Вероятностные тематические модели
Лекция 9.
Модели совстречаемости 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 • весна 2017

Содержание

Биграммы 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

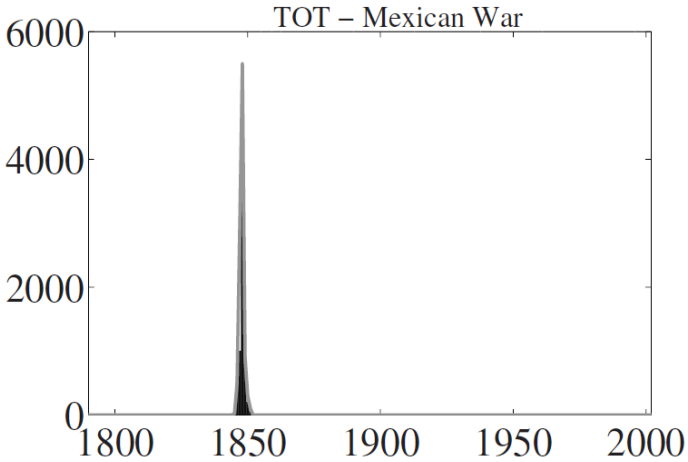
Коллекция 1000 статей конференций ММРО, ИОИ на русском

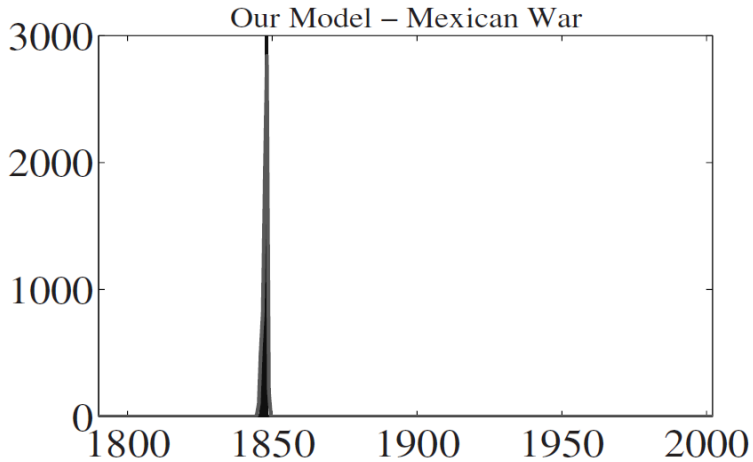
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	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 динамической и n-граммной модели

По коллекции выступлений президентов США





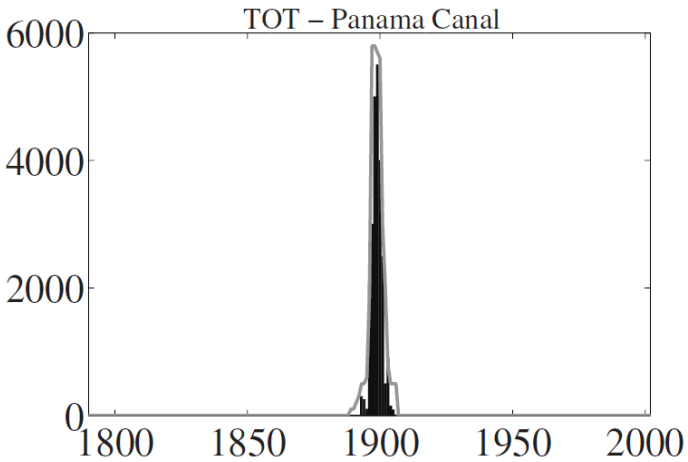
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

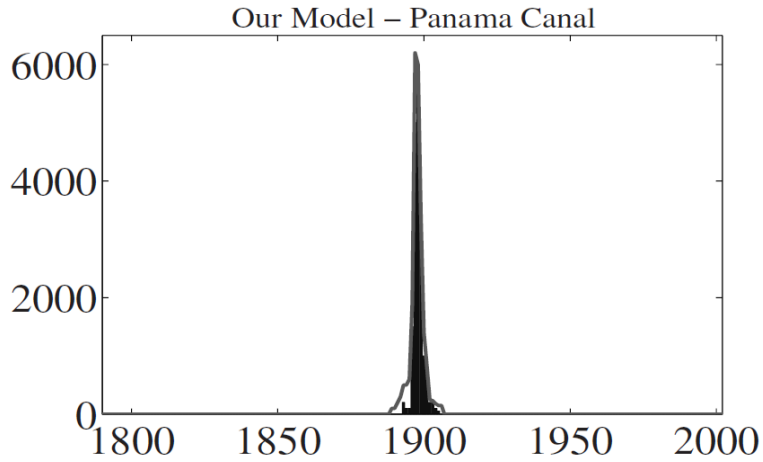
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

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

Совмещение динамической 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):

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

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 токенов $O(|W|^2)$.

