Вероятностные тематические модели
Лекция 5. Модели сочетаемости слов

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 26 марта 2020

Содержание

Несколько терминов из лингвистики

Единица языка, в зависимости от уровня членения текста — фонема, морфема, слово, словосочетание, фраза, предложение.

Сочетаемость (co-occurrence) — свойство языковых единиц сочетаться в речи, образуя единицы более высокого уровня.

Виды сочетаемости: *контактная* и *дистантная*.

n-грамма — последовательность из n единиц языка (n-грамма — это пример контактной сочетаемости).

Коллокация — *n-грамма* слов, встречающаяся в корпусе гораздо чаще, чем ожидается при их случайном соединении.

Словосочетание — *n-грамма* слов, связанных по смыслу и грамматически, служащая для обозначения единого понятия.

n-граммы радикально улучшают интерпретируемость тем

Коллекция 20Conf заголовков научных статей DBLP, тема «Information Retrieval»

Terms	*Phrases*
search	information retrieval
web	social networks
retrieval	web search
information	search engine
based	support vector machine
model	information extraction
document	web page
query	question answering
text	text classification
social	collaborative filtering
user	topic model

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.

Биграммы радикально улучшают интерпретируемость тем

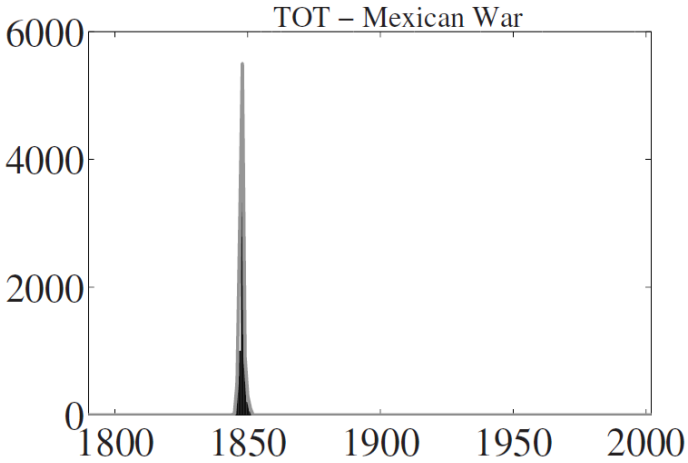
Коллекция 1000 статей конференций ММРО, ИОИ на русском

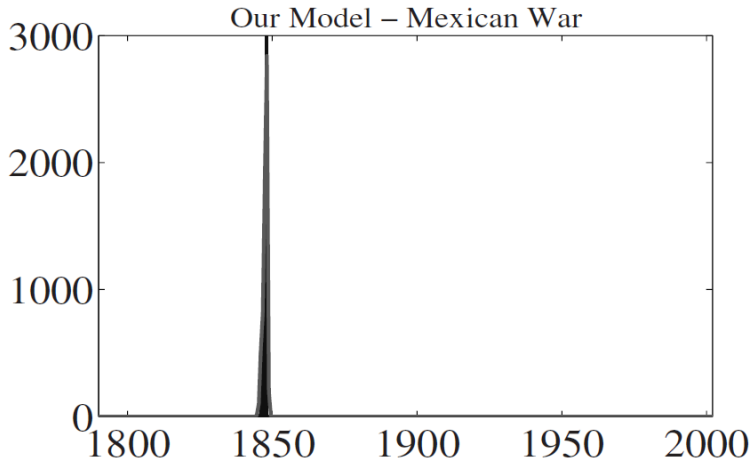
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Совмещение темпоральной и *n*-граммной модели

По коллекции выступлений президентов США





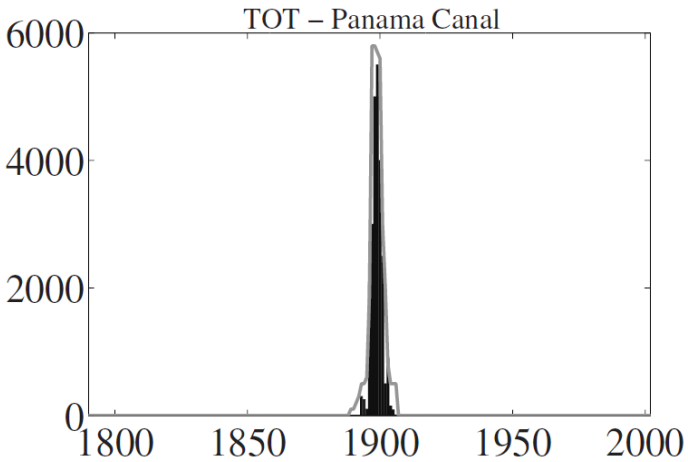
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

