

Вероятностные тематические модели

Лекция 11.

Суммаризация и именование тем

K. B. Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 28 апреля 2025

1 Суммаризация текстов

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

2 Автоматическое именование тем

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

3 Резюме по курсу

- Что работает в тематическом моделировании
- Нейросетевые тематические модели
- Задания по курсу

- ❶ Гарантированная интерпретируемость тем «из коробки»
 - проблема тематической несбалансированности коллекции
 - тематические модели внимания последовательного текста
 - измерение и оптимизация внутритекстовой когерентности
 - оптимизация числа тем и других гиперпараметров
 - обнаружение новых тем и трендов в потоке текстов
- ❷ Каждая тема должна уметь рассказать о себе
 - автоматическое именование и суммаризация тем
 - гибридная экстрактивно-абстрактивная суммаризация
 - ранжирование фраз темы для сценария суммаризации
 - пересказ темы с помощью LLM
- ❸ Тематическое моделирование в «Мастерской знаний»
 - визуализация тематики подборки, «карты знаний»
 - выявление нетривиальных междисциплинарных связей
- ❹ Тематическое моделирование в контент-анализе
 - поддержка экспертной разметки и отбора тем
 - использование категоризации фрагментов в ARTM
 - интеграция ARTM и LLM для поиска фрагментов

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

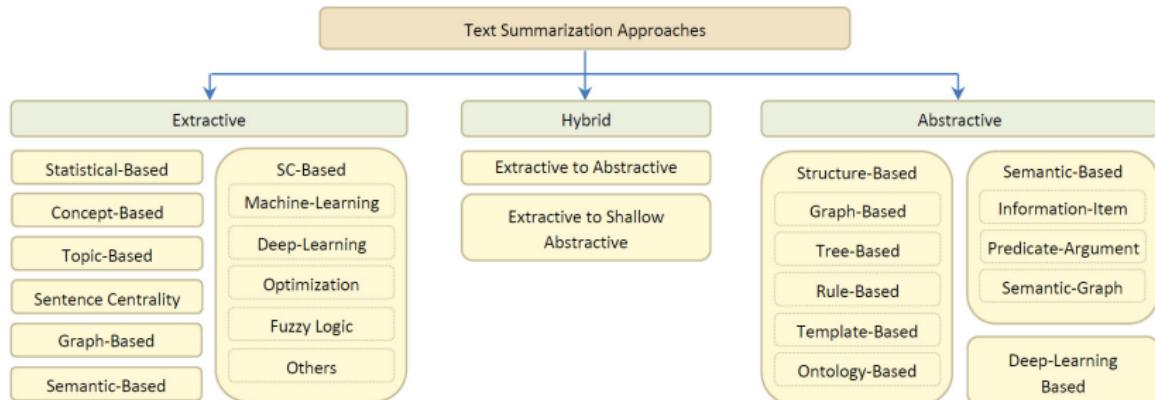
Основные типы задач суммаризации:

- *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

- ➊ Внутреннее представление текста
 - граф/кластеризация/тематизация предложений в тексте
 - вычисление важности и других признаков предложений
- ➋ Оценивание полезности (ранжирование) предложений
- ➌ Отбор предложений для реферата
 - оптимизация критериев релевантности + различности
 - оптимизация последовательности предложений
 - учёт целей и особенностей прикладной задачи
(статьи/патенты/новости/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

Граф: вершины — предложения; 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 s также похоже.

Вероятность попасть в s , случайно блуждая по графу:

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

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

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

$\delta = 0.85$ — 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

Для пары предложений:

- доля общих слов
- доля общих слов, за исключением слов общей лексики
- доля общих n -грамм
- сходство тематических распределений
- сходство векторных представлений (эмбедингов)

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.

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:

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

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

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

Тематическая модель предложений для суммаризации

S_d — множество предложений документа d ;

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

n_s — длина предложения s .

Отбор тем: $p(t|d) \rightarrow \max_{t \in T} k$ **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 терма w в предложении s .

