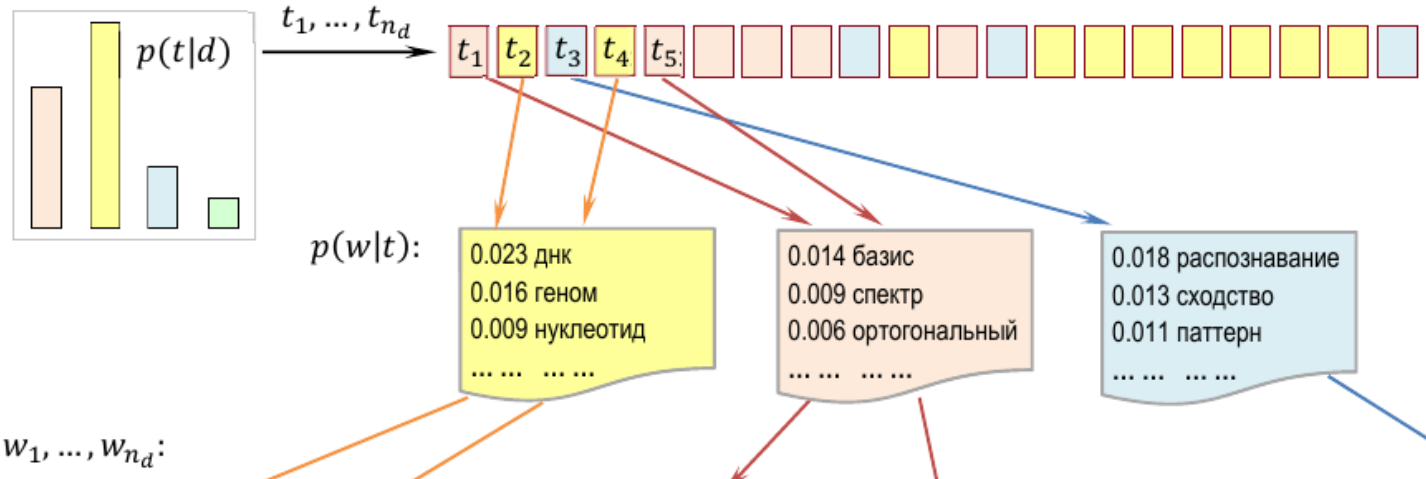
Машинное обучение
Тематическое моделирование



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

https://yandexdataschool.ru/edu-process/courses/machine-learning

Содержание лекции

- Постановка задачи
- Предыстория
- Латентный семантический анализ (LSA)
- Вероятностный LSA (PLSA)
- Латентное размещение Дирихле (LDA)

Предыстория

- Векторная модель документов:
  $d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$
- w_{ij} – вес i-того слова в j-том документе
- Методы взвешивания термов:
  - булевский вес (0,1)
  - Tf - term frequency (функция от количества вхождений слова в документ)
  - Tf-idf = TF*IDF
- Близость между документами (или запросом и документом) вычислялась про правилу косинуса:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

TF-IDF

- Мера важности слов в контексте документа: TF-IDF = TF*IDF

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k} \qquad \mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|}$$

  - n_i – число вхождений i-того термина в документ
  - $|d_i \supset t_i|$ - число документов с термином t_i
  - $|D|$ - количество документов