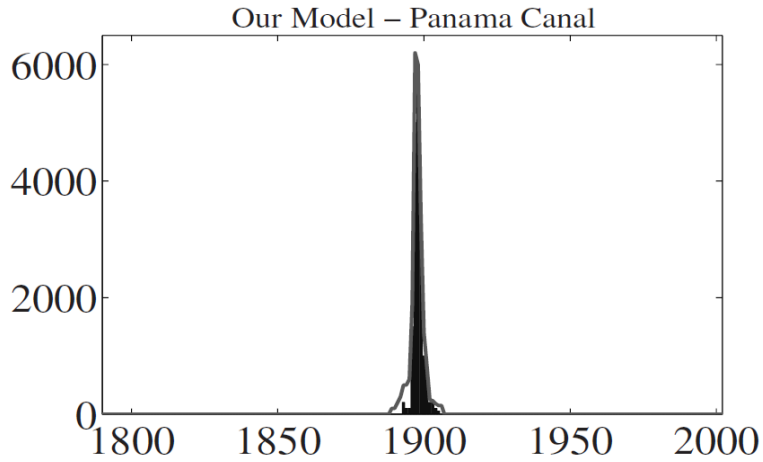
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Совмещение темпоральной и *n*-граммной модели

По коллекции выступлений президентов США





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Биграммная тематическая модель

n_{dvw} — частота пары слов «vw» в документе d
$\phi^v_{wt} = p(w|v,t)$ — распределение слов после слова v в теме t

Модель BTM (Bigram Topic Model):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \phi^v_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Формально, это мультимодальная модель:
$M = W$, каждому слову v соответствует отдельная модальность,
$W^v = W$ — все слова, которые могут следовать за v.

Недостатки биграммной модели BTM:

- все пары соседних слов образуют биграммы;
- модель не описывает отдельные слова (униграммы);
- общее число термов-биграмм $O(|W|^2)$.

Hanna Wallach. Topic modeling: beyond bag-of-words. ICML 2006

Объединение униграмм и биграмм в одной модели

Модель TNG (Topical n-grams):

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T} \underbrace{\big(x_{vwt}\phi^{v}_{wt} + (1-x_{vwt})\phi_{wt}\big)}_{p(w|v,t)} \theta_{td} \to \max_{\Phi,\Theta}$$

$x_{vwt} = \mathrm{P}($пара слов «vw» является биграммой в теме $t)$.

Частные случаи:

- $x_{vwt} = x_{vt}$ — матрица параметров в модели TNG.
- $x_{vwt} \equiv 1$ — модель BTM;
- $x_{vwt} =$ [пара слов «vw» из словаря биграмм];

Xuerui Wang, Andrew McCallum, Xing Wei. Topical *N*-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. 2007.

Мультимодальная мультиграммная ARTM

W^k — словари k-грамм, либо составленные экспертами, либо отобранные по совокупности синтаксических, статистических и тематических критериев.

Связь с моделью TNG: при $x_{vwt} = \lambda[vw \in W^2]$ максимизируем нижнюю оценку log-правдоподобия TNG:

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln \sum_{t\in T}\big(x_{vwt}\phi^v_{wt} + (1 - x_{vwt})\phi_{wt}\big)\theta_{td} =$$

$$\sum_{d\in D}\sum_{vw\in d} n_{dvw} \ln\Big(\lambda \sum_{t\in T} \phi^v_{wt}\theta_{td} + (1-\lambda)\sum_{t\in T}\phi_{wt}\theta_{td}\Big) \geqslant$$

$$\lambda \sum_{d,vw} n_{dvw} \ln \sum_{t\in T} \phi^v_{wt}\theta_{td} + (1-\lambda)\sum_{d,w} n_{dw} \ln \sum_{t\in T}\phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Задача автоматического выделения терминов

Термин — фраза (*n*-грамма) со следующим набором свойств:

1. *высокая частотность* (frequency):
   много раз встречается в коллекции;
2. *контактная сочетаемость слов* (collocation):
   состоит из слов, неслучайно часто встречающихся вместе;
3. *полнота* (completeness):
   является максимальной по включению цепочкой слов;
4. *синтаксическая связность* (syntactic connectedness):
   является грамматически корректным словосочетанием;
5. *тематичность* (topicality):
   часто встречается в небольшом числе тем.

Сумма технологий для ATE (Authomatic Term Extraction):
TopMine 1 2 3 + UDPipe 4 + BigARTM 5

Алгоритм TopMine: определения и основные идеи

- $C(a_1, \ldots, a_k)$ — хэш-таблица частот k-грамм.
- $A_{d,k}$ — множество позиций i в документе d, с которых начинаются все частые k-граммы:

$$C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k.$$

- Свойство антимонотонности:

$$C(a_1, \ldots, a_k) \geqslant C(a_1, \ldots, a_k, a_{k+1}).$$

- Основной шаг алгоритма: для всех $i = 1, \ldots, n_d$
  если $(i \in A_{d,k})$ **и** $(i+1 \in A_{d,k})$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015.

