Лекция 6. Тематическое моделирование и word2vec

Александр Юрьевич Авдюшенко

МКН СПбГУ

24 марта 2022

Факультет
математики
и компьютерных
наук
СПбГУ

Пятиминутка

- Какие основные недостатки архитектуры encoder-decoder?
- Выпишите формулу модели внимания $Attn(q, K, V)$
- Опишите два критерия для обучения BERT

План лекции

1. тематическое моделирование
   - Probabilistic latent semantic analysis (PLSA) и ARTM
   - Latent Dirichlet Allocation (LDA)
   - регуляризаторы ARTM
   - метрики качества
2. word2vec

Возможные определения

Тема —

1. специальная терминология предметной области
2. набор часто совместно встречающихся терминов
3. семантически однородный кластер текстов

Nature

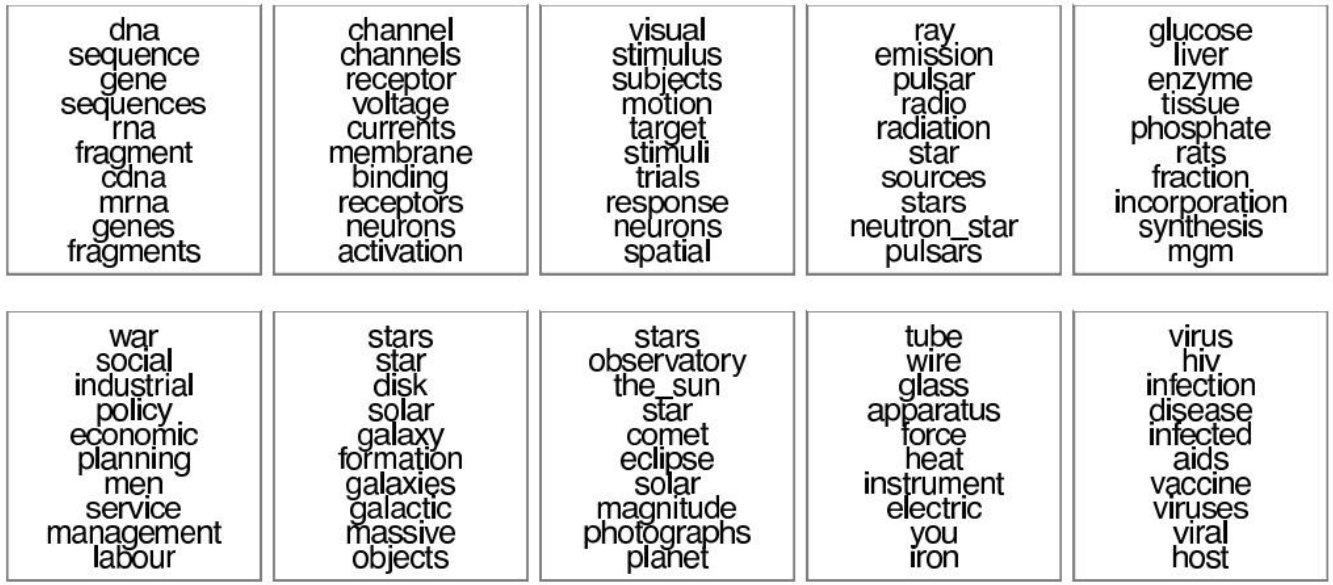


Figure 1: Posterior topics from the hierarchical Dirichlet process topic model on two large data sets. These posteriors were approximated using stochastic variational inference with 1.8M articles from the *New York Times* (top) and 350K articles from *Nature* (bottom). (See Section 3.3 for the modeling details behind the hierarchical Dirichlet process and Section 4 for details about the empirical study.) Each topic is a weighted distribution over the vocabulary and each topic's plot illustrates its most frequent words.