Недостатки векторной модели

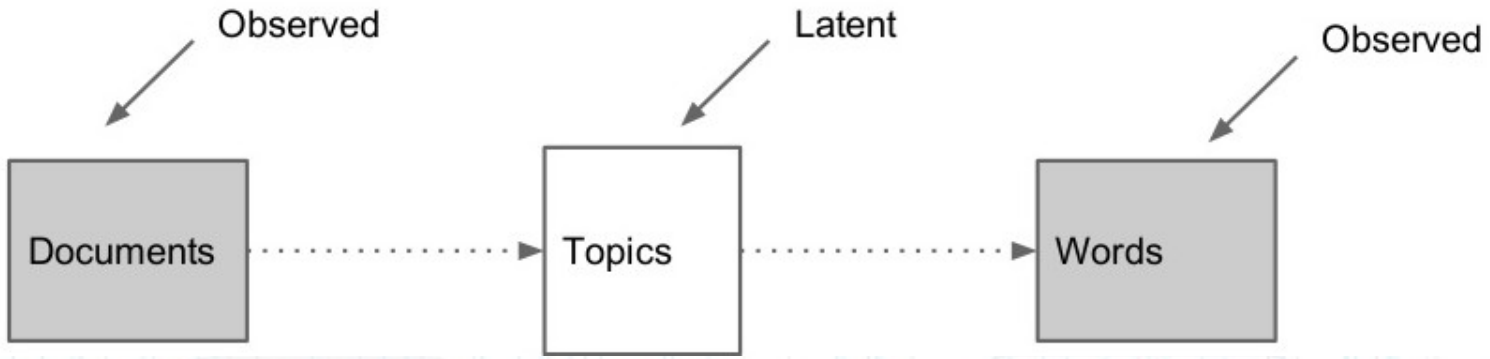
- Проблемы с большими документами (они дают приближенно равные маленькие скалярные произведения) – это “проклятие размерности”
- Поисковая система находит документы только со словами из запроса. Документы на ту же тему, но другими словами – не находятся

Тематическое моделирование

- Для анализа текстов удобно описывать их небольшим набором тем (математически: понизить размерность)
- Это позволит легко:
  - проводить категоризацию документов
  - аннотировать тексты
  - вычислять близость документов
- Сферы применения: информационный поиск, категоризация, поиск рецензентов/экспертов, рекомендательные системы, аннотирование изображений,...

Латентный семантический анализ

- Ключевая идея: любой текст является смесью небольшого количества скрытых (latent) составных элементов (тем)



- Пример: наша лекция сегодня состоит из линейной алгебры, теории вероятностей, моделирования
  Значит, она с большой вероятностью должна содержать слова:
  ЛА: вектор, скалярное произведение, ортогональный, SVD-разложение,...
  ТВ: вероятность, Байес, при условии, распределение,…
  М: модель, соответствие, проверка,...

Латентный семантический анализ

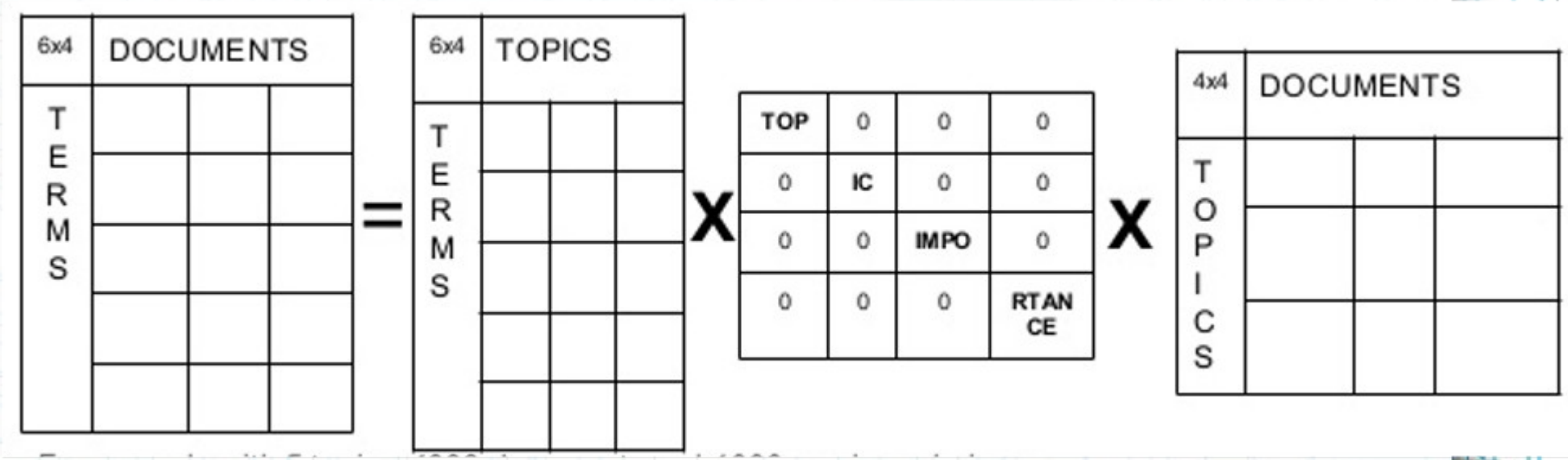
- [Deerwester и др. ‘90]:
- Считает частоты встречаемости слов в каждом документе
- Записывает их в term-document матрицу
- Понижает размерность по методы PCA: SVD-разложение + сокращение числа компонент
- Темы – небольшой набор ортогональных векторов, лучшим образом приближающий исходную линейную оболочку документов