Алгоритм TopMine: быстрый поиск всех частых k-грамм

Вход: коллекция D, пороги ε_k;
Выход: хэш-таблица частот $C(a_1, \ldots, a_k)$, $k = 1, \ldots, k_{\max}$;

$C(w) := n_w$ для всех $w \in W$;
$A_{d,0} := \{1, \ldots, n_d\}$;
для $k := 1, \ldots, k_{\max}$
  для всех $d \in D$
    $A_{d,k} := \left\{ i \in A_{d,k-1} \mid C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k \right\}$;
    для всех $i \in A_{d,k}$
      если $i+1 \in A_{d,k}$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$;
  оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$;

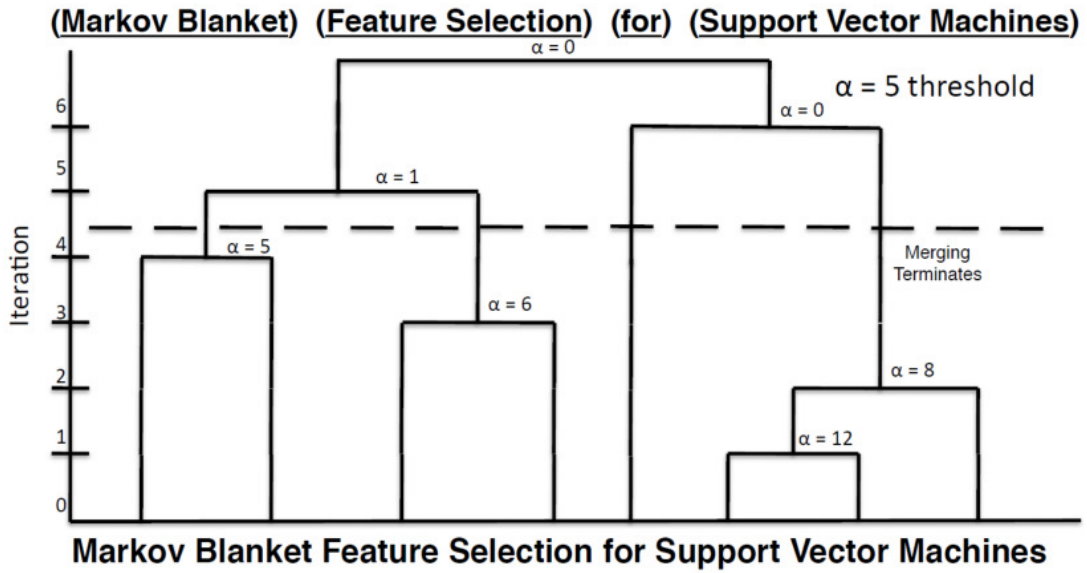
Преимущество алгоритма: линейная память и скорость.

Алгоритм TopMine: отбор фраз по частоте и полноте

Итеративное слияние фраз с понижением значимости α.

p_u — оценка вероятности встретить фразу u
p_{uv} — оценка вероятности встретить фразу uv

Критерии: $\text{SignificanceScore} = \dfrac{p_{uv} - p_u p_v}{\sqrt{p_{uv}}}$ или $\text{PMI} = \log \dfrac{p_{uv}}{p_u p_v}$



Синтаксические анализаторы (UDPipe, SyntaxNet)

Вход: список предложений

Выход, для каждого слова в каждом предложении:

- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- нормальная форма
- часть речи: `NOUN`, `VERB`, `ADJ`, `ADV`, ...
- член предложения: `nsubj`, `dobj`, `conj`, `cc`, `nmod`, ...

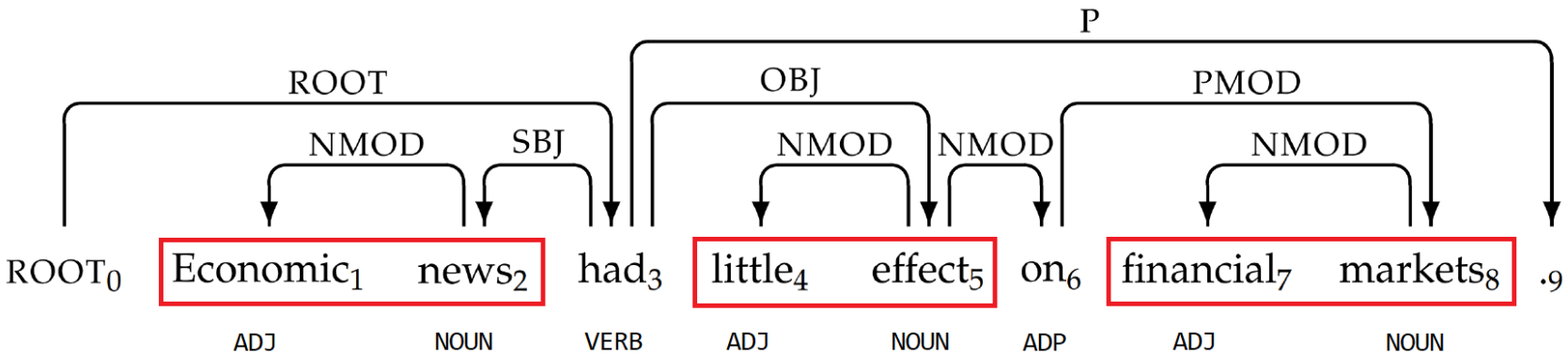
UDPipe (Universal Dependencies), 60 языков, включая русский

Google SyntaxNet — предобученная нейросеть поверх TensorFlow, поддерживает 40 языков, включая русский.

