# Кроссязычный поиск дубликатов

*Е. В. Тищенко*[1], *К. В. Воронцов*[2], *П. С. Потапова*[3]

В данной статье рассматривается задача кроссязычного поиска текстового плагиата. Современные методы векторизации документов и поиска совпадений преимущественно основываются на одном языке, что приводит к проблеме возникновения однообразных мультиязыковых коллекций документов.

Целью работы является получение модели, выделяющей информацию о распределении слов в тексте независимо от их языковой принадлежности, при этом ограниченной по размеру и времени обучения для ее практического использования.

## 1 Введение

Задача поиска и распознавания плагиата является особенно важной в эпоху информационных технологий. Невозможно проверить все документы общей тематики на плагиат в силу размеров глобальной сети. Задача становится особенно трудной, так как возникает множество документов, являющихся переводом исходных работ. Согласно исследованиям Donald L. McCabe, [1] 36% студентов американских университетов перефразировали или копировали информацию из всемирной паутины без ссылки на источник.

Современные методы векторизации документов для поиска плагиата [2, 3] преимущественно ограничиваются одним языком. В таком случае возникает проблема создания единообразной системы получения векторных эмбедингов мультиязыковой коллекции документов. Для возможности применения модели за пределами научных экспериментов ставятся технические ограничения ресурсами сервера, а именно на размер модели, временную сложность обучения.

Объектом исследования являются мультиязыковые тематические модели, алгоритмы поиска документов в текстовой коллекции по словам и документам, способы векторного представления слов и документов запроса и коллекции, применяемые для поиска дубликатов независимо от языковой принадлежности.

Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова или термины образуют каждую тему. Вероятностная тематическая модель описывает каждую тему дискретным распределением вероятностей слов, а каждый документ — дискретным распределением вероятностей тем. Тематическая модель преобразует любой текст в вектор вероятностей тем.

Для решения задачи была построена мультимодальная тематическая модель. Такая тематическая модель описывает документы, содержащие метаданные наряду с основным текстом. Метаданные позволяют более точно определять тематику документа. В качестве модальностей использовались 100 языков, а также научные рубрики. Использование языков в качестве модальностей позволяет получить векторное представление текста, независимое от оригинального языка текста, что позволяет решать проблему поиска плагиата без ограничения на язык статьи. Во время предобработки текста используется BPE токенизация — итеративная замена наиболее встречаемой пары символов на символ, который не встречается в слове. Это позволяет существенно уменьшить объем изначального словаря для практического применения модели.

Целью эксперимента является построение модели, исследование влияния регуляризации и предобработки текстовых данных на качество поиска, подбор разнообразных функций для сравнения тематических расстояних векторов а также поиск эвристик для улучшения точности предсказаний модели. В качестве обучающих данных используются статьи

с сайта Wikipedia, а также выборка научных статей из научной электронной библиотеки eLIBRARY.ru.

## 2 Постановка задачи

Пусть D − некоторая коллекция документов. Кандидатом на дубликат для документа $d \in D$ обозначим такой элемент коллекции f(d), что

$$f(d) = \arg\min_{d' \in D \setminus d} distance(m(d), m(d'))$$

В качестве функции *distance* может быть использована произвольная метрика векторного пространства. Функция $m(d)$− модель, осуществляющая преобразование документа в векторное пространство. Качество модели поиска дубликатов измеряется на тестовой выборке при помощи двух метрик сопоставления переводов:

1. Средняя частота, с которой документ-запрос попадает в топ 10%

2. Средний процент документов в топ 10% документов-переводов, которые имеют такую же рубрику, что у документа-запроса

Требуется построить тематическую модель, позволяющую получить векторное представление документов, при этом ошибка по метрикам качества должна превосходить 0.9 и 0.3 соответственно.

## Литература

[1] *Donald L. McCabe* Cheating among college and university students: A North American perspective// International Journal for Educational Integrity, 2005

[2] *Zdenek Ceska, Michal Toman, and Karel Jezek* Multilingual Plagiarism Detection// Artificial Intelligence: Methodology, Systems and Applications, 2008

[3] *Duygu Ataman, Jose G. C. de Souza, Marco Turchi, Matteo Negri* Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual Word Embeddings//

*Поступила в редакцию*