

Вероятностные тематические модели

Лекция 12.

Суммаризация и именоване 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)»

1 Суммаризация текстов

- Оценивание и отбор предложений для суммаризации
- Тематическая модель предложений для суммаризации
- Метрики качества суммаризации

2 Автоматическое именоване тем

- Формирование названий-кандидатов
- Максимизация функции релевантности
- Максимизация покрытия и различности

3 Резюме по курсу

- Что работает в тематическом моделировании
- Открытые задачи тематического моделирования
- Сборник заданий

Задача суммаризации (реферирования, аннотирования) текста

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

Полуавтоматическая суммаризация

- MAHS, machine aided human summarization
- HAMS, human aided machine summarization

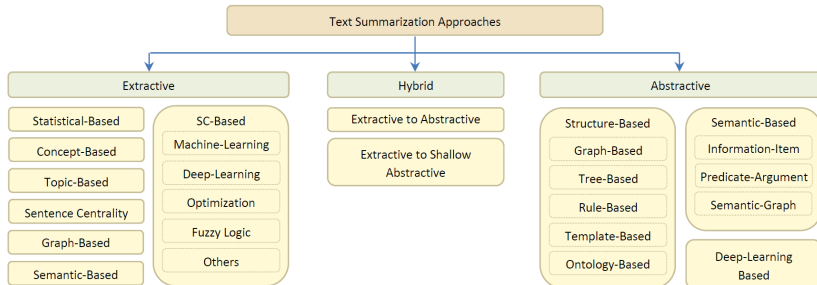
Основные типы задач суммаризации:

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

H.P.Luhn. The automatic creation of literature abstracts. 1958

Juan-Manuel Torres-Moreno. Automatic Text Summarization. 2014

Основные подходы и методы суммаризации



Основные подходы к суммаризации:

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

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, Hoda K. Mohamed.
Automatic text summarization: A comprehensive survey. 2021

Основные этапы выборочной (extractive) суммаризации

- 1 Внутреннее представление текста
 - граф/кластеризация/тематизация предложений в тексте
 - вычисление важности и других признаков предложений
- 2 Оценивание полезности (ранжирование) предложений
- 3 Отбор предложений для реферата
 - оптимизация критериев релевантности + 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

Y.Desai, P.Rokade. Multi document summarization: approaches and future scope. 2015

M.Gambhir, V.Gupta. Recent automatic text summarization techniques: a survey. 2016

TextRank — аналог ссылочного ранжирования PageRank

Текст — граф предложений; рёбра — похожие предложения.

Предложение $s \in S$ тем важнее,

- чем больше других предложений c , похожих на s ,
- чем важнее предложения c , похожие на s ,
- чем меньше предложений, на которые c также похоже.

Вероятность попасть в s , случайно блуждая по графу:

$$TR(s) = (1 - \delta) \frac{1}{|S|} + \delta \sum_{c \in S_s^{in}} \frac{TR(c)}{|S_c^{out}|},$$

$S_s^{in} \subset S$ — множество предложений c , похожих на s ,

$S_c^{out} \subset S$ — множество предложений, на которые похоже c ,

$\delta = 0.85$ — вероятность продолжать блуждания (damping factor)

S.Brin, L.Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. 1998
R.Mihalcea, P.Tarau. TextRank: Bringing Order into Text. EMNLP-2004

Определение сходства предложений

- Доля общих слов в двух предложениях
- Доля общих слов, за исключением слов общей лексики
- Доля общих n -грамм в двух предложениях
- Сходство тематических распределений двух предложений
- Сходство векторных представлений двух предложений

Другое применение TextRank — *извлечение ключевых слов* (keyword extraction) из отдельных документов.

В этом случае близость между словами (n -граммами) определяется по частоте их сочетаемости в окне ширины h

Покрытие терминологии и тематики документа

S_d — множество предложений документа d

$a \subset S_d$ — искомая суммаризация

Покрытие терминологии документа (lexicon coverage):

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

Покрытие тематики документа (topic coverage):

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

Избыточность суммаризации (redundancy):

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

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

Marina Litvak et al. 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:

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

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

Максимизация правдоподобия с регуляризацией:

$$WCov(a) + \tau_1 TCov(a) + \tau_2 Red(a) =$$

$$\sum_{w \in d} n_{dw} \ln \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s + \tau_1 \sum_{t \in T} \theta_{td} \ln \sum_{s \in d} \theta_{ts} \pi_s - \tau_2 \sum_{s, s' \in d} B_{ss'} \pi_s \pi_{s'} \rightarrow \max_{\{\pi\}}$$

Можно добавить регуляризатор разреживания:

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

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
- поощрять более простые (удобочитаемые) предложения
- 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 тем: $p(t|d) \rightarrow \text{top } k$ $t \in T$ и предложений: $p(s|t) \rightarrow \max_{s \in S_d}$

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 .

