Университет ИТМО, факультет инфокоммуникационных

технологий Отчетная работа по «Информатике»: аннотация к статье

Выполнил(а) Бархатова Н.А., № группы К3123, дата 27.09.2022, оценка

| **Название статьи/главы книги:** О поиске эквивалентных текстов | | |
| --- | --- | --- |
| **ФИО автора статьи:**  Бухаров Д. С. | **Дата публикации:**  2016 г**.** | **Размер статьи**  7 стр. |
| **Прямая полная ссылка на источник и сокращенная ссылка:**  https://elibrary.ru/item.asp?id=26236250  https://clck.ru/32AY8b | | |
| **Тэги, ключевые слова или словосочетания**  Проверка на плагиат, поиск эквивалента, сравнение текстов, антиплагиат | | |
| **Перечень фактов, упомянутых в статье:**  Решение проблемы поиска эквивалентного текста посредством создания поискового множества в качестве маски.  Поисковое множество создается наиболее оптимальным для последующей проверки образом. Его заполнение происходит с учетом нескольких факторов, характеризующих стиль написанного текста. Среди данных факторов выделяют следующие: исключение наиболее частотных слов из вспомогательного множества, отслеживание пунктуационных и орфографических ошибок, сокращение слов за счет удаления незначимых морфем и учёт артефактов текста. Технология сравнения некоторого исходного текста с созданным множеством предполагает двойную проверку символов на совпадение: с учетом регистра и без. Данный алгоритм реализован на языке программирования Java.  Для проверки данного способа обработки входных данных было реализовано 2 эксперимента, каждый из которых включал в себя 915 поисковых текстов, часть из них была модифицирована для усложнения задачи. Результаты экспериментов показали, что алгоритм нуждается в доработке. Технология была оптимизирована с помощью добавления базы данных синонимов. Представление текста в виде множеств позволяет избежать сбоев и обеспечивает стабильную, но ресурсоёмкую работу по проверке текста на антиплагиат. | | |
| **Позитивные следствия и/или достоинства описанной в статье**  **технологии**  - Алгоритм обладает широким рядом инструментов и функций  - Отсутствие сбоев во время работы алгоритма  - Вероятность верного определения плагиата равна 98,25% | | |
| **Негативные следствия и/или недостатки описанной в статье**  **технологии**  - Каждый вычислительный эксперимент занимает большое количество  времени  - Технология требует использования многопоточных вычислений  - Вероятность верной работы алгоритма снижается, если на вход подается  текст, модифицированный синонимами слов | | |
| **:)** | | |