Пример

$$\begin{array}{cc} & \mathbf{d}_j \\ & \downarrow \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

SVD-разложение



Числа в диагональной матрице имеют смысл “важностей” тем в нашей коллекции документов

Пример

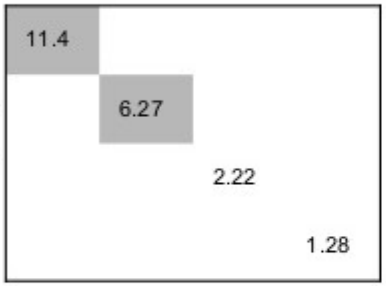
3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

=

	To1	To2	To3	To4
Te1	-0.33	-0.53	0.37	-0.14
Te2	-0.32	-0.54	-0.49	0.35
Te3	-0.62	-0.10	0.26	-0.14
Te4	-0.38	0.42	0.30	-0.24
Te5	-0.36	0.25	-0.68	-0.47
Te6	-0.37	0.42	0.02	0.75

X

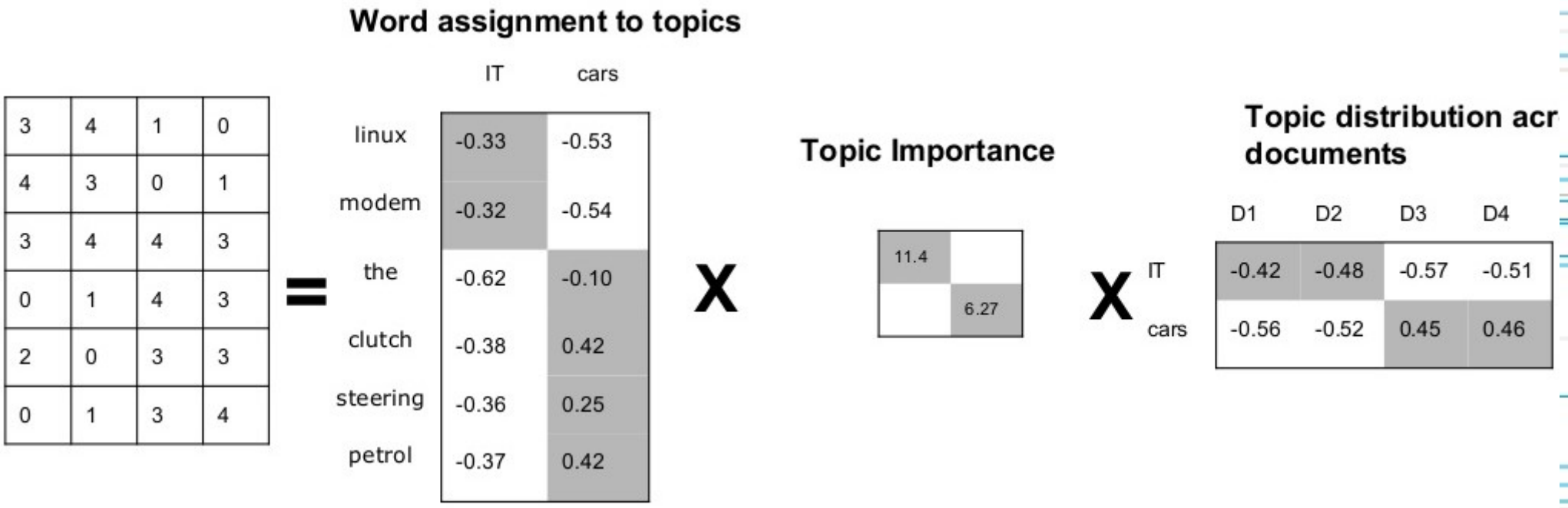
Topic Importance



X

	D1	D2	D3	D4
To1	-0.42	-0.48	-0.57	-0.51
To2	-0.56	-0.52	0.45	0.46
To3	-0.65	0.62	0.28	-0.35
To4	-0.30	0.34	-0.63	0.63

Пример



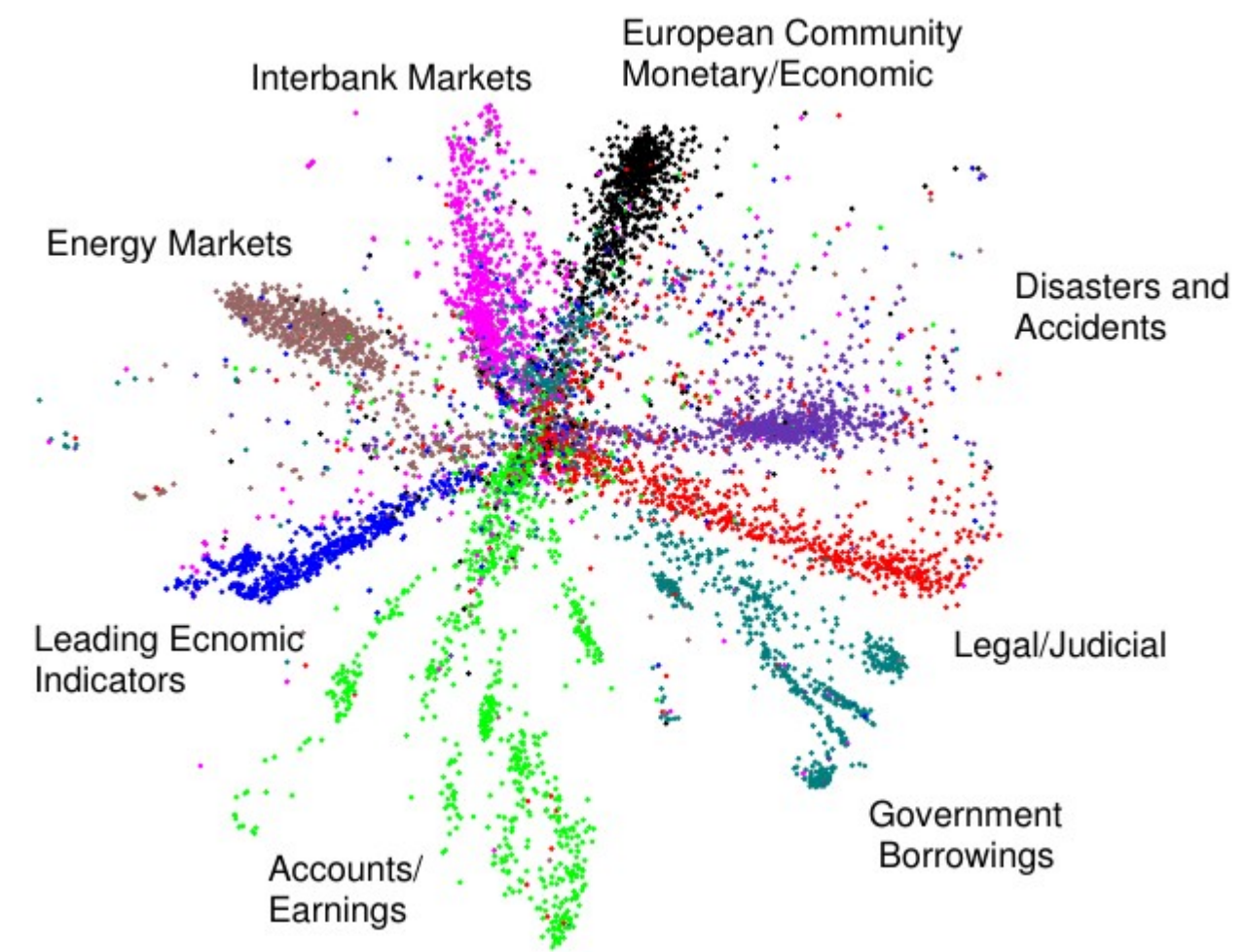
3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