Темы в документах $d \in D$ нуждаются в согласовании:

$$p(w|t, d) = \sum_{s \in S_d} p_{ws} \psi_{st} \rightarrow p(w|t) = \phi_{wt}$$

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

BSTM — Bayesian Sentence-based Topic Model

Критерий максимума регуляризованного правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{s \in S_d} p_{ws} \sum_{t \in T} \psi_{st} \theta_{td} + R(\Psi, \Theta) \rightarrow \max_{\Psi, \Theta}$$

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

Согласованная ТМ для много-документной суммаризации

Строим одновременно две тематические модели, в которых ковариационным регуляризатором приближаем распределения $p(w|t, d) = \sum_s p_{ws} \psi_{st}$ к $p(w|t) = \phi_{wt}$ для всех $d \in D$:

$$\sum_{d \in D} \sum_{t \in T} \sum_{w \in W} \phi_{wt} \sum_{s \in S_d} p_{ws} \psi_{st} \rightarrow \max_{\Phi, \Psi}$$

Критерий максимума регуляризованного правдоподобия:

$$\begin{aligned} & \sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \\ & + \tau \sum_{d,w} n_{dw} \ln \sum_{s \in S_d} p_{ws} \sum_{t \in T} \psi_{st} \theta_{td} + \\ & + \mu \sum_{d,w} \sum_{s \in S_d} \sum_{t \in T} \phi_{wt} p_{ws} \psi_{st} + R(\Phi, \Psi, \Theta) \rightarrow \max_{\Phi, \Psi, \Theta} \end{aligned}$$

где τ, μ — коэффициенты регуляризации

Согласованная ТМ для много-документной суммаризации

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{l}
 \text{E-шаг:} \\
 \text{M-шаг:}
 \end{array}
 \left\{ \begin{array}{l}
 p_{tdw} \equiv p(t|d, w) = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\
 p_{stdw} \equiv p(s, t|d, w) = \mathop{\text{norm}}_{s, t \in S_d \times T} (p_{ws} \psi_{st} \theta_{td}) \\
 p_{wst} = p_{ws} \phi_{wt} \psi_{st} \\
 \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \mu \sum_{d \in D} \sum_{s \in S_d} p_{wst} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\
 \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \tau \sum_{s \in S_d} \sum_{w \in W} n_{dw} p_{stdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \\
 \psi_{st} = \mathop{\text{norm}}_{s \in S_d} \left(\tau \sum_{w \in W} n_{dw} p_{stdw} + \mu \sum_{w \in W} p_{wst} + \psi_{st} \frac{\partial R}{\partial \psi_{st}} \right)
 \end{array} \right.$$

Можно ли было сделать проще?

S_d — множество предложений документа d ;

n_{sw} — частота термина w в предложении s ;

n_s — длина предложения s .

Отбор тем: $p(t|d) \rightarrow \operatorname{top} k_{t \in T}$ и предложений: $p(s|t) \rightarrow \max_{s \in S_d}$

Тематическая модель предложения s :

$$p(s|t) = p(t|s) \frac{p(s)}{p(t)} = \sum_{w \in S} p(t|w) p(w|s) \frac{p(s)}{p(t)} = \sum_{w \in S} \phi_{wt} \frac{p(s)}{p(w)} p_{ws}$$

Кажется, можно обойтись обычной тематической моделью, либо моделью линейной однопроходной тематизации.

Но есть сомнение: темы не оптимизируются под суммаризацию.

Какой подход для отбора предложений лучше?

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

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 :

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

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

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

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

$JS(p(w|s), p(w|R))$ — 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.

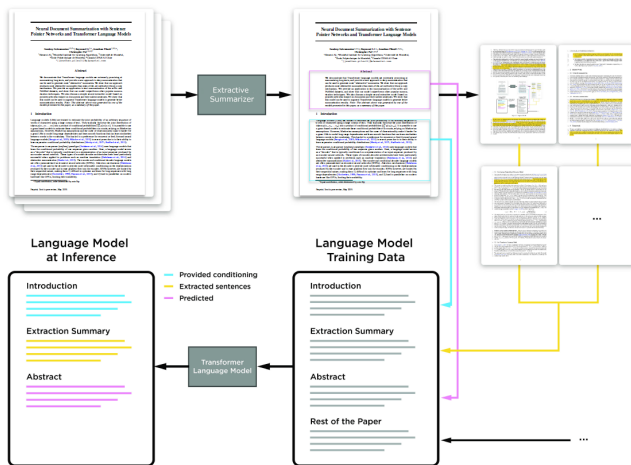
Абстрактивная суммаризация на основе трансформеров

Abstract

We present a method to produce abstractive summaries of long documents that exceed several thousand words via neural abstractive summarization. We perform a simple extractive step before generating a summary, which is then used to condition the transformer language model on relevant information before being tasked with generating a summary. We show that this extractive step significantly improves summarization results. We also show that this approach produces more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. *Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper.*

S.Subramanian, R.Li, J.Pilault, C.Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. 2019.

