Вероятностные тематические модели
Лекция 10.
Сегментация, суммаризация, автоматическое именование тем

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

Проблема коротких текстов

Короткие тексты (short text): Twitter и другие микроблоги, социальные медиа, заголовки статей и новостных сообщений.

Типичная задача — «классифицировать иголки в стоге сена»
Поиск заданной тематики: персоны, организации, технологии, товары, этничности, болезни, лекарства, аварии, терроризм, ...

Основные проблемы коротких сообщений:

- огромный объём ($\sim 10^9$ твитов в день)
- опечатки и намеренное искажение слов языка
- концентрация распределения $p(t|d)$ в одной теме
- слишком много сена (life style, личное, репосты)
- появление новых тем, необходимо их раннее обнаружение

Тривиальные подходы и их недостатки

- Считать каждое сообщение отдельным документом
  - для коротких сообщений $p(t|d)$ оценивается не надёжно
- Разреживать $p(t|d)$ вплоть до единственной темы
  - определение темы статистически не надёжно
  - тема может охватывать цепочку сообщений
- Объединить сообщения по автору (времени, региону и т.п.)
  - появится дисбаланс документов по длине
  - появятся тематически неоднородные документы
- Объединить посты с комментариями
  - комментарии могут отсутствовать у большинства постов
- Дополнить коллекцию длинными текстами (Википедия и др.)
  - часть тем может не покрываться внешней коллекций
  - лексикон социальной сети может существенно отличаться

Модель Twitter-LDA

Предположения:

1. Каждый автор $a \in A$ написал множество сообщений $d \in D_a$.
2. Каждое сообщение d относится к одной теме $p(t|d) \in \{0, 1\}$.
3. Есть фоновая тема $b \in T$ с распределением $p(w|b)$.
4. Вероятность фона одинакова для документов, $p(b|d) = \pi$.

Порождающий процесс:

Вход: распределения $p(w|t)$, $p(t|a)$
для всех авторов $a \in A$
  для всех сообщений $d \in D_a$ автора a
    выбрать тему t из $p(t|a)$, кроме фоновой, $t \neq b$;
    для всех позиций слов $i = 1, \ldots, n_d$ в сообщении d
      выбрать слово w_i из $(1 - \pi)p(w|t) + \pi p(w|b)$;

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models // ECIR 2011.

Модель смеси униграмм для сегментированного текста

S_d — множество сегментов, на которые разбит документ d;

n_s — длина сегмента s;

n_{sw} — число вхождений термина w в сегмент s.

Тематическая модель монотематичного сегмента:

$$p(s|d) = \sum_{t \in T} p(t) \prod_{w \in s} p(w|t)^{n_{sw}} = \sum_{t \in T} \pi_t \prod_{w \in s} \phi_{wt}^{n_{sw}}$$

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d \in D} \sum_{s \in S_d} \ln \sum_{t \in T} \pi_t \prod_{w \in s} \phi_{wt}^{n_{sw}} + R(\Phi, \pi) \to \max_{\Phi, \pi}$$

K.Nigam, A.K.McCallum, S.Thrun, T.Mitchell. Text classification from labeled and unlabeled documents using EM. 2000.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Модель смеси униграмм для сегментов текста

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{s\in S_d}\ln\sum_{t\in T}\pi_t\prod_{w\in s}\phi_{wt}^{n_{sw}} + R(\Phi,\pi) \to \max_{\Phi,\pi}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:}\quad p_{tds} \equiv p(t|d,s) = \operatorname*{norm}_{t\in T}\Big(\pi_t \prod_{w\in s}\phi_{wt}^{n_{sw}}\Big);$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\Big); & n_{wt} = \sum\limits_{d\in D}\sum\limits_{s\in S_d} n_{sw}\,p_{tds} \\ \pi_t = \operatorname*{norm}_{t\in T}\Big(n_t + \pi_t\dfrac{\partial R}{\partial \pi_t}\Big); & n_t = \sum\limits_{d\in D}\sum\limits_{s\in S_d} n_s\,p_{tds} \end{cases}$$

Вместо θ_{td}: $p(t|d) = \sum\limits_{s\in S_d} p(s|d)p(t|d,s) = \sum\limits_{s\in S_d}\frac{n_s}{n_d}p(t|d,s).$

Контекстная документная кластеризация (CDC)

n_{uw} — со-встречаемость слов u и w;

$p(u|w) = \frac{n_{uw}}{n_w}$ — контекст слова w;

$H(w) = -\sum_{u \in W} p(u|w) \log p(u|w)$ — энтропия контекста слова w.

Узкий контекст — контекст с низкой энтропией, аналог темы, подмножество неслучайно часто со-встречающихся слов.

Метод CDC — Contextual Document Clustering:

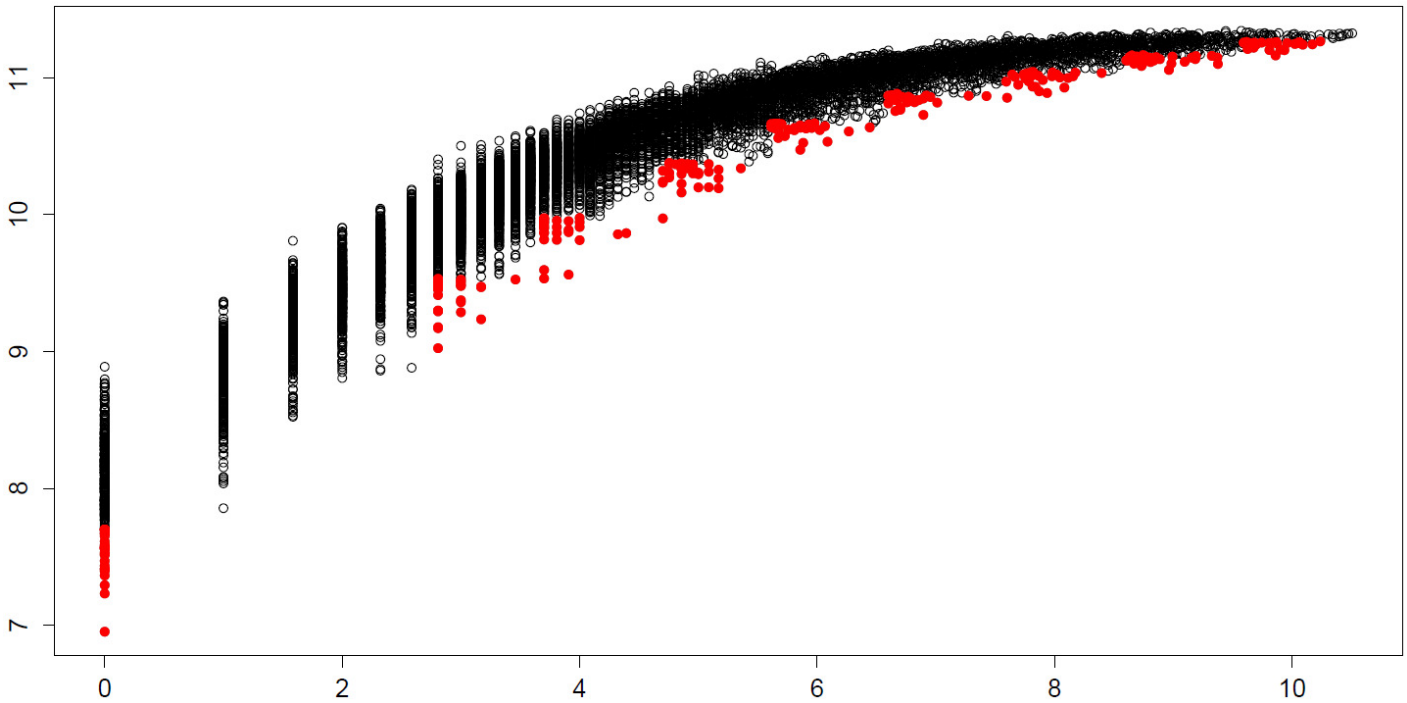
1. разбить документы на однородные сегменты (абзацы)
2. выделить слова с узкими контекстами
3. кластеризовать узкие контексты (найти темы)
4. отнести каждый сегмент к ближайшей теме

V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.

Выделение слов, имеющих узкие контексты

Оригинальный CDC: диапазон $\log_2 N_w$ разбивается на интервалы, в каждом интервале отбираются слова с наименьшими $H(w)$:

Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$

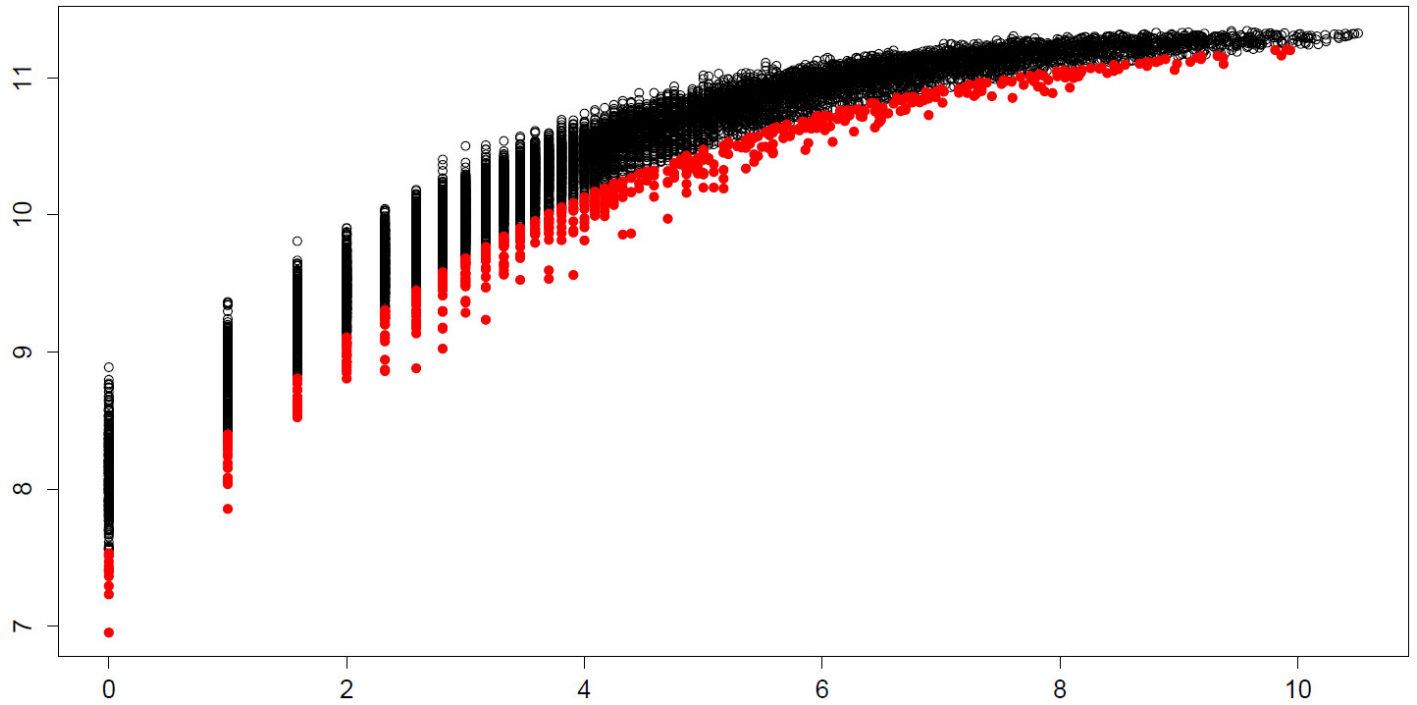


V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.

Выделение слов, имеющих узкие контексты

Более аккуратный отбор локальных контекстов
с помощью квантильной регрессии (отсекаем 5% снизу).