Пример работы
Матрица: слова-темы

Блоки – темы. Представлены слова с максимальными координатами в темах.

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

Пример работы 2
Визуализация документов

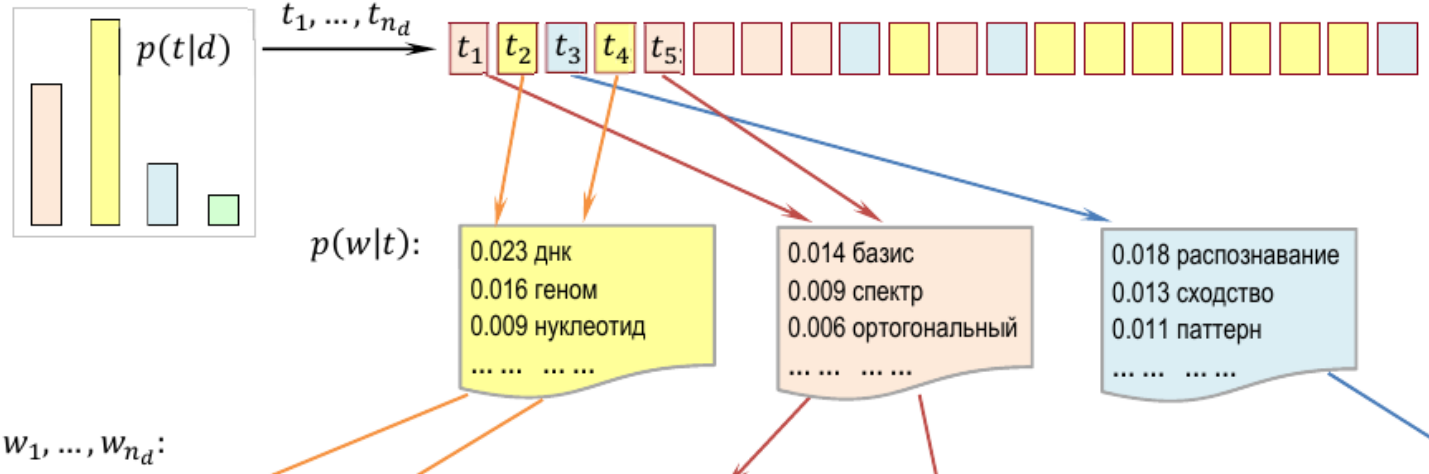


Вероятностный подход к LSA (PLSA)

- Вместо сокращения размерности документов предположим, что документы – случайны и порождены совместными вероятностными распределениеми:
  (слова,темы) и (темы,документы)
- Найдем параметры этого распределения
- Математически получается та же формула, что и в SVD. Но SVD находит темы оптимизируя евклидово расстояние, а PLSA - правдоподобие

Случайный процесс порождения документов в модели PLSA

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



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Восстановление плотности p(w|t) и p(t|d)

- Дано: n_{dw} – количество вхождений слова w в документ d, n_d – размер документа

$$\frac{n_{dw}}{n_d} \approx p(w|d)$$

- Найти: вероятности терминов в темах, вероятности тем в документах

$$\phi_{wt} = p(w|t)$$
$$\theta_{td} = p(t|d)$$

так, чтобы выполнялось равенство:

$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

Принцип максимума правдоподобия

- Правдоподобие коллекции документов:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

- Максимизация логарифма правдоподобия

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

эквивалентна максимизации функционала:

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

EM-алгоритм

- E-шаг 1:
  Оцениваем число слов в документе d, порожденных темой t. По формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$$

