Вероятностные тематические модели контекста для больших коллекций документов

Анна Потапенко

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

12 марта 2015

Задача выделения тем в коллекции документов

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах: $p(w|t)$ — вероятность встретить термин w в теме t.

Дана коллекция текстовых документов

Темы документа мы не видим, а видим только его слова:
n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Задача – определить латентные темы:

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

Приложения тематического моделирования

- Тематический поиск документов и объектов
  по тексту любой длины или по любому объекту
- Поиск научных статей, экспертов, рецензентов, проектов
- Выявление трендов и фронта исследований
- Суммаризация и аннотирование текстовых документов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики
- ...

Требования к тематической модели:

- Интерпретируемость выделяемых тем
- Обработка больших объемов данных

Вероятностная формализация постановки задачи

Базовые предположения:

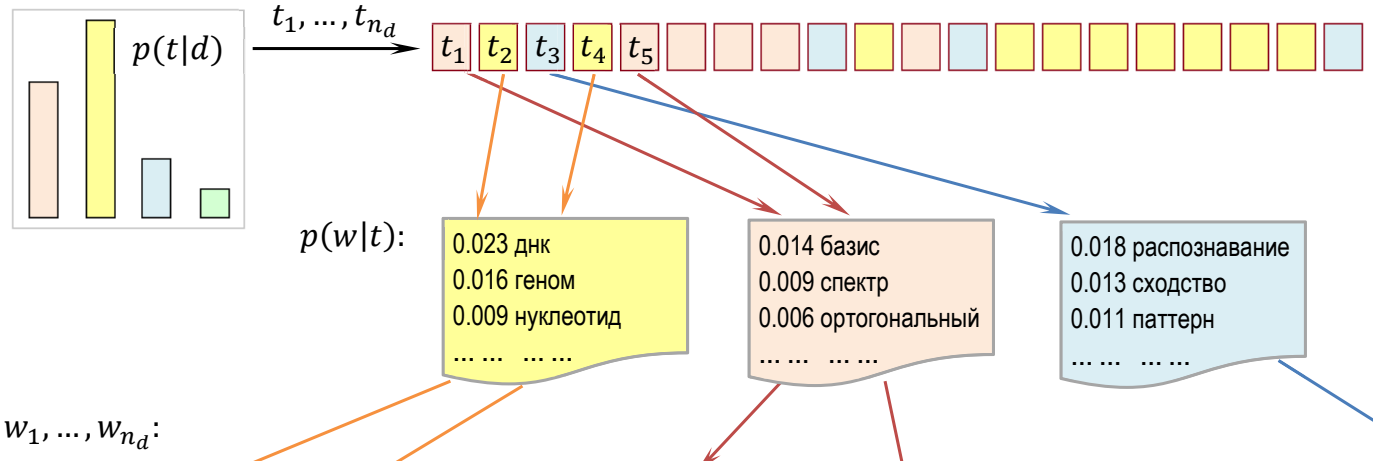
- порядок документов в коллекции не важен
- порядок слов в документе не важен
- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^{n} \sim p(d, w, t)$
- d_i, w_i — наблюдаемые,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

Вероятностная модель порождения документа:

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

Вероятностная модель порождения документа:

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



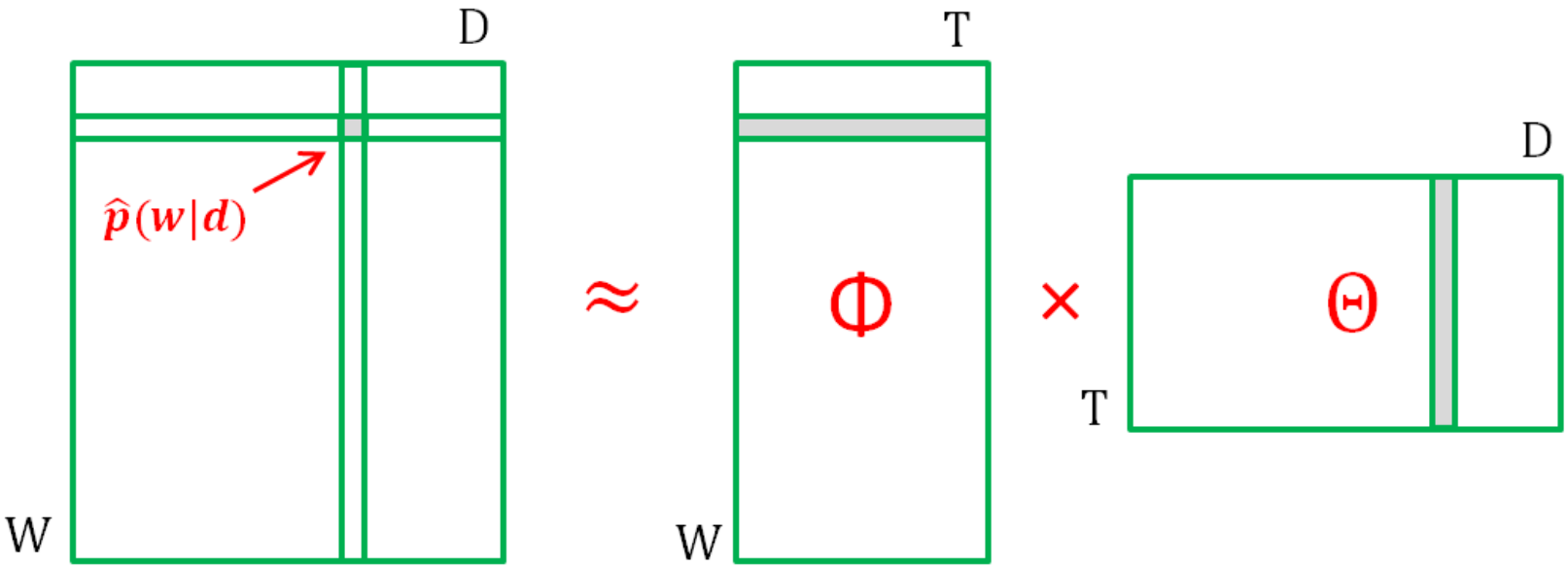
Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным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

Задача построения тематической модели:

Частотное распределение слов в документе $\hat{p}(w|d) = n_{dw}/n_d$

приблизить модельным: $\hat{p}(w|d) \approx p(w|d) = \sum_{t \in T} \phi_{wt}\, \theta_{td}$, где

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



PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Принцип максимума правдоподобия:

$$L = \ln \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1$$

Это задача максимизации неполного правдоподобия, решается EM-алгоритмом.

- **E-шаг:** распределение на скрытые переменные $p(T|D, W)$
- **M-шаг:** максимизация $\mathbb{E}_{p(T|D,W)} L(\Phi, \Theta) \to \max_{\Phi, \Theta}$

EM-алгоритм

E-шаг. Выразим $p(t|d,w)$ через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}$$

$n_{dwt} = n_{dw}p(t|d,w)$ — оценка числа троек (d,w,t) в коллекции

M-шаг. Частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{dt}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

или краткая запись:

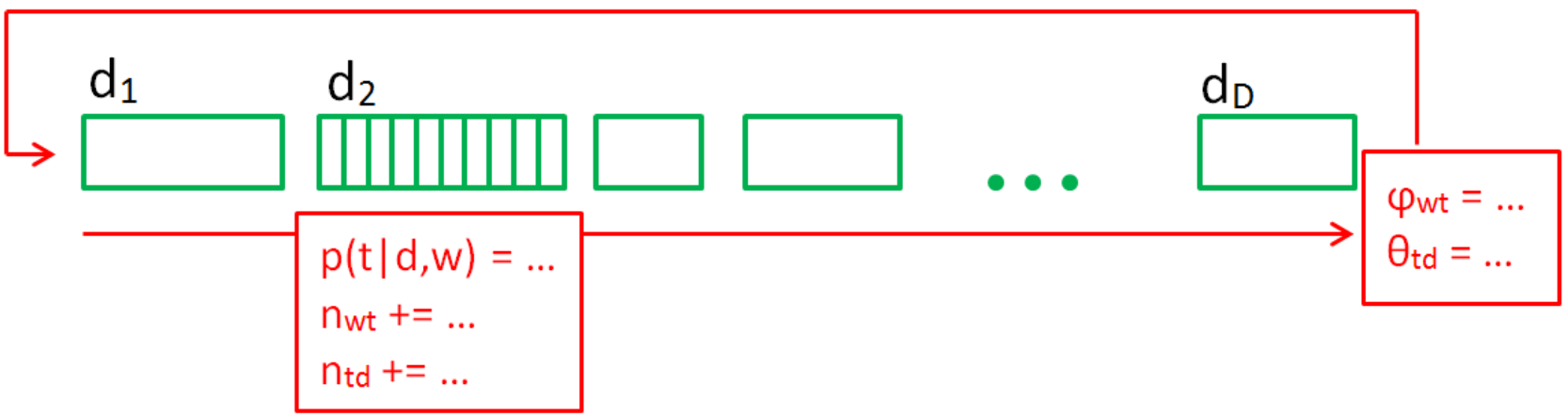
$$\phi_{wt} \propto n_{wt} \qquad \theta_{td} \propto n_{dt}$$