Matthew D. Hoffman , David M. Blei, Chong Wang, John Paisley. Stochastic Variational Inference, https://arxiv.org/abs/1206.7051

Сферы применения тематического моделирования

- анализ и агрегирование новостных потоков
- рубрикация документов, изображений, видео, музыки
- рекомендательные сервисы
- поиск экспертов, рецензентов или проектов
- выявление трендов и фронтов исследований

Математическое определение темы

- Тема — условное дискретное вероятностное распределение на множестве терминов
  $p(w|t)$ — вероятность термина w в теме t
- Тематический профиль документа — условное распределение
  $p(t|d)$ — вероятность темы t в документе d

Порождающая модель

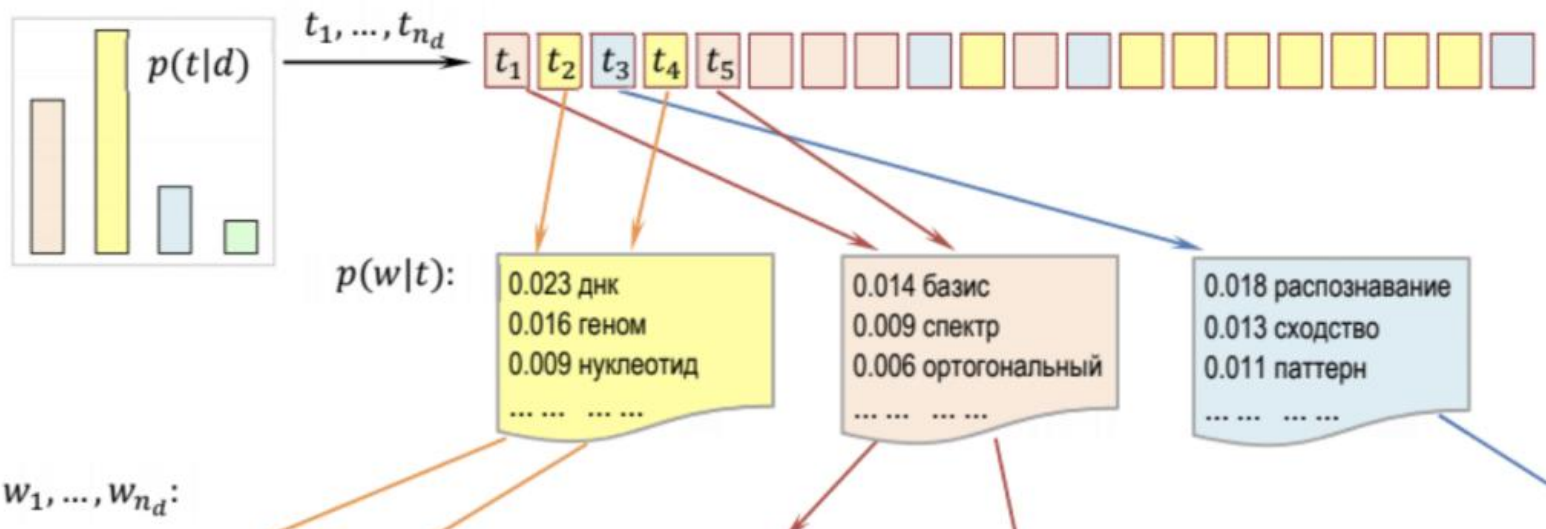
W — конечное множество слов (терминов, токенов)
D — конечное множество текстовых документов
T — конечное множество тем
$W \times D \times T$ — дискретное вероятностное пространство

Порождающая модель

W — конечное множество слов (терминов, токенов)
D — конечное множество текстовых документов
T — конечное множество тем
$W \times D \times T$ — дискретное вероятностное пространство

Гипотезы

- порядок слов в документе не важен (bag of words)
  Коллекция — это выборка $(w_i, d_i, t_i)_{i=1}^n \sim p(w, d, t)$
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$
  $p(w|d) = \sum_{t \in T} p(t|d)p(w|t)$



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).

