

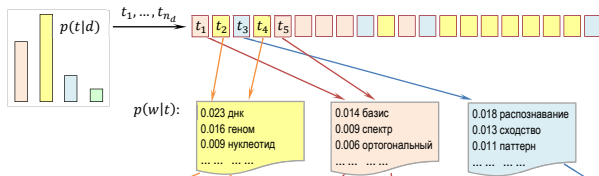
# Сглаживание, разреживание и декоррелирование тематических моделей

Анна Потапенко

научный семинар ШАД, 30 сентября 2014

# Специфика текста на естественном языке

Термины предметных областей образуют ядра тем.  
Предложения содержат большую долю нетематических слов.



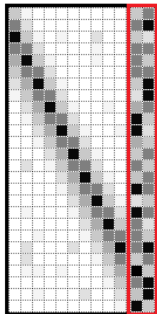
$w_1, \dots, w_{n_d}$ :

Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в **геномных** последовательностях. Метод основан на **разномасштабном** оценивании **сходства нуклеотидных** последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим **ортогональным базисам**. Найлены 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).

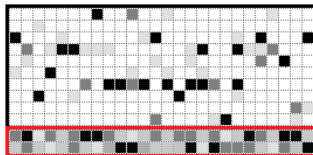
## Гипотезы о структуре интерпретируемых тем

- 1 Предметные темы разреженные, существенно различные, состоят из терминов, образующих ядро темы
- 2 Фоновые темы плотные, содержат слова общей лексики

$W \times T$ -матрица  $\Phi$



$T \times D$ -матрица  $\Theta$



## Разреженность предметных тем

**Гипотеза:** Предметная тема состоит из небольшого числа слов; документ относится к небольшому числу предметных тем.

**Регуляризатор:** Максимизируем дивергенцию между искомыми распределениями  $\phi_{wt}$ ,  $\theta_{td}$  и равномерными распределениями  $\beta_w$ ,  $\alpha_t$ :

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

**Формула М-шага:**

$$\phi_{wt} \propto (n_{wt} - \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{dt} - \alpha_0 \alpha_t)_+.$$

## Различность предметных тем

**Гипотеза:** Предметные темы существенно различны между собой, т.к. описывают отдельные области.

**Регуляризатор:** Минимизируем ковариации между вектор-столбцами  $\phi_t$ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max,$$

**Формула М-шага:**

$$\phi_{wt} \propto \left( n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+.$$

## Сглаженность фоновых тем

**Гипотеза:** Фоновые темы содержат все слова коллекции и присутствуют в каждом документе.

**Регуляризатор:** Минимизируем дивергенцию между искомыми распределениями  $\phi_{wt}$ ,  $\theta_{td}$  и равномерными распределениями  $\beta_w$ ,  $\alpha_t$ :

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \rightarrow \max.$$

**Формула М-шага:**

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \quad \theta_{td} \propto n_{dt} + \alpha_0 \alpha_t.$$

## Меры интерпретируемости

Как измерить интерпретируемость темы автоматически?

**Когерентность:** тема интерпретируемая, если её наиболее вероятные слова часто встречаются в текстах «рядом».

Для оценки используем топ- $k$  слов темы:

$$TC\_PMI = \frac{2}{k(k-1)} \sum_{j=2}^k \sum_{i=1}^j \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

где вероятности оцениваются по частоте документов, в которых встречается одно из слов или оба слова сразу.

Когерентность хорошо коррелирует с человеческими оценками [Automatic evaluation of topic coherence, Newman et al., 2010].

## Чистота и контрастность тем

**Гипотеза:** тема интерпретируема, если она содержит с большими вероятностями отличительные слова, которые практически не встречаются в других темах.

**Ядро** темы  $t$  – это множество слов, для которых  $p(t|w) \propto \phi_{wt} n_t$  превышает заданный порог  $\kappa = 0.25$ .

Три показателя качества на основе понятия ядра  $W_t$ :

- размер ядра:  $|W_t|$ ;
- чистота темы:  $\sum_{w \in W_t} p(w|t)$ ;
- контрастность темы:  $|W_t|^{-1} \sum_{w \in W_t} p(t|w)$ .