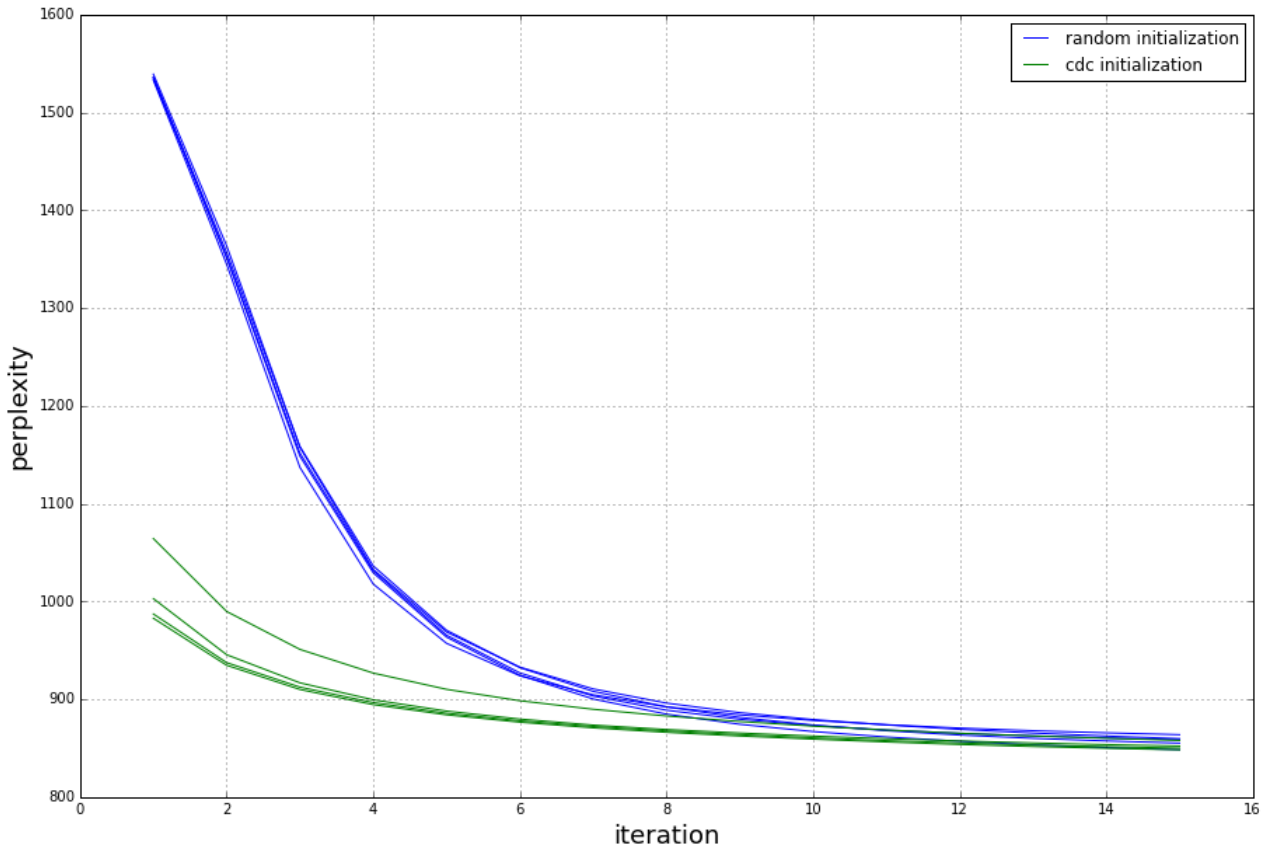
Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



А.Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей. 2015. МФТИ.

Инициализация тематической модели с помощью CDC

Зависимость перплексии от числа итераций (коллекция MMPO)



А.Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей. 2015. МФТИ.

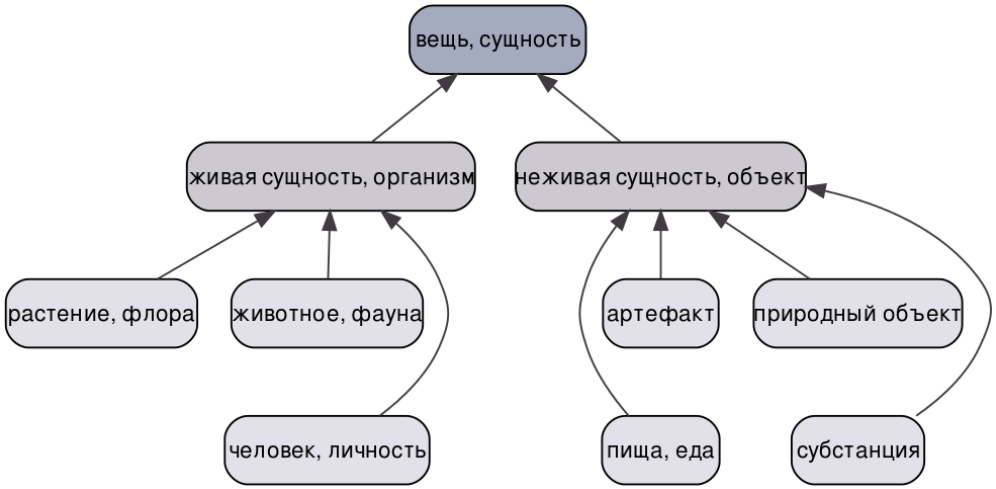
Семантическая сеть WordNet

117K *наборов синонимов* (synset), 155K слов, с определениями и примерами, связанных семантическими отношениями:

- *гипероним* — более общее (родовое) понятие
- *гипоним* — частное (видовое) понятие
- *холоним* — объемлющее целое
- *мероним* — составная часть

Словари разделены по частям речи:

- существительные
- глаголы
- прилагательные
- наречия



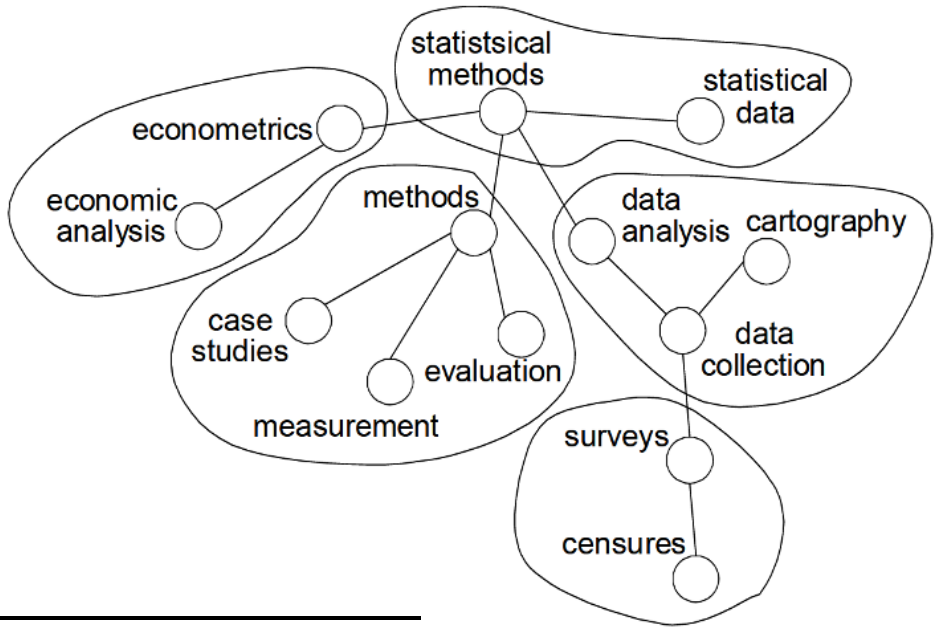
http://wordnet.princeton.edu

Метод лексических цепочек (Lexical Chains)

