• Вероятностные языковые модели •
Лекция 9.
Открытые проблемы, проект A*RTM,
...а также про всякую кластеризацию

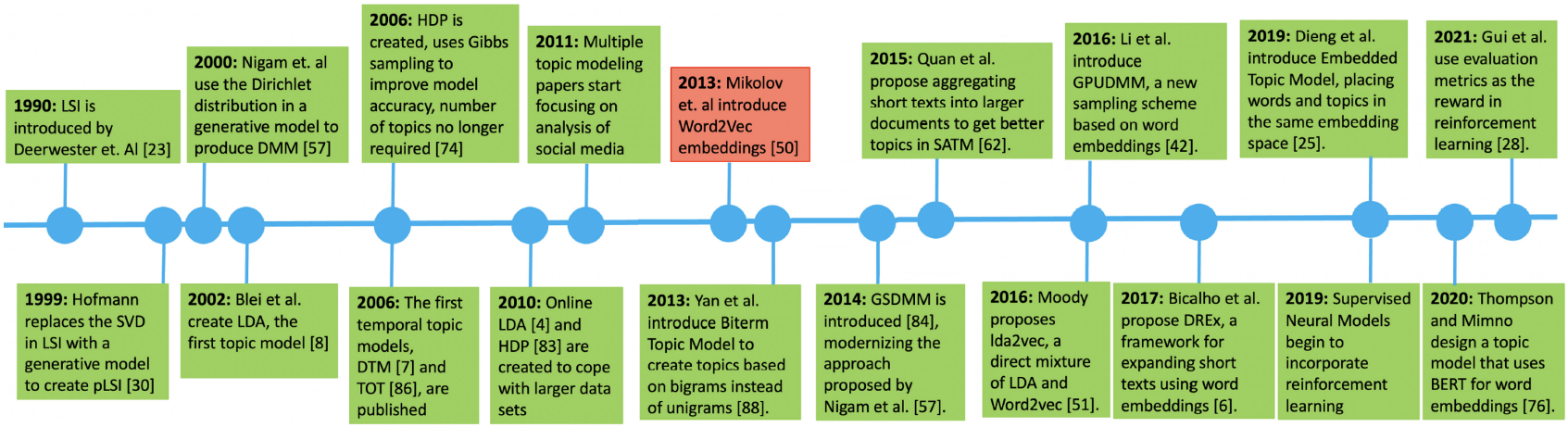
Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 27 апреля 2026

Содержание

Эволюция тематического моделирования



1999 PLSA — Probabilistic Latent Semantic Analysis

2001 LDA — Latent Dirichlet Allocation

200x мультимодальные, темпоральные, иерархические модели

2013 модели битермов и WNTM — аналоги word2vec

2016 тематические модели на основе предобученных word2vec

2020 BERTopic — TM на основе предобученного BERT

202x огромное разнообразие NTMs — Neural Topic Models...

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. 2022.

Нейросетевые и тематические языковые модели

Преимущества нейросетевых языковых моделей

- *генеративность:* способны порождать связный текст
- *универсальность:* решают широкий класс задач NLP/NLU
- *предобученность:* «знают всё о языке» (и о мире)

Преимущества вероятностных тематических моделей:

- *интерпретируемость* тематических эмбедингов
- *скорость* на узком классе задач «мягкой кластеризации»
- *полнота* тематической кластерной структуры коллекции

Как «объединить лучшее от двух миров»?

Что объединяет PTM и LLM, и что их разобщает:

⊕ обе — вероятностные языковые модели,

⊕ обе — автокодировщики, векторные представления текста

⊖ PTM: ~~мешок слов~~, архитектура матричного разложения, ~~байесовское обучение~~, трудности предобучения и др.

Открытые проблемы PTM и семейство моделей A*RTM

Цель: «Make Topic Modeling Great Again», а именно,
создать новый стандарт PTM, объединив всё лучшее от:

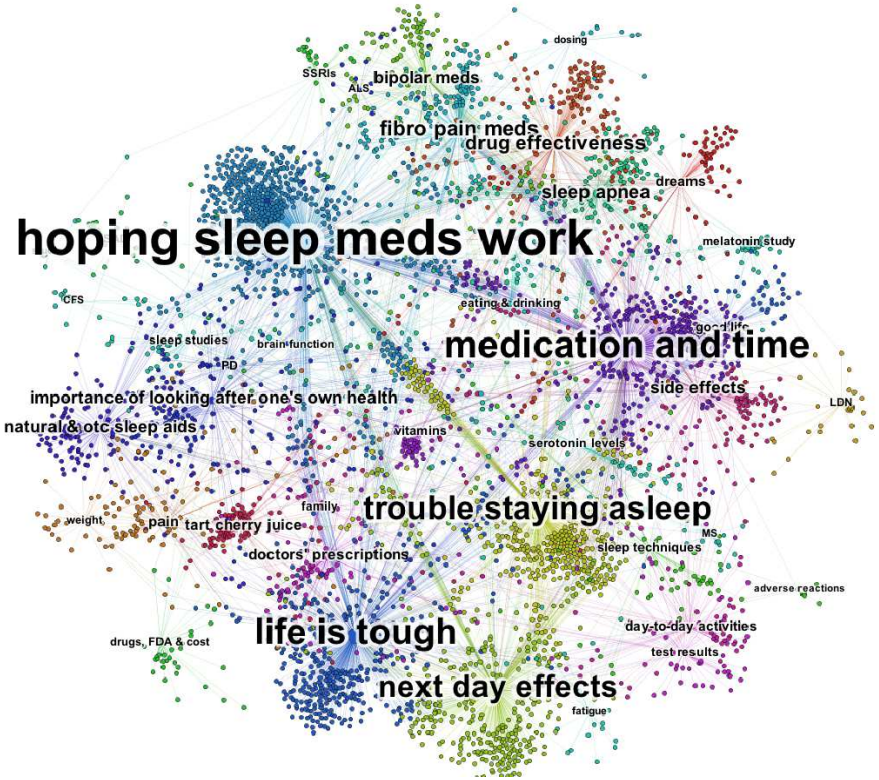
1. ARTM: регуляризация, модальности, иерархии, транзакции
2. BigARTM: батчи, параллельность, скорость, лёгкость
3. LLM: локальные контексты вместо «мешков слов»
4. LLM: модель внимания и её параметризация
5. NTM: каждая тема должна уметь «рассказать о себе»
6. NTM: согласование тем с предобученными эмбедингами
7. AutoML: настройка гиперпараметров, даже в потоке данных
8. МатСтат: тесты однородности и согласованности тем
9. решение проблем тематической несбалансированности, плохо интерпретируемых, дублирующих и мусорных тем

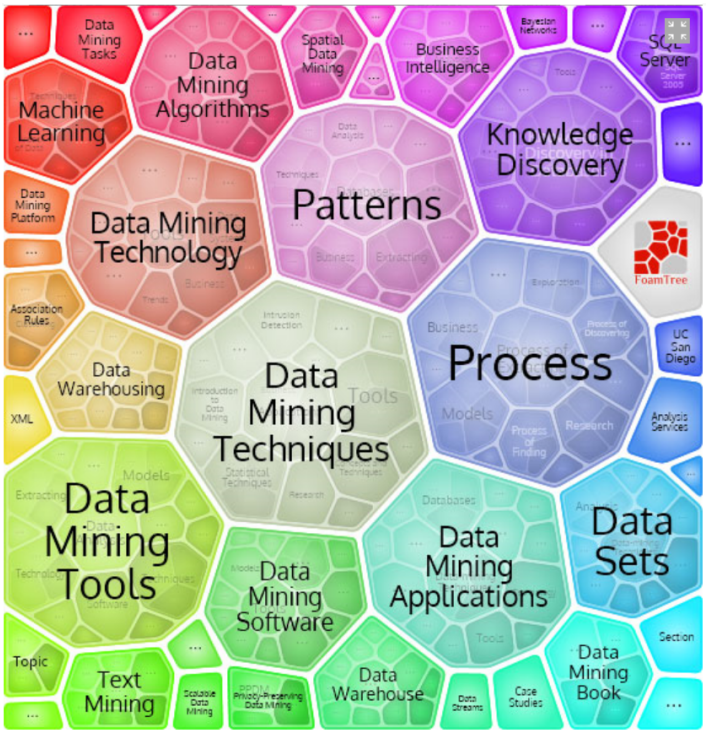
A*RTM означает: Attentive, Apprehensive, Aware, Adaptive, Automated, Available, etc... Additively Regularized TM

