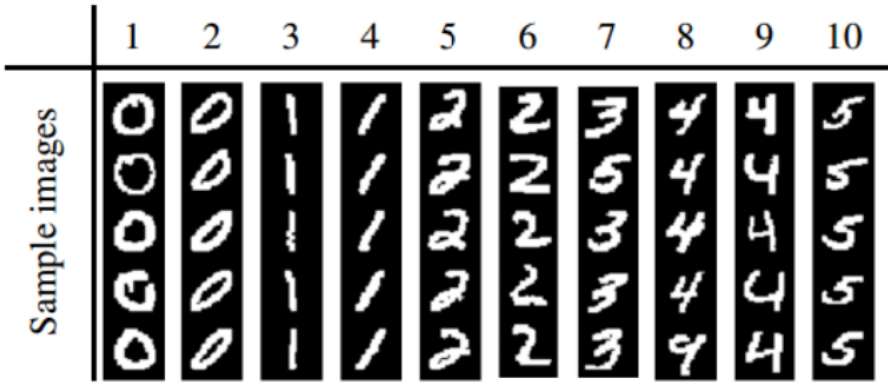
В литературе по тематическим моделям понятие *темы* (**topic**) определяется по-разному, в зависимости от научной школы: «скрытые паттерны», «компактные описания смысла документов», «вероятностные (нечёткие) кластеры семантически связанных терминов», «связующее звено между терминами и другими объектами (документами, авторами, организациями, конференциями, и т.д.), которое позволяет находить скрытые ассоциативные связи между ними».

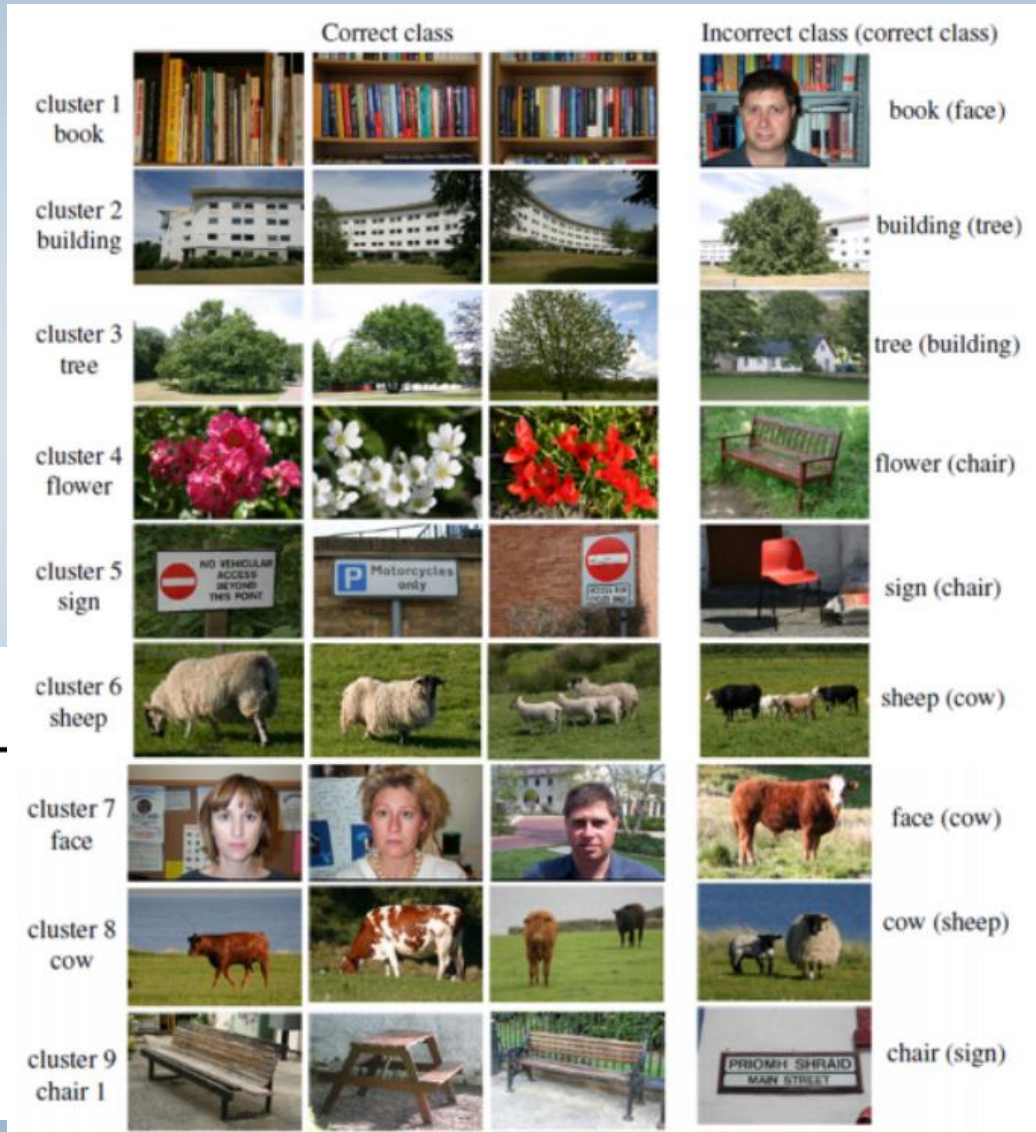
Parallel Computing	
parallel	0.067
processor	0.026
communication	0.024
performance	0.024
application	0.015
implementation	0.013
computation	0.012
parallelism	0.012
cluster	0.011
high_performance	0.011
workstation	0.011
multiprocessor	0.010
shared_memory	0.010
system	0.010
machine	0.010
distributed_memory	0.009
message_passing	0.009

Probabilistic Models	
model	0.047
estimation	0.033
parameter	0.024
estimator	0.018
distribution	0.018
data	0.017
regression	0.016
method	0.015
bayesian	0.015
estimate	0.014
statistical	0.013
error	0.013
estimates	0.012
sample	0.010
variance	0.009
prior	0.009
density	0.008

Тематическое моделирование

Тема
Вместо набора документов могут быть использованы коллекции изображений. Соответственно под темой могут пониматься сходные изображения. Под тематическим моделирование понимается процесс нахождение сходных изображений.





