

# Мультиязычный тематический поиск и категоризация научных публикаций

Евгений Владимирович Тищенко

Московский физико-технический институт

*Эксперт:* К. В. Воронцов

*Консультант:* П. С. Потапова

2021

# Цель исследования

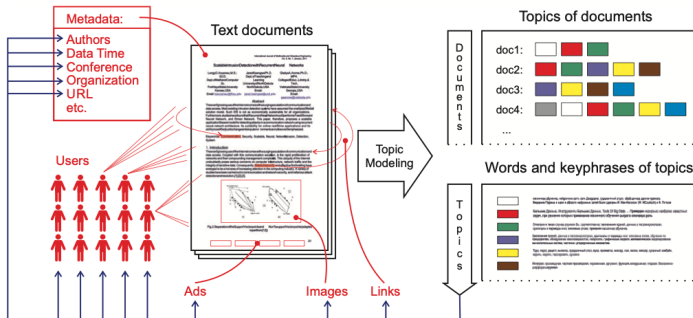
Решается задача кроссязычного поиска текстового плагиата.

Целью работы является получение модели, выделяющей информацию о распределении слов в тексте независимо от их языковой принадлежности, при этом удовлетворяющей ограничениям по размеру модели в 100 Гб и максимальным временем обучения не более 24 часов для ее практического использования.

Для преобразования текста в векторное представление предлагается использовать тематическое моделирование.

# Использование мультимодальной тематической модели

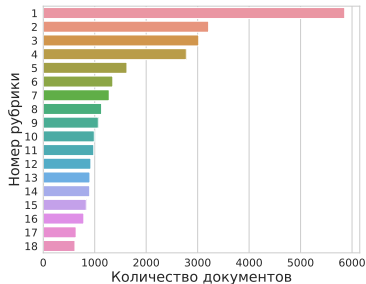
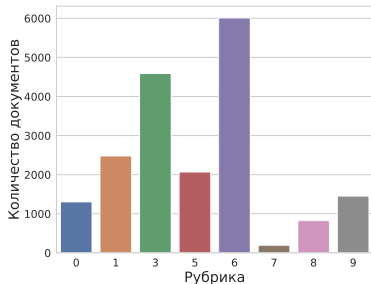
В качестве модальностей используются **100** языков, а также научные рубрики ГРНТИ и УДК. В качестве предобработки используется **ВРЕ-токенизация**.



- ▶ Воронцов К. В. Обзор вероятностных тематических моделей.
- ▶ Zdenek Ceska, Michal Toman, and Karel Jezek Multilingual Plagiarism Detection
- ▶ Duygu Ataman, Jose G. C. de Souza, Marco Turchi, Matteo Negri Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual

# Описание данных

- ▶ Статьи из научной электронной библиотеки elibrary.ru, а также статьи с сайта wikipedia.org
- ▶ Данные на 100 языках
- ▶ Для большинства статей известны коды рубрик УДК (10 рубрик) и ГРНТИ (70 рубрик)
- ▶ Машинный перевод статей на 42 языка



# Оценка качества

- ▶ Для каждого документа известен его перевод
- ▶ Поиск перевода осуществляется не по всей поисковой коллекции, а по подвыборке, в которой:
  - ▶ 10% документов с таким же УДК как у документа-запроса
  - ▶ 90% документов с другими УДК
- ▶ Метрики качества кроссязыкового поиска:
  - ▶ Средняя частота, с которой документ-запрос попадает в топ 10% документов отранжированной поисковой коллекции
  - ▶ Средний процент документов в топ 10% документов-переводов, которые имеют такую же рубрику, что у документа-запроса
- ▶ В качестве метрики в векторном пространстве используется метрика аналогий

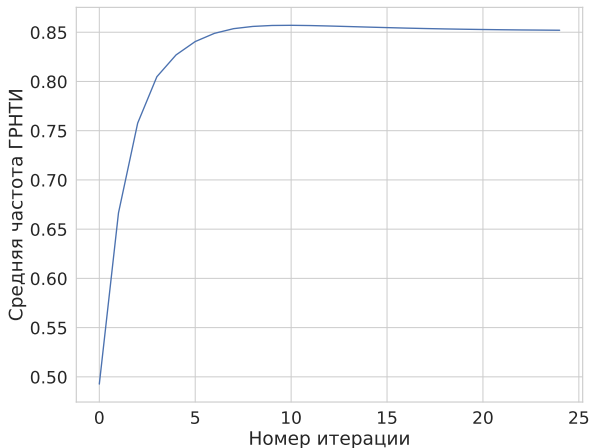
$$\rho(d, d') = 1 - \cos(d', d - a + b)$$

## Решение

- ▶ Выделение 125 отдельных тем
- ▶ Исключение стоп-слов из 15 основных языков
- ▶ Генерация на каждой итерации алгоритма подвыборки, имеющей равномерное распределение рубрик ГРНТИ
- ▶ Расширение словаря до 11 тыс. токенов на язык, используемых при ВРЕ-токенизации для улучшения выразительности модели
- ▶ Отказ от использовании рубрики "нет" из-за различной тематики документов, представленных в этой рубрике

Обучение занимает 25 итераций, суммарное время обучения - 21 час, размер модели составляет 96 Гб.

# Вычислительный эксперимент



Метод быстро сходится к оптимальному значению.



## Сравнение моделей

Название модели	Средняя частота УДК	Средний процент УДК	Средняя частота ГРНТИ	Средний процент ГРНТИ
Базовая тематическая модель	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.878	0.250	0.860	0.361
Итоговая тематическая модель	0.995	0.225	0.852	0.366

- ▶ предложен метод, получения векторного представления текста при помощи тематического моделирования
- ▶ построена модель для поиска мультязычных дубликатов
- ▶ достигнуты ограничения для практического использования