Лексическая цепочка — множество терминов:

- пары терминов связаны тезаурусными связями
- соседние термины на расстоянии не более 2 предложений
- возможна транзитивная связь через третий термин

Сильная цепочка — (почти) все слова связаны (клика)



Jane Morris, Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. 1991.

Пример выделения лексических цепочек

Пример использования русскоязычного тезауруса РуТез

О порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы.

Во исполнение Закона Российской Федерации "О статусе военнослужащих" и в целях обеспечения прав на жилище военнослужащих и граждан, уволенных с военной службы, Правительство Российской Федерации постановляет:
1. Утвердить прилагаемое Положение о порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы.
2. Министерству обороны Российской Федерации и иным федеральным органам исполнительной власти, в которых предусмотрена военная служба:
в месячный срок разработать и утвердить формы и перечень документов, необходимых для принятия решения об оказании военнослужащим безвозмездной

Лаборатория анализа информационных ресурсов, НИВЦ МГУ
http://www.labinform.ru/info/onthology

Применение TM для построения LC без тезауруса

LDA Mode Method (LDA-MM):

- тема каждого термина: $t(w) = \arg\max_t p(t|d,w)$
- термины с одинаковыми $t(w)$ образуют цепочку
- возможен учёт второй темы t' при $p(t'|d,w) > \varepsilon$

LDA Graph Method (LDA-GM):

- граф близостей всех терминов документа по $p(t|d,w)$
- максимальные клики этого графа образуют цепочки

LDA Top-N Method (LDA-TM):

- для каждого d выбираем top-N тем из $p(t|d)$
- для каждой t выбираем top-M терминов из $p(w|t)$
- все такие термины из d образуют цепочку

Steffen Remus. Automatically Identifying Lexical Chains by Means of Statistical Methods — A Knowledge-Free Approach. 2012.

Измерение качества построения лексических цепочек

Эксперты выделяли термины и лексические цепочки:

- по принципу однородности тематики
- повторения терминов, синонимы, коллокации, меронимы, гиперонимы, антонимы

	LDA-MM	LDA-GM	LDA-TM	S&M	G&M	Anno A	Anno B
avg. num. of lexical items per doc.	38.20	29.32	30.82	14.40	15.29	38.66	38.96
avg. num. of chains per doc.	13.80	9.12	7.32	5.83	5.71	11.25	7.38
avg. num. of links per doc.	8.60	2.06	1.44	–	–	5.47	2.41
avg. size lexical chains	2.82	3.41	4.61	2.48	2.68	3.69	5.57
avg. num. of merged lexical chains	5.76	7.06	5.98	–	–	6.10	4.99
avg. size merged lexical chains	8.29	4.45	5.57	–	–	7.60	8.91

Результаты:

- тематические модели сравнимы с экспертами
- тематические модели лучше семантических сетей

Steffen Remus. Automatically Identifying Lexical Chains by Means of Statistical Methods — A Knowledge-Free Approach. 2012.

Метод тематической сегментации TopicTiling

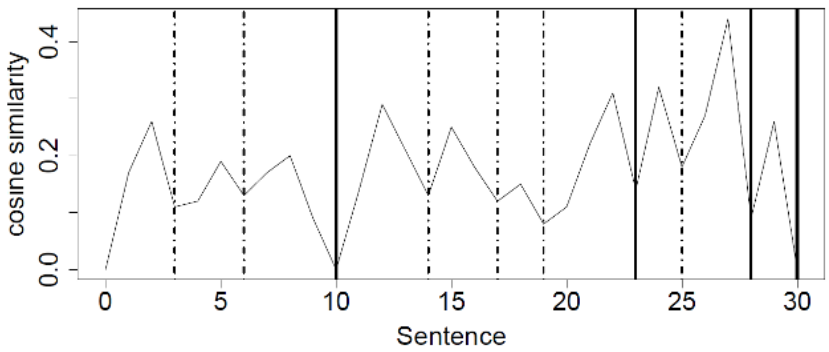
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

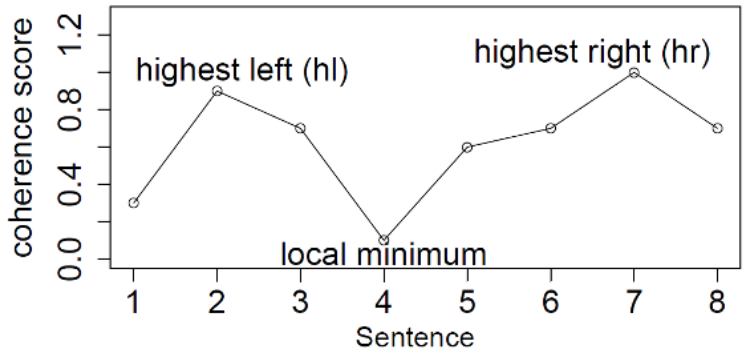
$p(t|d,s) = \frac{1}{|s|}\sum_{w\in s} p(t|d,w)$ — тематика предложения s

$p_j = \big(p(t|d,s_j)\big)_{t\in T}$ — тематический вектор предложения s_j

$c_j = \cos(p_{j-1}, p_j)$ — *coherence score*, оценка близости соседних предложений (чем глубже провал, тем чётче граница)

$d_j = \frac{1}{2}\big(\mathrm{hl}_j + \mathrm{hr}_j - 2c_j\big)$ — *depth score*, оценка глубины провала





Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Эвристики для TopicTiling

Эвристики для определения числа сегментов:

- заданное число провалов с наибольшей глубиной d_j
- провалы с глубиной более $\text{avr}\{d_j\} + \delta\,\text{stdev}\{d_j\}$, $\delta = 0{,}5..1{,}2$

Дополнительные эвристики и параметры:

- filter: игнорировать короткие предложения (менее 5 слов)
- игнорировать стоп-слова
- использовать фоновые темы и игнорировать их в p_j
- использовать $p(t|d,w)$ или $\arg\max_t p(t|d,w)$
- подбирать число итераций
- подбирать число тем $|T|$ в тематической модели LDA
- подбирать параметры α и β в LDA
- подбирать число предложений слева и справа от j

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Измерение качества сегментации

Базовые методы сегментации по векторам $p(w|s_j)$ и $p(t|s_j)$

- TT и TT-LDA — TextTiling (Hearst, 1997)
- C99 и C99-LDA — кластеризация предложений (Choi, 2000)

Коллекции для сравнения методов сегментации:

- *Choi dataset*: синтетический корпус, 700 документов по 10 сегментов, нарезанных из «Brown corpus»
- *Galley dataset*: синтетический корпус, 500 документов по 4–22 сегментов, нарезанных из «WSJ corpus»

Метрики для сравнения методов сегментации:

- Precision/Recall не учитывают границы между сегментами
- P_k (Beeferman et al., 1997)
- WD, WindowDiff (Pevzner and Hearst, 2002)

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Метрики для сравнения методов сегментации

Все метрики основаны на сравнении с идеальной сегментацией, т.н. «золотым стандартом» (gold standard).

- P_k (Beeferman et al., 1997) — чем меньше, тем лучше:
  $B_i = [\text{словопозиции } i \text{ и } i+k-1 \text{ лежат в одном сегменте}]$
  B_i^0 — то же самое для идеальной сегментации
  P_k — доля позиций, для которых $B_i \neq B_i^0$
- WD, WindowDiff (Pevzner and Hearst, 2002)
  $C_i = (\text{число сегментов между позициями } i \text{ и } i+k-1)$
  C_i^0 — то же самое для идеальной сегментации
  WD — доля позиций, для которых $C_i \neq C_i^0$

Doug Beeferman, Adam Berger, John Lafferty. Statistical models for text segmentation. 1999.

Lev Pevzner, Marti Hearst. A critique and improvement of an evaluation metric for text segmentation. 2002.

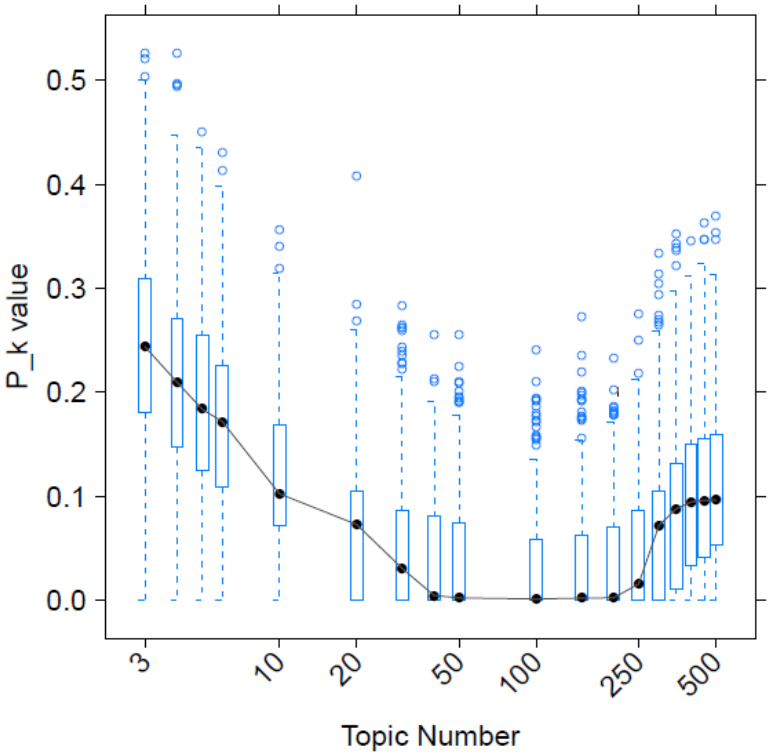
Результаты сравнения методов сегментации (Choi dataset)

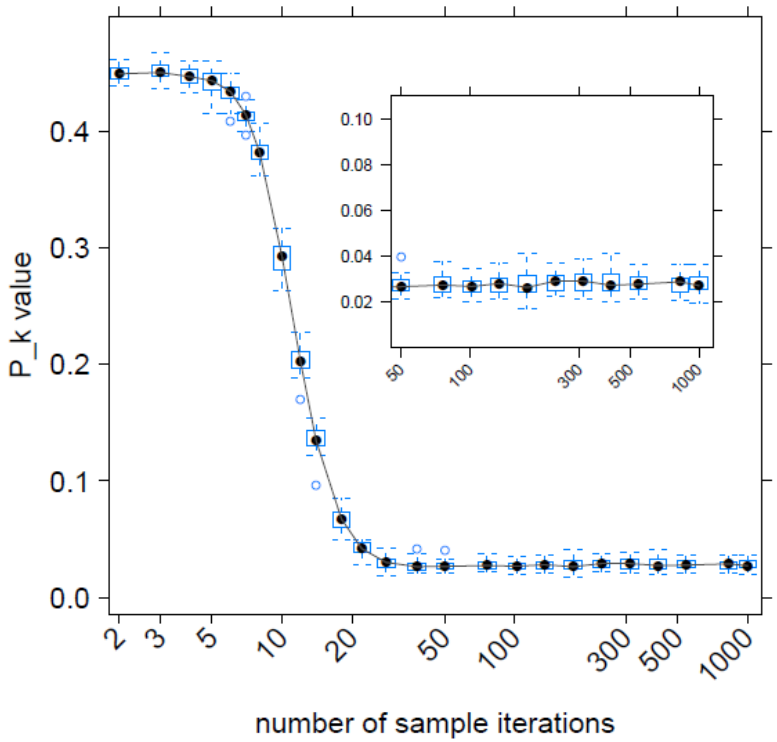
Method	Segments provided		Segments unprovided	
	P_k	WD	P_k	WD
C99	11.20	12.07	12.73	14.57
C99LDA	4.16	4.89	8.69	10.52
TT	44.48	47.11	49.51	66.16
TTLDA	1.85	2.10	16.41	21.40
TopicTiling	2.65	3.02	4.12	5.75
TopicTiling (filtered)	**1.50**	**1.72**	**3.24**	**4.58**

- Тематические модели лучше
- Лидирует TopicTiling с фильтрацией коротких предложений
- «Segments provided» — число сегментов известно (на реальных данных это нереалистичное предположение)

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимости P_k ($k = 6$) от параметров модели

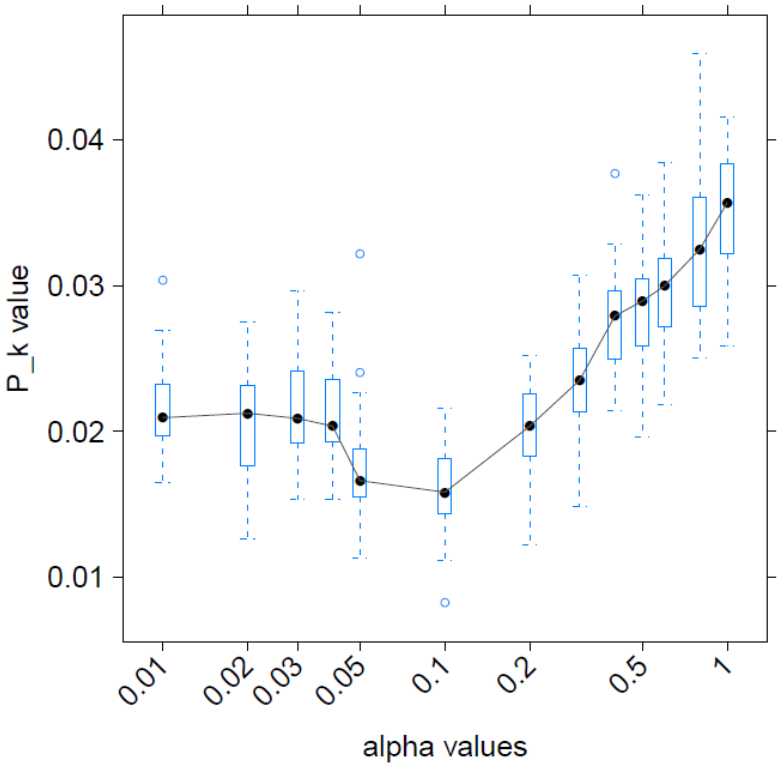


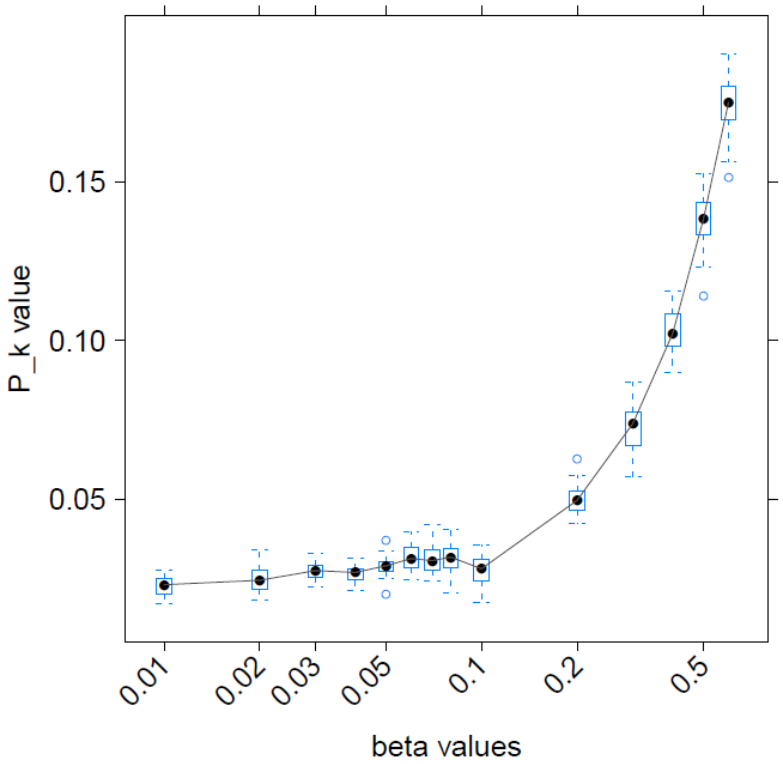


- Качество сегментации сильно зависит от $|T|$
- оптимальный диапазон $|T| = 50..150$ достаточно широк
- при $|T| = 100$ сходимость за 20–30 итераций

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимости P_k ($k = 6$) от параметров α, β модели LDA

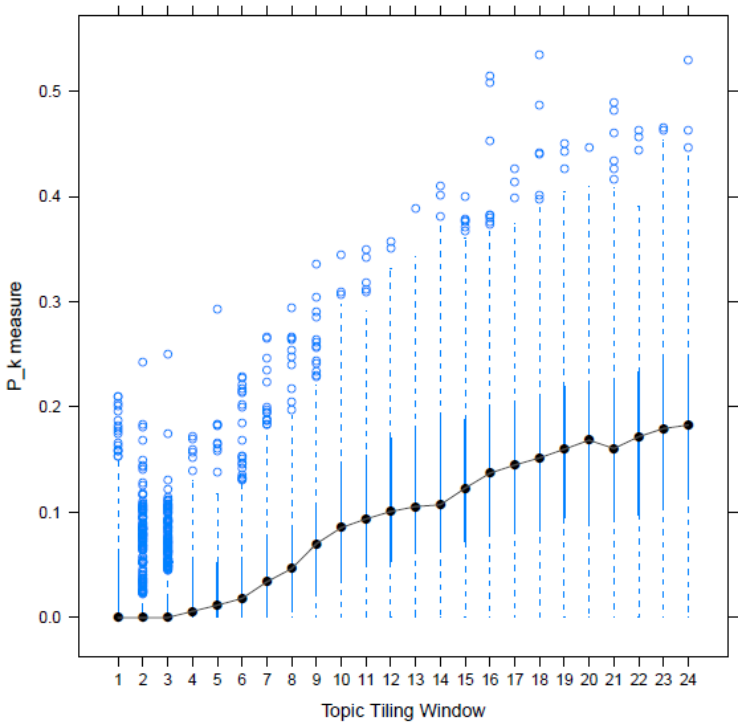




- Разреживать надо, но матрицу Θ — не слишком сильно
- параметры α, β менее критичны, чем число тем

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Зависимость P_k ($k = 6$) от ширины окна w (window)



фиксированное число сегментов:

seg. size	3-5		6-8		9-11		3-11	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
d=false,w=1	2.71	3.00	3.64	4.14	5.90	7.05	3.81	4.32
d=true,w=1	3.71	4.16	1.97	2.23	2.42	2.92	2.00	2.30
d=false,w=2	1.46	1.51	1.05	1.20	1.13	1.31	1.00	1.15
d=true,w=2	**1.24**	**1.27**	**0.76**	**0.85**	**0.56**	**0.71**	**0.95**	**1.08**
d=false,w=5	2.78	3.04	1.71	2.11	4.47	4.76	3.80	4.46
d=true,w=5	2.34	2.65	1.17	1.35	4.39	4.56	3.20	3.54

определяемое число сегментов:

seg. size	3-5		6-8		9-11		3-11	
	P_k	WD	P_k	WD	P_k	WD	P_k	WD
d=false,w=1	**2.39**	**2.45**	4.09	5.85	9.20	15.44	4.87	6.74
d=true,w=1	3.54	3.59	1.98	2.57	3.01	5.15	2.04	2.62
d=false,w=2	15.53	15.55	0.79	0.88	1.98	3.23	1.03	1.36
d=true,w=2	14.65	14.69	**0.62**	**0.62**	**0.67**	**0.88**	**0.66**	**0.78**
d=false,w=5	21.47	21.62	16.30	16.30	6.01	6.14	14.31	14.65
d=true,w=5	21.57	21.67	17.24	17.24	6.44	6.44	15.51	15.74

- Оптимальная ширина окна $w = 2$–3 предложения
- «d=true»: усреднение $\arg\max_t p(t|d, w)$ по каждому w
- Почему они не догадались использовать $p(t|d, w)$?

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Эксперименты на более реалистичных данных Galley's WSJ

фиксированное число сегментов:

Parameters	All words		Filtered	
	P_k	WD	P_k	WD
d=false,w=1	37.31	43.20	37.01	43.26
d=true,w=1	35.31	41.27	33.52	39.86
d=false,w=2	22.76	28.69	21.35	27.28
d=true,w=2	21.79	27.35	19.75	25.42
d=false,w=5	14.29	19.89	12.90	18.87
d=true,w=5	**13.59**	**19.61**	**11.89**	**17.41**
d=false,w=10	14.08	22.60	14.09	22.22
d=true,w=10	13.61	21.00	13.48	20.59

определяемое число сегментов:

Parameters	All words		Filtered	
	P_k	WD	P_k	WD
d=false,w=1	53.07	72.78	52.63	72.66
d=true,w=1	53.42	74.12	51.84	72.57
d=false,w=2	46.68	65.01	44.81	63.09
d=true,w=2	46.08	64.41	43.54	61.18
d=false,w=5	30.68	43.73	28.31	40.36
d=true,w=5	28.29	38.90	26.96	36.98
d=false,w=10	19.93	32.98	18.29	29.29
d=true,w=10	**17.50**	**26.36**	**16.32**	**24.75**

- Качество сегментации сильно зависит от коллекции
- Определять число сегментов стало труднее
- Окно пришлось расширить до $w = 5$–10 предложений
- Здесь «filtered» — учитывать только существительные, прилагательные и глаголы — помогает, но не сильно

Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Задача суммаризации (аннотирования, реферирования) текста

Автоматическая суммаризация — *краткий* текст, построенный по одному или нескольким документам и *наиболее полно* передающий их содержание.

Полуавтоматическая — HAMS, human aided machine summarization

Основные типы задач суммаризации:

- *one-document* — на входе один документ $d \in D$
- *multi-document* — на входе набор документов $D' \subseteq D$
- $\oplus$ *topic* — на входе набор фрагментов темы $p(d, s|t)$

Основные подходы к суммаризации:

- *extractive* — выбор некоторых предложений целиком
- *abstractive* — генерация текста на естественном языке

H.P.Luhn. The automatic creation of literature abstracts. 1958.
Juan-Manuel Torres-Moreno. Automatic Text Summarization. 2014.

Основные этапы выборочной (extractive) суммаризации

1. Внутреннее представление текста
   - выявление тематики текста и отдельных предложений
   - вычисление признаков предложений
2. Оценивание полезности (ранжирование) предложений
3. Отбор предложений для реферата
   - оптимизация критериев информативности и различности
   - оптимизация последовательности предложений
   - учёт целей и особенностей прикладной задачи (новости/статьи/веб-страницы/посты/мэйлы)

D.Das, A.Martins. A survey on automatic text summarization. 2007.

A.Nenkova, K.McKeown. A survey of text summarization techniques. 2012.

Yogita Desai, Prakash Rokade. Multi Document Summarization: Approaches and Future Scope. 2015.

Mahak Gambhir, Vishal Gupta. Recent automatic text summarization techniques: a survey. 2016.

Покрытие терминологии и тематики документа

S_d — множество предложений документа d
$a \subset S_d$ — искомая суммаризация

Покрытие терминологии документа (lexicon coverage):

$$\text{WCov}(a) = \text{KL}\big(p(w|d)\|p(w|a)\big) \to \min_{a \subset S_d}$$

Покрытие тематики документа (topic coverage):

$$\text{TCov}(a) = \text{KL}\big(p(t|d)\|p(t|a)\big) \to \min_{a \subset S_d}$$

Избыточность суммаризации (redundancy):

$$\text{Red}(a) = \sum_{s,s' \in a} B_{ss'} \to \min_{a \subset S_d}, \qquad B_{ss'} = \text{sim}\big(p(w|s), p(w|s')\big),$$

где sim — одна из мер сходства: cos, JS, Jaccard и т.п.

Marina Litvak, Natalia Vanetik, Chunlei Liu, Lemin Xiao, Onur Savas. Improving Summarization Quality with Topic Modeling. 2015.

Задача многокритериальной дискретной оптимизации

Метод релаксации: вместо $a \subset S_d$ ищем $\pi_s = p(s|a)$, где $s \in S_d$.
В релаксированной задаче:

$$p(w|a) = \sum_{s \in d} p(w|s)p(s|a) = \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s$$

$$p(t|a) = \sum_{s \in d} p(t|s)p(s|a) = \sum_{s \in d} \theta_{ts} \pi_s$$

Сумма трёх критериев $\mathrm{WCov}(a) + \tau_1 \mathrm{TCov}(a) + \tau_2 \mathrm{Red}(a)$:

$$\sum_{w \in d} n_{dw} \ln \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s + \tau_1 \sum_{w \in d} \theta_{td} \ln \sum_{s \in d} \theta_{ts} \pi_s - \tau_2 \sum_{s,s' \in d} B_{ss'} \pi_s \pi_{s'} \to \max_{\{\pi\}}$$

Максимизация покрытия — это максимизация правдоподобия!

Можно добавить регуляризатор разреживания:

$$R(\pi) = -\tau_3 \sum_{s \in S_d} \ln \pi_s \to \max_{\{\pi\}}$$

Оценка полезности предложений

Дополнительные признаки для отбора предложений:

- *SumBasic* — средняя частота слов, исключая стоп-слова
- *Centriod* — средний TF-IDF слов, превышающий порог
- *LexicalChain* — число слов сильных лексических цепочек
- *ImpactBased* — число слов из ссылающихся контекстов
- *TopicBased* — число слов из запроса пользователя

Стратегии отбора предложений:

- по одному top-предложению от каждой из top-тем
- поощрять выбор соседних предложений
- штрафовать предложения с анафорой и эллипсисом

A.Nenkova, K.McKeown. A survey of text summarization techniques. 2012.

Тематическая модель предложений для суммаризации

S_d — множество предложений документа d;
n_{sw} — частота термина w в предложении s;
n_s — длина предложения s.

Отбор предложений для суммаризации: $p(s|t) \to \max\limits_{s \in S_d}$

Тематическая модель сегментированного текста:

$$p(w|d) = \sum_{s \in S_d} p(w|s) \sum_{t \in T} p(s|t)p(t|d) = \sum_{s \in S_d} p_{ws} \sum_{t \in T} \psi_{st}\theta_{td}$$

где $p_{ws} \equiv p(w|s) = \frac{n_{ws}}{n_s}$ — частота термина w в предложении s.

Вместо ϕ_{wt} нельзя взять $p(w|t) = \sum\limits_{d \in D} \sum\limits_{s \in S_d} p_{ws}\psi_{st}$. Почему?

Dingding Wang, Shenghuo Zhu, Tao Li, Yihong Gong. Multi-document summarization using sentence-based topic models // ACL-IJCNLP 2009.

BSTM — Bayesian Sentence-based Topic Models

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{s\in S_d} p_{ws} \sum_{t\in T} \psi_{st}\theta_{td} \to \max_{\Phi,\Theta}$$

- Авторы утверждают, что модель переходит в обычную $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, если предложение $\equiv$ слово
- Это не так, если предложения уникальны: $S_d \cap S_{d'} = \varnothing$
- Модель разваливается на независимые модели документов (Litvak, 2015) такую LDA строят явно, это тоже работает!
- Но это не будет работать для multi-document summarization!
- А то, что модель «Bayesian», вообще не имеет значения ;)

Dingding Wang, Shenghuo Zhu, Tao Li, Yihong Gong. Multi-document summarization using sentence-based topic models // ACL-IJCNLP 2009.

Идея обобщения для много-документной суммаризации

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + \tau \sum_{d,w} n_{dw} \ln \sum_{s\in S_d} p_{ws} \sum_{t\in T} \psi_{st}\theta_{td} + R \to \max_{\Phi,\Psi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad \begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\big(\phi_{wt}\theta_{td}\big) \\ p_{stdw} \equiv p(s,t|d,w) = \underset{s,t\in S_d\times T}{\text{norm}}\big(p_{ws}\psi_{st}\theta_{td}\big) \end{cases}$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ \psi_{st} = \underset{s\in S_d}{\text{norm}}\Big(\sum_{w\in S_d} n_{dw}p_{stdw} + \psi_{st}\frac{\partial R}{\partial \psi_{st}}\Big) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \tau\sum_{w\in d}\sum_{s\in S_d} n_{dw}p_{stdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

$r \in R$ — множество рефератов, написанных людьми

s — суммаризация, построенная системой

Чем больше, тем лучше — для всех метрик семейства ROUGE

Доля n-грамм из рефератов, вошедших в суммаризацию s:

$$\text{ROUGE-}n(s) = \frac{\sum\limits_{r \in R} \sum\limits_{w} [w \in s][w \in r]}{\sum\limits_{r \in R} \sum\limits_{w} [w \in r]}$$

Доля n-грамм из самого близкого реферата, вошедших в s:

$$\text{ROUGE-}n_{\text{multi}}(s) = \max_{r \in R} \frac{\sum\limits_{w} [w \in s][w \in r]}{\sum\limits_{w} [w \in r]}$$

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. 2004.

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

$r \in R$ — множество рефератов, написанных людьми
s — суммаризация, построенная системой
Чем больше, тем лучше — для всех метрик семейства ROUGE

ROUGE-L(s) максимальная общая подпоследовательность s, r
ROUGE-W(s) штрафует за пропуски в подпоследовательности
ROUGE-S(s) аналог ROUGE-2(s) для биграмм с пропусками
ROUGE-SU-m(s) для биграмм с пропусками не длиннее m

$\mathrm{JS}\big(p(w|s), p(w|R)\big)$ — лучше всего коррелирует с экспертными оценками качества суммаризации (Lin, 2006).

Готовые пакеты для вычисления метрик: pyRouge и др.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. 2004.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, Jian-Yun Nie.
An Information-Theoretic Approach to Automatic Evaluation of Summaries. 2006.

Резюме

- Модели предложений — основа сегментации и суммаризации
- TopicTiling — метод тематической сегментации
- WindowDiff — мера качества сегментации
- Качественная суммаризация — открытая проблема NLP
- Тематические модели суммаризации развиты слабо
- ROUGE — семейство мер качества суммаризации, характеризуют далеко не все аспекты качества