Тематическое моделирование (Latent Dirichlet allocation)

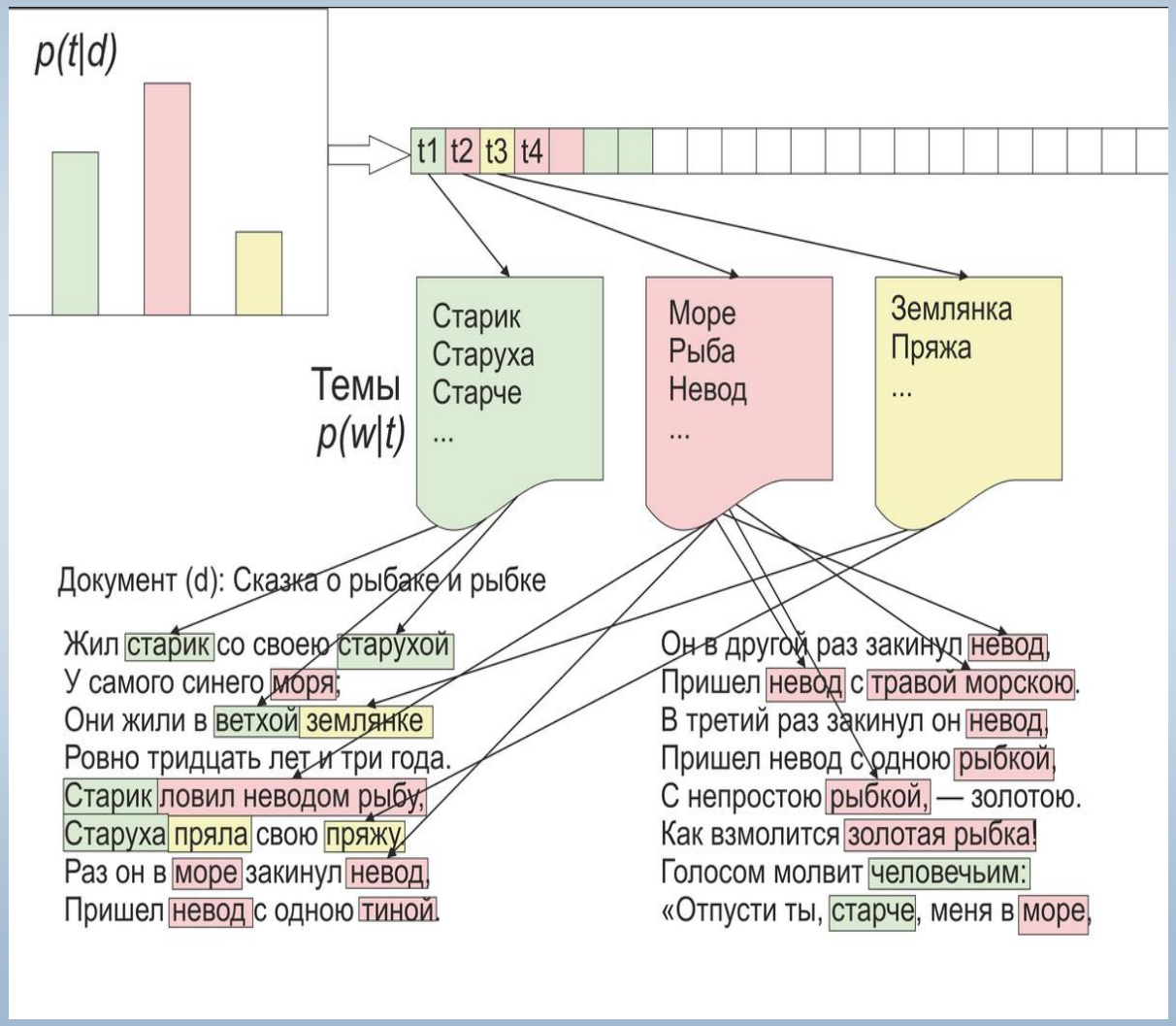
Основное предположение тематической модели Latent Dirichlet Allocation состоит в том, что каждый документ с некоторой вероятностью может принадлежать множеству тематик. Тема эта совокупность слов, где каждое слово имеет некоторую вероятность принадлежности к данной тематике.

Формально тема определяется как дискретное (мультиномиальное) вероятностное распределение в пространстве слов заданного словаря.

Тематическим моделированием называется решение задачи, обратной классификации . Каждый документ в корпусе текстов рассматривается как наблюдаемая случайная независимая выборка слов (мешок слов), порождённая некоторым, скрытым (латентным) множеством тем. По этим данным требуется восстановить вероятностные распределения всех тем в корпусе и определить, каким именно подмножеством тем порождён каждый документ.

Тематическое моделирование основано на применении формулы Байеса, в которой распределение слов и тем выражено в виде смеси плотностей распределений слов и документов.

Тематическое моделирование



Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем.