Обратная задача

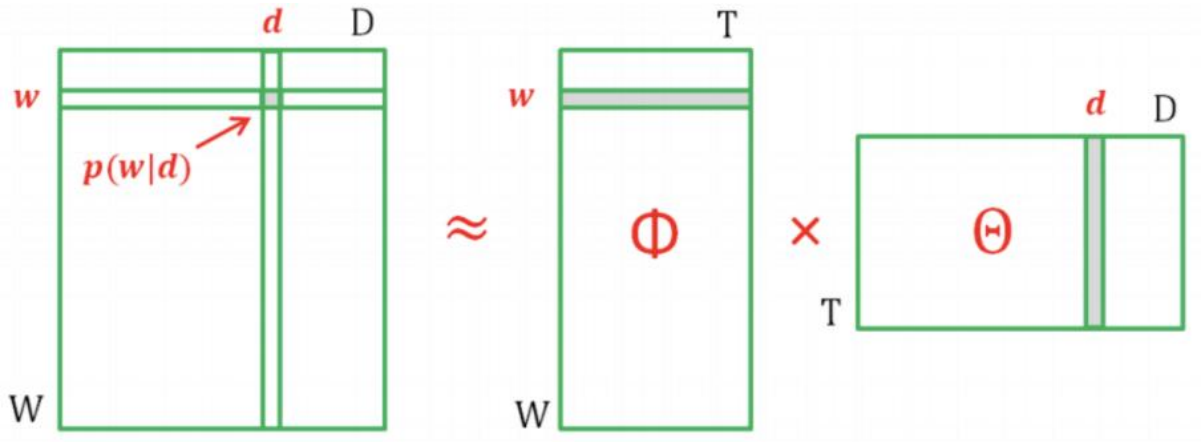
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения



Принцип максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

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

Максимизация логарифма правдоподобия

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

приводит к задаче математического программирования:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки:

$$\phi_{wt} \geq 0, \sum_{w \in W} \phi_{wt} = 1,$$

Бесконечность решений задачи

Задача матричного разложения **некорректно поставлена**:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta), rankS = |T|$
- $\mathcal{L}(\Phi', \Theta') = \mathcal{L}(\Phi, \Theta)$
- $\mathcal{L}(\Phi', \Theta') \leq \mathcal{L}(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью дополнительных критериев

Приницип максимума правдоподобия (PLSA и ARTM)

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} E-\text{шаг} \\ M-\text{шаг} \end{cases} \begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d \in D} n_{dw}p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w \in d} n_{dw}p_{tdw} \end{cases}$$

где $\underset{t \in T}{\text{norm}}(x_t) = \frac{\max(x_t,0)}{\sum_{s \in T}\max(x_s,0)}$ — операция нормировки вектора

Если $R(\Phi, \Theta) = 0$, то перед нами **Probabilistic latent semantic analysis (PLSA)**. Если нет, то **адаптивная регуляризация тематических моделей (ARTM)**.

Вырожденность тем и документов

Тема t вырождена

Если для всех терминов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leq 0$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$