D.Andor, C.Alberti, D.dWeiss, A.Severyn, A.Presta, K.Ganchev, S.Petrov, M.Collins. Globally Normalized Transition-Based Neural Networks. 2016.

Использование дерева зависимостей для отбора терминов

Пример дерева зависимостей:



Варианты стратегий отбора терминов-кандидатов:

- брать все поддеревья
- брать все именные группы (корень — NOUN)
- не брать CONJ, SCONJ, DET, AUX, INTJ, PART, PUNCT, SYM

Announcing SyntaxNet: the world's most accurate parser goes open source.
https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

Денис Кирьянов. Изучаем синтаксические парсеры для русского языка. 2018.
https://habr.com/ru/company/sberbank/blog/418701

Критерии тематичности фраз

Насколько далеко $p(t|w) = \phi_{wt}\frac{n_t}{n_w}$ от равномерного $p_0(t) = \frac{1}{|T|}$.

Дивергенция Кульбака-Лейблера:

$$\mathrm{KL}(w) = \mathrm{KL}(p_0\|p) = \sum_{t\in T}\frac{1}{|T|}\ln\frac{\frac{1}{|T|}}{p(t|w)} \to \max$$

Дивергенция Йенсена-Шеннона (метрика, не имеет проблем с нулевыми вероятностями), где $\bar{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|})$:

$$\mathrm{JS}(w) = \frac{1}{2}\mathrm{KL}(p_0\|\bar{p}) + \frac{1}{2}\mathrm{KL}(p\|\bar{p}) \to \max$$

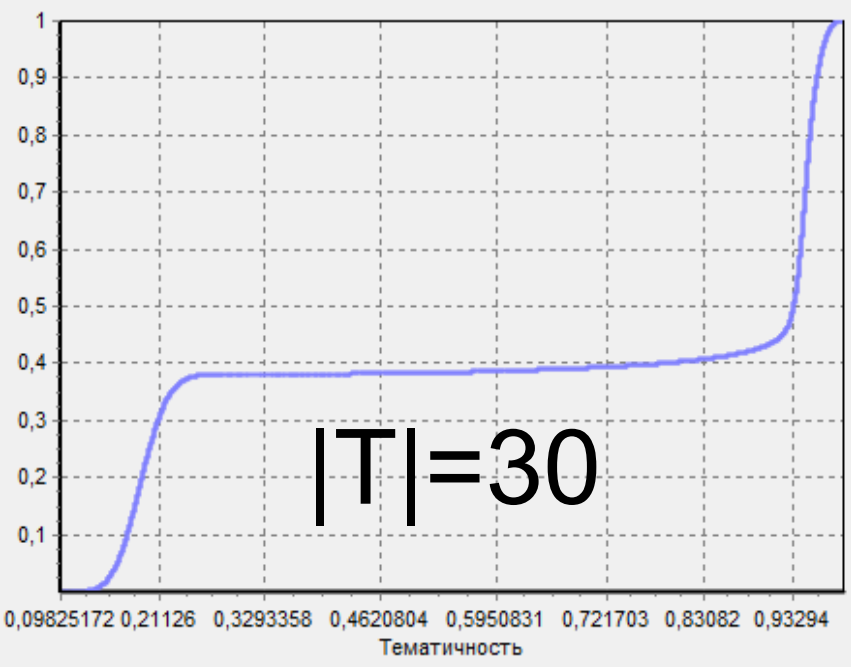
Нормированная сумма степенных функций, $\gamma > 1$:

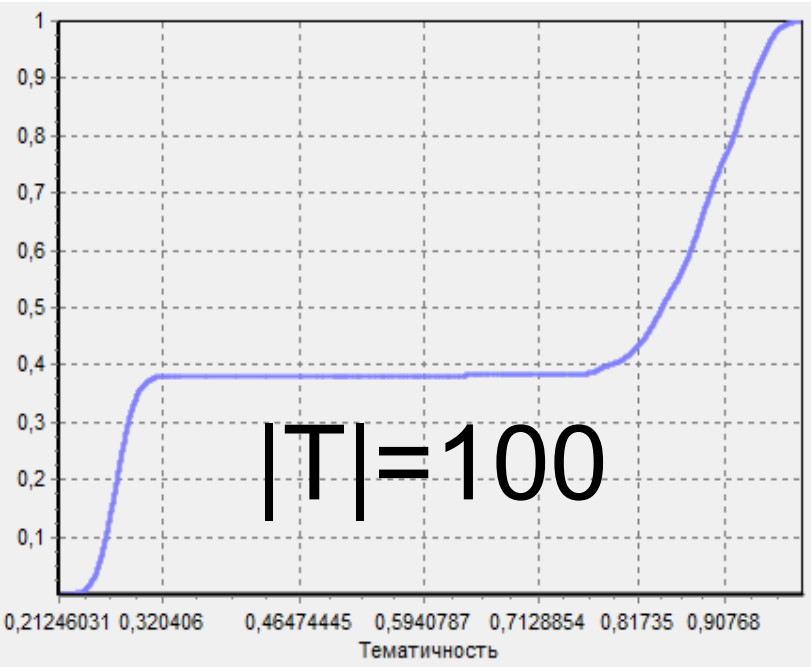
$$\text{Тематичность}(w) = |T|^{\gamma-1}\sum_{t\in T}p(t|w)^{\gamma} \to \max$$

Фразы чётко разделяются на тематичные и нетематичные

$|W| = 46\,000$ фраз из $|D| = 600$ документов коллекции SyntagRus, тематические модели LDA на 30 и 100 тем.

Распределение фраз по нормированной тематичности:





Пограничный слой между тематичными и нетематичными фразами очень узкий $\approx \frac{200}{46\,000}$ и слабо зависит от числа тем.

Основной эксперимент ATE: SyntaxNet + TopMine + BigARTM

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого *случайного* подмножества (2000 n-грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
  логистическая регрессия, градиентный бустинг

Владимир Полушин. Тематические модели для ранжирования рекомендаций текстового контента. Бакалаврская диссертация, ВМК МГУ, 2017.

Сравнение методов автоматического отбора терминов

Найти *как можно больше терминов* — полнота важнее точности

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	**0.95**	**0.38**	**0.91**	**0.97**	**0.41**	**0.99**

Стат < Син < Син+Стат < Тем < Стат+Тем, Син+Тем < Стат+Син+Тем

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

Проблема коротких текстов

Короткие тексты (short text):

- Twitter и другие микроблоги
- социальные медиа
- заголовки статей и новостных сообщений

Тривиальные подходы:

- считать каждое сообщение отдельным документом
- разреживать $p(t|d)$ вплоть до единственной темы
- объединить сообщения по автору/времени/региону/и т. п.
- объединить посты с комментариями
- дополнить коллекцию длинными текстами (Википедия и др.)

Более интересная идея:

- использовать сочетаемость пар слов в сообщениях

Дистрибутивная гипотеза и виды семантической близости слов

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
со-встречаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

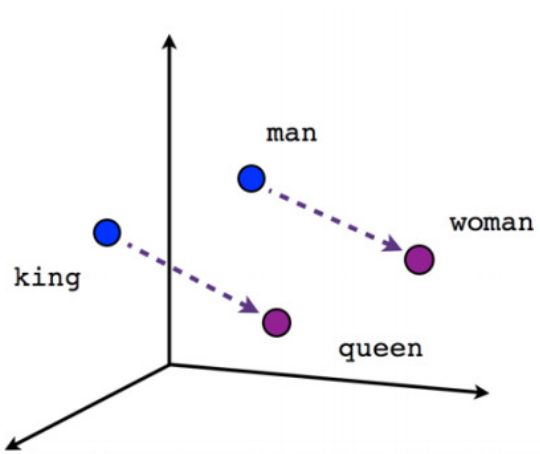
здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.D.Turney, P.Pantel. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR). 2010.

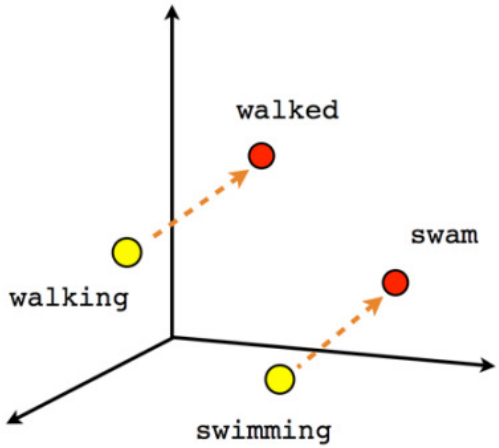
Задача семантического векторного представления слов

Задача: по наблюдаемой синтагматической близости слов построить *векторные представления слов* (word embedding) $x_w \in \mathbb{R}^T$, $w \in W$, так, чтобы парадигматически близкие слова имели близкие векторы.