Тематическое моделирование

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:

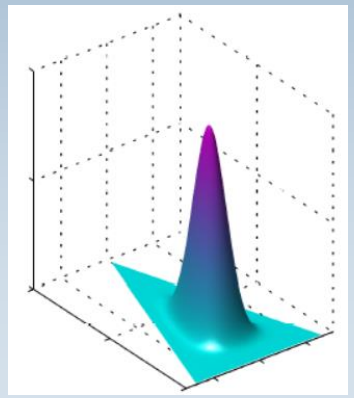
$$p(w|d) = \sum_{t \in T} p(w|d, t)\, p(t|d) = \boxed{\sum_{t \in T} p(w|t)\, p(t|d)}$$

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

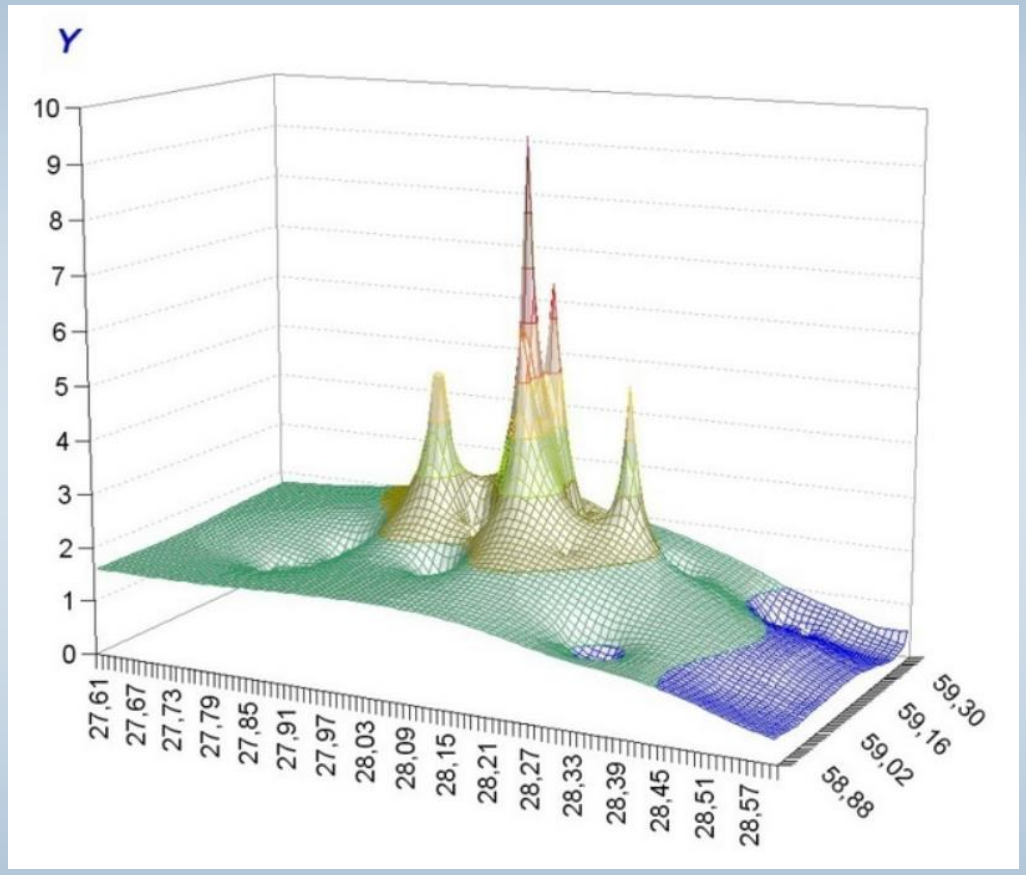
- $\phi_{wt} \equiv p(w|t)$ – распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ – распределение тем в документах $d \in D$.