Документ d вырожден

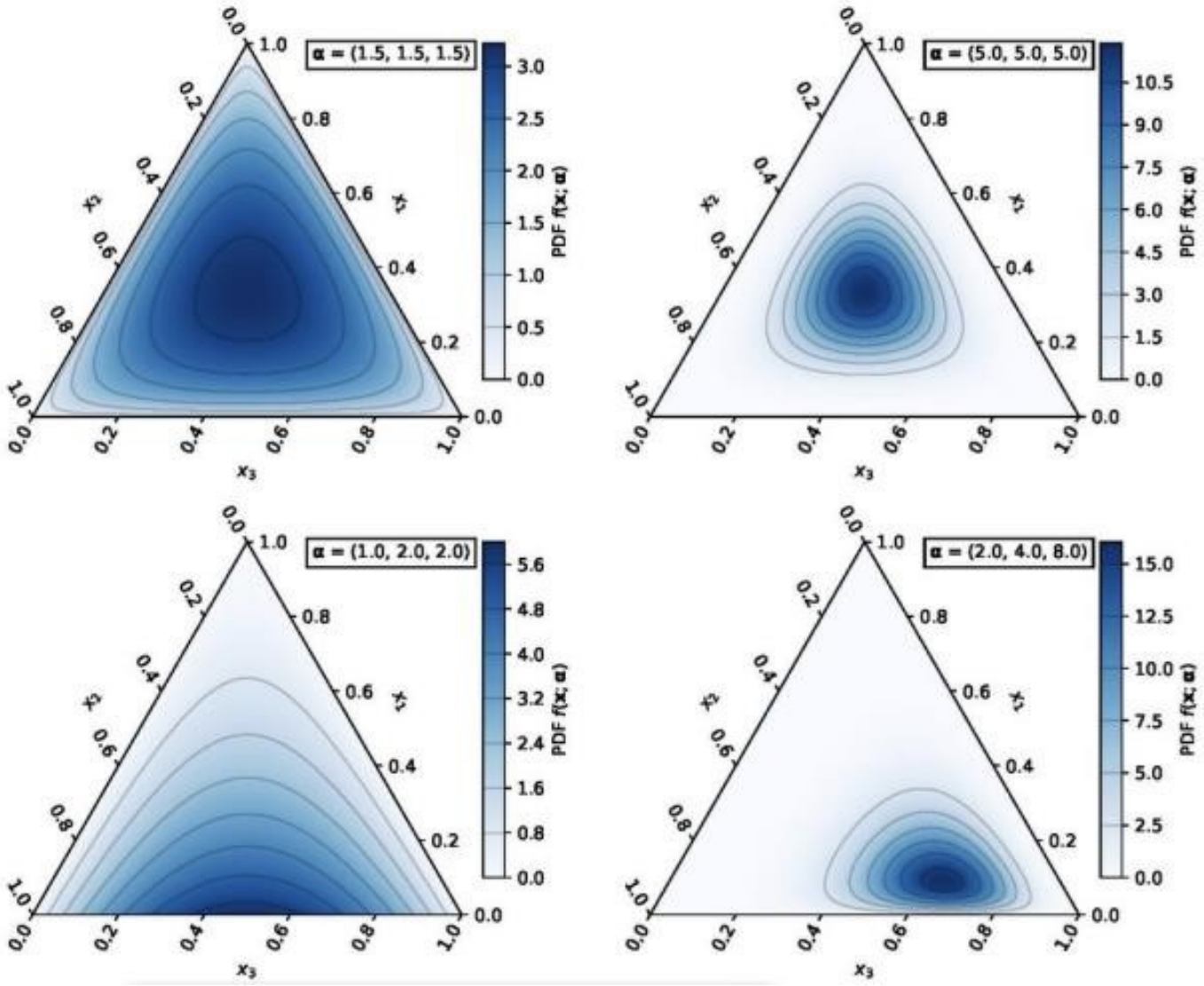
Если для всех тем $t \in T$

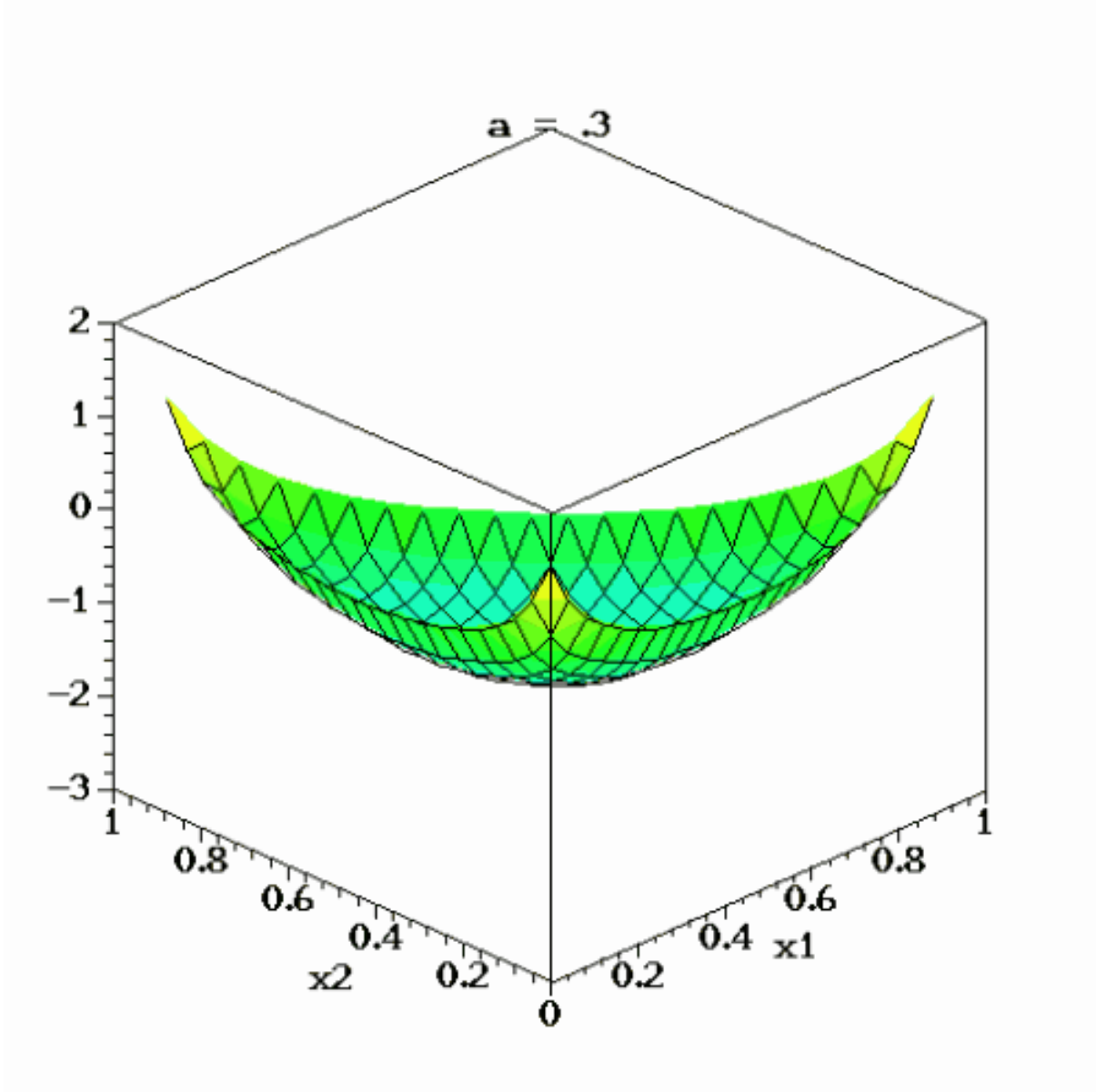
$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leq 0$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$