Автоматическое именование тем для визуализации

Пример 1: тематика обсуждений на www.PatientsLikeMe.com

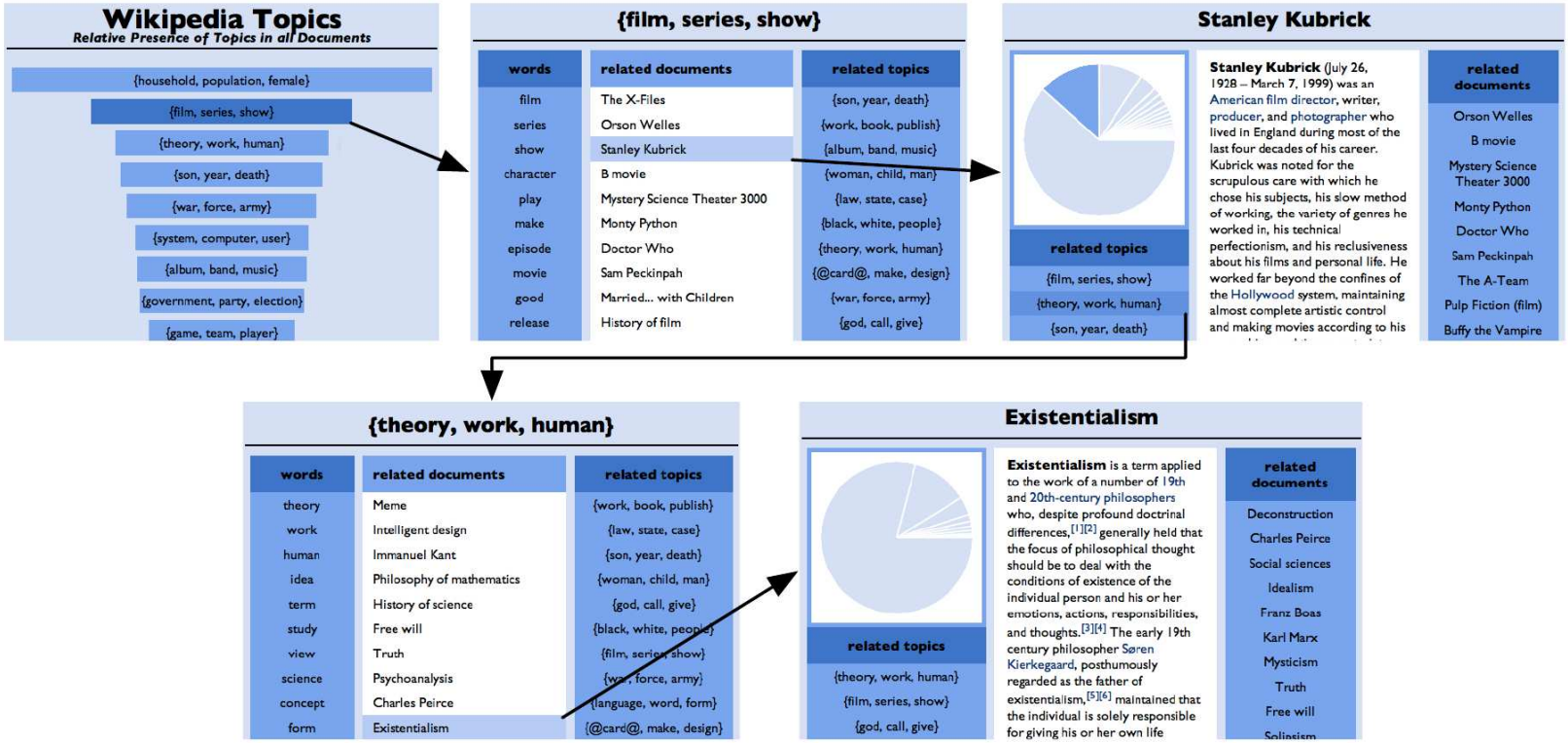
Пример 2: иерархическая карта *Data Mining*





Система TMVE — Topic Model Visualization Engine

Три топовых слова темы — слишком 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 $\ell_1, \ldots, \ell_m$
- построение (обучение) функции релевантности $\text{rel}(\ell, t)$
- 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

Отбираются вероятно специфичные для данной темы:

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

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:

$$\text{rel}(\ell, t) = \sum_{w \in \ell} \log \frac{p(w|t)}{p(w)} \to \max$$

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

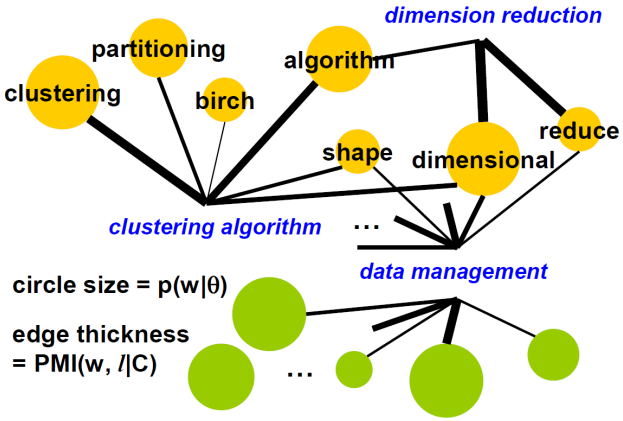
$$\text{rel}(\ell, t) = \sum_{w \in C} p(w|t) \underbrace{\log \frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\text{PMI}(w,\ell|C)} \to \max$$

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 название ℓ так, чтобы оно было

- максимально релевантно, $\text{rel}(\ell, t) \to \max$
- максимально не похоже на названия ℓ' остальных тем:

$$\text{rel}(\ell, t) + \lambda \min_{\ell'} \text{KL}(\ell' \| \ell) \to \max$$

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 $\text{rel}'(\ell, t)$:

- максимизирует релевантность своей темы, $\text{rel}(\ell, t) \to \max$
- минимизирует релевантность других тем, $\text{rel}(\ell, t') \to \min$

$$\text{rel}'(\ell, t) = \text{rel}(\ell, t) - \mu \sum_{t' \in T \setminus t} \text{rel}(\ell, t') \to \max$$

где параметр μ подбирается эмпирически.

Методика оценивания качества именования тем:

- 3 асессора, каждый асессор видит для каждой темы:
  — список топ-слов темы, список топ-документов темы
  — варианты названия, сгенерированные разными методами
- асессор ранжирует методы $0, 1, 2, \ldots$ (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 100 статей начиная с 2007 г.
- Важно при разработке приложений для пользователей
- Требования *полноты* и *неизбыточности* в иерархиях: названия дочерних тем должны акцентировать их различия, не повторяя родительскую тему
- Задача существенно облегчается с появлением LLM

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.

Задача суммаризации (реферирования, аннотирования) текста

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

Основные типы задач суммаризации:

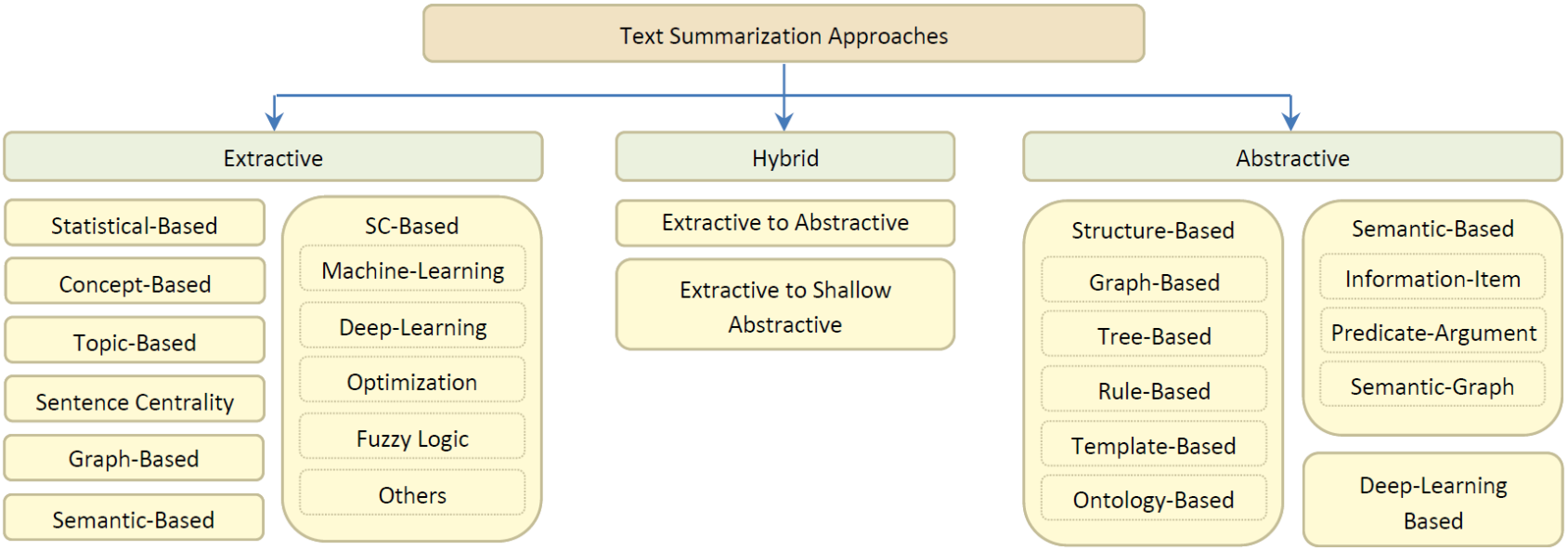
- *one-document* — на входе один документ $d \in D$
- *multi-document* — на входе набор документов $D' \subseteq D$
- ⊕ *topic* — на входе набор сегментов темы $p(d, s|t)$

