

# Вероятностные тематические модели

## Лекция 10.

### 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
- Эволюция тематического моделирования

## Задача суммаризации (реферирования, аннотирования) текста

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

*Полуавтоматическая* — MAHS, machine aided human 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  $p(d, s|t)$

Основные подходы к суммаризации:

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

---

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

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

## Основные этапы выборочной (extractive) суммаризации

- 1 Внутреннее представление текста
  - 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:

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

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

$S_c^{\text{out}} \subset S$  — 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  $\text{sim}$  — 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:

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

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

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 эллипсом

---

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  $s$ .

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

---

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

## BSTM — Bayesian Sentence-based Topic Models

**Критерий** максимума регуляризованного правдоподобия:

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

- Авторы утверждают, что модель переходит в обычную  $p(w|d) = \sum_t \phi_{wt} \theta_{td}$ , если предложение  $\equiv$  слово
- Это не так, 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 это не будет работать для multi-document summarization!
- А то, что модель «Bayesian», вообще не имеет значения ;)

---

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

## Идея обобщения для много-документной суммаризации

**Критерий** максимума регуляризованного правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

## ROUGE: Recall-Oriented Understudy for Gisting Evaluation

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

$s$  — суммаризация, построенная системой

Чем больше, тем лучше — для всех метрик семейства ROUGE

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

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

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

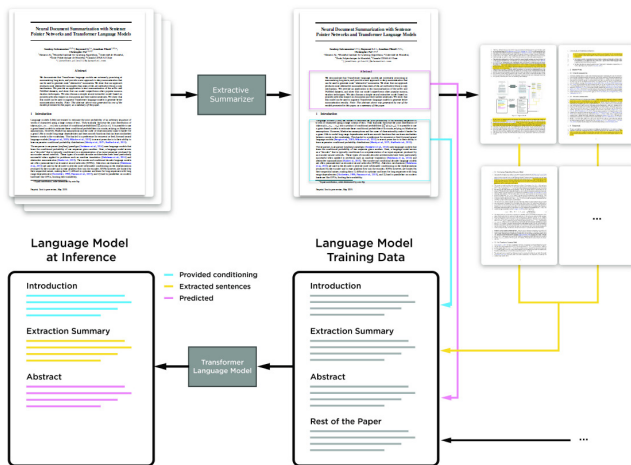
### Abstract

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.



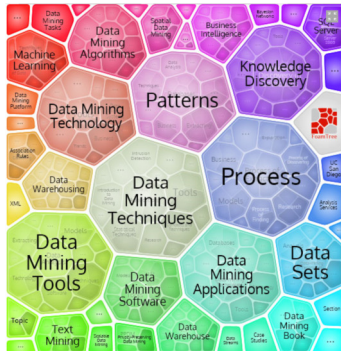
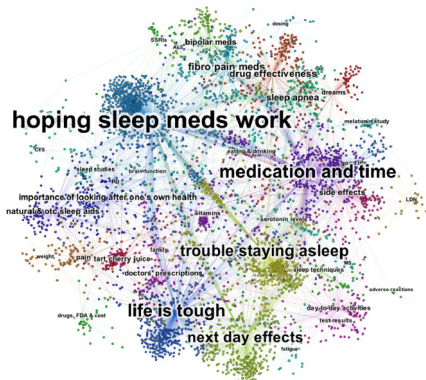
## Резюме по суммаризации

- 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 именоване тем

## Автоматическое именовании тем для визуализации

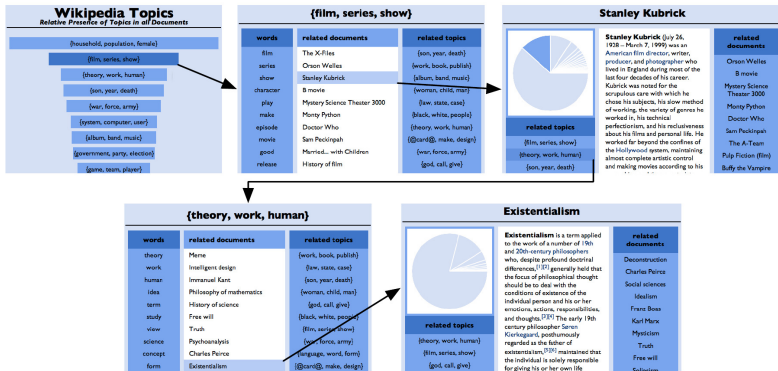
**Пример 1:** тематика обсуждений на [www.PatientsLikeMe.com](http://www.PatientsLikeMe.com)

**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 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 (Цяо Чжу 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  $\ell$ :

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

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

---

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

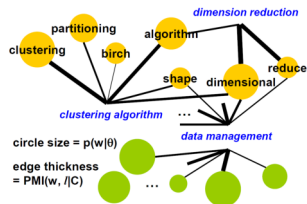
## Проблема названий, подходящих для нескольких тем

**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  $\ell'$  остальных тем:

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

где параметр  $\lambda$  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  $\mu$  подбирается эмпирически.

### Методика оценивания качества именования 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
$\theta$	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	<b>1.49</b>
SIGMOD	5	0.36	1.15	<b>1.51</b>
AP	1	0.97	0.99	<b>1.02</b>
AP	5	0.85	0.66	<b>1.48</b>

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

## 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.

## Что точно работает в тематическом моделировании

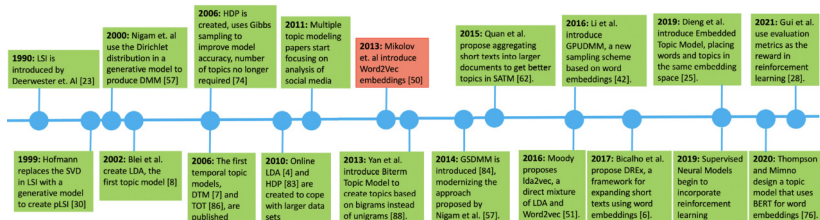
...или «узнав об этом, по другому уже не захочется»

- 1 многокритериальная регуляризация, ARTM и BigARTM
- 2 многокритериальное оценивание
- 3  $n$ -граммы
- 4 декоррелирование
- 5 модальности
- 6 однопроходный EM-алгоритм
- 7 тематические иерархии
- 8 спектр тем
- 9 реализация ARTM на PyTorch
- 10 установка BigARTM под Python до 3.11 включительно

## Открытые задачи тематического моделирования

- 1 ТЕМАТИЗАТОР =)
- 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

## 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)

*Rob Churchill, Lisa Singh.* The Evolution of Topic Modeling. November, 2022.  
*He Zhao et al.* Topic Modelling Meets Deep Neural Networks: A Survey. 2021.