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

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

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

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

М-шаг:
$$\begin{aligned} \phi_{wt} &= \underset{w \in W}{\text{norm}} \left(\sum_{d \in D} n_{dw} p_{tdw} + \mu \sum_{d \in D} \sum_{s \in S_d} p_{wst} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} &= \underset{t \in T}{\text{norm}} \left(\sum_{w \in s} n_{dw} p_{tdw} + \tau \sum_{s \in S_d} \sum_{w \in s} n_{dw} p_{stdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \\ \psi_{st} &= \underset{s \in S_d}{\text{norm}} \left(\tau \sum_{w \in s} n_{dw} p_{stdw} + \mu \sum_{w \in s} p_{wst} + \psi_{st} \frac{\partial R}{\partial \psi_{st}} \right) \end{aligned}$$

$$R(\Psi) = -\tau \sum_{t \in T} \ln \psi_{st} — \text{для разреживания распределений } p(t|s)$$

Нельзя ли сделать проще?

S_d — множество предложений документа d ;

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

n_s — длина предложения s .

Отбор тем: $p(t|d) \rightarrow \max_{t \in T} k$ **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 для сегментации.

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

Йенсен-Шенон 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.

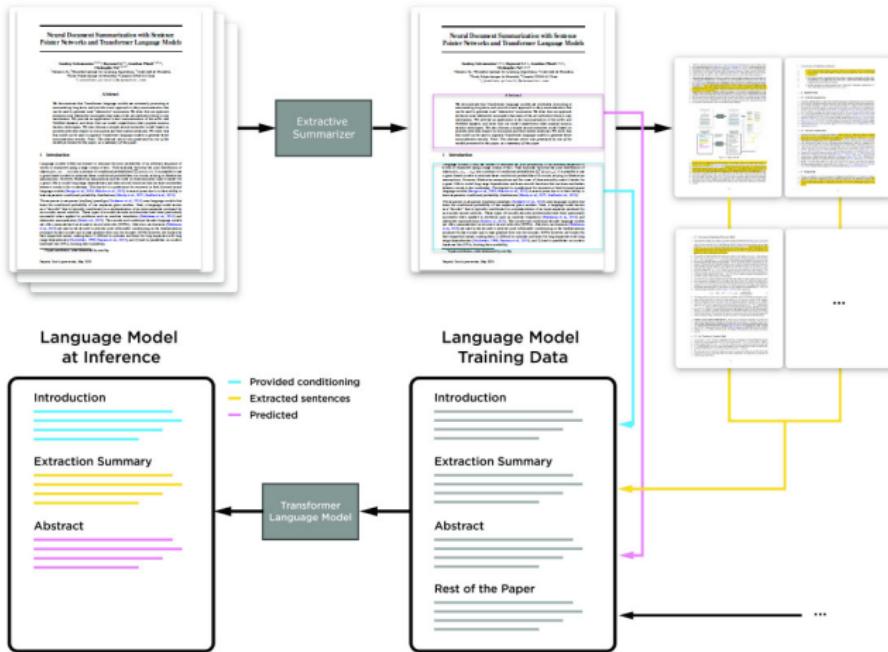
Абстрактивная суммаризация на основе трансформеров

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.

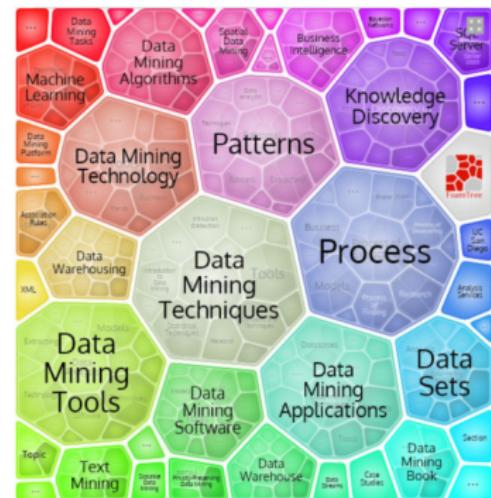
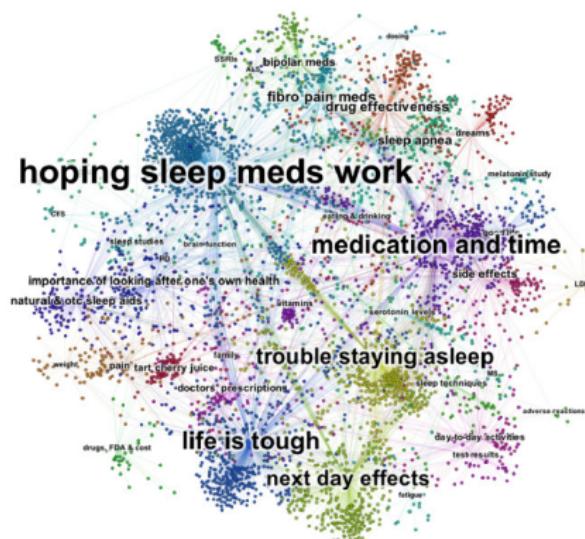
Резюме по суммаризации