Тематическое моделирование

$$P(w|t) =$$



Функция Дирихле

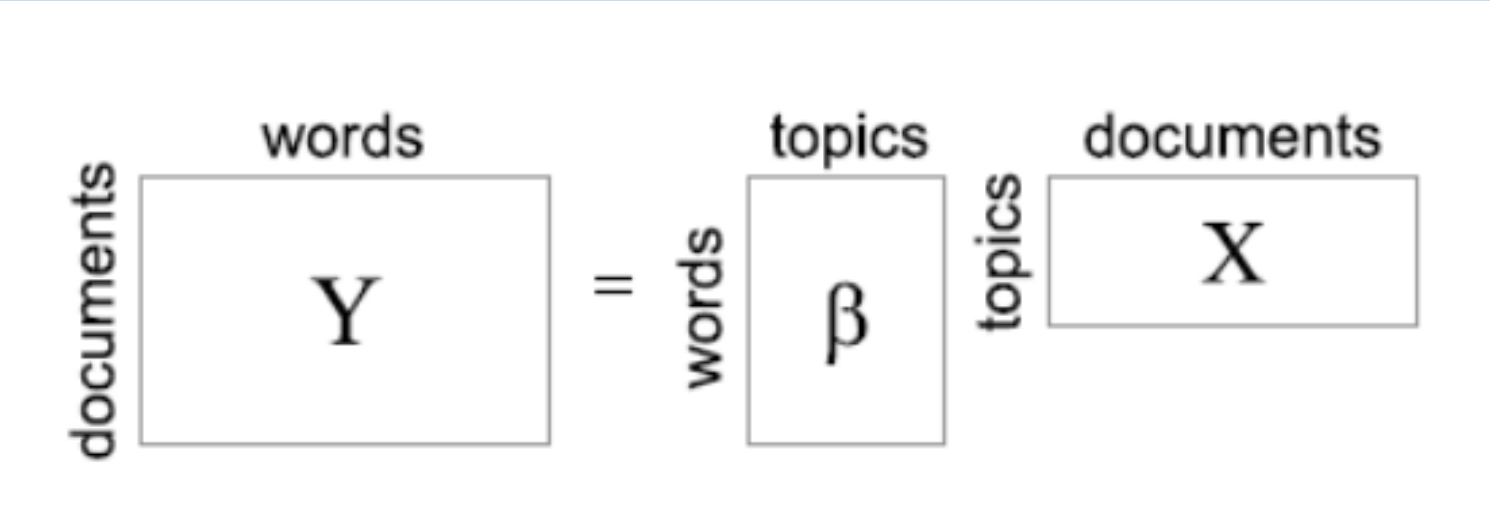


$$P(w|d) =$$

Вероятность того что слово **w** вынутое из документа **d** принадлежит теме **T** описывается очень сложной функцией (произведение функций Дирихле)

Тематическое моделирование

Тематическое моделирование в терминах матричного анализа, заключается в том, что очень большая матрица **Документы – Слова** аппроксимируется двумя маленькими матрицами:
1. **Слова – Темы**. 2. Матрица **Документы – Темы**.



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

Проблема аппроксимации:

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

Latent Dirichlet Allocation(Gibbs sampling)

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

$C^{WT}_{m,j}$ - Матрица; ячейка: сколько раз слово **w** связанно с темой **t**,

$C^{DT}_{d,j}$ - Матрица; ячейка: : сколько раз слово **w** в документе **d** связанно с темой **t**,

$\sum_{m'} C^{WT}_{m',j} = n_t$ - Вектор; ячейка: число слов связанное с темой **t**,

$C^{DT}_{d,j'} = n_d$ Длина документа **d** в словах

Results of simulation:

1. Матрица распределение слов по темам. 2. Матрица распределение документов по темам.

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

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

Тематическое моделирование

Words with high probability

	1	2	3	4	5	6
1	сша: 0,035673	страна: 0,054558	музыка: 0,036338	бог: 0,031730	деревня: 0,017067	исследо
2	американский: 0,018040	китай: 0,033951	группа: 0,020940	мир: 0,017793	земля: 0,016213	мозг: 0,(
3	сирия: 0,017584	китайский: 0,028159	концерт: 0,020643	дух: 0,011312	хозяйство: 0,011423	ученый
4	страна: 0,014949	мир: 0,022707	песня: 0,019742	душ: 0,009062	сельский: 0,007545	экспери
5	война: 0,013610	день: 0,017546	песнь: 0,015207	сила: 0,008775	лошадь: 0,006973	резуль
6	военный: 0,012463	индий: 0,012635	петь: 0,014799	земля: 0,007723	корова: 0,006612	исследо
7	оружие: 0,010839	китаец: 0,010613	музыкальный: 0,013924	слово: 0,007278	крестьянин: 0,006183	универ
8	россия: 0,010036	африка: 0,008368	альбом: 0,012364	вера: 0,006758	поле: 0,005733	показы
9	обама: 0,009962	южный: 0,008270	играть: 0,011076	жизнь: 0,006364	садиться: 0,005573	проводи
10	президент: 0,009602	виза: 0,007607	музыкант: 0,010281	великий: 0,006246	хлеб: 0,005144	ген: 0,0(
11	оон: 0,008692	месяц: 0,006855	танец: 0,009457	зло: 0,005670	зерно: 0,004677	обнару
12	международный: 0,007694	европа: 0,006792	сцена: 0,009172	человеческий: 0,00531	время: 0,004572	теория:
13	химический: 0,007621	америка: 0,006510	слушать: 0,009104	сердце: 0,004777	жить: 0,004517	тест: 0,(
14	мир: 0,007442	бразилия: 0,006392	зал: 0,007056	свет: 0,004659	урожай: 0,004441	группа:
15	сирийский: 0,007218	азия: 0,006365	звук: 0,007022	грех: 0,004234	свинья: 0,004139	время: (
16	иран: 0,007070	индийский: 0,005734	голос: 0,006754	добро: 0,004150	хороший: 0,004029	вывод: (
17	американец: 0,006866	житель: 0,005589	выступать: 0,006686	высокий: 0,004106	работа: 0,003958	челове
18	сила: 0,005999	остров: 0,005150	выступление: 0,005547	истина: 0,004052	ферм: 0,003659	клетка:
19	америка: 0,005332	сша: 0,004946	инструмент: 0,005436	божий: 0,003881	скот: 0,003592	жизнь: (
20	правительство: 0,005282	граница: 0,004840	исполнять: 0,005270	смерть: 0,003702	фермер: 0,003558	научны
21	безопасность: 0,005121	австралия: 0,004723	исполнение: 0,005122	господь: 0,003645	место: 0,003533	считать
22	сторона: 0,004995	нова: 0,004707	танцевать: 0,005096	путь: 0,003556	пол: 0,003529	приводи
23	асад: 0,004882	восточный: 0,004644	рок: 0,005007	религия: 0,003494	село: 0,003441	называ
24	вашингтон: 0,004437	япония: 0,004405	гитара: 0,004956	верить: 0,003430	работать: 0,003340	анализ:
25	заявлять: 0,004370	пекин: 0,004272	артист: 0,004956	ангел: 0,003318	сельскохозяйственный	поведе
26	государство: 0,004056	мексика: 0,004241	шоу: 0,004531	враг: 0,003169	маленький: 0,003234	информ
27	конфликт: 0,004007	провинция: 0,004217	звучать: 0,004433	любовь: 0,003154	га: 0,003226	образ: (
28	операция: 0,003905	корея: 0,004014	время: 0,004352	рай: 0,003028	кукуруза: 0,003108	метод: (