Абстрактивная суммаризация использует экстрактивную



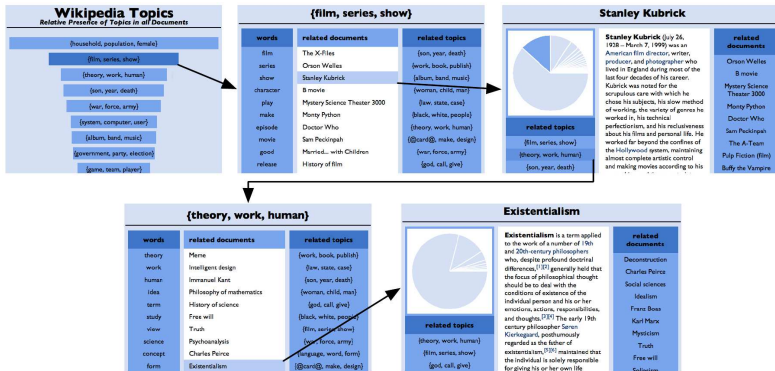
S.Subramanian, R.Li, J.Pilault, C.Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. 2019.

Резюме по суммаризации

- Тематические модели используются в суммаризации, чтобы выделить и покрыть наиболее важные темы
- Суммаризация темы — открытая проблема ТМ, её не только не решали, но даже и не ставили!
«Let the topics tell about themselves!»
- ROUGE — семейство мер качества суммаризации, характеризуют далеко не все аспекты качества
- BLUE — аналогичные метрики, но precision-based
- Для визуализации нужны суммаризация и именование тем

Система TMVE — Topic Model Visualization Engine

Три топовых слова темы — самая простая модель именования:



<https://github.com/ajbc/tmv>

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

Задача автоматического именованя тем (topic labeling)

Требования к названию темы (topic label):

- интерпретируемость и грамматическая корректность
- точность представления семантики темы
- полнота представления семантики темы
- непохожесть на названия других тем, включая похожие

Гипотеза: все названия уже придуманы, осталось их найти.

Подзадачи

- формирование названий-кандидатов l_1, \dots, l_m
- построение (обучение) функции релевантности $s(l, t)$
- выбор названия с учётом названий похожих тем

Qiaozhu Mei (ЦяоЧжу Мэй), Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Способы формирования названий-кандидатов

Специфичные для данной темы:

- топовые n -граммы данной темы
- синтаксические ветки наиболее тематичных предложений
- тематичные именные группы (вырезанные OpenNLP chunker)
- тематичные фразы «объект, субъект, действие»
- заголовки тематичных документов или их фрагменты
- метаданные (теги, категории) тематичных документов

Общие для всех тем:

- n -граммы из внешней коллекции, например, Википедии
- заголовки статей или категорий Википедии
- термины из внешних тезаурусов:
WordNet, PyТез, Викисловарь, и др.

Функция релевантности (relevance score)

Релевантность нулевого порядка:

$$s(\ell, t) = \sum_{w \in \ell} \log \frac{p(w|t)}{p(w)} \rightarrow \max$$

Релевантность первого порядка: слова темы t неслучайно часто появляются рядом (в одном контексте C) с названием ℓ :

$$s(\ell, t) = \sum_{w \in C} p(w|t) \log \underbrace{\frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\text{PMI}(w, \ell|C)} \rightarrow \max$$

где C — релевантный теме контекст, в котором ожидается появление как слов темы t , так и названия ℓ целиком (Например, статья или категория Википедии).

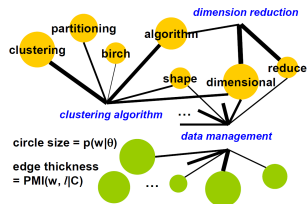
Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Проблема названий, подходящих для нескольких тем

Пример: оранжевая тема покрывается двумя названиями:

- *clustering algorithm*
- *dimension reduction*

но название *data management* неудачно, конкурирует с другой темой



Выбирать каждое следующее название, чтобы оно было

- максимально релевантно, $s(\ell, t) \rightarrow \max$,
- максимально не похоже на названия ℓ' остальных тем:

$$s(\ell, t) + \lambda \max_{\ell'} KL(\ell' || \ell) \rightarrow \max$$

где параметр λ подбирается эмпирически.