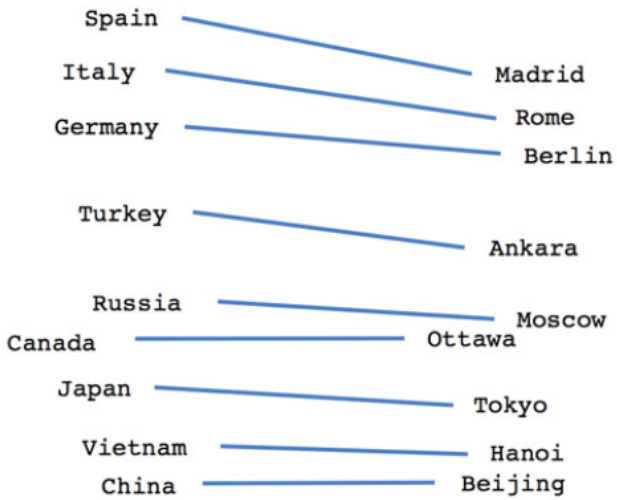
Способ проверки — задача семантической аналогии слов: по трём словам угадать четвёртое.



Male-Female



Verb tense



Country-Capital

Формализация дистрибутивной гипотезы в программе word2vec

Дано: n_{uw} — частота пары слов u, w в окне $\pm h$ слов

Найти: векторные представления слов x_w и контекстов y_u

Модель: вероятность слова w в контексте слова u:

$$p(w|u) = \operatorname*{SoftMax}_{w \in W} \langle x_w, y_u \rangle = \operatorname*{norm}_{w \in W} \big(\exp \langle x_w, y_u \rangle \big)$$

Критерий максимума log-правдоподобия:

$$\sum_{w,u \in W} n_{wu} \ln p(w|u) \;\to\; \max_{\{x_w, y_u\}}$$

и его аппроксимация SGNS (Skip-Gram Negative Sampling):

$$\sum_{w,u \in W} n_{wu} \Big(\ln \sigma \langle x_w, y_u \rangle + \sum_{v \in V_k(u)} \ln \sigma \big(-\langle x_v, y_u \rangle \big) \Big) \;\to\; \max_{\{x_w, y_u\}}$$

где $V_k(u) \subset W$ — случайные k слов не из контекста u.

T.Mikolov, K.Chen, G.Corrado, J.Dean. Efficient estimation of word representations in vector space, 2013.

Связь word2vec с матричными разложениями

T — размерность векторов слов x_w и контекстов y_u
$X = (x_w)_{W \times T}$ — матрица векторов слов
$Y = (y_u)_{W \times T}$ — матрица векторов контекстов

SGNS строит матричное разложение $P \approx XY^{\mathrm{T}}$ матрицы Shifted PMI (Point-wise Mutual Information):

$$P_{wu} = \ln \frac{n_{wu} n}{n_w n_u} - \ln k,$$

n_{wu} — частота пары слов w, u в окне $\pm h$ слов,
n_w, n_u — число пар с участием слова w и u соответственно,
n — число всех пар слов в коллекции.

В качестве эвристики используют также Shifted Positive PMI:

$$P^+_{wu} = \left(\ln \frac{n_{wu} n}{n_w n_u} - \ln k \right)_+ .$$

O.Levy, Y.Goldberg. Neural word embedding as implicit matrix factorization, 2014.

Модели векторных представлений для текстов и графов

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных *n*-грамм
`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: эмбединги фраз и предложений от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

Недостаток: координаты векторов не интерпретируемы

Преимущества и недостатки нетематических эмбедингов

- ⊕ Удивительно высокое качество на задачах семантической аналогии и близости слов.
- ⊕ Возможность нейросетевой реализации методом SG.
- ⊕ Имеются готовые реализации от Google, Facebook и др.
- ⊕ Имеются готовые векторы слов, предобученные по большим текстовым коллекциям на разных языках
- ⊖ Неинтерпретируемые компоненты векторов
- ⊖ Не ясно, почему XY^{T}, а не XX^{T} (обычно Y игнорируют)

Тематические модели BitermTM, WordNetworkTM, WordTM обучаются по частотам сочетания слов, аналогично word2vec.

Битермы: модель сочетаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm Topic Model):

$$p(u,v) = \sum_{t\in T} p(u|t)p(v|t)p(t) = \sum_{t\in T} \phi_{ut}\phi_{vt}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{vt} \geqslant 0; \quad \textstyle\sum_v \phi_{vt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts. WWW 2013.

Необходимые условия точки максимума правдоподобия

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{u,v} n_{uv} \ln \sum_{t} \phi_{ut}\phi_{vt}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

n_{uv} — частота битерма (u,v) в документах коллекции.

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tuv} \equiv p(t|u,v) = \underset{t\in T}{\text{norm}}\left(\phi_{ut}\phi_{vt}\pi_t\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{vt} = \underset{v\in W}{\text{norm}}\left(n_{vt} + \phi_{vt}\dfrac{\partial R}{\partial \phi_{vt}}\right), \quad n_{vt} = \sum\limits_{u\in W} n_{uv}p_{tuv} \\ \pi_t = \underset{t\in T}{\text{norm}}\left(n_t + \pi_t\dfrac{\partial R}{\partial \pi_t}\right), \quad n_t = \sum\limits_{u,v\in W} n_{uv}p_{tuv} \end{cases}$$

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt} \to \max.$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big); \quad p_{tuw} = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

Это эквивалентно обработке *псевдо-документов* d_u, где каждый d_u объединяет все контексты слова u, причём $\theta_{tu} \propto n_t \phi_{ut}$;
n_{uw} — число вхождений слова w в псевдо-документ d_u.