Полуавтоматическая суммаризация

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

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

Абстрактивная суммаризация на основе трансформеров

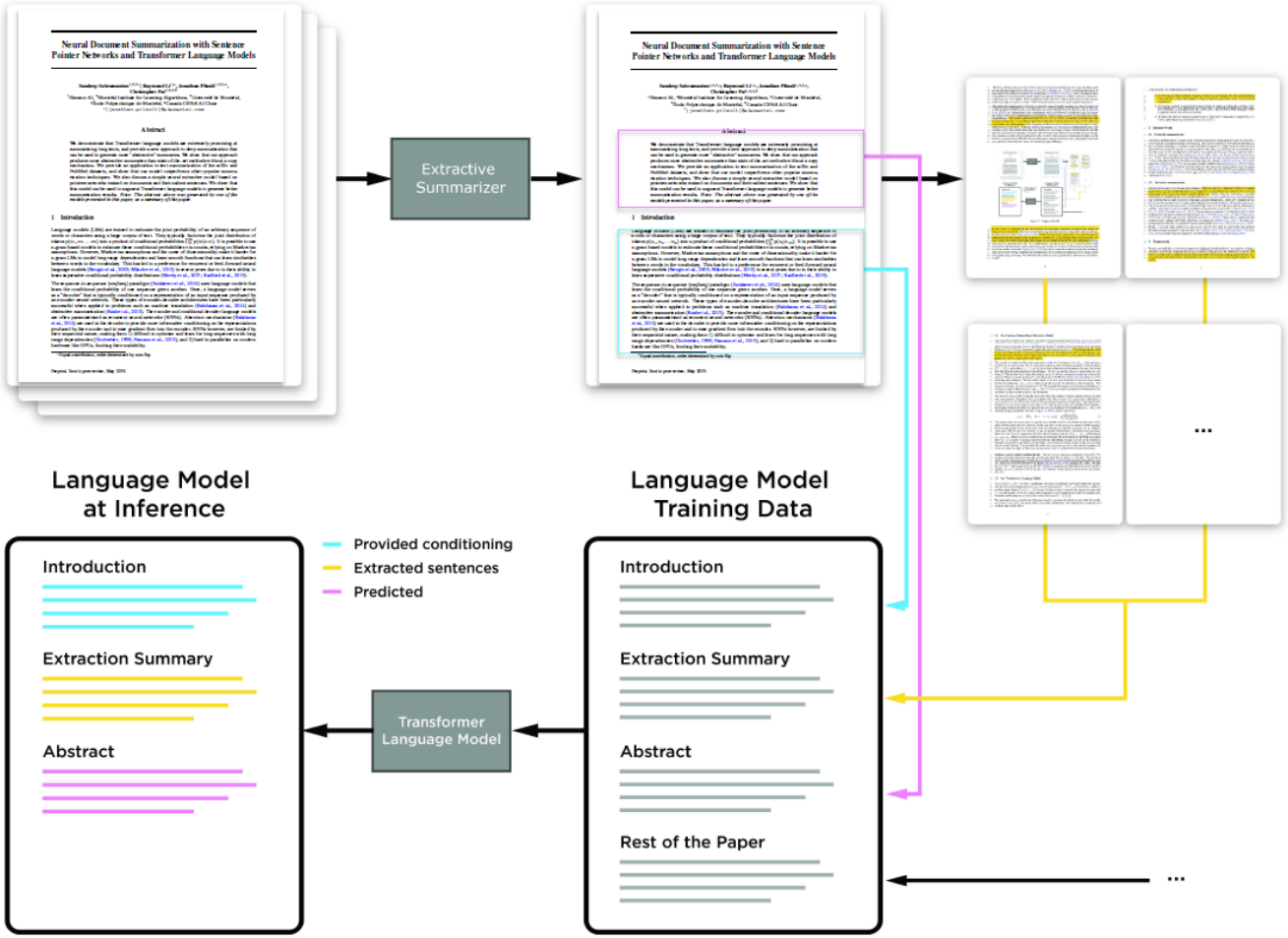
Первая в истории статья со сгенерированной аннотацией

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.

Задача суммаризации темы

S_d — множество предложений документа d

$p(t|s) = \sum_{w \in s} p(t|w)p(w|s)$ — тематика предложения $s \in S_d$

$\text{rel}(s,t) = \sum_{w \in s} \log \frac{p(w|t)}{p(w)}$ — оценка релевантности s теме t

1. отбор предложений: $p(t|s), \text{rel}(s,t) \to \max$, $s \in D_d$, $d \in D$
2. кластеризация предложений, чтобы сгруппировать похожие
3. выбор лучших предложений в каждом кластере
4. сортировка списка предложений в порядке упоминания (по времени публикации, по важности, и т.п.)
5. генерация суммаризации LLM по сортированному списку

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, Yihong Gong. Integrating clustering and multi-document summarization to improve document understanding. CIKM 2008

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:

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

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

Йенсен-Шеннон $\mathrm{JS}\big(p(w|s), p(w|R)\big)$ 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 по суммаризации

- Для визуализации тем нужны именование и суммаризация
- Тематические модели в экстрактивной суммаризации — для выделения и покрытия наиболее важных тем
- «Let the topics tell about themselves»
  Суммаризация темы — открытая проблема TM,
  до LLM её не только не решали, но даже и не ставили
- Задача существенно облегчается с появлением LLM
- ROUGE — семейство мер качества суммаризации,
  характеризуют далеко не все аспекты качества
- BLUE — аналогичные метрики, но precision-based
- Теперь и оценивать качество суммаризации лучше с LLM

Напоминание. Тематическая модель «мешка термов»

Дано: коллекция текстовых документов D, словарь W;
n_{dw} — частота терма $w \in W$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

Критерий: $\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ & \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases}$$

Напоминание. Тематическая модель локальных контекстов

Дано: последовательность $w_1, \ldots, w_n$ термов словаря W;
$C_i \subset \{1, \ldots, n\}$ — локальный контекст терма w_i, $i = 1, \ldots, n$;
α_{ci} — коэффициент внимания, вес терма w_c из C_i для w_i.

Найти: вер. языковую модель $p(w|C_i) = \sum\limits_{t \in T} \phi_{tw} \frac{p(w)}{p(t)} p(t|C_i)$
с параметрами $\phi_{tw} = p(t|w)$

Критерий: $\sum\limits_{i=1}^{n} \ln \sum\limits_{t \in T} \phi_{tw_i} \frac{p(w_i)}{p(t)} \sum\limits_{c \in C_i} \alpha_{ci} \phi_{tw_c} + R(\Phi) \to \max\limits_{\Phi}$

EM-алгоритм (после некоторых насильственных упрощений):

$$\text{E-шаг:} \begin{cases} p_{ti} = \underset{t \in T}{\text{norm}} \left(\frac{\phi_{tw_i}}{p(t)} \sum\limits_{c \in C_i} \alpha_{ci} \phi_{tw_c} \right), \quad p(t) = \sum\limits_{w \in W} \phi_{tw} p(w) \\ \phi_{tw} = \underset{t \in T}{\text{norm}} \left(n_{tw} + \phi_{tw} \frac{\partial R}{\partial \phi_{tw}} \right), \qquad n_{tw} = \sum\limits_{i=1}^{n} p_{ti} [w_i = w] \end{cases} \text{M-шаг:}$$

EM-алгоритм для модели Attentive ARTM

вход: коллекция, число тем $|T|$, параметры K, L, τ;
выход: матрица Φ, векторы $\{p_{ti}\colon t \in T, i = 1..n\}$;

инициализация ϕ_{tw}; $p(t) := 1/|T|$ для всех $w \in W$, $t \in T$;
для всех итераций $k = 1..K$ (проходов по всей коллекции)
  инициализация $n_{tw} := 0$ для всех $w \in W$, $t \in T$;
  для всех документов $d \in D$
    $p_{ti} := \phi_{tw_i}$ для всех $t \in T$, $i \in I_d := \{i\colon d_i = d\}$;
    для всех $l = 1..L$ (аналог L блоков внимания)
      $\theta_{ti} := \sum\limits_{c \in C_i} \alpha_{ci} p_{tc}$ для всех $t \in T$, $i \in I_d$;
      $p_{ti} := \underset{t \in T}{\text{norm}}\big(p_{ti}\theta_{ti}/p(t)\big)$ для всех $t \in T$, $i \in I_d$;
    $n_{tw_i} := n_{tw_i} + p_{ti}$; для всех $t \in T$, $i \in I_d$;
  $\phi_{tw} := \underset{t \in T}{\text{norm}}\left(n_{tw} + \phi^*_{tw} \frac{\partial R(\Phi^*)}{\partial \phi_{tw}}\Big|_{\phi^*_{tw} = \frac{n_{tw}}{n_w}}\right)$ для всех $t \in T$, $w \in W$;
  $p(t) := \sum\limits_{w \in W} \phi_{tw} p(w)$ для всех $t \in T$;

Быстрое вычисление взвешенных средних по контексту

Двунаправленные экспоненциальные скользящие средние, время линейное (вместо квадратичного) по длине контекста

функция $\theta_{ti} := \sum_{c \in C_i} \alpha_{ci} p_{tc}$ *для всех* $t \in T$, $i \in I_d$

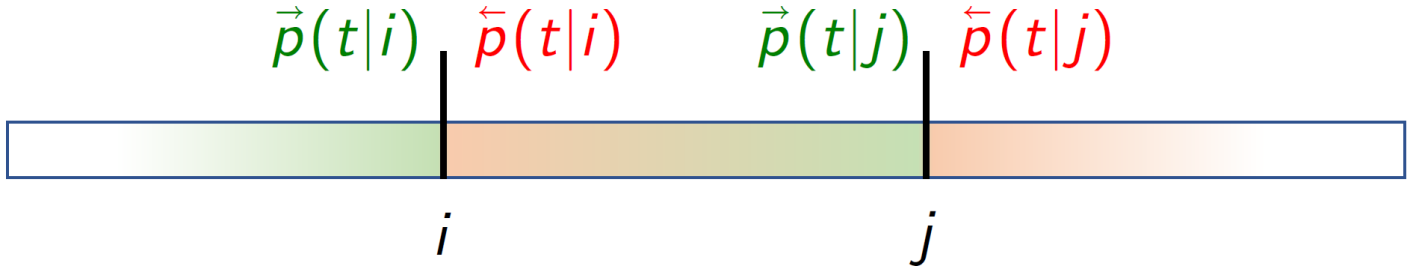
$$\vec{\theta}_i := \vec{\gamma}_i p_i + (1-\vec{\gamma}_i)\,\vec{\theta}_{i-1} \quad \text{для всех } i = 1..n_d, \quad \vec{\gamma}_1 = 1;$$
$$\overleftarrow{\theta}_i := \overleftarrow{\gamma}_i p_i + (1-\overleftarrow{\gamma}_i)\,\overleftarrow{\theta}_{i+1} \quad \text{для всех } i = 1..n_d, \quad \overleftarrow{\gamma}_{n_d} = 1;$$
$$\theta_i := \beta\vec{\theta}_i + (1-\beta)\overleftarrow{\theta}_i \quad \text{для всех } i = 1..n_d;$$