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Максимизация различности названий различных тем

Модифицированная функция релевантности $s'(l, t)$:

- максимизирует релевантность своей темы, $s(l, t) \rightarrow \max$
- минимизирует релевантность других тем, $s(l, t') \rightarrow \min$

$$s'(l, t) = s(l, t) - \mu \sum_{t' \in T \setminus t} s(l, t') \rightarrow \max$$

где параметр μ подбирается эмпирически.

Методика оценивания качества именованя тем:

- 3 ассессора, каждый ассессор видит для каждой темы:
 - список топ-слов темы, список топ-документов темы
 - варианты названия, сгенерированные разными методами
- ассессор ранжирует методы $0, 1, 2, \dots$ (чем выше, тем лучше)

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Оценивание качества именованя тем

Две коллекции: научная (SIGMOD), новостная (Assoc.Press)
 Автоматические и ассессорские названия тем, SIGMOD:

Auto Label	clustering algorithm	r tree	data streams	concurrency control
Man. Label	clustering algorithms	indexing methods	Stream data management	transaction management
θ	clustering clusters video dimensional cluster partitioning quality birch	tree trees spatial b r disk array cache	stream streams continuous monitoring multimedia network over ip	transaction concurrency transactions recovery control protocols locking log

Победил выбор n -грамм по релевантности 1-го порядка,
 но он всё ещё заметно хуже человеческого именованя тем:

Baseline v.s. Zero-order v.s. First-order				
Dataset	#Label	Baseline	Ngram-0-B	Ngram-1
SIGMOD	1	0.76	0.75	1.49
SIGMOD	5	0.36	1.15	1.51
AP	1	0.97	0.99	1.02
AP	5	0.85	0.66	1.48

System v.s. Human			
Dataset	#Label	Ngram-1	Human
SIGMOD	1	0.35	0.65
SIGMOD	5	0.25	0.75
AP	1	0.24	0.76
AP	5	0.21	0.79

Qiaozhu Mei et al. Automatic labeling of multinomial topic models. KDD 2007.

Резюме по автоматическому именованию тем

- *Automatic Topic Labeling* — очень узкое направление, около 50 статей начиная с 2007 г.
- Важно для автоматизации создания приложений
- Близко к задаче суммаризации темы
- Для иерархических моделей добавляется специфичное требование *полноты*: названия дочерних тем должны адекватно описывать разделение родительской темы

Alex Yoo. Automatic topic labeling in 2018: history and trends.

<https://medium.com/datadriveninvestor/automatic-topic-labeling-in-2018-history-and-trends-29c128cec17>

A.Gourru et al. United we stand: Using multiple strategies for topic labeling. 2018.

Ciprian-Octavian Truicam And Elena-Simona Apostol TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition. 2021.

Supriya Kinariwala, Sachin Deshmukh Onto_TML: Auto-labeling of topic models. 2021.

M.Allahyari, S.Pouriyeh, K.Kochut, H.R.Arabnia. A knowledge-based topic modeling approach for automatic topic labeling. 2017.

Обсуждение. Главные выводы по всему курсу

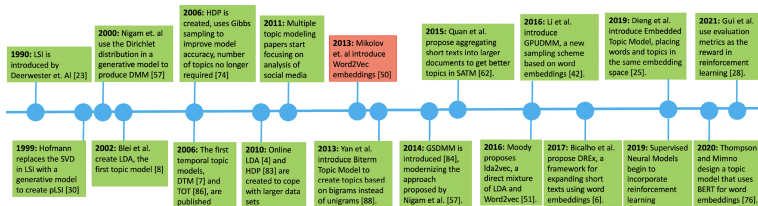
Что точно работает в тематическом моделировании
...или «узнав об этом, по другому уже не захочется»

- 1 лемма о максимизации функции на единичных симплексах
- 2 аддитивная регуляризация вместо байесовского обучения
- 3 библиотека BigARTM: скорость + функциональность
- 4 многокритериальное оценивание
- 5 словари терминов n -грамм
- 6 декоррелирование тем
- 7 мультимодальные модели
- 8 однопроходный EM-алгоритм
- 9 тематические иерархии
- 10 спектр тем (ранжирование тем по взаимной близости)
- 11 реализация ARTM на PyTorch