  отсюда:

$$n_{td} = \sum_w n_{wd} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$$

- M-шаг 1:
  оценка вероятности темы в документе

$$\theta_{td} = p(t|d) = \frac{n_{td}}{n_d}$$

EM-алгоритм

- E-шаг 2:
  Оцениваем количество вхождений слова w в тему t

$$n_{wt} = \sum_d n_{wd} \frac{\phi_{wt}\theta_{td}}{\sum_t \phi_{wt}\theta_{td}}$$

- M-шаг 2:
  оценка вероятности вхождения слова в тему

$$\phi_{wt} = p(w|t) = \frac{n_{wt}}{n_t}$$

Неединственность решения

- Если задача разрешима, то решений бесконечно много:

$$\underset{W\times D}{\left(\frac{n_{dw}}{n_d}\right)} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta} = (\Phi S)(S^{-1}\Theta) = \underset{W\times T}{\Phi'} \cdot \underset{T\times D}{\Theta'}$$

S – произвольная невырожденная матрица

Недостатки PLSA

- Переобучение
- Некоторые документы имеют пару выраженных тем, а для многих – почти все темы по чуть-чуть присутствуют
- Если уменьшать число тем – получится плохая модель, если увеличивать – много документов с тематической неопределенностью
- Как сделать тем много и решить задачу с условием: в каждом документе не более 3-5 тем?
  Именно для этого служит LDA.

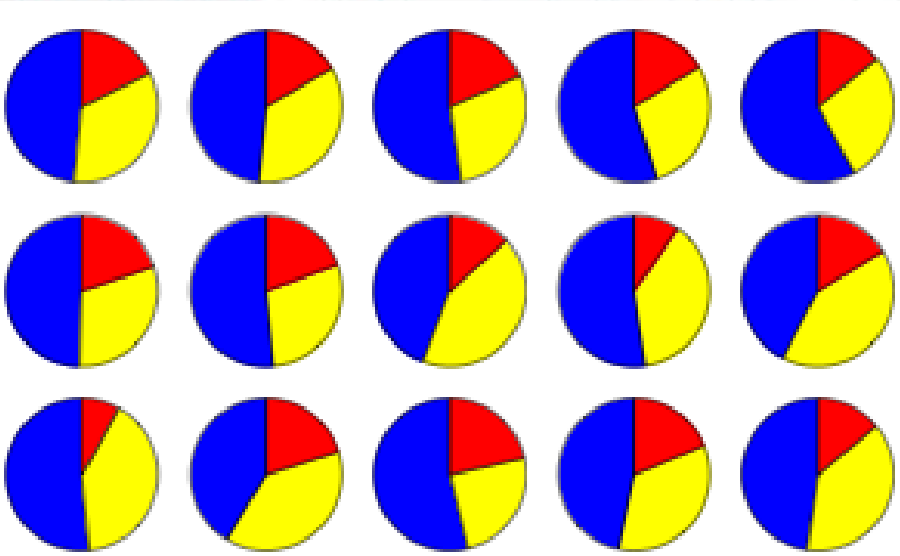
Распределение Дирихле

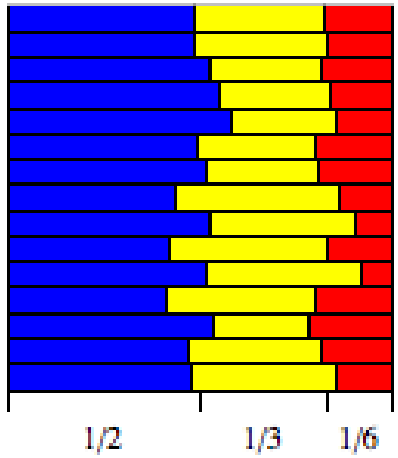
- Пусть отрезок длины 1 разрезан на K частей. Рассмотрим эксперимент с фиксированным исходом: случайно (равномерно) бросили несколько точек на отрезок и в каждой части оказалось ровно α_i-1 штук. Какими могут быть длины частей?
- Ответ: они распределены по закону Дирихле!
- Вероятность того, что вероятность каждого из K взаимоисключающих событий равна x_i при условии, что каждое событие наблюдалось α_i-1 раз

$$f(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

$$\sum_{i=1}^{K} x_i = 1$$

Наглядная трактовка





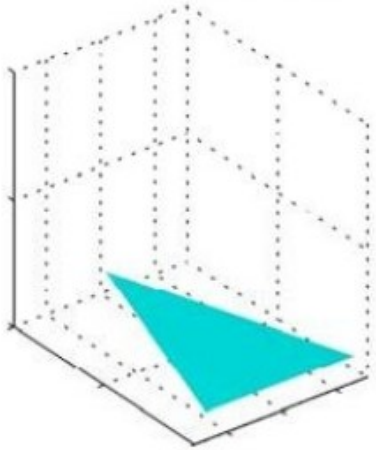
Другая трактовка: урны и шары

- Рассмотрим урну с шарами К различных цветов. Изначально в ней α_1 шаров цвета 1, α_2 – цвета 2,…
- Возьмем случайно из урны шар и положим его назад вместе с шаром того же цвета
- Если повторять это бесконечно много раз, то пропорции цветов в урне будут подчинены распределению Дирихле $\mathrm{Dir}(\alpha_1,\ldots,\alpha_K)$

Распределение Дирихле

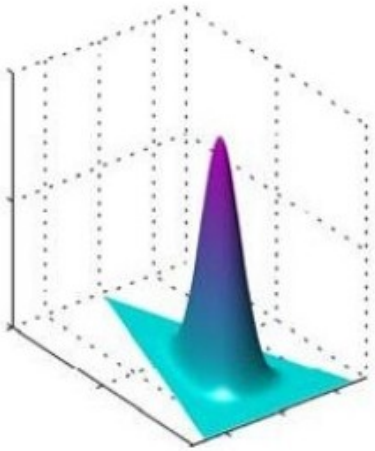
N=3:

Params = [1, 1, 1]



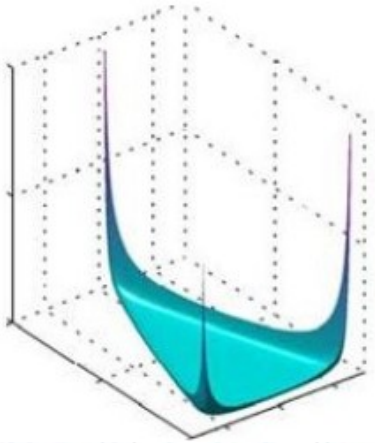
Bigger than 1

Params = [10, 10, 10]