Тематическое моделирование

Documents - Topics Distribution

	ID	Orig text	N	U	D	T	1	2	3	4	5	6	7	8	9	10	1
1							0.00120	0.01803	0.00120	0.00120	0.01082	0.00120	0.00120	0.11178	0.00120	0.00120	0
2	1	Посмотрите фотографию. Это нижегородец Ник					0.00051	0.00051	0.00154	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0
3	2	Оригинал взят у в Верной дорогой идете, товар					0.00943	0.06604	0.02830	0.00943	0.00943	0.00943	0.00943	0.00943	0.00943	0.00943	0
4	3	Вопрос оказался суперлегким. Вместо ответа пр					0.00514	0.00171	0.00171	0.00171	0.01199	0.02226	0.00514	0.00171	0.00171	0.01884	0
5	4	В России, как это широкомасштабно известно, пр					0.00109	0.03611	0.00328	0.01422	0.02079	0.00109	0.00109	0.00328	0.00109	0.00985	0
6	5	Оригинал взят у в ЩеглишкоТак близко щеглы					0.00311	0.00311	0.00311	0.00311	0.01553	0.00311	0.00311	0.00311	0.00311	0.00311	0
7	6	Митинги и акцииМировой судья Тверского район					0.00030	0.42280	0.00517	0.00030	0.00030	0.00152	0.00030	0.00030	0.00091	0.00213	0
8	7	Ну в общем я думаю так: менты на втором видео					0.00050	0.00050	0.00050	0.00050	0.00050	0.00050	0.00050	0.00050	0.00050	0.00050	0
9	8	Закон бутерброда Даже не намазанный маслом,					0.00095	0.00285	0.03321	0.00095	0.00095	0.00285	0.00095	0.00095	0.00664	0.01803	0
10	9	Внеочередное заседание коллегии МВД в Казани					0.00362	0.03986	0.00362	0.00362	0.00362	0.00362	0.00362	0.01087	0.01087	0.00362	0
11	10	Вчера в центральную больницу города Гатчина					0.00338	0.00338	0.00338	0.00338	0.00338	0.00338	0.01689	0.01014	0.00338	0.01014	0
12	11	Что не говори, над программой Пусть говорят					0.00376	0.01128	0.00376	0.00376	0.03383	0.01128	0.04135	0.00376	0.00376	0.00376	0
13	12	в 19 отчаянных В Нижнем Новгороде сломали пу					0.00051	0.00153	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0
14	13	Посмотрите фотографию. Это – нижегородец Ни					0.00053	0.00159	0.00053	0.00053	0.00053	0.00053	0.00053	0.00053	0.00053	0.00053	0
15	14	Пытками в казанском отделе "Дальний" занялись					0.01471	0.04412	0.01471	0.01471	0.01471	0.10294	0.01471	0.01471	0.01471	0.01471	0
16	15	Начальник полиции Казани Рустем Кадыров обяз					0.00355	0.05319	0.00355	0.00355	0.00355	0.00355	0.03901	0.00355	0.01064	0.00355	0
17	16	Дело в отношении сотрудников отдела полиции					0.00141	0.04366	0.00141	0.00141	0.00423	0.00141	0.00141	0.01268	0.00141	0.00423	0
18	17	Оригинал взят у в «19 отчаянных» В Нижнем Но					0.00152	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0.00051	0.00152	0.00051	0
19	18	Пытками в казанском отделе Дальний занялись					0.01786	0.01786	0.01786	0.01786	0.01786	0.01786	0.01786	0.01786	0.01786	0.01786	0

Тематическое моделирование. Проблемы определения качества и количества тем.

1. Количество тем.

Как и в кластерном анализе существует проблема выбора числа тем. Эта проблема может решатся разными методами, но задача до конца не решена.

А. Можно увеличивать (уменьшать) число тем и смотреть на получаемый результат.

Б. Можно использовать такие методы как метод скачков (по аналогии в кластерном анализе).

2. Качество тем.

Проблема заключается в том что получаемые темы могут быть плохо интерпретируемы (так называемые мусорные темы). Как правило, это решается за счет экспертной оценки. Есть методики, которые хорошо себя зарекомендовали в поисковых системах, например, можно слова в каждой теме, отсортированные по вероятности, умножать на коэффициент TF-IDF. Соответственно можно посчитать сумму величин в каждой теме и таким образом выделить плохие и хорошие темы.

Пример плохих и хороших тем

Средне интерпр.

демократия: 0.019804
они: 0.010584
статья: 0.009047
остров: 0.008279
фашизм: 0.006166
слово: 0.006166
борьба: 0.005590
фейхтвангер: 0.004629
остер: 0.004437
человек: 0.004245
рабочий: 0.004053
социализм: 0.003861
создавать: 0.003669
парень: 0.003669
фронт: 0.003477
франций: 0.003477