Схема итерационного процесса

Бегаем по коллекции документов, повторяя до сходимости:

- $p(t|d,w) \propto \phi_{wt}\theta_{td}$
- $\phi_{wt} \propto n_{wt}, \quad n_{wt} = \sum_{d \in D} n_{dw} p(t|d,w)$

  $\theta_{td} \propto n_{td}, \quad n_{td} = \sum_{w \in W} n_{dw} p(t|d,w)$

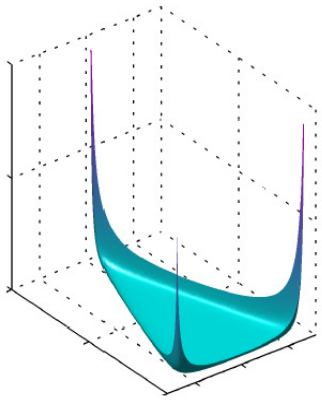


Модель LDA (Latent Dirichlet Allocation, [Blei, Ng, Jordan 2003])

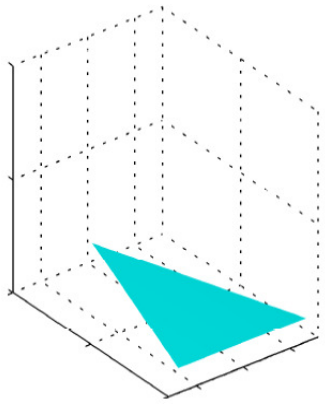
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \ \beta_t \geqslant 0;$$

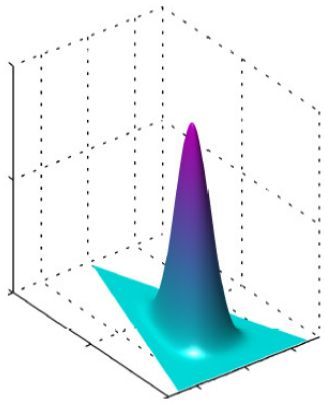
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t \geqslant 0;$$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$



$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Отличие моделей PLSA и LDA в оценивании Φ и Θ:

- в PLSA — точечные оценки максимума правдоподобия:

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

- в LDA — апостериорные распределения методами приближенного байесовского вывода (Variational Bayes, Gibbs Sampling):

$$\phi_{wt} \sim Dir\left(\phi_t | \frac{n_{wt} + \beta_w}{n_t + \beta_0}\right), \qquad \theta_{td} \sim Dir\left(\theta_d | \frac{n_{td} + \alpha_t}{n_d + \alpha_0}\right).$$

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

Potapenko A. A., Vorontsov K. V. Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

Проблемы моделей PLSA и LDA

Задача построения тематической модели

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

это задача матричного разложения:

$$\underset{W \times D}{\hat{F}} \approx \underset{W \times D}{F} = \underset{W \times T}{\Phi}\, \underset{T \times D}{\Theta}$$

Матричное разложение неединственно:

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

Задача является некорректно поставленной и нуждается в регуляризации!

Подход аддитивной регуляризации:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где $R_i(\Phi,\Theta)$ – регуляризаторы, накладывающие дополнительные требования на распределения,
τ_i – коэффициенты регуляризации,
устанавливающие баланс между требованиями.

В EM-алгоритме меняется M-шаг:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \quad \theta_{td} \propto \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$$

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks, and Texts AIST-2014. – LNCS, Springer.