Основное свойство: если $\gamma_i = \gamma$, то $\alpha_{ci} = \gamma(1-\gamma)^{|i-c|}$

Как задать $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — коэффициенты сглаживания в позиции i:

- $\gamma_i \approx \frac{1}{h}$, где h — ширина окна, размер контекста
- $\gamma_i = 1$, если надо забыть контекст, сменить документ
- $\gamma_i = 0$, если надо проигнорировать терм w_i
- γ_i можно умножать на оценку важности терма w_i

Использование двунаправленных векторов контекста



Через *двунаправленные тематические векторы* определяется:

- $\vec{\theta}_{ti} = \vec{p}(t|i)$ — тематика левого контекста терма w_i
- $\overleftarrow{\theta}_{ti} = \overleftarrow{p}(t|i)$ — тематика правого контекста терма w_i
- $\theta_{ti} = \frac{1}{2}\left(\vec{\theta}_{ti} + \overleftarrow{\theta}_{ti}\right)$ — тематика двустороннего контекста w_i
- $p(t|i\,..\,j) = \frac{1}{2}\left(\overleftarrow{\theta}_{ti} + \vec{\theta}_{tj}\right)$ — тематика сегмента $[i\,..\,j]$
- $\overleftarrow{\theta}_{ti} \approx \vec{\theta}_{tj}$ — однородность тематики сегмента $[i\,..\,j]$
- $\max_i \left\|\vec{\theta}_{ti} - \overleftarrow{\theta}_{ti}\right\|$ — граница i между сегментами
- при различных γ_i — короткие и длинные контексты

Аналогия с моделями языка GCNN, Attention, Transformer

Модель внимания (self-attention) Query–Key–Value

Входные векторы слов (эмбединги)

$$X = (x_1, \ldots, x_n) \in \mathbb{R}^T$$

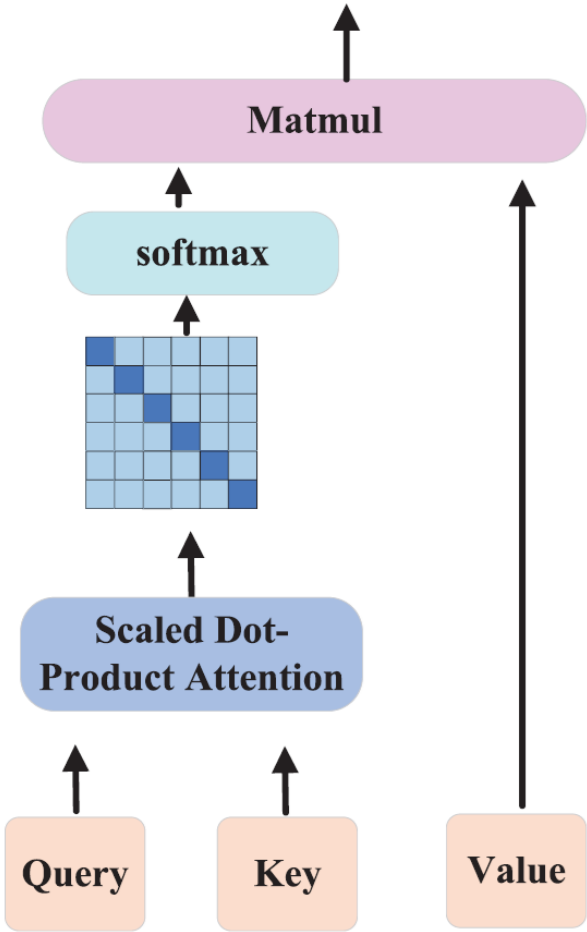
трансформируются в векторы слов, зависящие от контекстов C_i:

$$H = (h_1, \ldots, h_n) \in \mathbb{R}^d$$

Модель внимания (self-attention):

$$h_i = \sum_{c \in C_i} W_v x_c \operatorname*{SoftMax}_{c \in C_i} \langle W_k x_c, W_q x_i \rangle$$

$W_v x_c$ — вектор-значение (value)
$W_k x_c$ — вектор-ключ (key)
$W_q x_i$ — вектор-запрос (query)
W_q, W_k, W_v — обучаемые параметры



A. Vaswani et al. Attention is all you need. 2017.

Аналогия Attentive ARTM с моделью само-внимания

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\Big(\sum_{c\in C_i} \phi_{tw_c} \alpha_{ci} \frac{1}{p(t)} \phi_{tw_i} \Big)$$

Контекстный вектор на выходе модели само-внимания:

$$h_i = \sum_{c\in C_i} W_v x_c \, \alpha_{ci} = \sum_{c\in C_i} W_v x_c \, \underset{c\in C_i}{\text{SoftMax}}\langle W_k x_c, W_q x_i \rangle$$

Сходство:

- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов термов w_c из его контекста,
- наиболее схожих с ним по тематике

Отличия локализованного E-шага:

- адамарово умножение вектора ϕ_{w_c} на вектор-фильтр ϕ_{w_i}
- нет обучаемых матриц W_q, W_k, W_v как у модели внимания
- проецирование итогового вектора на единичный симплекс

Свёрточная нейросеть GCNN (Gated Convolutional Network)

Входные векторы слов (эмбединги)

$$X = (x_1, \ldots, x_n) \in \mathbb{R}^T$$

трансформируются в векторы слов, зависящие от контекстов C_i:

$$H = (h_1, \ldots, h_n) \in \mathbb{R}^d$$

через адамарово произведение:

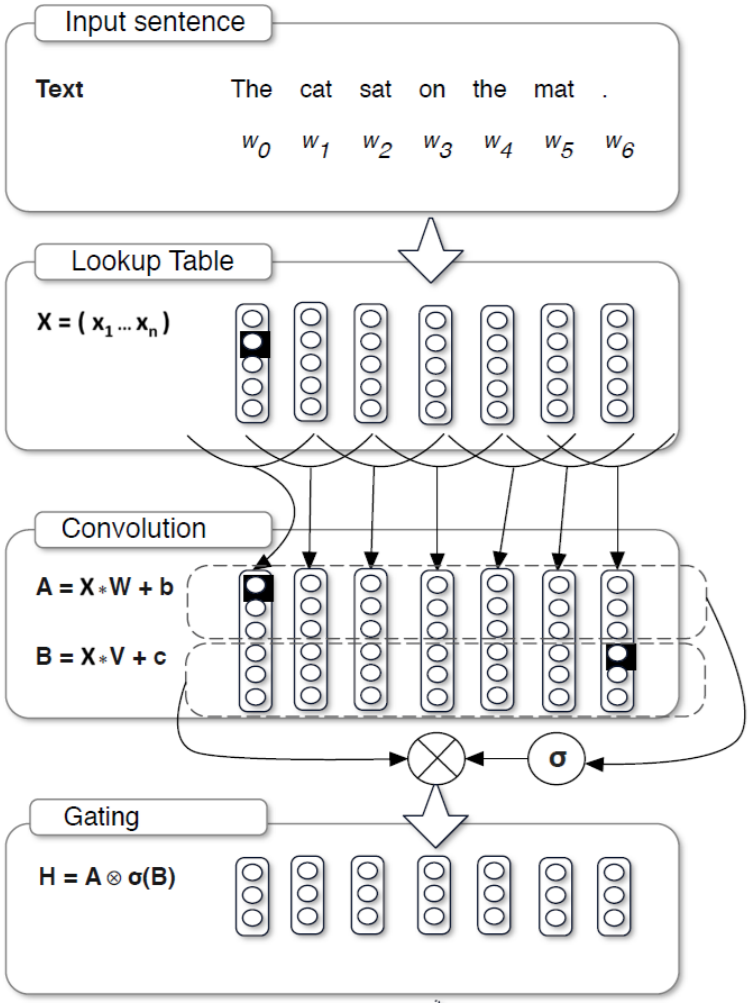
$$h_i = a_i \otimes \sigma(b_i), \quad \text{где}$$

$a_i = \sum_{c \in C_i} W_c x_c$ — свёртка-контекст,

$b_i = \sum_{c \in C_i} V_c x_c$ — свёртка-фильтр,

W_c, V_c — матрицы размера $d \times T$, обучаемые параметры модели,

$\sigma(x) = \frac{1}{1+e^{-x}}$ — функция сигмоида



Yann N. Dauphin et al. Language modeling with gated convolutional networks, 2017.

Аналогия Attentive ARTM с моделью GCNN

Контекстный тематический вектор на выходе E-шага:

$$p(t|C_i, w_i) \equiv p_{ti} = \underset{t\in T}{\text{norm}}\Big(\sum_{c\in C_i} \alpha_{ci} \phi_{tw_c} \frac{1}{p(t)} \phi_{tw_i} \Big)$$

Контекстный вектор на выходе модели GCNN:

$$h_i = \Big(\sum_{c\in C_i} W_c x_c \Big) \otimes \sigma \Big(\sum_{c\in C_i} V_c x_c \Big)$$

Сходство:

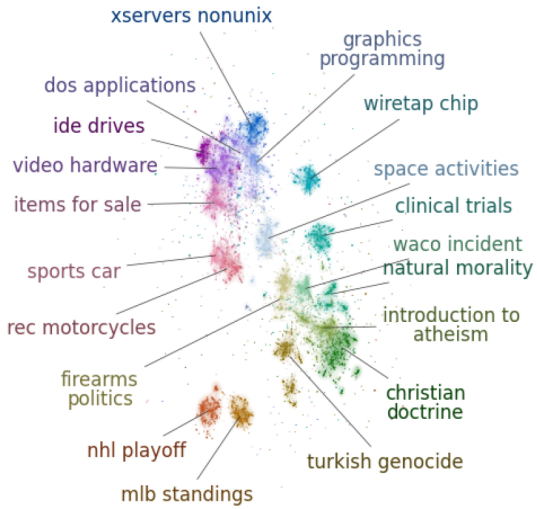
- вектор терма w_i трансформируется в контекстный вектор
- путём усреднения векторов ϕ_{w_c} его контекста,
- семантически схожих с вектором терма w_i, фильтруемых адамаровым умножением на неотрицательный вектор