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 отобранные по совокупности синтаксических, статистических и тематических критериев, либо составленные экспертами.

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. *совстречаемость 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):
   имеет пиковую тему в тематической модели.

Сумма технологий для ATE (Authomatic Term Extraction):
TopMine + SyntaxNet + BigARTM

Алгоритм TopMine: определения и основные идеи

1. Хэш-таблица $C(a_1, \ldots, a_k)$ счётчиков частых k-грамм, инициализируется для всех униграмм a с частотой $n_a \geqslant \varepsilon_1$
2. Свойство антимонотонности:
$$C(a_1, \ldots, a_k) \geqslant C(a_1, \ldots, a_k, a_{k+1}).$$
3. $A_{d,k}$ — множество позиций i в документе d таких, что
$$C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k,$$
инициализируется для всех частых униграмм.
4. Основной шаг алгоритма:
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 высокочастотных k-грамм

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

$A_{d,1} := \{1, \ldots, n_d\}$;
$C(w) := n_w$ для всех $w \in W$ таких, что $n_w \geqslant \varepsilon_1$;
для $k := 2, \ldots, k_{\max}$ **пока** $D \neq \varnothing$
  для всех $d \in D$
    $A_{d,k} := \left\{ i \in A_{d,k-1} \;\middle|\; C(w_{d,i}, \ldots, w_{d,i+k-2}) \geqslant \varepsilon_k \right\}$;
    если $A_{d,k} = \varnothing$ **то** $D := D \backslash \{d\}$;
    для всех $i \in A_{d,k}$
      если $i + 1 \in A_{d,k}$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k-1})$;
  оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$;

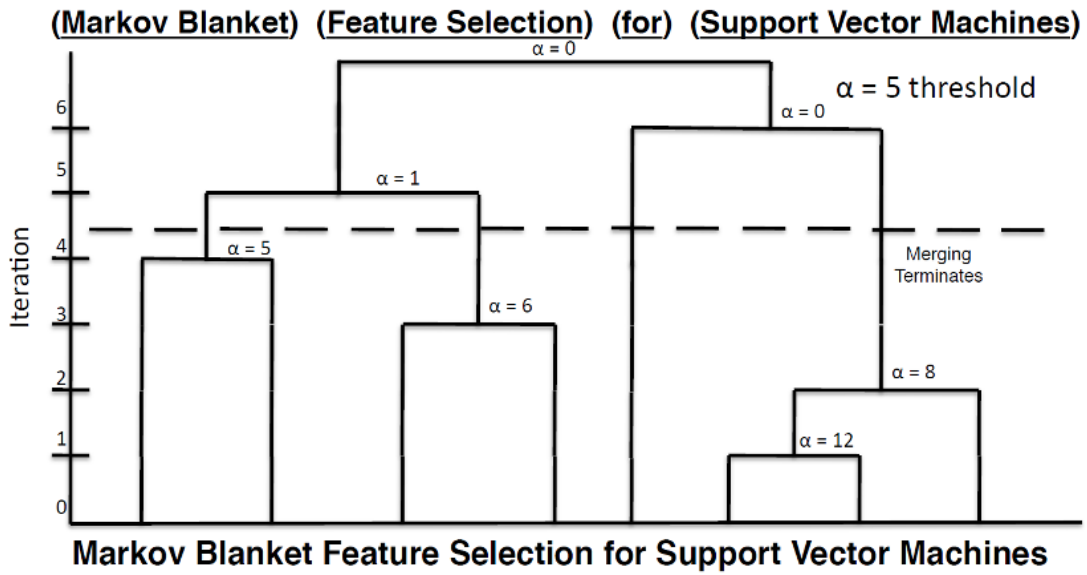
Преимущество алгоритма: линейная память и скорость.

Алгоритм TopMine: отбор фраз по совстречаемости 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 $\text{PMI} = \log \dfrac{p_{uv}}{p_u p_v}$



Синтаксический анализатор Google SyntaxNet

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

Вход:

- список предложений

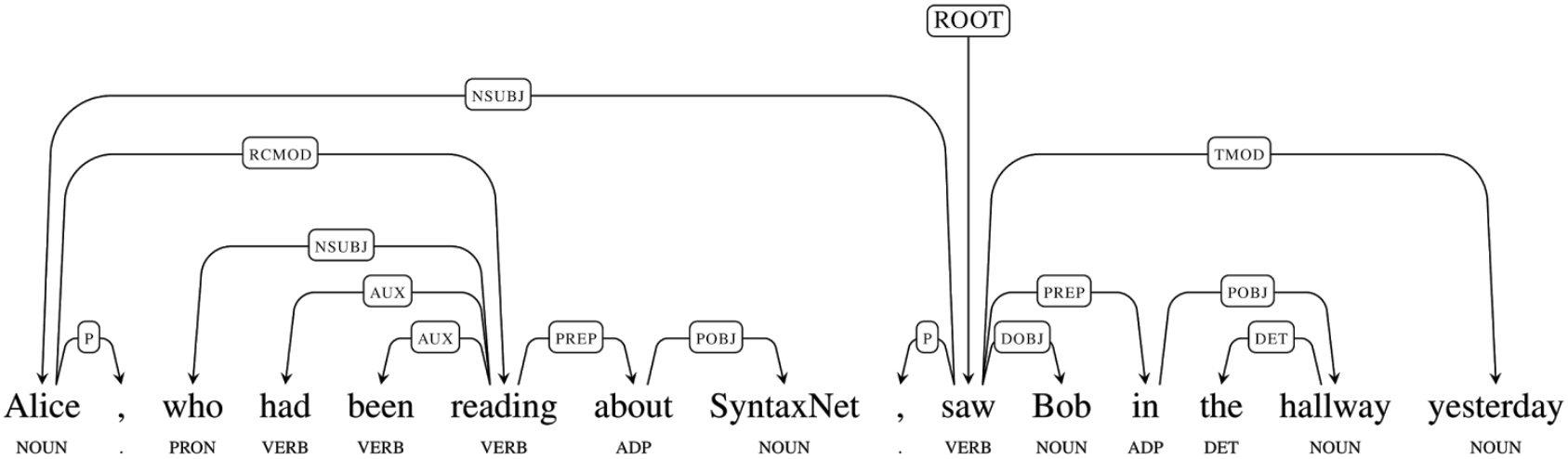
Выход, для каждого предложения:

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

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

Синтаксический анализатор Google SyntaxNet

Пример дерева зависимостей:



Варианты стратегий отбора терминов-кандидатов:

- брать все поддеревья
- брать все именные группы (корень — существительное)

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

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 $\bar{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|})$:

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

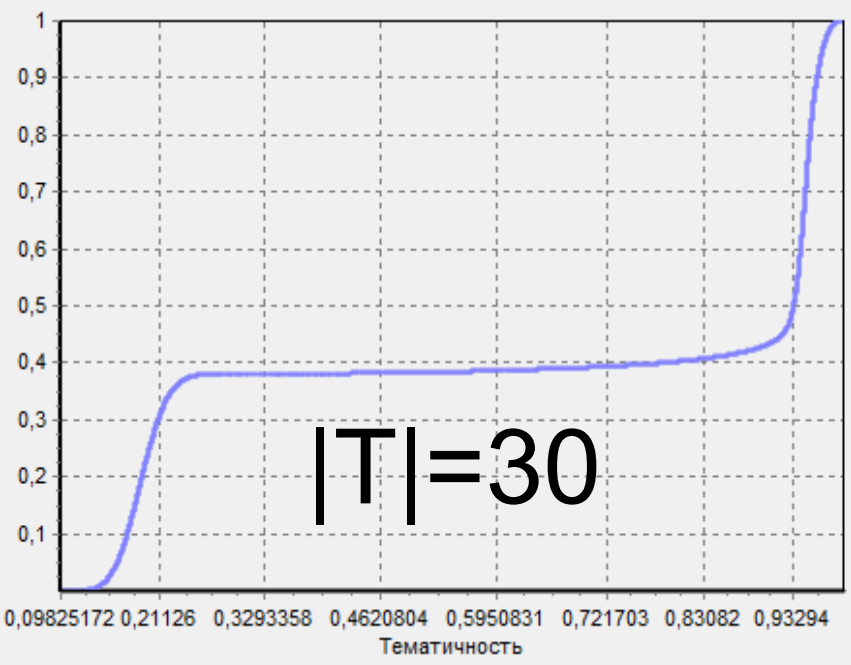
Нормированная сумма степенных функций, $\gamma > 1$:

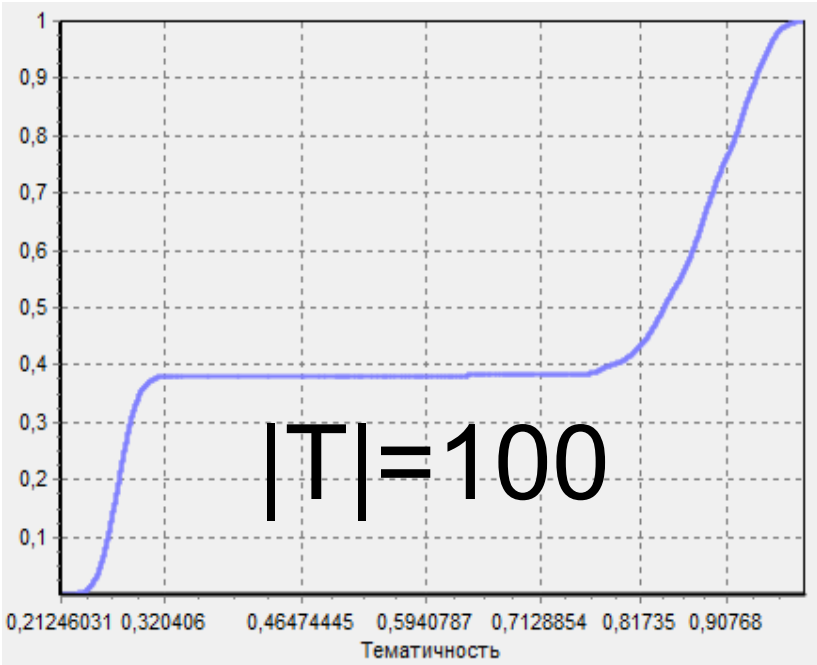
$$\text{Тематичность}(w) = |T|^{\gamma-1}\sum_{t\in T}p(t|w)^{\gamma} \to \max$$

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 — около 200 слов из 46 000.

Число тем почти не влияет на тематичность

$|W| = 46\,000$ фраз из $|D| = 600$ документов коллекции SyntagRus,

Число фраз, которые переходят из тематичной в нетематичную при изменении числа тем $T_{\text{в строке}} \to T_{\text{в столбце}}$

|T|	5	10	20	30	50
5	0	96	7	1	0
10	831	0	13	1	0
20	1119	390	0	10	0
30	1250	515	147	0	0
50	1320	585	208	71	0
100	1365	630	253	116	45

30 тем вполне достаточно для определения тематичности.

Открытая задача: оценить долю терминов среди тематичных и нетематичных фраз.

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.)

Дистрибутивная гипотеза

- 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.

Битермы: модель совстречаемости слов в коротких текстах

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

Тематическая модель битермов (Biterm topic model):

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

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

Критерий максимума логарифма правдоподобия:

$$\sum_{u,w} n_{uw} \ln \sum_t \phi_{wt}\phi_{ut}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

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:

$$\sum_{u,w} n_{uw} \ln \sum_{t} \phi_{wt}\phi_{ut}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

n_{uw} — частота битерма (u,w) в документах коллекции.