- Тематические модели в экстрактивной суммаризации
 - для выделения и покрытия наиболее важных тем
- 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

Автоматическое именование тем для визуализации

Пример 1: тематика обсуждений на www.PatientsLikeMe.com

Пример 2: иерархическая карта Data Mining



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 именования 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 $s(\ell, t)$
- выбор названия с учётом названий похожих тем

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

Способы формирования названий-кандидатов

Отбираются вероятно 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, РуТез, 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 :

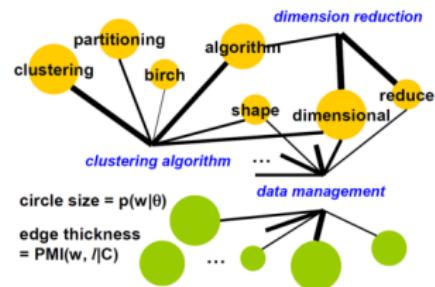
$$s(\ell, t) = \sum_{w \in C} p(w|t) \underbrace{\log \frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\text{PMI}(w, \ell|C)} \rightarrow \max$$

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:

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

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 $s'(\ell, t)$:

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

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

где параметр μ подбирается эмпирически.

Методика оценивания качества именования тем:

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

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

Оценивание качества именования тем

Две коллекции: научная (SIGMOD), новостная (Assoc. Press)

Автоматические и асессорские 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 именования 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.

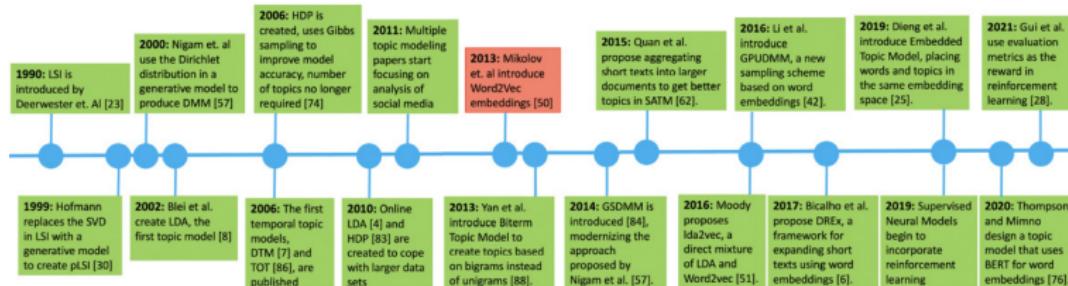
M.Allahyari, S.Pouriyeh, K.Kochut, H.R.Arabnia. A knowledge-based topic modeling approach for automatic topic labeling. 2017.

Начинаем подводить итоги по всему курсу

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

- ❶ лемма о максимизации функции на единичных симплексах
- ❷ аддитивная регуляризация вместо байесовского обучения
- ❸ библиотека BigARTM: скорость + функциональность
- ❹ однопроходный EM-алгоритм, тематическое внимание
- ❺ многокритериальное оценивание
- ❻ словари терминов n -грамм
- ❼ декоррелирование тем
- ❼ мультимодальные модели
- ❼ тематические иерархии
- ❽ спектр тем (ранжирование тем по взаимной близости)
- ❾ реализация ARTM на PyTorch

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, гомогенизация, 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.