Latent Dirichlet Allocation (LDA)

Распределение на симплексе




a = .3
2
1
0
−1
−2
−3
1
0.8
0.6
0.4
0.2
0
x2
0
0.2
0.4
0.6
0.8
1
x1

Распределение Дирихле

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

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\sum\limits_w \beta_w)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0, \quad \beta_w > 0$$

$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\sum\limits_t \alpha_t)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0, \quad \alpha_t > 0$$

Максимум апостериори

Совместное правдоподобие данных и модели

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

Регуляризатор — логарифм априорного распределения:

$$R(\Phi, \Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}$$

M-шаг — сглаженные или разреженные частотные оценки:

$$\phi_{wt} = \underset{w \in W}{\mathrm{norm}} \left(n_{wt} + \beta_w - 1\right), \theta_{td} = \underset{t \in T}{\mathrm{norm}} \left(n_{td} + \alpha_t - 1\right)$$

при $\beta_w > 1, \alpha_t > 1$ — сглаживание,
при $0 < \beta_w < 1, 0 < \alpha_t < 1$ — слабое разреживание

Преимущества распределения Дирихле

- сопряженное к мультиномиальному
- как следствие, если априорное распределение обозначено как $\mathrm{Dir}(\alpha)$, то $\mathrm{Dir}(\alpha + \gamma)$ есть апостериорное распределение после серии наблюдений с гистограммой γ
- это очень удобно для байесовского вывода
- ещё можем управлять разреженностью через коэффициенты

Группа К.В. Воронцова (библиотека ARTM) предложила просто эмпирически строить разные регуляризаторы, возможно без вероятностного обоснования.
В нашем курсе подробно про это не будем, но вообще так можно

- строить модели мультимодального тематического моделирования (например, текст, картинка и время в каждом документе)
- решать задачи классификации, регрессии на документах одновременно с решением задачи выделения тем (не последовательно)
- выделять фоновые темы (содержащие слова общей лексики)
- усиливать различность тем

Метрики качества в тематическом моделировании

Правдоподобие и перплексия (perplexity, implicit мера)

Правдоподобие языковой модели $p(w|d)$ — чем больше, тем лучше:

$$\mathcal{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d), \quad p(w|d) = \sum_t \phi_{wt}\theta_{td}$$

Перплексия языковой модели $p(w|d)$ — чем меньше, тем лучше:

$$\mathcal{P}(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\right), \quad n = \sum_{d\in D}\sum_{w\in d} n_{dw}$$