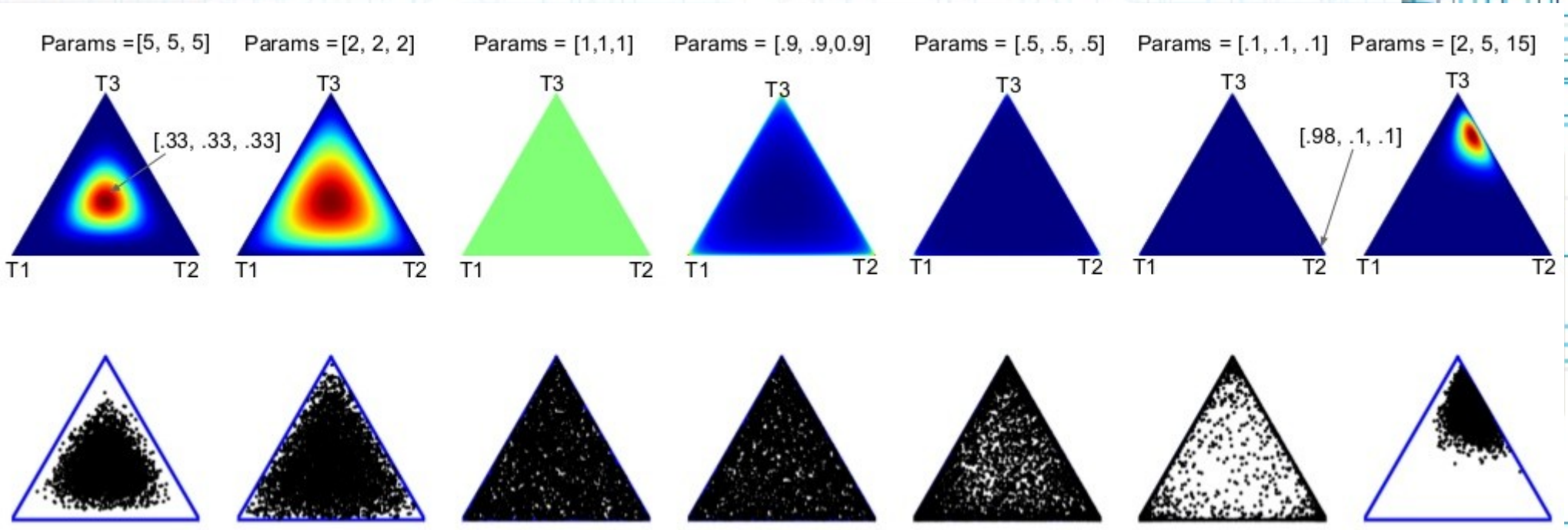
Less than 1

Params = [.1, .1, .1]



Распределение Дирихле

- Распределение тем по документам и слов по темам можно моделировать распределением Дирихле!



Латентное размещение Дирихле

- Upgade PLSA:
  приближенно оценим (или зададим) среднее число тем в документе K и среднее число ключевых слов в теме V
- Смоделируем распределение тем по документам и слов по темам размещениями Дирихле с параметрами:
  K-мерный вектор α (чем меньше α, тем меньше выраженных тем в документе)
  V-мерный вектор β (чем меньше β, тем меньше слов, характеризующих тему)
- Обычно все координаты векторов α и β берут одинаковыми

Случайный процесс порождения документов в модели LDA

- Дано: количество тем K в документе и слов в теме V, параметры α и β
- Для каждого документа генерируем вероятности тем из распределения Дирихле с параметром α
- Для каждой темы генерируем вероятности слов из распределения Дирихле с параметром β
- Для каждой позиции в документе
  - Выбираем случайно тему согласно сгенерированным вероятностям
  - Вибираем случайно слово, согласно вероятностям слов в теме

Принцип максимума апостериорной вероятности

- Принцип максимума правдоподобия модели $f(x|\theta)$ для случайной величины x:

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} f(x|\theta)$$

- Если параметр θ – случайная величина с известным априорным распределением g, то по формуле Байеса можно вычислить апостериорное распределение θ.
- Принцип максимума апостериорной вероятности:

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \frac{f(x|\theta)\, g(\theta)}{\int_{\Theta} f(x|\theta')\, g(\theta')\, d\theta'} = \arg\max_{\theta} f(x|\theta)\, g(\theta)$$

Принцип максимума апостериорной вероятности

$$\ln \prod_{d \in D} \prod_{w \in d} p(d,w)^{n_{dw}} \prod_{t \in T} \mathrm{Dir}(\phi_t | \beta) \prod_{d \in D} \mathrm{Dir}(\theta_d | \alpha) \to \max_{\Phi, \Theta}$$

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} +$$

$$+ \sum_{t \in T} \sum_{w \in W} \ln \phi_{wt}^{\beta_w - 1} + \sum_{d \in D} \sum_{t \in T} \ln \theta_{td}^{\alpha_t - 1} \to \max_{\Phi, \Theta}$$

Регуляризованный EM-алгоритм

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\ln \text{ правдоподобия } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \tilde{\beta}_w \ln \phi_{wt} + \sum_{d,t} \tilde{\alpha}_t \ln \theta_{td}}_{\text{критерий регуляризации } R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

Если коэффициенты регуляризации >0, тогда чем больше логарифмы тем лучше. А когда мы берем α и β < 1, то получается чем больше нулевых вероятностей, тем лучше!

- В PLSA:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- В LDA:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0} \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}$$