Задания по курсу

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

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	5X
обзор по NeuralTM	5X
интеграция ARTM в pyTorch	5X
участие в одном из проектов	10X
работа над открытой проблемой	10X

где X — оценка за вид деятельности по 5-балльной шкале.

Итоговая оценка: $\min(5, \lfloor \text{score}/10 \rfloor)$ по 5-балльной шкале.

Теоретическое задание к лекции 2

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 или основную лемму.

3. (более творческое задание)

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.

Теоретическое задание к лекции 3

4. Заменим \ln гладкой 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. Заменим \ln гладкой 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 М-шаг и воздействие регуляризатора на модель?

6. Какому регуляризатору соответствует формула М-шага

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

Теоретическое задание к 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? Почему это плохо и как этого избежать?
9. Предложите способ инициализации Φ для новой модели.

Теоретическое задание к лекции 4

10. Выведите EM-алгоритм с локализованным E-шагом (слайд 13) для локализованной тематической модели.

Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw} ?

11. Предложите параметризацию для тематической модели внимания (слайд 21) Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательское задание к лекции 4

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный Е-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i, \hat{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i, \hat{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i, \hat{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайнового EM-алгоритма
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на Е-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}, \hat{\theta}_{ti}$ »

Теоретическое задание к лекции 5

- 12.** 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 .

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

- 14.** 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 $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 5

Участие в проекте «Мастерская знаний»

Дано:

- подборки, сгенерированные SciRus по одной статье
- асессорская разметка статей подборки по релевантности
- несколько вариантов токенизации
 - в том числе с автоматическим выделением терминов

Найти:

- тематическую модель
- модель ранжирования подборки по релевантности
- оптимальные: токенизацию, число тем, регуляризаторы
- распределение терминов по тематичности

Критерий:

- качество ранжирования
- (визуально) интерпретируемость тем
 - в том числе автоматического именования тем

Исследовательские задания к лекции 6

- Проблема несбалансированности тем
 - генераторы синтетических несбалансированных коллекций
 - модели локального контекста лишены этой проблемы?
 - регуляризаторы декоррелирования + семантической однородности
- Семейство средневзвешенных статистик
 - генераторы синтетических коллекций, удовлетворяющих гипотезе условной независимости
 - как (и нужно ли) определять пороги для построения статистических тестов условной независимости?
 - как ослабить проверку гипотезы условной независимости в модели локального контекста?
 - как перестраивать несогласованные темы?
- Критерий внутритекстовой когерентности
 - найти лучший вариант критерия с помощью калибровки по размеченным тематическим цепочкам
 - вычисление критерия должно естественным образом встраиваться в модель локального контекста

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

- «Мастерская знаний» для научного поиска
 - пользователь строит тематические подборки статей,
 - поисковая выдача формируется моделью SciRus.
 - задача: показать пользователю тематику подборки
 - понадобится автоматическое выделение терминов,
 - выделение тематических фраз из документов,
 - автоматическое именование и суммаризация тем
 - конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
 - пользователь задаёт грубый фильтр текстового потока
 - задача: «классифицировать иголки в стоге сена»,
 - разделив темы на информативные и мусорные,
 - выделив аспекты и тональности в каждой теме
 - конечная цель: q&q аналитика проблемной среды

- 1 Проблема несбалансированности тем в коллекции
- 2 Обеспечение 100%-й интерпретируемости тем
- 3 Тематические модели внимания последовательного текста
- 4 Обнаружение новых тем или трендов в потоке текстов
- 5 Автоматическое именование и аннотирование тем
- 6 Обзор подходов в нейросетевых тематических моделях
- 7 Обеспечение полноты и устойчивости множества тем
- 8 Автоматический подбор гиперпараметров, AutoML
- 9 Оптимизация гиперпараметров в потоковом режиме
- 10 Проблема несбалансированности текстов по длине
- 11 Бережное слияние моделей нескольких коллекций
- 12 Гиперграфовые тематические модели в RecSys