Готовые реализации тематических моделей:

- **Gensim** – библиотека для обработки текстов на Python, radimrehurek.com/gensim
- **MALLET** – библиотека для обработки текстов на Java, mallet.cs.umass.edu
- **Vowpal Wabbit** – библиотека машинного обучения с быстрой реализацией online LDA, hunch.net/ vw
- **GibbsLDA++** – C++ реализация обучения LDA методом сэмплирования Гиббса, gibbslda.sourceforge.net
- **От авторов** – реализации LDA (D. Blei) и online-LDA (M. Hoffman), www.cs.princeton.edu/ blei/topicmodeling.html
- **bigARTM** – библиотека тематического моделирования на основе аддитивной регуляризации. C++, распределенная, интерфейсы под C++ и Python, bigartm.org

Задача перехода от мешка слов к последовательному тексту

- Использование информации о контексте слов в последовательном тексте документов позволит улучшить интерпретируемость тем и применить тематические модели для более широкого круга приложений.

Идея дистрибутивной семантики

Frith (1957): "You shall know a word by the company it keeps."

Дистрибутивная гипотеза: два слова семантически близки, если они встречаются в схожих контекстах.

Различия моделей дистрибутивной семантики:

- Типы контекста: документ, окно заданной ширины, синтаксическая фраза, ...
- Оценка частоты встречаемости: абсолютная частота, TF-IDF, PMI, ...
- Мера расстояния между векторами: косинусная, KL-дивергенция, ...
- Метод уменьшения размерности: SVD, ...

The TOEFL synonym match task

- *You will find the office at the main **intersection**.*
  (a) **place**
  (b) **crossroads**
  (c) **roundabout**
  (d) **building**

- Humans:
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%

- Machines:
  - Classic LSA: 64.4%
  - Padó and Lapata's dependency-filtered model: 73%
  - Rapp's SVD-based model trained on lemmatized BNC: 92.5%

Контекст в тематическом моделировании

Когерентность – автоматическая мера интерпретируемости.

Идея: тема интерпретируема, если ее топ-слова часто встречаются в одном и том же *контексте*:

$$TCPMI = \sum_{i=1}^{10} \sum_{j=i+1}^{10} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

Newman et al. – Automatic evaluation of topic coherence, 2010;
Mimno et al. – Optimizing Semantic Coherence in Topic Models, 2011;
Aletras, Stevenson – Evaluating Topic Coherence Using Distributional Semantics, 2013

Когерентность по **окну из 10 слов** лучше коррелирует с оценками экспертов, чем когерентность по контексту длиной в **документ**!

Word Intrusion

1 / 10
floppy alphabet computer processor memory disk

2 / 10
molecule education study university school student

3 / 10
linguistics actor film comedy director movie

4 / 10
islands island bird coast portuguese mainland

Topic Intrusion

6 / 10

DOUGLAS_HOFSTADTER

Douglas Richard Hofstadter (born February 15, 1945 in New York, New York) is an American academic whose research focuses on consciousness, thinking and creativity. He is best known for ", first published in

Show entire excerpt

student	school	study	education	research	university	science	learn
human	life	scientific	science	scientist	experiment	work	idea
play	role	good	actor	star	career	show	performance
write	work	book	publish	life	friend	influence	father

Методика word intrusion, Boyd-Graber et al. – Reading Tea Leaves: How Humans Interpret Topic Models, NIPS-2009.

Тематическая модель на мешке контекстов

- *Задача:* определение смысла слова
- *Данные:* наборы контекстов для нескольких слов
- *Модель:* $p(w|c) = \sum_{s=1}^{S} p(w|s)p(s|c)$
- *Обучение:* сэмплирование Гиббса
- Учет дополнительных признаков как «псевдослов»
- Хорошая F-мера на данных SemEval-2007 (35 слов)

Senses of *drug* (WSJ)
1. U.S., administration, federal, against, war, dealer
2. patient, people, problem, doctor, company, abuse
3. company, million, sale, maker, stock, inc.
4. administration, food, company, approval, FDA

Brody, Lapata – Bayesian Word Sense Induction, 2009

Марковские тематические модели последовательного текста

- **HMMLDA:** скрытая марковская модель для связи слов в тексте, один из её классов – тематическая модель.
- **HMTM:** скрытая марковская модель, классы – темы.
- **BTM:** трехмерная матрица Φ вероятностей слова, при условии темы и предыдущего слова.
- **LDACOL:** бинарные переменные для каждого слова «образовать биграмму с предыдущим / не образовать» + матрица переходов из слова в слово для биграмм.
- **TNG:** бинарные переменные и матрица переходов получают дополнительную размерность по темам.