Обсуждение. Открытые задачи тематического моделирования

- 1 ТЕМАТИЗАТОР: встраивание в Orange, PolyAnalyst
- 2 проблема несбалансированности тем
- 3 доля интерпретируемых тем должна быть 100%
- 4 автоматическое именование и аннотирование тем
- 5 автоматическое обнаружение новых тем в пакетах
- 6 автоматическое разделение тем на подтемы
- 7 автоматический подбор гиперпараметров, AutoML
- 8 оптимизация гиперпараметров в потоковом режиме
- 9 тематическая модель внимания (локальных контекстов)
- 10 обеспечение полноты и устойчивости множества тем
- 11 бережное слияние моделей нескольких коллекций
- 12 развитие нейросетевых тематических моделей

Обсуждение. Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гоменизация, локализация (контекст и внимание)

X. Wu et al. A survey on neural topic models: methods, applications, and challenges. 2023.

Rob Churchill, Lisa Singh. The evolution of topic modeling. 2022.

He Zhao et al. Topic modelling meets deep neural networks: a survey. 2021.

Напоминание. Практические задания по курсу

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: сотворить нечто общественно полезное

виды деятельности	оценка
теоретические задачи	$\sum_i A_i X_i$
решение прикладной задачи	10X
обзор по NeuralTM	12X
реализация ARTM для pyTorch	16X
участие в проекте Тематизатор	20X
работа над открытой проблемой	20X

где X — оценка за вид деятельности по 5-балльной шкале,
 A_i — сложность задачи = числу «звёздочек»

Итоговая оценка: $\min(10, \lfloor \text{score}/10 \rfloor)$ по 10-балльной шкале.

Некоторые датасеты для заданий по спецкурсу

- Научные статьи: arXiv, PubMed, Semantic Scholar
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- Википедия
- Акты арбитражных судов РФ
- Данные социальных сетей: Twitter, VK, LJ,...
- TechCrunch (английский)
- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Транзакции клиентов Sberbank DSD 2016

<http://bigartm.org>

<http://drive.google.com/drive/folders/1PPnw6aZOJAJolRYuwdGm437RssV-XQx0>

Два упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

Найти параметры модели ξ_{dw} .

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

Найти параметры модели ξ_w .

Подсказка: применить условия ККТ.

Третье упражнение в продолжение — более творческое:

3. Предложите модель, определяющую роли слов в текстах:

— тематические слова

— специфичные слова документа (шум)

— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т, ш, ф}\}$.

Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

4. Заменяем \log другой монотонно возрастающей функцией μ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \mu \left(\sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Как изменится EM-алгоритм? Возможно ли подобрать функцию μ так, чтобы сократился объём вычислений?

5. Заменяем \log монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \mu(\phi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_t \mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

6*. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \text{norm}_w(n_{wt} [n_{wt} > \gamma n_t])$$

Подсказка: см. слайд 12 лекции №3.

Аналитик построил тематическую модель Φ^0, Θ^0 и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

7. Предложите регуляризаторы для этого.

8. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi_{wt}^0$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности?

9. Почему это плохо? Как этого избежать?

10. Предложите способ инициализации Φ для новой модели.

11. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = (p(s|t))$, гарантирующий, что

- 1) у каждой родительской темы будет хотя бы одна дочерняя;
- 2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу M-шага для матрицы Ψ .

12*. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s , то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$.

13.** Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$

с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$

14. Выведите EM-алгоритм для тематической модели битермов (Biterm Topic Model) из лекции №8.

15. Выведите EM-алгоритм для тематической модели с гладким регуляризатором $R(\Phi, \Theta)$ и суммой L_1 -регуляризаторов

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) - \sum_{j \in J} \lambda_j |R_j(\Phi, \Theta)|.$$

Подсказка: ввести дополнительные неотрицательные переменные, чтобы избавиться от негладкой функции модуля, затем применить условия Каруша–Куна–Таккера.

16. Запишите формулы M-шага для частного случая — L_1 -сглаживания $p(i|t) = \phi_{it}$ в соседних интервалах $i, i-1$:

$$R_{\text{сгл}}(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} |\phi_{it} - \phi_{i-1,t}| \rightarrow \max.$$

- 17.** Для тематической модели предложений с фоновой темой (слайд 19) выведите решение для переменной x_{dsw} .
- 18.** Для тематической модели предложений с фоновой темой выведите формулы M-шага в случаях: $x_{dsw} = x$, $x_{dsw} = x_d$.
- 19***. Выведите EM-алгоритм с локализованным E-шагом (слайд 31) для локализованной тематической модели. Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw} ?
- 20**.** Предложите параметризацию для тематической модели внимания (слайд 35). Используя «основную лемму», получите уравнения для новых параметров модели.