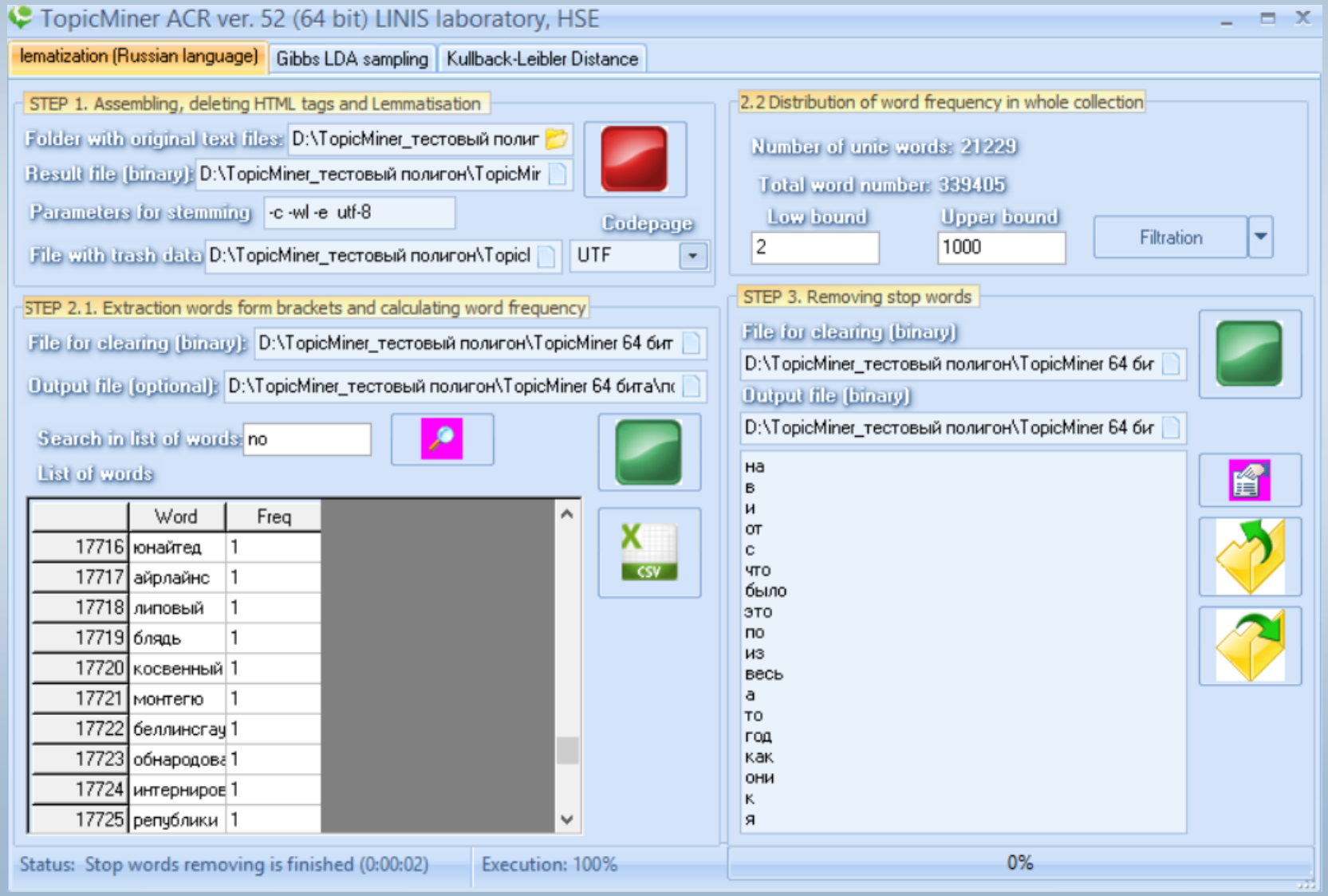
Интерпр.

война: 0.016583
армия: 0.014445
самолет: 0.012520
американский: 0.011344
военком: 0.010382
американец: 0.007816
летний: 0.007709
со: 0.007067
боец: 0.006212
экипаж: 0.006105
офицер: 0.005891
быль: 0.005464
машин: 0.005357
советский: 0.005143
сша: 0.005143

Не интерпр.

можно: 0.015012
из: 0.012960
палочка: 0.008540
или: 0.008382
сделать: 0.006961
ванная: 0.005856
змей: 0.005067
очень: 0.005067
животное: 0.004594
для: 0.004594
цвет: 0.003962
помощь: 0.003962
бумага: 0.003646
струйка: 0.003489
температура: 0.003331
полагать: 0.003331

TopicMiner. 1. Препроцессинг данных

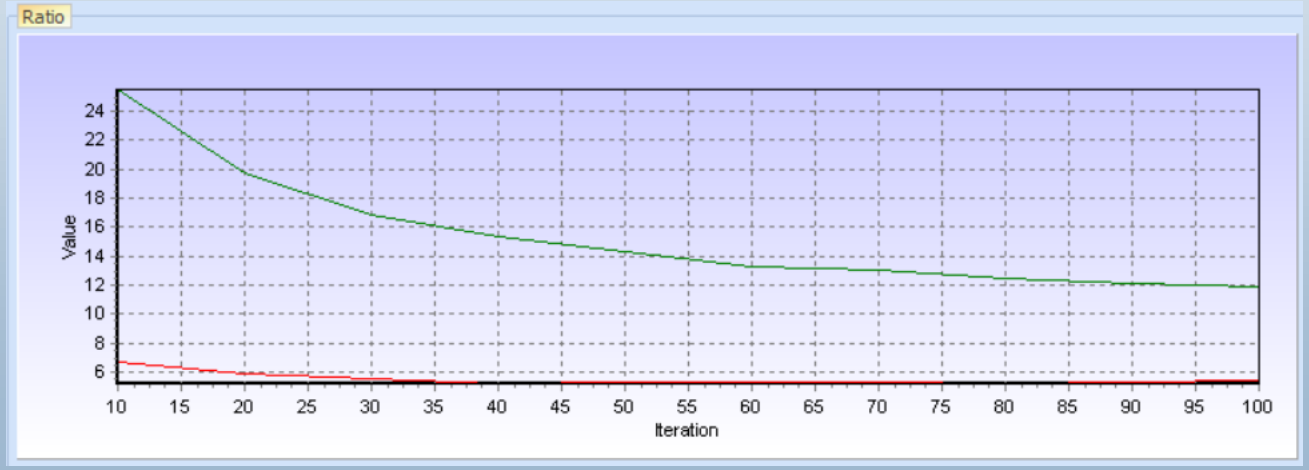


TopicMiner. 2. Параметры моделирования

Number of topics – число тем Number of iteration– число итераций

Alpha, Beta– параметры, которые определяют размер (толщину) функций Дирихле

Save step – параметр, который определяет какую итерацию визуализировать.