Отличия локализованного E-шага:

- нет обучаемых матриц W_c, V_c как у модели GCNN
- вектор-фильтр ϕ_{w_i} без усреднения по контексту C_i
- проецирование итогового вектора на единичный симплекс

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо PTM — конвейер 8 технологий:



1. векторизация токенов (Sentence-BERT)
2. векторизация предложений скользящим окном в 50 токенов (mean pooling)
3. понижение размерности векторов (UMAP)
4. иерархическая кластеризация (hDbscan), автоматическое определение числа тем
5. иерархичекое укрупнение тем слиянием мелких кластеров с ближайшими соседями (Top2Vec)
6. разбиение документа на монотематические сегменты
7. $p(t|d)$ = доля векторов данной темы в документе
8. именование тем: поиск фраз, ближайших к центроиду темы

Dimo Angelov. Top2vec: Distributed representations of topics. 2020.
D.Angelov, D.Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

Нейросетевая тематическая модель Contextual-Top2Vec

Недостатки:

- это не единая модель, а конвейер эвристических моделей
- долго-дорого, особенно на больших коллекциях
- инкрементное добавление документов не предполагается

Достоинства — что хотелось бы перенять и встроить в ARTM:

- модель внимания, локальные контексты вместо документов
- отбор n-грамм и релевантных фраз по каждой теме
- именование и суммаризация тем на основе этих фраз
- инициализация тем по предобученным эмбедингам BERT, чтобы обеспечить качество тем даже на малых коллекциях
- автоматическое определение числа тем
- разбиение документа на монотематичные сегменты

Dimo Angelov. Top2vec: Distributed representations of topics. 2020.
D.Angelov, D.Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

Контекстная документная кластеризация CDC

n_{uw} — частота сочетания пары слов u, w в некотором окне

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

$H(w) = -\sum_u p(u|w) \log p(u|w)$ — энтропия контекста слова w

Узкий контекст — контекст с низкой энтропией, аналог темы, слова u, неслучайно часто встречающиеся рядом со словом w

Метод CDC — Contextual Document Clustering:

1. выделить «тематичные» слова с узкими контекстами
2. кластеризовать узкие контексты (найти кластеры-темы)
3. разбить документы на однородные сегменты (абзацы)
4. отнести каждый сегмент к ближайшей теме
5. $p(t|d)$ = доля сегментов темы t в документе

Vladimir Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. 2004.
D.Patterson, N.Rooney, V.Dobrynin, M.Galushka. SOPHIA: A novel approach for textual case-based reasoning. 2005.

Сравнение CDC [2004] и Contextual-Top2Vec [2024]

Сходство:

- основаны на общих лингвистических допущениях:
  — гипотеза дистрибутивной семантики
  — каждая фраза — это сообщение о взаимодействии двух агентов [Тён ван Дейк], значит, относится к 1–2 темам
  — фразы одного абзаца имеют общую основную тему
- тема — это кластер узких локальных контекстов
- документ разбивается на монотематичные сегменты
- $p(t|d)$ = доля сегментов данной темы в документе
- фразы отбираются для именования и суммаризации тем

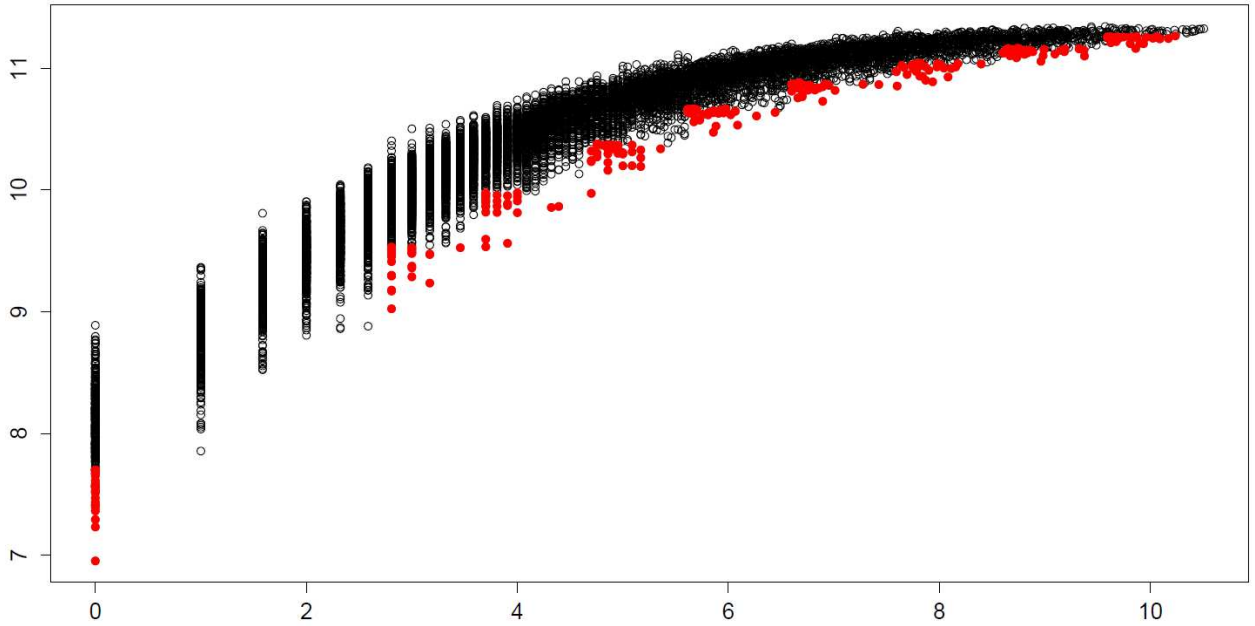
Отличия Contextual-Top2Vec:

- использование предобученных эмбедингов BERT
- есть этапы векторизации и понижения размерности
- нет отбора узких тематичных контекстов

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 $\log_2 N_w$



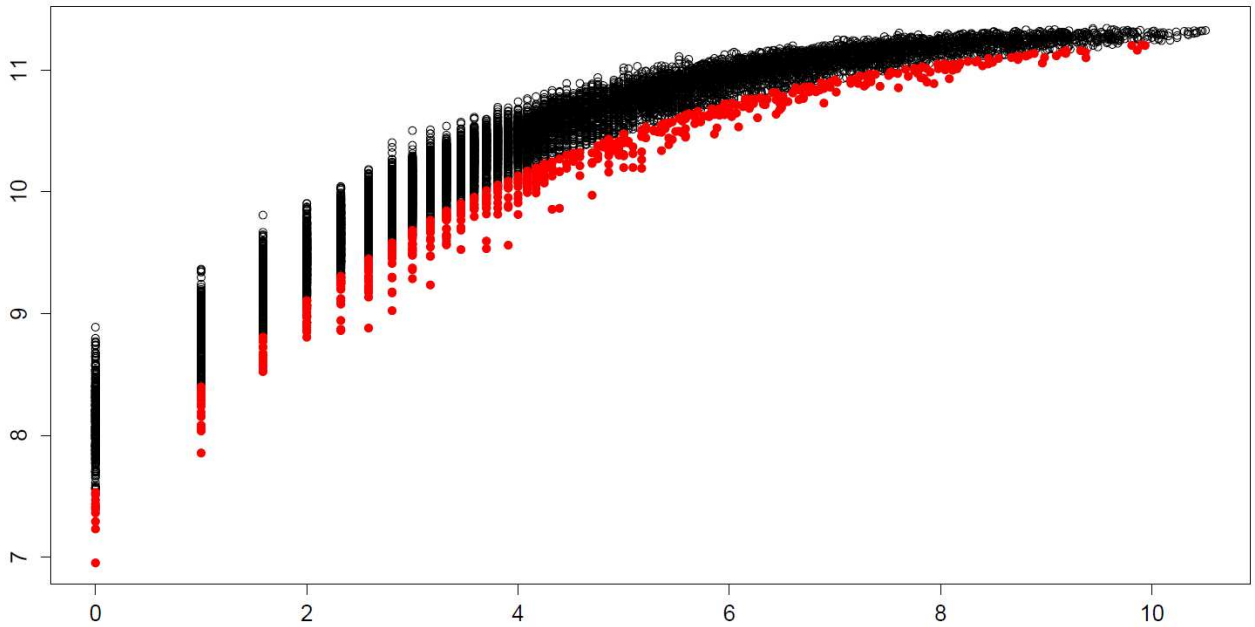
Недостаток: из-за разбиения на интервалы значительная часть узких контекстов пропускается (предвзятый отбор)

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

Выделение слов, имеющих узкие контексты

Закон Хипса $\Rightarrow$ $H(w)$ зависит от $\log_2 N_w$ логарифмически
Отбор локальных контекстов 5%-квантильной регрессией:

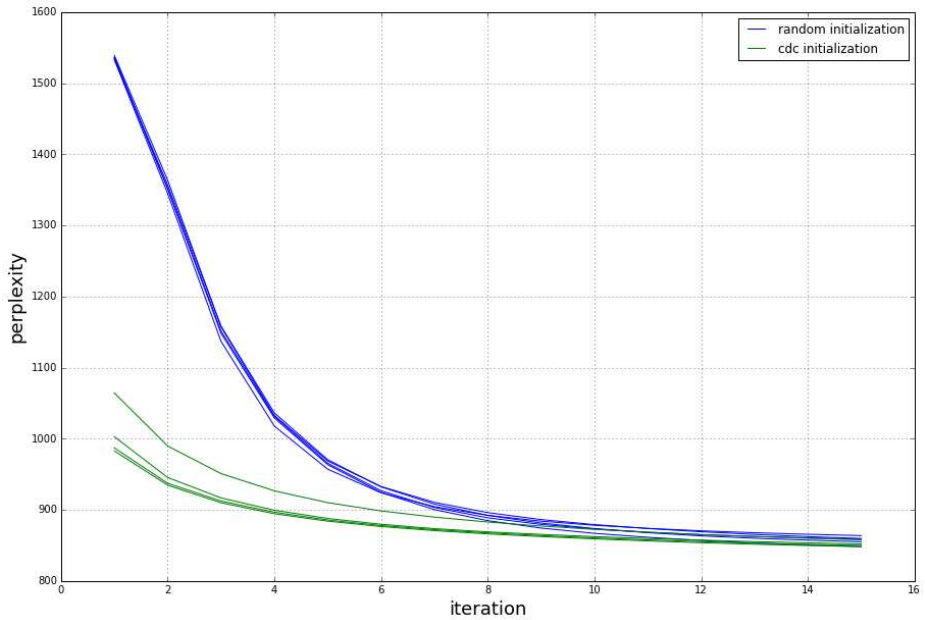
Зависимость энтропии $H(w)$ от документной частоты $\log_2 N_w$



V.Dobrynin, D.Patterson, N.Rooney. Contextual document clustering. ECIR, 2004.
Алексей Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей // ВКР бакалавра, МФТИ. 2015.