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера различности или неопределённости слов в тексте
- коэффициент ветвления (branching factor) текста

Перплексия тестовой (отложенной) выборки

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n''}\sum_{d\in D'}\sum_{w\in d''} n_{dw}\ln p(w|d)\right), \quad n'' = \sum_{d\in D'}\sum_{w\in d''} n_{dw}$$

- $d = d' \cap d''$ — случайное разбиение тестового документа на две половины равной длины
- параметры ϕ_{wt} оцениваются по обучающей коллекции D
- параметры θ_{td} оцениваются по первой половине d'
- перплексия вычисляется по второй половине d''

Интерпретируемость

explicit мера

Экспертные оценки

- интерпретируемость темы по балльной шкале
- каждую тему оценивают несколько экспертов

Метод интрузий (intrusion)

- в список топовых слов внедряется лишнее слово
- измеряется доля ошибок экспертов при его определении

Когерентность

implicit мера

Когерентность (согласованность) темы t по k топовым словам:

$$\text{PMI}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_i, w_j)$$

где w_i — i-й термин в порядке убывания ϕ_{wt}

$\text{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information)
N_{uv} — число документов, в которых термины u, v хотя бы один раз встречаются рядом (в окне 10 слов)
N_u — число документов, в которых u встретился хотя бы один раз

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100-108.

word 1	*word 2*	*count word 1*	*count word 2*	count of co-occurrences	PMI
puerto	rico	1938	1311	1159	10.0349081703
hong	kong	2438	2694	2205	9.72831972408
los	angeles	3501	2808	2791	9.56067615065
carbon	dioxide	4265	1353	1032	9.09852946116
prize	laureate	5131	1676	1210	8.85870710982
san	francisco	5237	2477	1779	8.83305176711
nobel	prize	4098	5131	2498	8.68948811416
ice	hockey	5607	3002	1933	8.6555759741
star	trek	8264	1594	1489	8.63974676575
car	driver	5578	2749	1384	8.41470768304
it	the	283891	3293296	3347	-1.72037278119
are	of	234458	1761436	1019	-2.09254205335

Word2Vec

Мотивация

One-hot encoding представления

- никак не отражают смысловую близость слов
- имеют слишком большую размерность

Например,
скалодром [0, 1, 0, 0, 0, 0]
диван [0, 0, 1, 0, 0, 0]
отдых [0, 0, 0, 1, 0, 0]

Идея — смотреть контекст

«You shall know a word by the company it keeps»
(J.R. Firth, 1957)

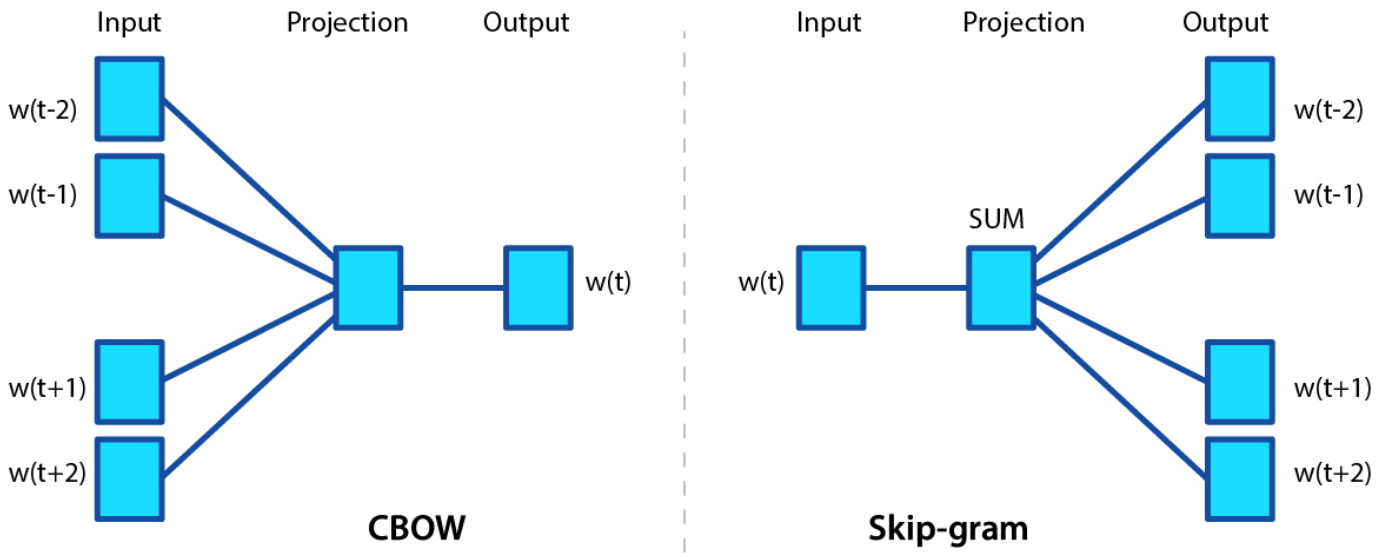
Пример

- из всех домашних животных ??? лучше всего справляются с ловлей мышей

Continuous Bag of Words (CBOW) and Skip-gram

$$\frac{1}{T}\sum_{t=1}^{T}\log(p(w_t|w_{t-m},\ldots,w_{t-1},w_{t+1},\ldots,,w_{t+m}))$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m,j\neq 0}\log(p(w_{t-j}|w_t))$$



Mikolov, Corrado, Chen, Dean «Efficient Estimation of Word Representations

Модель CBOW

Вероятность слова w_t в заданном контексте
$C_t = (w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, , w_{t+m})$

$$p(w_t = w | C_t) = \text{SoftMax}_{w \in W} \left\langle u_w, v^{-t} \right\rangle$$

$v^{-t} = \frac{1}{2m} \sum_{w \in C_t} v_w$ — средний вектор слов из контекста C_t
v_w — векторы предсказывающих слов,
u_w — вектор предсказываемого слова, в общем случае $u_w \neq v_w$

Критерий максимума лог-правдоподобия, $U, V \in \mathbb{R}^{|W| \times d}$:

$$\sum_{t=1}^{n} \log p(w_t | C_t) \max_{U,V}$$

Skip-gram: как считать вероятности

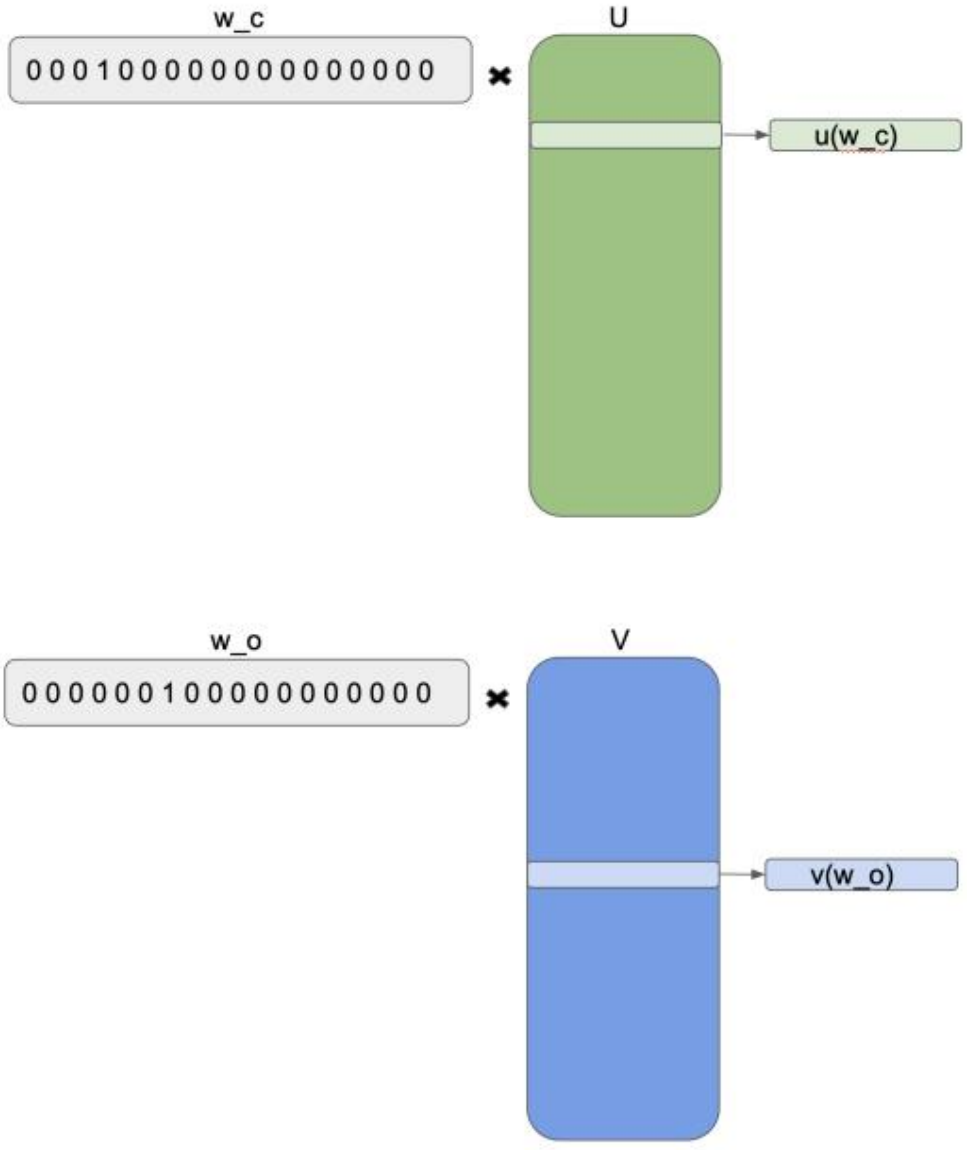
$$p(w_o|w_c) = \frac{\exp[v(w_o)u^T(w_c)]}{\sum_{w=1}^{W} \exp[v(w)u^T(w_c)]}$$