Ratio – величина, которая характеризует процесс схождения расчета.

TopicMiner. 3. Результаты моделирования

Результатами моделирования являются две матрицы:

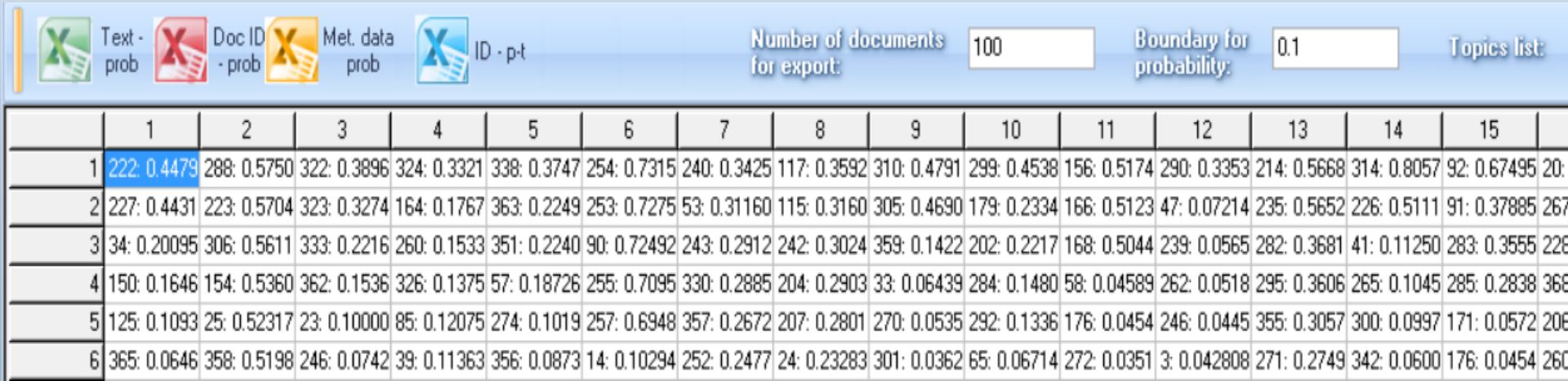
1. Матрица распределение слов по темам.

1. Матрица распределение документам по темам.

Можно посмотреть отсортированные матрицы (сортировка по вероятности):

Отсортированная матрица слов по темам

Отсортированная матрица документам по темам



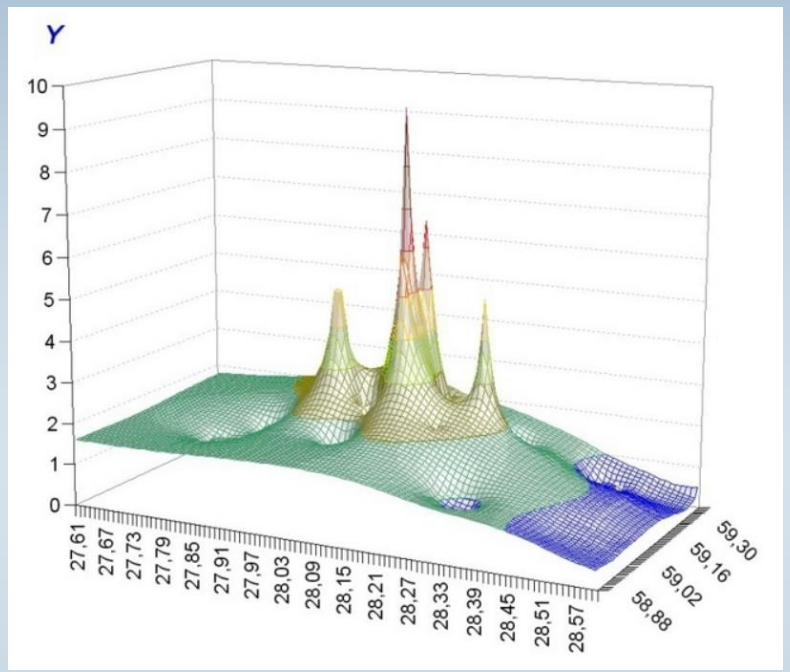
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	222: 0.4479	288: 0.5750	322: 0.3896	324: 0.3321	338: 0.3747	254: 0.7315	240: 0.3425	117: 0.3592	310: 0.4791	299: 0.4538	156: 0.5174	290: 0.3353	214: 0.5668	314: 0.8057	92: 0.67495	20:
2	227: 0.4431	223: 0.5704	323: 0.3274	164: 0.1767	363: 0.2249	253: 0.7275	53: 0.31160	115: 0.3160	305: 0.4690	179: 0.2334	166: 0.5123	47: 0.07214	235: 0.5652	226: 0.5111	91: 0.37885	267
3	34: 0.20095	306: 0.5611	333: 0.2216	260: 0.1533	351: 0.2240	90: 0.72492	243: 0.2912	242: 0.3024	359: 0.1422	202: 0.2217	168: 0.5044	239: 0.0565	282: 0.3681	41: 0.11250	283: 0.3555	228
4	150: 0.1646	154: 0.5360	362: 0.1536	326: 0.1375	57: 0.18726	255: 0.7095	330: 0.2885	204: 0.2903	33: 0.06439	284: 0.1480	58: 0.04589	262: 0.0518	295: 0.3606	265: 0.1045	285: 0.2838	368
5	125: 0.1093	25: 0.52317	23: 0.10000	85: 0.12075	274: 0.1019	257: 0.6948	357: 0.2672	207: 0.2801	270: 0.0535	292: 0.1336	176: 0.0454	246: 0.0445	355: 0.3057	300: 0.0997	171: 0.0572	208
6	365: 0.0646	358: 0.5198	246: 0.0742	39: 0.11363	356: 0.0873	14: 0.10294	252: 0.2477	24: 0.23283	301: 0.0362	65: 0.06714	272: 0.0351	3: 0.042808	271: 0.2749	342: 0.0600	176: 0.0454	260