Действительно ли CDC — по сути тематическая модель?

Зависимость перплексии от числа итераций (коллекция MMPO) при инициализации тематической модели с помощью CDC:



- инициализация CDC ускоряет сходимость,
- но она не идеальна, и EM-алгоритм заметно улучшает её

Алексей Гринчук. Использование контекстной документной кластеризации для улучшения качества тематических моделей // ВКР бакалавра, МФТИ. 2015.

Задача кластеризации (clustering)

Дано: $X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка, $x_i \in \mathbb{R}^n$

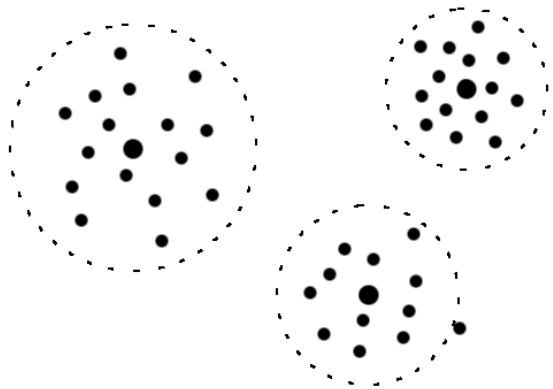
Найти:

— центры кластеров — параметры $\mu_a \in \mathbb{R}^n$, $a = 1, \ldots, K$

— какому кластеру принадлежит каждый объект $a_i \in \{1, \ldots, K\}$

Критерий: минимум суммы внутрикластерных расстояний

$$\sum_{i=1}^{\ell} \rho(x_i, \mu_{a_i})^2 \to \min_{\{a_i\}, \{\mu_a\}}$$



Метрика, как правило, евклидова (но может быть и другая):

$$\rho(x, \mu_a)^2 = \sum_{d=1}^{n} \big(f_d(x) - \mu_{ad}\big)^2$$

Метод K-средних (K-means) для кластеризации

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \rho(x_i, \mu_{a_i})^2 \to \min_{\{a_i\},\,\{\mu_a\}}, \quad \rho(x_i, \mu_a)^2 = \sum_{j=1}^{n} \big(f_j(x_i) - \mu_{aj}\big)^2$$

Алгоритм Ллойда (сильно упрощённый EM-алгоритм)

вход: X^ℓ, K; **выход:** центры μ_a, $a \in \{1, \ldots, K\}$;

$\mu_a :=$ начальное приближение центра, $a \in \{1, \ldots, K\}$;

повторять

отнести каждый x_i к ближайшему центру:

$a_i := \arg\min\limits_{a \in Y} \rho(x_i, \mu_a)$, $i = 1, \ldots, \ell$;

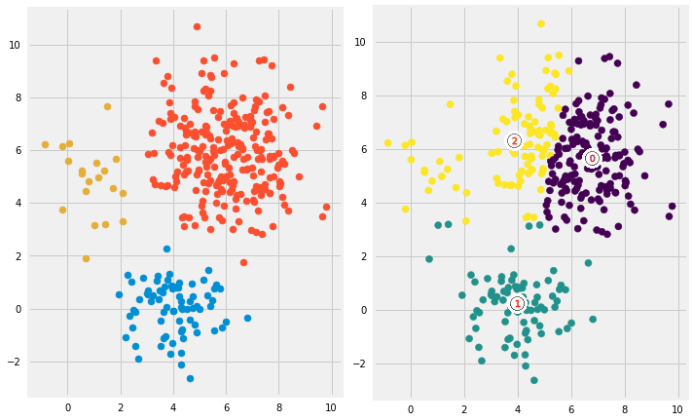
вычислить новые положения центров:

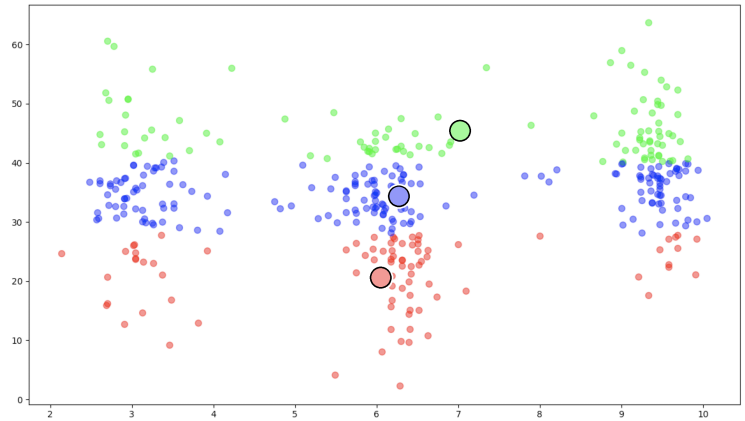
$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in \{1, \ldots, K\};$$

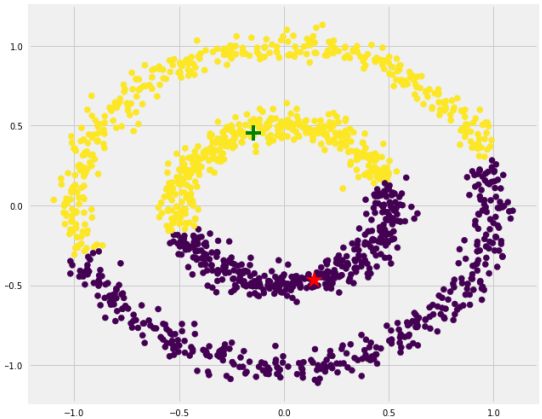
пока a_i не перестанут изменяться;

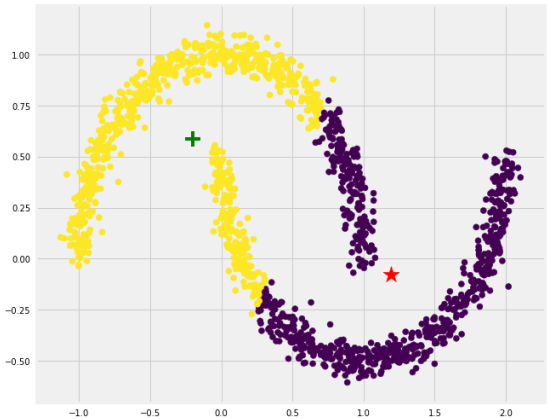
Примеры неудачной кластеризации k-means

Причина — неудачное начальное приближение или форма кластеров, существенно отличная от сферической









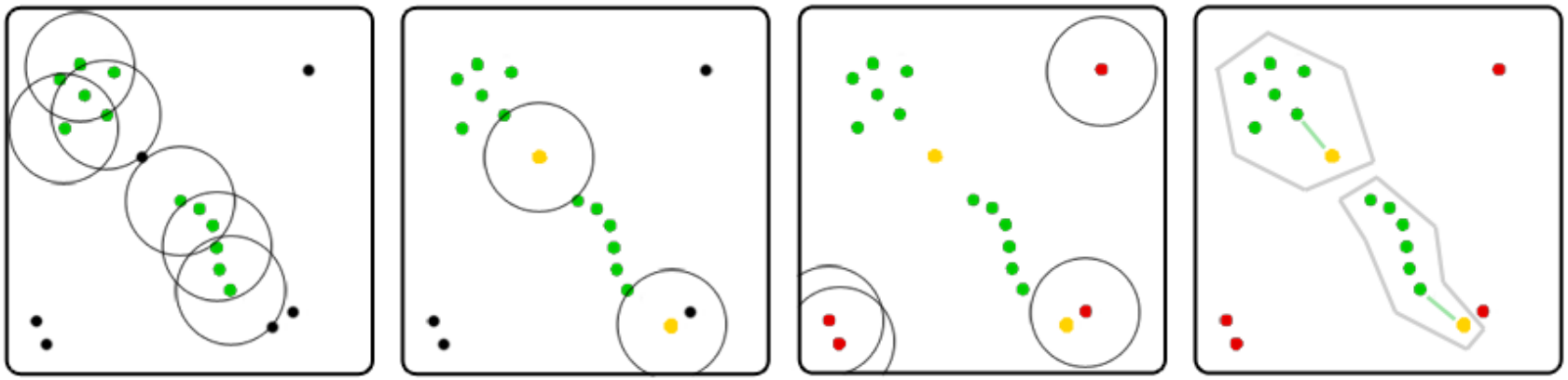
Алгоритм кластеризации DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)

Объект $x \in U$, его ε-окрестность $U_\varepsilon(x) = \{u \in U : \rho(x,u) \leqslant \varepsilon\}$