## Условия экспериментов

### Данные: NIPS

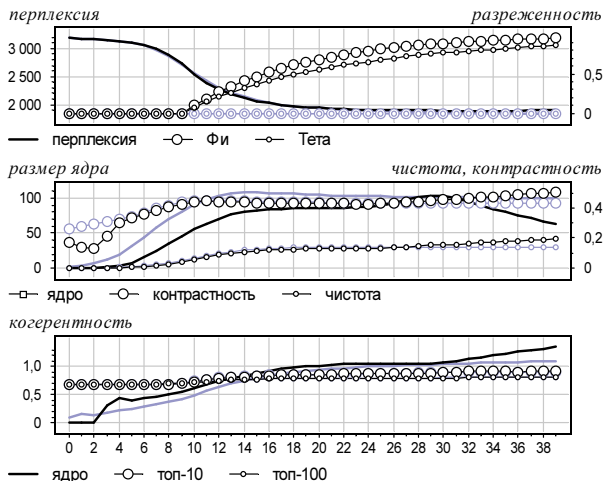
- $|D| = 1566$  статей конференции NIPS на английском языке;
- суммарной длины  $n \approx 2.3 \cdot 10^6$ ,
- словарь  $|W| \approx 1.3 \cdot 10^4$ .
- Контрольная коллекция:  $|D'| = 174$ .

### Меры качества:

- перплексия:  $\mathcal{P} = \exp(-\mathcal{L}/n)$
- разреженность матриц  $\Phi$  и  $\Theta$  (доля нулей)
- средний размер ядра, чистота и контрастность тем
- средняя когерентность: по топ-10, по топ-100, по ядру

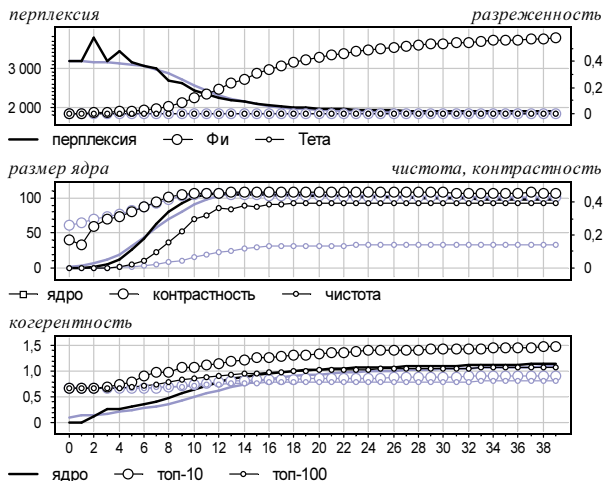
Число тем 100 (из них 10 фоновых), число итераций 40.

## Разреживание предметных тем



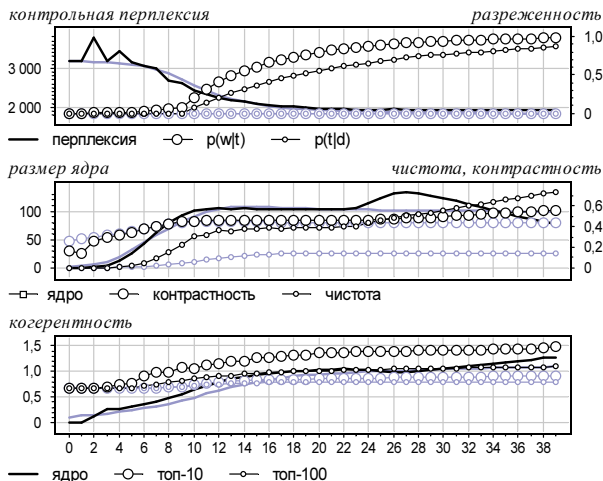
Голубой: PLSA, **черный**: разреживание (параметры: 10, 0.08, 0.1.).

## Декоррелирование предметных тем



Голубой: PLSA, черный: декоррелирование ( $\tau = 200000$ ).

# Совмещение разреживания и декоррелирования



Голубой: PLSA, черный: разреживание + декоррелирование.

## Примеры тем (топ-20 наиболее вероятных слов)

- **PLSA, предметная тема:** face, images, faces, recognition, set, image, based, hme, facial, representation, view, figure, model, experts, network, human, expert, space, examples, system
- **ARTM, предметная тема:** face, faces, facial, cottrell, pentland, gesture, lane, emotion, person, steering, appearance, baluja, setpoint, camera, tracking, pose, pomerleau, mouth, darrell, lighting
- **ARTM, фоновая тема:** model, data, models, parameters, noise, neural, mixture, prediction, set, gaussian, likelihood, networks, test, figure, training, performance, network, number, input, results

Красным отмечены слова, входящие в ядра тем.

Регуляризованная модель, комбинирующая критерии разреживания, сглаживания, декоррелирования:

- улучшает когерентность, чистоту и контрастность тем;
- группирует нетематические слова в фоновых темах;
- строит сильно разреженные предметные темы, содержащие специфические термины отдельных областей.