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \to \max$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big);$$

$$p_{tuw} \equiv p(t|u,w) = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

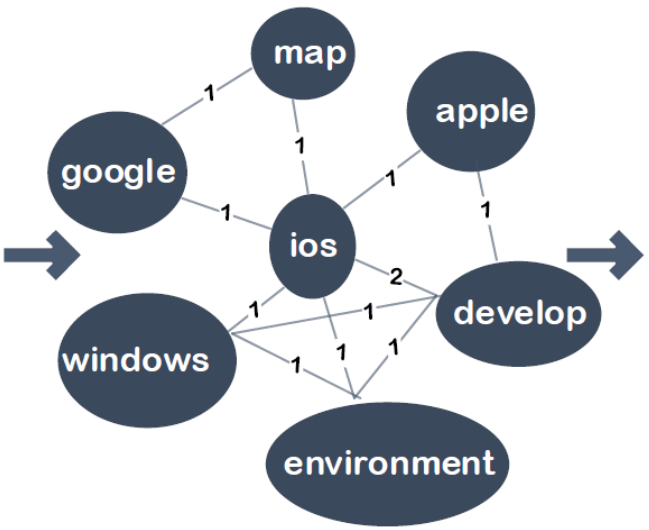
2. Регуляризатор разреживания для матрицы Θ:

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

Модель сети слов WNTM для коротких текстов

Идея: моделировать не документы, а связи между словами.
d_w — псевдо-документ, объединение всех контекстов слова w.
n_{wu} — число вхождений слова u в псевдо-документ d_w.
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 и WTM (Word Topic Model)

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

$$p(u|d_w) = \sum_{t \in T} p(u|t)p(t|d_w) = \sum_{t \in T} \phi_{ut}\theta_{tw},$$

где d_w — псевдо-документ слова w.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{wu} \log \sum_{t \in T} \phi_{ut}\theta_{tw} \to \max_{\Phi,\Theta},$$

где n_{wu} — совстречаемость слов w, u.

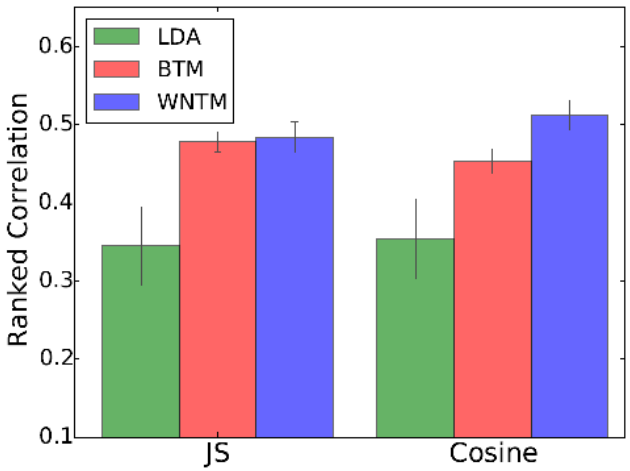
Отличие от модели битермов: там $\Theta = \mathrm{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathrm{T}}$.

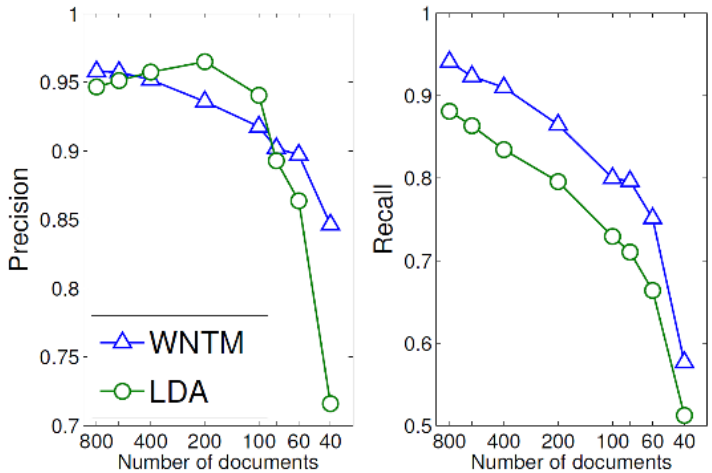
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

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 Biterm TM; 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.

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, Pp. 100–108.

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, Pp. 100–108.

Когерентность как внутренняя мера интерпретируемости

Когерентность (согласованность) темы t по k топовым словам:

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

где w_i — i-й термин 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 термины 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, Pp. 100–108.

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. — Pp. 262–272.

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 совстречаемости.

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 совстречаемости:

- выделение фраз на этапе предобработки
- выделение фраз во время тематического моделирования
- выделение фраз на этапе постобработки
- тематические модели дистрибутивной семантики (BitermTM, WTM, WNTM, регуляризаторы когерентности)

Сети слов (WNTM)

- лучший способ тематизации коротких текстов
- легко реализовать в BigARTM, переразбив коллекцию на псевдо-документы — локальные контексты слов