Каждый объект может быть одного из трёх типов:

- корневой: имеющий плотную окрестность, $|U_\varepsilon(x)| \geqslant m$
- граничный: не корневой, но в окрестности корневого
- шумовой (выброс): не корневой и не граничный



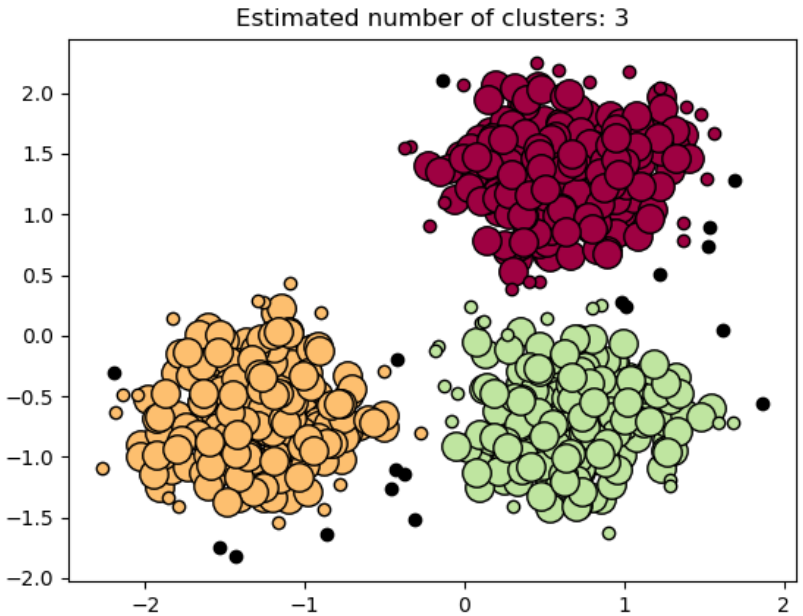
Ester, Kriegel, Sander, Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-1996.

Алгоритм кластеризации DBSCAN

вход: выборка $X^\ell = \{x_1, \ldots, x_\ell\}$; параметры ε и m;
выход: разбиение выборки на кластеры и шумовые выбросы;
$U := X^\ell$ — непомеченные; $a := 0$;
пока в выборке есть непомеченные точки, $U \neq \varnothing$:
  взять случайную точку $x \in U$;
  если $|U_\varepsilon(x)| < m$ **то**
    пометить x как, возможно, шумовой;
  иначе
    создать новый кластер: $K := U_\varepsilon(x)$; $a := a + 1$;
    для всех $x' \in K$, не помеченных или шумовых
      если $|U_\varepsilon(x')| \geqslant m$ **то** $K := K \cup U_\varepsilon(x')$;
      иначе пометить x' как граничный кластера K;
    $a_i := a$ для всех $x_i \in K$;
    $U := U \setminus K$;

Преимущества алгоритма DBSCAN

- быстрая кластеризация больших данных:
  $O(\ell^2)$ в худшем случае,
  $O(\ell \ln \ell)$ при эффективной реализации $U_\varepsilon(x)$;
- кластеры произвольной формы (долой центры!);
- деление объектов на корневые, граничные, шумовые.



Вероятностная модель векторного представления соседства

Stochastic Neighbor Embedding (SNE); развитие: tSNE, UMAP

Дано: исходные точки $x_i \in \mathbb{R}^n$, $i = 1, \ldots, \ell$

Найти: точки на карте-проекции $z_i \in \mathbb{R}^d$, $i = 1, \ldots, \ell$, $d \ll n$

Критерий: расстояния $\|z_i - z_j\|$ близки к исходным $\|x_i - x_j\|$

Вероятностная модель события «j является соседом i»
на основе перенормированных гауссовских распределений:

$p(j|i) = \underset{j \neq i}{\text{norm}} \exp\left(-\frac{1}{2\sigma_i^2}\|x_i - x_j\|^2\right)$ — в исходном пространстве;

$q(j|i) = \underset{j \neq i}{\text{norm}} \exp\left(-\|z_i - z_j\|^2\right)$ — в пространстве проекции;

где $p(j) = \underset{j}{\text{norm}}(z_j) = \frac{z_j}{\sum_k z_k}$ — операция нормировки вектора.

Максимизация правдоподобия (стохастическим градиентом):

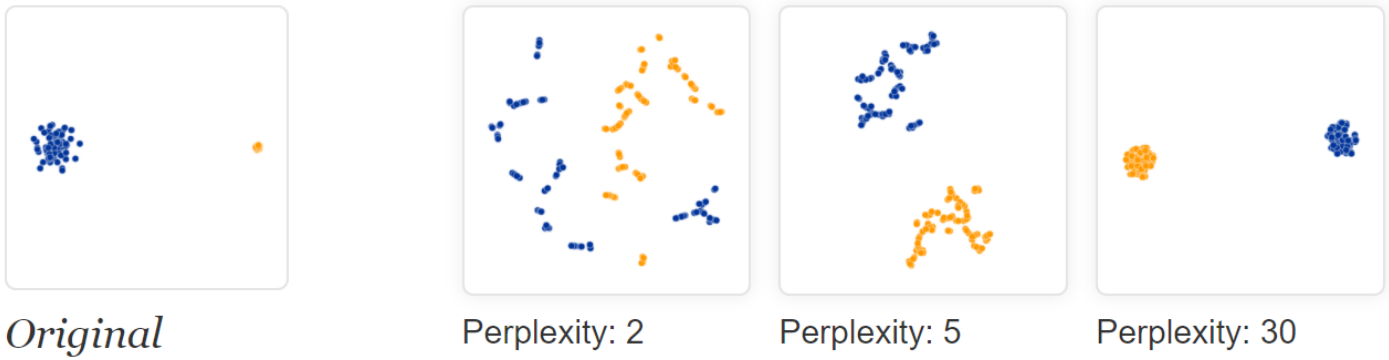
$$\sum_i \sum_{j \neq i} p(j|i) \ln q(j|i) \;\to\; \max_{\{z_i\}}$$

Преимущества метода SNE

- Преобразование расстояний в вероятности устраняет дисбалансы между большими и малыми расстояниями
- Дисбаланс между точками с большой и малой плотностью соседей выравнивается настройкой σ_i по перплексии

$H(i) = -\sum_j p(j|i) \log_2 p(j|i)$ — энтропия распределения $p(j|i)$;
$2^{H(i)}$ это *перплексия* = «эффективное число соседей у x_i»
(если $p(j|i) = \frac{1}{k}$, то $2^{H(i)} = k$); обычно перплексия = 5..50.

Выбор перплексии может существенно влиять на вид проекции:

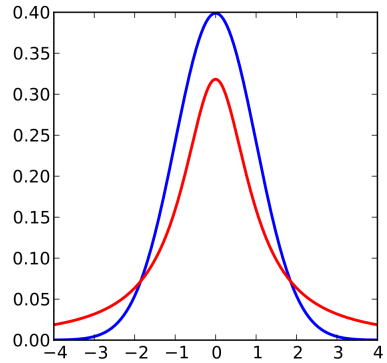


G.E.Hinton, S.T.Roweis. Stochastic Neighbor Embedding. 2002.

Вероятностная модель t-SNE: два усовершенствования SNE

Проблема скученности в SNE: окрестность вмещает гораздо больше точек в n-мерном пространстве, чем в d-мерном

- Использование t-распределения Стьюдента с более тяжёлым хвостом и симметричного совместного распределения $q(i,j)$:



$$q(i,j) = \underset{(i,j):\, i \neq j}{\operatorname{norm}} \left(1 + \|z_i - z_j\|^2\right)^{-1}$$

- Использование совместного распределения $p(i,j)$:

$$p(i,j) = \tfrac{1}{2\ell}\left(p(j|i) + p(i|j)\right)$$

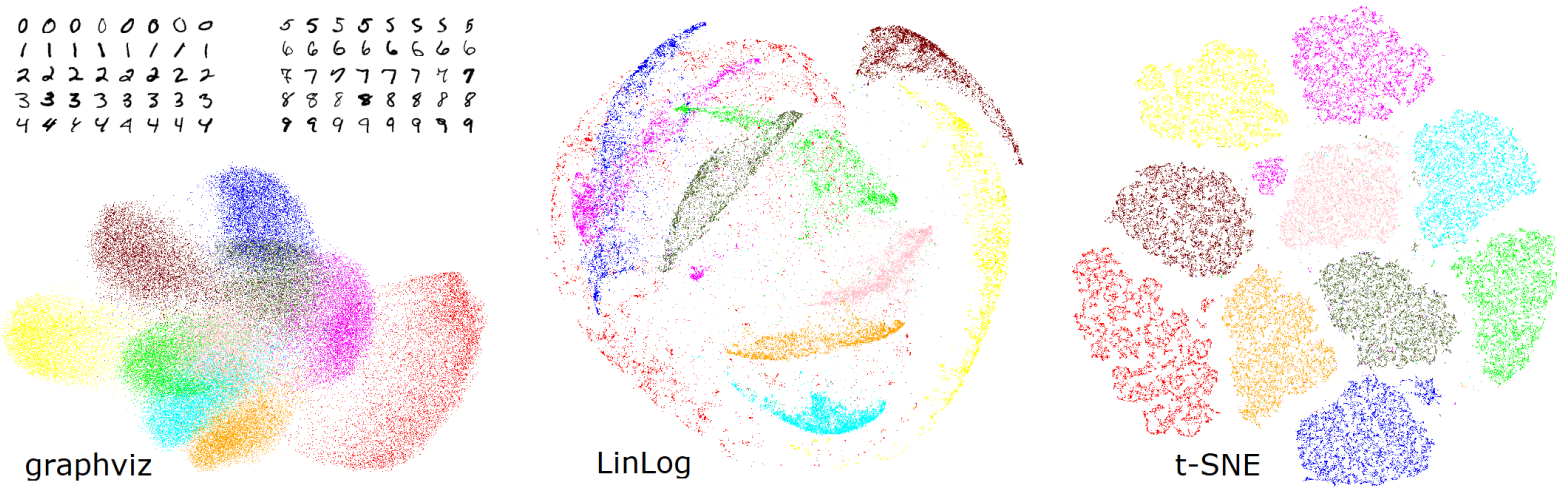
Максимизация правдоподобия (стохастическим градиентом):

$$\sum_{(i,j):\, j \neq i} p(i,j) \ln q(i,j) \to \max_{\{z_i\}}$$

L.J.P. van der Maaten, G.Hinton. Visualizing data using t-SNE. 2008

Преимущества и недостатки t-SNE

Лучшее представление структур сходства по сравнению с другими методами многомерного шкалирования (mnist)