Griffiths, Steyvers – Integrating Topics and Syntax, 2007
Wang, McCallum – A Note on Topical N-grams, 2005
Lau, Baldwin, Newman – On Collocations and Topic Models, 2010

На практике – выделение *n*-грамм в предобработке

Схема: найти в текстах n-граммы, преобразовать в специальные слова, обучить модель мешка слов.

Как выделять n-граммы:

- статистический критерий (например, t-критерий Стьюдента), проверяющий гипотезу о том, что слова встречаются вместе чаще, чем встречались бы случайно
- привлечение морфологии и синтаксиса

Lau, Baldwin, Newman – On Collocations and Topic Models, 2010

- **Цель диссертационной работы** – разработка вычислительно эффективных способов использования контекста для построения многофункциональных тематических моделей больших текстовых коллекций.

- **Предполагаемый подход** – аддитивная регуляризация тематических моделей:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta},$$

где $R_i(\Phi,\Theta)$ – регуляризаторы, накладывающие дополнительные требования на искомые распределения, τ_i – коэффициенты регуляризации.

Обучение – модифицированный EM-алгоритм.

Регуляризатор повышения когерентности:

$$R(\Phi, \Theta) = \tau \sum_{t \in T} \sum_{(u,v) \in W \times W} \frac{N_{uv}}{N_v} n_{ut} \ln \phi_{vt} \to \max,$$

где N_u – число документов, в которых встречается термин u;
N_{uv} – число документов, в которых встречаются термины u и v.

Более понятная запись:

$$R(\Phi, \Theta) = \tau \sum_{t} n_t KL_u(\hat{p}(u|t)||\phi_{ut}) \to \max,$$

где $\hat{p}(u|t)$ вычисляется через слова, когерентные для u:

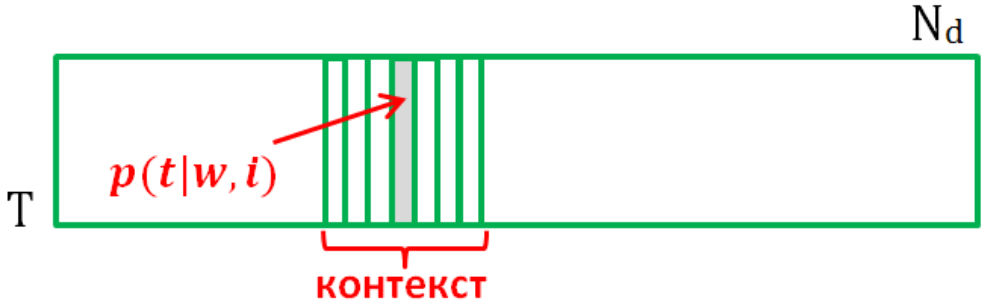
$$\hat{p}(u|t) = \sum_{u \in W} \hat{p}(v|u)\hat{p}(u|t) = \sum_{u \in W} \frac{N_{uv}}{N_u} \hat{p}(u|t)$$

Формула M-шага:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W \setminus w} \frac{N_{uw}}{N_u} n_{ut}$$

Важные особенности метода:

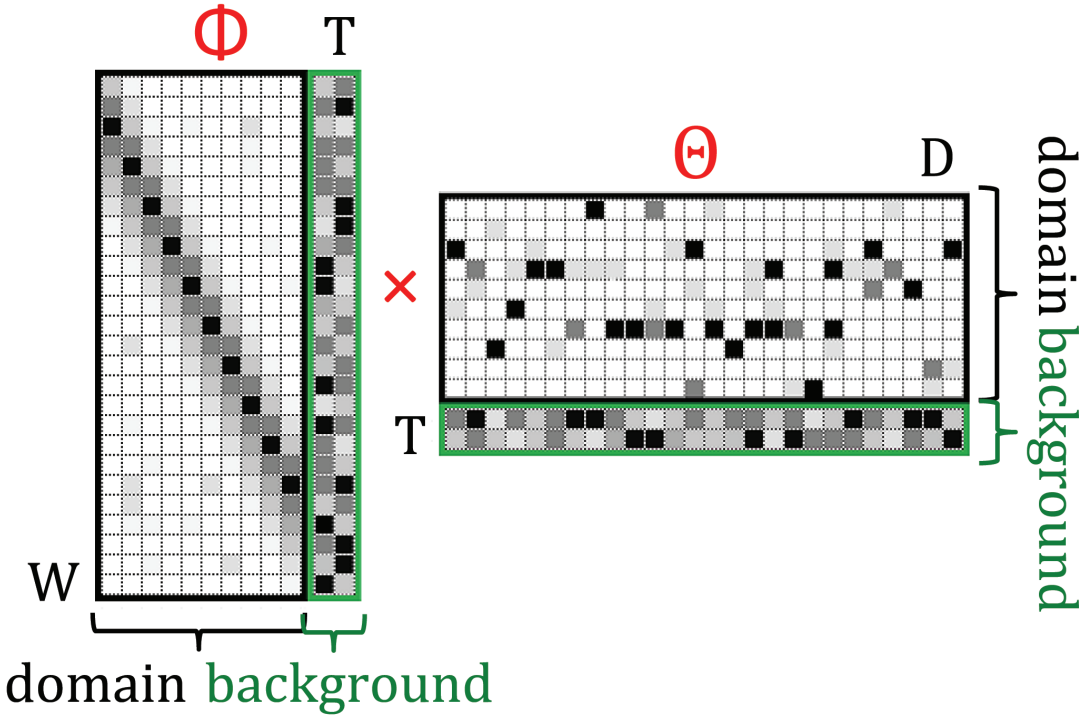
1. Учет контекста «на лету» при проходе по документу без хранения дополнительных матриц большого размера.



2. Свободное комбинирование регуляризаторов анализа контекста с другими модификациями. Уже изучены:

- повышение разреженности и различности тем
- разделение терминов предметных областей и нетематической общеупотребительной лексики
- выбор числа тем за счет исключения из модели незначимых и линейно-зависимых тем

Модель с предметными и фоновыми темами



- Два типа тем: сильно разреженные, различные предметные темы + сглаженные фоновые темы
- Достигается набором из 5 регуляризаторов
- Улучшение интерпретируемости тем по набору мер

Теоретические вопросы:

- Вероятностное обоснование модели, модифицирующей тематическую принадлежность слов на основе контекста
- Разработка гибридных подходов на стыке тематического моделирования и дистрибутивной семантики
- Исследование сходимости регуляризованных алгоритмов и способов подбора коэффициентов регуляризации

Решение задач анализа текстов:

- Разрешение лексической многозначности
- Выделение составной терминологии предметных областей
- Тематическая сегментация текста документов

Применение в конечных приложениях:

- Социологический анализ блогов (на данных ЖЖ)
- Выявление структуры и динамики интересов пользователей вопросно-ответных систем (на данных ответы@mail.ru)

Список публикаций

1. *Воронцов, Потапенко* – Робастные разреженные ВТМ // ИОИ-2012.
2. *Воронцов, Потапенко* – Регуляризация, робастность и разреженность ВТМ // Компьютерные исследования и моделирование, 2012.
3. *Potapenko, Vorontsov* – Robust PLSA Performs Better Than LDA // 35th European Conference on Information Retrieval (ECIR-2013).
4. *Воронцов, Потапенко* – Модификации EM-алгоритма для ВТМ // Машинное обучение и анализ данных, 2013.
5. *Потапенко* – Разреживание вероятностных тематических моделей // Математические методы распознавания образов (ММРО-2013).
6. *Потапенко* – Регуляризация вероятностных тематических моделей для выделения ядер тем // Ломоносов-2014.
7. *Воронцов, Потапенко* – Многокритериальная регуляризация вероятностных тематических моделей для улучшения интерпретируемости тем и определения числа тем // Диалог-2014.
8. *Vorontsov, Potapenko* – Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST-2014.
9. *Potapenko, Vorontsov* – Additive Regularization of Topic Models // Machine Learning Journal, Special Issue «Data Analysis and Intelligent Optimization».