

# Автоматическое выделение терминов для тематического моделирования

Никитина Мария

Московский физико-технический институт  
Факультет прикладной математики и информатики  
Кафедра интеллектуальных систем

Консультант: Потапова Полина  
Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович

4 мая 2023 года

- Создание алгоритма автоматического выделения терминов в коллекции документов.
- Сравнение результатов работы вероятностной тематической модели и нейронной сети.

## Вероятности

$p_{\omega d}$  – вероятность появления термина  $\omega$  в документе  $d$

$\phi_{\omega t}$  – вероятность того, что терм  $\omega$  относится к теме  $t$

$\theta_{td}$  – вероятность встречи темы  $t$  в документе  $d$

## Матрицы

$P = (p_{\omega d})_{W \times D}$  – матрица частот термов в документах

$\Phi = (\phi_{\omega t})_{W \times T}$  – матрица термов тем

$\Theta = (\theta_{td})_{T \times D}$  – матрица тем документов

## Множества

$W, D, T$  – множества всех термов, документов и тем соответственно

# Постановка задачи

## Задача

Найти разложение  $P = \Phi\Theta$  с помощью максимизации логарифма правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{wd} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

## Проблема

В общем случае решение задачи не единственно, поэтому задача считается некорректно поставленной по Адамару.

## Решение

Добавление регуляризаторов  $R_i(\Phi, \Theta)$  с неотрицательными коэффициентами регуляризации  $\tau_i$ ,  $i = 1, \dots, k$ . Новая задача:

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

$$R(\Phi, \Theta) = \sum_{i=1}^k \tau_i R_i(\Phi, \Theta)$$

## Декоррелирование

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{\omega \in W} \phi_{\omega t} \phi_{\omega s}$$

Отвечает за степень различия между темами

## Сглаживание-разреживание

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{\omega \in W} \beta_{\omega t} \ln \phi_{\omega t} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$$

Сглаживание применяется для фоновых тем, куда собираются слова, не имеющие определённой темы. Разреживание – для предметных тем.

# Критерии качества

$TP$  – истинно-положительное решение

$FP$  – ложноположительное решение

$FN$  – ложноотрицательное решение

Precision – точность

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall – полнота

$$\text{Recall} = \frac{TP}{TP+FN}$$

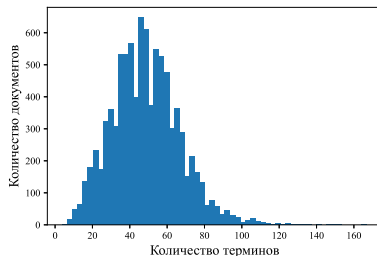
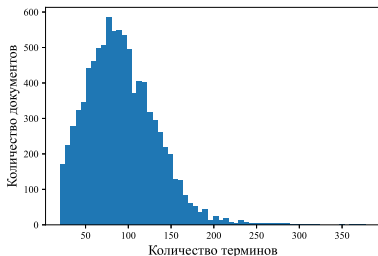
F1

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Среднее гармоническое между Precision и Recall

## ACL RD-TEC

Для обучения модели используется открытый датасет ACL RD-TEC, в котором собраны статьи на английском языке с 1965 по 2006 год из области компьютерной лингвистики. Для проведения эксперимента из него удаляются документы, содержащие менее 20 терминов. В результате получается датасет из 9,095 статей.



Из текста удаляются заголовки, литература, числа и стоп-слова. Также документы подвергаются стеммингу – отбрасыванию окончаний слов.

До

The paper also presents an evaluation of the system which shows that the system successfully retrieves the identification numbers of approximately 80% of the parts.

После

paper present evalu system show system success retriev identif  
number approxim part



Текст обрабатывается двумя алгоритмами: TopMine и вероятностной моделью с помощью библиотеки BigARTM. После чего берётся пересечение результатов работы.

## TopMine

Оценка частота вхождения термина в документ. Позволяет оценивать частоту термина, неслучайность последовательности слов в термине.