Анализ ррезультатов моделирования

Проблема стабильности воспроизводства

1. Вероятность выбора слова **w** из случайного документа **d** и принадлежащего теме **T** является функцией с множеством локальных минимумов.
2. В основе сэмплирования лежит случайный выбор слова.

Все это приводит к тому что содержимое тем меняется от запуска к запуску.



Выбор важных тем

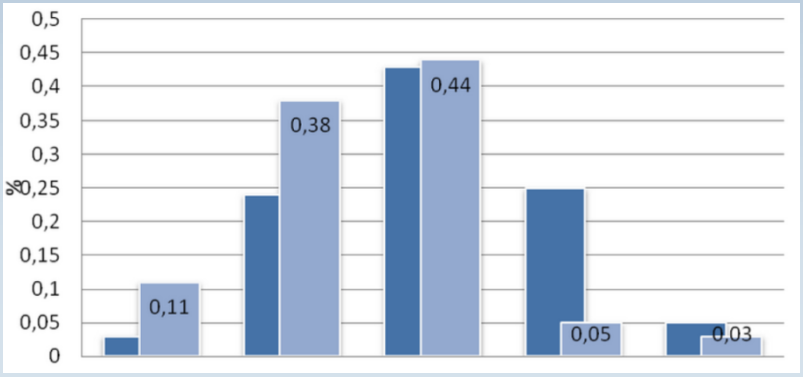
Можно выбирать темы, чей вес большой по сравнению с другими темами. Под весом Подразумевается сумма вероятностей по каждой теме.

ID document	Number of words	Topic(1)	Topic(2)	Topic(3)	Topic(4)	Topic(5)	Topic(6)
		2.86429	**15.86236**	**1.42353**	**1.72578**	**1.76096**	**5.59065**
1	950	0.00051	0.00051	0.00154	0.00051	0.00051	0.00051
2	33	0.00943	0.06604	0.0283	0.00943	0.00943	0.00943
3	275	0.00514	0.00171	0.00171	0.00171	0.01199	0.02226
4	449	0.00109	0.03611	0.00328	0.01422	0.02079	0.00109
5	174	0.00311	0.00311	0.00311	0.00311	0.01553	0.00311
6	1691	0.0003	0.4228	0.00517	0.0003	0.0003	0.00152
7	981	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005

Оценка качества LDA при помощи меры Kullback–Leibler

Мера Kullback - Leibler используется когда надо рассчитать уровень похожести между двумя распределениями.

$$K = 0.5 \sum_{k}^{W} \Phi_k^1 \log\left(\frac{\Phi_k^1}{\Phi_k^2}\right) + 0.5 \sum_{k}^{W} \Phi_k^2 \log\left(\frac{\Phi_k^2}{\Phi_k^1}\right)$$



IF K=0, два распределения идентичны. IF K=Max – два распределения максимально не похоже друг на друга..

Однако прямой расчет меры K-L не удобен так как длинный хвост распределения оказывает сильное влияние.

KL-based similarity metric

$$Kn = \left(1 - \frac{K}{Max}\right) * 100$$

IF Kn=100%, then две темы идентичны. IF K=0 тогда темы полностью различны.

Порог сходства тем

Уровень 90 - 93% означает, что первые 30- 50 слов одинаковы.

Similarity 0.935			
USA	0.04734	USA	0.03567
American	0.02406	American	0.01804
Syria	0.02082	Syria	0.01758
Obama	0.01374	country	0.01495
weapon	0.01343	war	0.01361
war	0.01309	military	0.01246
president	0.01169	weapon	0.01084
UN	0.01018	Russia	0.01004
military	0.01014	Obama	0.00996
country	0.01005	president	0.0096
chemical	0.00944	UN	0.00869
Syrian	0.00851	international	0.00769

Слова отсортированы по вероятности

Порог сходства тем

Similarity 0.854			
USA	0.04734	water	0.01758
American	0.02406	help	0.01296
Syria	0.02082	city	0.01262
Obama	0.01374	far	0.01199
weapon	0.01343	house	0.01064
war	0.01309	east	0.0104
president	0.01169	region	0.00945
UN	0.01018	dam	0.0091
military	0.01014	flood	0.00904
country	0.01005	resident	0.00839
chemical	0.00944	injured	0.00714
Syrian	0.00851	FRS	0.00698

Темы, чья похожесть порядка 85% или ниже совершенно различны.

Тесты на стабильность воспроизводства результатов моделирования

В экспериментах, мы запускали программу 5 раз на **K = 120** тем, датасет - **298 967 документов,** словарь **154 000 уникальных слов**. После этого мы сравнили 5 тематических решений между собой при помощи нормализованной меры K-L.