W — множество всех слов словаря

w_c — центральное слово

w_o — слово контекста

$u(\cdot)$ и $v(\cdot)$ — вектора параметров (эмбединги), которые скалярно перемножаются



Вопрос

В чем проблема такого подхода?

Вопрос

В чем проблема такого подхода?

В знаменателе, конечно, где все слова словаря.

Вопрос

В чем проблема такого подхода?

В знаменателе, конечно, где все слова словаря.

Mikolov предложил

Иерархический софтмакс

Моделируем вероятность эффективнее. Строим дерево Хаффмана на словах, после чего:

$1 - \sigma(x) = \sigma(-x)$

$p(w_o|w_c) = \prod_{n \in \text{Path}(w_o)} \sigma(d_{nw_o} v(n) u^T(w_c))$

Здесь уже $v(n)$ — обучаемый вектор в узле дерева

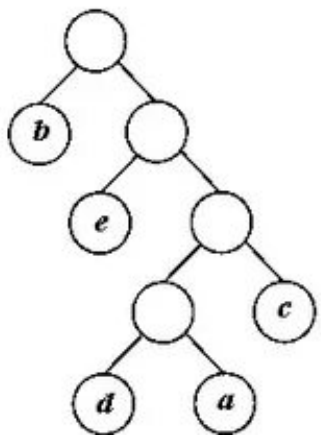
$d_{nw_o} = 1$, если w_o в правом поддереве

$d_{nw_o} = -1$ иначе

Вероятности появления символов

Символ	*Вероятность*
a	0,1
b	0,4
c	0,2
d	0,05
e	0,25

Кодовое дерево



Оптимальные префиксные коды

Символ	*Код*
a	1101
b	0
c	111
d	1100
e	10

doc2vec

на коленке

$$u'(\text{doc}) = \sum_{w \in \text{doc}} \omega_w u(w)$$

В качестве весов слов ω_w логично взять TF-IDF (term frequency / inverse document frequency)

Резюме

- математическая постановка задачи тематического моделирования
- латентное размещение Дирихле (LDA)
- правдоподобие и перплексия языковой модели
- word2vec — модель получения векторных представлений слов

Резюме

- математическая постановка задачи тематического моделирования
- латентное размещение Дирихле (LDA)
- правдоподобие и перплексия языковой модели
- word2vec — модель получения векторных представлений слов

Что ещё можно посмотреть?

- Лекция К.В. Воронцова про тематическое моделирование
- Open source фреймворк для работы с текстами