## ARTM

Решение задачи стохастического матричного разложения с добавлением дополнительных регуляризаций.

## Фоновая тема

'lie', 'apl', 'ion', 'tire', 'aud', 'tha', 'thc', 'arc', 'rel', 'lit'

## Предметные темы

['candidate', 'model', 'distance', 'method', 'measure', 'probability',  
'language', 'approach', 'based', 'performance']

['document', 'term', 'relevant', 'topic', 'approach', 'information',  
'relevance', 'result', 'query', 'number']

Для сравнения используется предобученная модель BERT из библиотеки transformers.

Таблица: Пример данных для обучения

sentence_id	words	labels
1	customer	O
1	service	O
1	allowing	O
1	users	O
1	retrieve	O
1	identification	B
1	numbers	O

Пометка 'O' — слово не является термином.

Пометка 'B' — слово является термином.

Таблица: Результаты работы алгоритмов

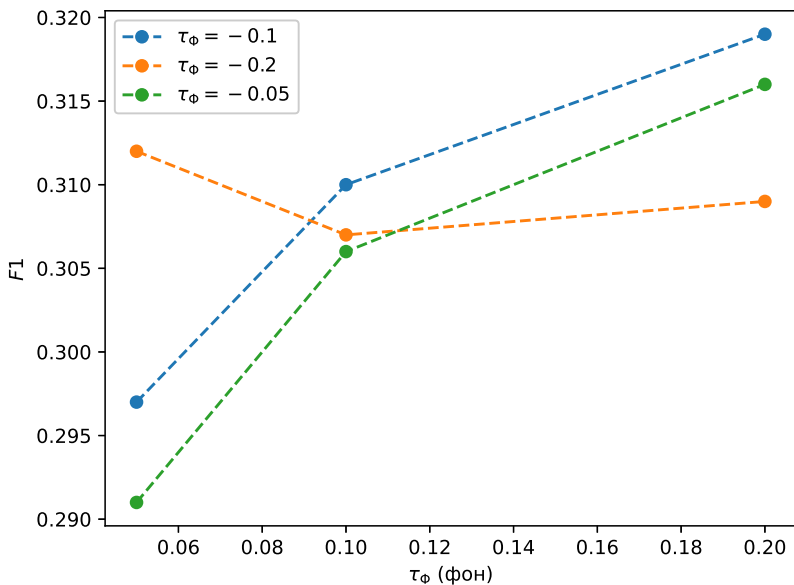
	Precision	Recall	F1
TopMine	0.063	0.946	0.117
Вероятностная модель	0.426	0.281	0.338
BERT	0.636	0.732	0.681






## Выводы

Вероятностная модель хорошо работает в сравнении с TopMine, но BERT показывает результат лучше. Однако недостатком нейронной сети является большое потребление оперативной памяти и памяти графического процессора. Также, в отличие от вероятностной модели, она неинтерпретируема.

- ❶ Использование датасета ACTER, для которого имеется большое количество результатов применения различных методов. Его минус в том, что он небольшой.
- ❷ Анализ термов из нескольких слов.
- ❸ Анализ информации о документе: авторы, год издания, заголовок, литература.

# Результаты



-  ВОРОНЦОВ К.В., *Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM.*
-  AHMED EL-KISHKY AND YANGLEI SONG AND CHI WANG AND CLARE R. VOSS AND JIAWEI HAN, *Scalable topical phrase mining from text corpora.*
-  НИКОЛАЙ ШАТАЛОВ, *Методы обучения без учителя для автоматического выделения составных терминов в текстовых коллекциях.*
-  TRAN, HANH THI HONG AND MARTINC, MATEJ AND CAPORUSSO, JAYA AND DOUCET, ANTOINE AND POLLAK, SENJA, *The Recent Advances in Automatic Term Extraction: A survey.*
-  BEHRANG Q. ZADEH AND SIEGFRIED HANDSCHUH, *The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics.*