2. Регуляризатор разреживания для матрицы Θ:

$$R(\Theta) = -\tau' \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Модель сети слов WNTM для коротких текстов

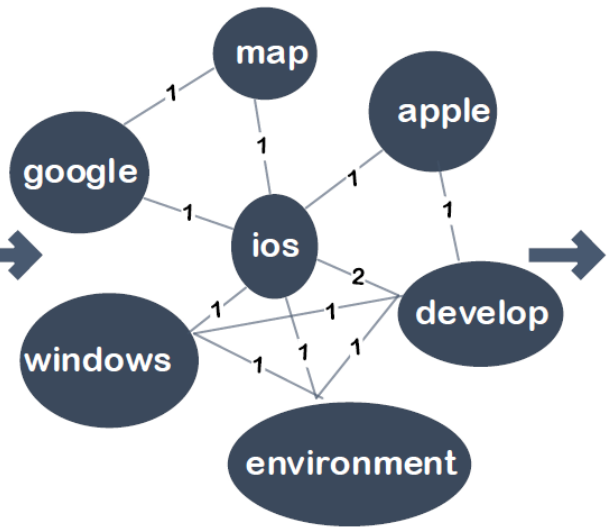
Идея: моделировать не документы, а связи между словами.

d_u — псевдо-документ, объединение всех контекстов слова u.

n_{uw} — число вхождений слова w в псевдо-документ d_u.

Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM (Word Network) и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

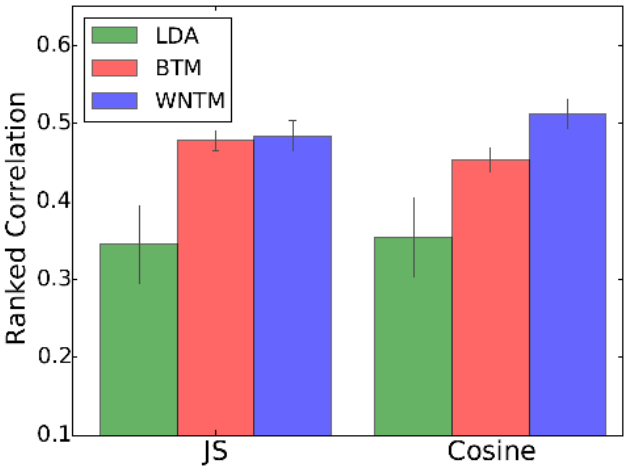
где n_{uw} — частота сочетания пары слов (w, u).

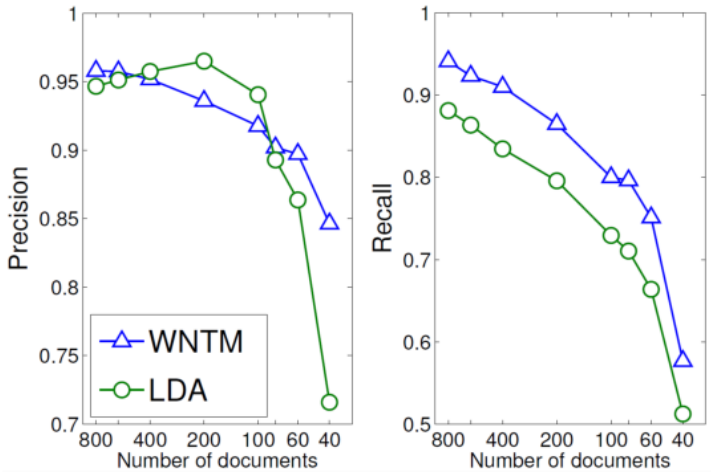
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. ACM Trans., 2009.

Результаты оценивания модели WNTM

- Когерентность на коротких текстах лучше, чем у LDA и BitermTM; на длинных текстах преимуществ нет.
- *Слева:* оценивание семантической близости слов по $p(t|w)$, корреляция с 10-балльными экспертными оценками.
- *Справа:* полнота и точность распознавания новой темы в зависимости от числа документов.





Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

WN-ARTM на задачах семантической аналогии слов

Два подхода к синтезу векторных представлений слов:

- **WN-ARTM**: интерпретируемые разреженные компоненты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат WN-ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

word2vec и WN-ARTM на задачах семантической близости слов

Дамп Википедии 2016-01-13, $|W| = 100$K, разреженность 93%.

Конкуренты: LDA, SVD-PPMI, SGNS (word2vec).
Варианты WN-ARTM: offline, online, online-with-sparsing.

	WordSim similarity	WordSim relatedness	WordSim joint	Bruni et al. MEN	Radinsky m.turk
LDA	0.530	0.455	0.474	0.583	0.483
SVD-PPMI	0.711	0.648	0.672	0.236	0.616
SGNS	**0.752**	0.632	0.666	**0.745**	**0.661**
ARTM off	0.701	0.615	0.647	0.707	0.613
ARTM on	0.718	**0.673**	**0.685**	0.669	0.639
ARTM on-sp	0.728	0.672	0.680	0.675	0.635

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Сравнение word2vec и WN-ARTM по интерпретируемости тем

SGNS (word2vec) — нет интерпретируемости:

- avg hearth soc protector decomposition whip stochastic sewer splinter accessory howie thief thermodynamic boltzmann equilibrium kingship unconscious
- rainy miocene snowy horner cfb triassic eleventh amadeus dams tenth mesozoic fourteenth thirteenth ninth diaries bight demographics seventh almanac eocene
- gnis usda bloomberg usgs regulator nhk gerd magnetism capacitor fed classifies capacitance stadt bipolar multilateral trpod kunst reciprocal smiths potassium

WN-ARTM — есть интерпретируемость:

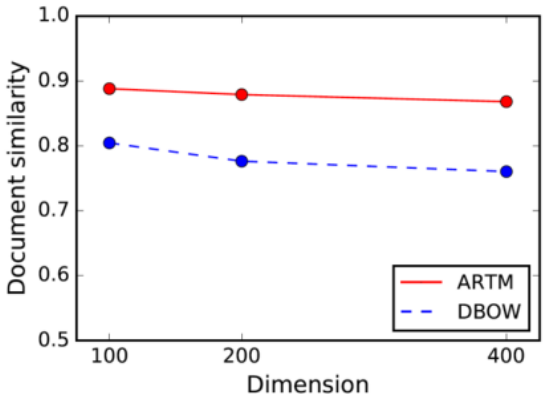
- scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus
- game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash zelda
- election party elected elections parliament assembly seats members minister legislative electoral liberal council representatives parliamentary democratic

---

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

WN-ARTM на задачах семантической близости документов

ArXiv triplets dataset [Dai et. al, 2015]: 20K троек статей:
⟨ статья A, схожая статья B, непохожая статья C ⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент: DBOW paragraph2vec [Dai et. al, 2015]

WN-ARTM превосходит модель DBOW (distributed bag-of-words)

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Интерпретируемости и когерентность

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- 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 его при определении

Нужна автоматически вычисляемая мера интерпретируемости, коррелирующая с экспертными оценками.

Ею оказалась *когерентность* (согласованность, coherence).

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

Эксперимент. Связь когерентности и интерпретируемости

Измерялась ранговая корреляция Спирмена между 15 метрикам и экспертными оценками интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

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

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

где w_i — i-е слово темы, в порядке убывания ϕ_{wt}.

$\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ — поточечная взаимная информация (pointwise mutual information),

N_{uv} — число документов, в которых слова u, v хотя бы один раз встречаются рядом (в окне 10 слов),

N_u — число документов, в которых u встретился хотя бы 1 раз.

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

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\text{norm}}\left(n_{wt} + \tau \sum_{u\in W} C_{uw} n_{ut}\right).$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models. EMNLP-2011.

Альтернативный регуляризатор когерентности

Квадратичный регуляризатор Quad-Reg:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max,$$

где $C_{uv} = N_{uv} \big[\mathrm{PMI}(u,v) > 0 \big]$ — оценка сочетаемости пары слов.

Подставляем в формулу M-шага, снова получаем сглаживание:

$$\phi_{wt} = \underset{w}{\mathrm{norm}} \left(n_{wt} + \tau \phi_{wt} \frac{\sum\limits_{(u,w) \in Q} C_{uw} \phi_{ut} + \sum\limits_{(w,v) \in Q} C_{wv} \phi_{vt}}{\sum\limits_{(u,v) \in Q} C_{uv} \phi_{ut} \phi_{vt}} \right).$$

В литературе пока не выработан окончательный вариант регуляризатора когерентности.

Newman D., Bonilla E. V., Buntine W. L. Improving topic coherence with regularized topic models. 2011.

Резюме

Различные способы учёта совместной встречаемости:

- выделение фраз на этапе предобработки
- выделение фраз вместе с тематическим моделированием
- тематические модели дистрибутивной семантики (BitermTM, WTM, WNTM, когерентность)

Модели сочетаемости преобразуют исходные данные о синтагматической близости слов n_{uv} в парадигматические близости слов (u, v) как распределений $p(t|u)$ и $p(t|v)$.

Сети слов (WNTM)

- лучший способ тематизации коротких текстов
- легко реализовать в BigARTM, переразбив коллекцию на псевдо-документы — локальные контексты слов
- даёт векторные представления слов, аналогично word2vect, но разреженные и интерпретируемые