Ложные кластерные структуры при низкой перплексии

Размеры кластеров и расстояния между ними неинформативны

Трудно отличить реальные структуры от артефактов метода

Нет ясного критерия качества для подбора перплексии

M.Wattenberg, F.Viegas, I.Johnson (Google). How to use t-SNE effectively. 2016. https://distill.pub/2016/misread-tsne

Резюме

Наша цель — «Make Topic Modeling Great Again», а именно, создать новый стандарт тематического моделирования:

1. от байесовского обучения — к аддитивной регуляризации
2. от мешка слов — к локальным контекстам
3. от BigARTM — к новой эффективной библиотеке ARTM
4. интерпретируемость тем при дисбалансе тем в коллекции

Что хорошего сохранить от BigARTM:

1. гибкость: регуляризации, модальности, иерархии, транзакции
2. технологичность: батчи, параллельность, скорость, метрики

Что нового привнести в ARTM от Attention и LLM:

1. встроенная токенизация, включая выделение терминов
2. модель тематического внимания, затем её параметризация
3. каждая тема должна уметь «рассказать о себе»
4. пред-обучение тем по эмбедингам из LLM

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3*. Творческое задание (возможны разные решения).
Предложите модель, разделяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксический анализ не нужен;
  зависимость качества ATE от объёма коллекции
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Теоретическое задание к лекции 3

Выведете EM-алгоритм для тематической языковой модели:

5. $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя в качестве исходных данных последовательность $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}.
Докажите эквивалентность обычному EM-алгоритму ARTM.

6. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$, $\theta_{td} = p(t|d)$ — параметры модели.

7. $p(w|d) = \sum_t \phi_{tw}\frac{p(w)}{p(t)}\theta_{td}$, где $p(t)$ фиксировано,
$\phi_{tw} = p(t|w)$ — параметры модели, $\theta_{td} = \sum_w \frac{n_{dw}}{n_d}\phi_{tw}$.

8*. Фиксация $p(t)$ как внешнего параметра упрощает выкладки, но может нарушать условия целостности модели:

$$p(t) = \sum_t \phi_{tw}p(w), \qquad p(t) = \sum_t \theta_{td}p(d).$$

Как обеспечить выполнение этих условий в EM-алгоритме?

Теоретическое задание к лекции 4

9. Докажите, что необходимым условием максимума

$$\sum_{i=1}^{n} \ln \sum_{t \in T} p(w_i, t|i, \Omega) \;\to\; \max_{\Omega}$$

для языковой модели со скрытыми переменными $t \in T$ (не обязательно темами) и параметрами $\Omega = (\omega_{kj})$ — набором неотрицательных нормированных векторов, является система

$$\text{E-шаг:} \quad p(t|w_i, i) = \operatorname*{norm}_{t \in T} p(w_i, t|i, \Omega)$$

$$\text{M-шаг:} \quad \omega_{kj} = \operatorname*{norm}_{k} \left(\sum_{i=1}^{n} \sum_{t \in T} p(t|w_i, i)\, \omega_{kj} \frac{\partial}{\partial \omega_{kj}} \ln p(w_i, t|i, \Omega) \right)$$

10. Выведете отсюда EM-алгоритм для частных случаев:

1) $p(w, t|i, \Omega) = \phi_{wt}\theta_{td_i}$;
2) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{w \in d_i} \frac{n_{d_i w}}{n_{d_i}} \phi_{tw}$;
3) $p(w, t|i, \Omega) = \phi_{tw} \frac{p(w)}{p(t)} \sum_{c \in C_i} \alpha_{ci} \phi_{tw_c}$.

11. Творческое задание.** Предложите способ ввести обучаемые параметры в тематическую модель внимания.

Исследовательская задача к лекции 4

Реализуйте EM-алгоритм для модели локального контекста (или воспользуйтесь готовой реализацией)

Исследуйте зависимость метрик качества модели

- перплексия: $\mathscr{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} p(w|C_i)\right)$
- разреженность, различность, когерентность тем
- дефекты целостности модели:
  $\left\|p(t) - \frac{n_t}{n}\right\|$, $\left\|p(t) - \sum_t \phi_{tw} p(w)\right\|$, $\left\|p(t) - \sum_t \theta_{td} p(d)\right\|$

от номера итерации и от параметров модели:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 5

12. Найдите дискретное распределение $P = (p_i)_{i=1}^n$ в задаче $\sum_i n_i \mu(p_i) \to \max$ с гладкой монотонно возрастающей $\mu(p)$.
Отдельно рассмотрите случаи $\mu(p) = p^s$, $s = 1$, $s \to 0$.

13. Выведете EM-алгоритм в случае, когда ln заменён гладкой монотонно возрастающей функцией μ:

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

Подумайте, какие замены логарифма полезны, и почему.

14. Простейшая идея разреживания — обнуление малых вероятностей. Чтобы обосновать эту эвристику, найдите, какому регуляризатору соответствует формула M-шага

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

Подсказка: с учётом подстановки несмещённой оценки ϕ^*_{wt}

Теоретическое задание к лекции 5

Проект «Тематизатор». Аналитик построил модель $\Phi^0\Theta^0$ 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 $t \in T_-$.

15. Предложите регуляризаторы для этого.

16. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

17. Предложите способ инициализации Φ для новой модели.

Исследовательская задача к лекции 5

Продолжение исследования по автоматическому выделению научных терминов (Automatic Term Extraction, ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  — оптимальную стратегию регуляризации на основе декоррелирования и сглаживания фоновых тем
  — рекомендации по управлению относительными коэффициентами регуляризации
  — критерий тематичности терминов по расстоянию между распределениями $p(t|w)$ и $p_0(t) = \frac{1}{|T|}$, позволяющий наиболее чётко отличать термины от фоновой лексики
- Критерий:
  — максимум доли терминов в предметных темах
  — минимум доли терминов в фоновых темах

Исследовательская задача к лекции 5

Продолжение исследования модели локального контекста (можно воспользоваться готовой реализацией EM-алгоритма)

Исследуйте устойчивость модели в сравнении с ARTM

- без регуляризации
- с регуляризатором декоррелирования, при различных значениях относительного коэффициента регуляризации

Как на устойчивость модели влияют её параметры:

- $|T|$ — число тем
- L — число проходов
- τ — вес N_{tw} в формуле M-шага, особый случай $\tau = 0$
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$, β — баланс левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, секций
- опция «$i \in C_i$ или $i \notin C_i$»

Теоретическое задание к лекции 6

18. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, 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 М-шага для матрицы Ψ.

19. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

20. 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 6

Проект «Мастерская знаний». Нужна тематическая модель подборок научных статей и/или поисковой выдачи.

Дано:

- 1000 подборок, в каждой по 1000 аннотаций научных статей, ранжированные по сходству с аннотацией-запросом по эмбедингам модели SciRus (эмбединги тоже даны)

Найти:

- метод согласования тематической модели с эмбедингами
- метод выделения терминов (Automatic Term Extraction)
- метод отбора терминов по тематичности
- метод отсева тематически нерелевантных аннотаций

Критерии:

- согласованность тематической модели с эмбедингами
- интерпретируемость тем
- качество выделения терминов

Теоретическое задание к лекции 7

21. Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t — константы (внешние параметры, не зависящие от Φ, Θ).
Докажите, что подстановка этого регуляризатора в M-шаг эквивалентна введению мультипликативной поправки $(1 + \tau\beta_{dw})$ в критерий log-правдоподобия.

22.** Выведите EM-алгоритм с регуляризатором семантической однородности, предполагая, что n_{tdw} и n_t выражаются через параметры модели Φ, Θ.

23*. Предложите формулу средневзешенных статистик S_* для тематической модели локальных контекстов.
Проверьте, что полученная формула совпадает с введённой на лекции, если контекстом является весь документ.

Исследовательская задача к лекции 7

Исследование EM-алгоритма для модели локального контекста

- Оценивание внутритекстовой когерентности
  - реализуйте вычисление средневзвешенной когерентности
  - подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
  - какие эвристики в модели локального контекста улучшают внутритекстовую когерентность?
  - воспроизводимо ли это улучшение на разных коллекциях?
- Оценивание средневзвешенных статистик
  - реализуйте вычисление S_t, S_{wt}
  - как зависит вид распределения $\{S_t\}$ от числа тем?
  - есть ли корреляция между S_t и когерентностью coh_t?
  - предложите способ разделения темы с большим S_t на подтемы и их инициализацию термами с большими S_{wt}
- Оценивание несбалансированности тем
  - реализуйте генератор коллекций с заданным дисбалансом тем
  - как дисбаланс влияет на число разделённых и слитых тем?
  - модели локального контекста лишены этой проблемы?
  - уменьшает ли регуляризатор семантической однородности число разделённых и слитых тем?

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (для хронокарты науки)

- Википедия
- Викиновости (1.5M статей, проект закрыт 30/03/2026)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «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: кол./кач. анализ предметной области,
  — реализация данного сценария как модуля в среде Orange

- «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: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys