

Определение дубликатов текста с помощью Sentence Transformers

Анишкова Анастасия Сергеевна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Цель исследования: разработка и реализация метода определения дубликатов текста с использованием семантических эмбедингов, полученных с помощью предобученной модели трансформеров distiluse-base-multilingual-cased-v2 из библиотеки sentence-transformers. Для реализации выбран облачный сервис Google Colab, предоставляющий доступ к вычислительным ресурсам, включая GPU, и позволяющий эффективно работать с моделями машинного обучения. Результат может быть применён в обучающих целях.

Ключевые слова: Sentence Transformers, дубликат текста, Google Colab

Detecting duplicate text using Sentence Transformers

Anishkova Anastasia Sergeevna

Sholom-Aleichem Priamurskiy State University

Student

Abstract

The purpose of the study: to develop and implement a method for detecting duplicate texts using semantic embeddings obtained using a pre-trained model of transformers distiluse-base-multilingual-cased-v2 from the library sentence-transformers. The Google Colab cloud service has been selected for implementation, which provides access to computing resources, including a GPU, and allows you to work effectively with machine learning models. The result can be used for educational purposes.

Key words: Sentence Transformers, duplicate text, Google Colab

1 Введение

1.1 Актуальность

Актуальность определения дубликатов текста с помощью Sentence Transformers обусловлена необходимостью эффективного выявления семантических повторений в больших объемах текстовых данных. Традиционные методы сравнения, основанные на точном совпадении слов, не всегда позволяют обнаружить перефразированные или частично измененные тексты. Модели на основе трансформеров, такие как distiluse-base-multilingual-cased-v2, обеспечивают высокую точность анализа за счёт преобразования текстов в семантические эмбединги и вычисления их косинусного сходства, что делает подход актуальным для задач обработки естественного языка,

включая детекцию плагиата, фильтрацию контента и улучшение качества данных.

1.2 Обзор исследований

В. Л. Дашонок, А. Д. Хомоненко, В. А. Боброва рассмотрели подход к устранению избыточности семантически близкой текстовой информации в хранилищах и базах данных [1]. Подход к обработке текста на естественном языке описал Р. А. Вартанян [2]. В. Девараджан, Р. Субраманьян провели эффективное изучение данных в веб-масштабе посредством семантической дедупликации [3]. Анализ семантического сходства текстовых документов с целью выявления практически совпадений провели В. Девараджан, Р. Субраманьян [4]. И. М. Нгуэна протестировали методы быстрого обнаружения семантических дубликатов в базах данных [5].

1.3 Цель исследования

Цель исследования: разработка и реализация метода определения дубликатов текста с использованием семантических эмбедингов, полученных с помощью предобученной модели трансформеров `distiluse-base-multilingual-cased-v2` из библиотеки `sentence-transformers`.

2 Материалы и методы

Использовалась модель `distiluse-base-multilingual-cased-v2` из библиотеки `sentence-transformers`.

3 Результаты

Установим библиотеку `sentence-transformers` для работы с предобученными моделями трансформеров, позволяющими получать эмбединги предложений и вычислять их семантическое сходство (рис.1).

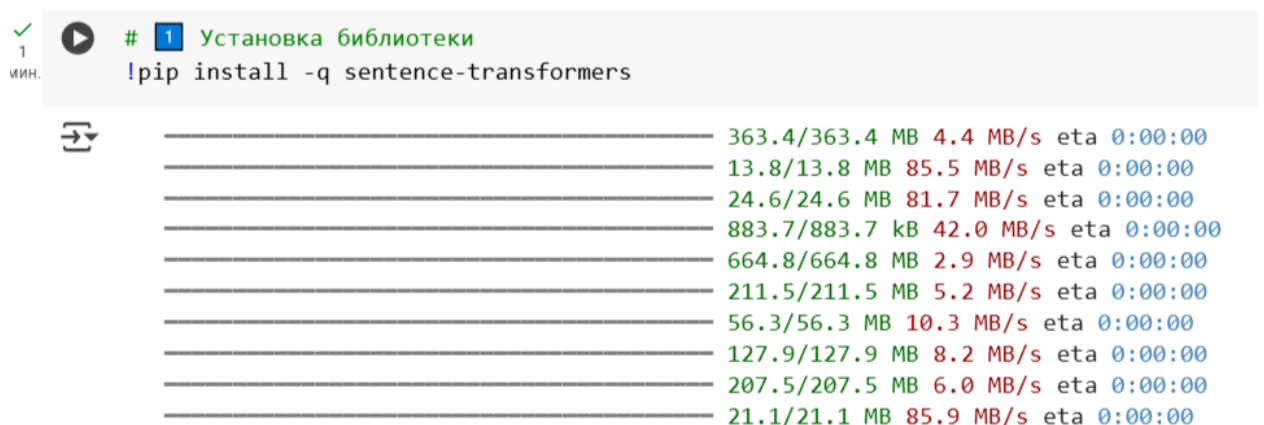


Рисунок 1. Установка библиотеки

Импортируем необходимые модули (рис.2): `SentenceTransformer` для загрузки модели, `util` для вычисления косинусного сходства, и `torch` для работы с тензорами.

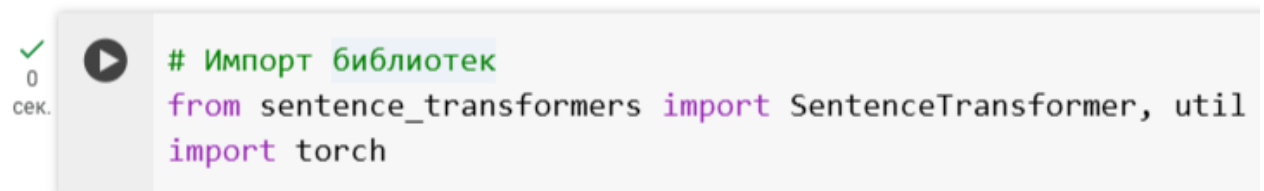


Рисунок 2. Импорт библиотек

Создадим список текстовых примеров на русском языке для последующего анализа на наличие дубликатов с помощью эмбедингов (рис.3).

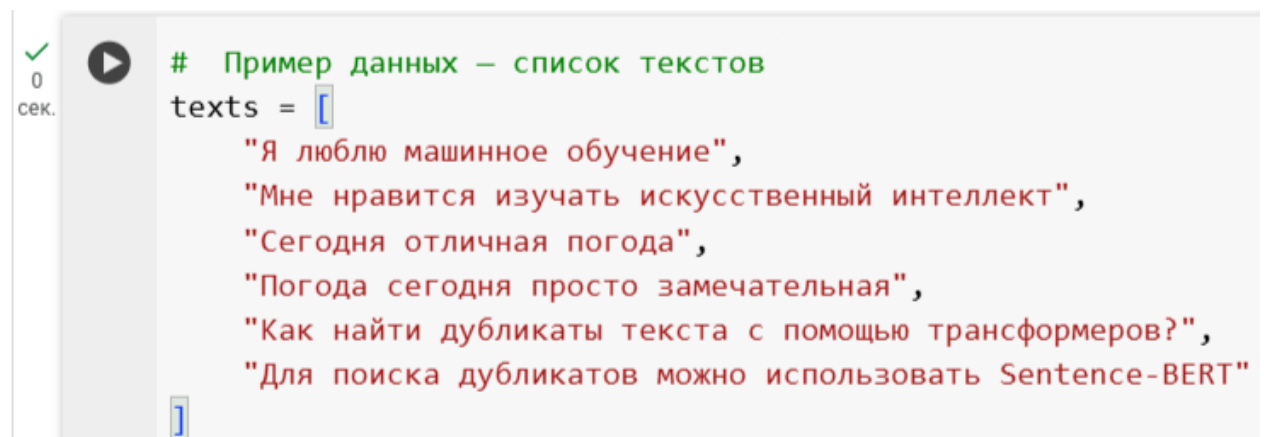


Рисунок 3. Пример данных

Загрузим предобученную многоязычную модель `distiluse-base-multilingual-cased-v2`, которая позволяет преобразовывать тексты в семантические эмбединги для анализа схожести (рис.4).

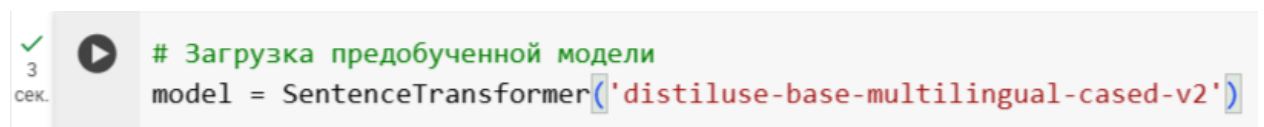


Рисунок 4. Загрузка предобученной модели

Преобразуем тексты в семантические эмбединги с помощью модели, результат сохраняется в виде тензора для дальнейшего вычисления схожести (рис.5).

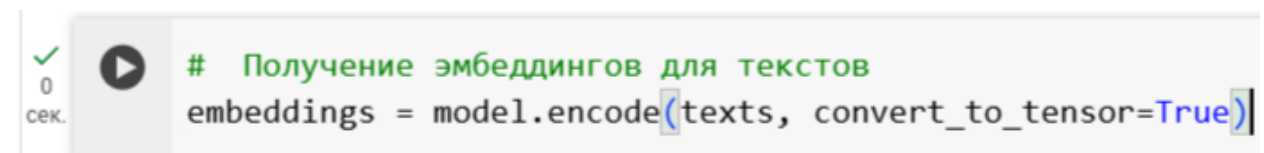


Рисунок 5. Получение эмбедингов для текстов

Вычислим матрицы косинусных сходств между всеми парами текстов на основе их эмбедингов для определения семантической близости (рис.6).

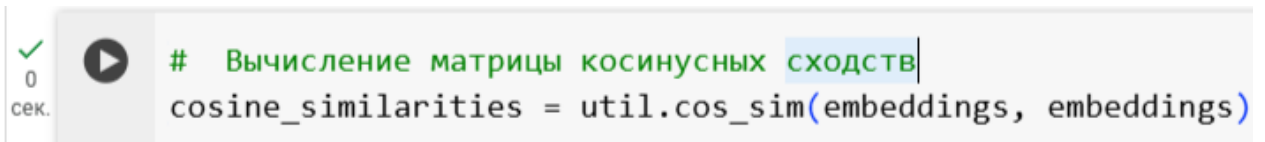


Рисунок 6. Получение эмбедингов для текстов

Установим порог схожести (0.8), выше которого тексты считаются потенциально дублирующими (рис.7).

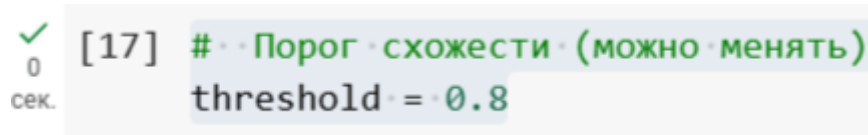


Рисунок 7. Порог схожести

Найдем все пары текстов, косинусное сходство между которыми превышает заданный порог, и сохранение их в список дубликатов с указанием индексов и уровня схожести (рис.8).

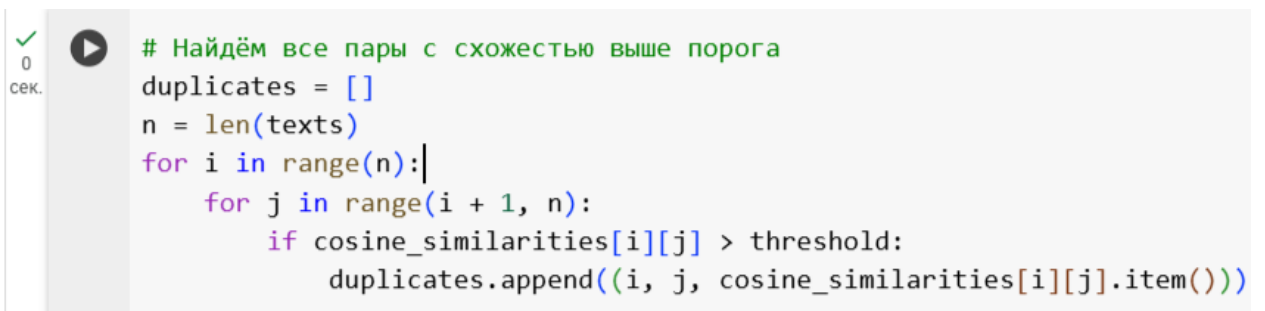


Рисунок 8. пары с схожестью выше порога

Выводим найденные пары текстов, которые превышают заданный порог схожести, с отображением индексов, уровня схожести и самих текстов для анализа возможных дубликатов (рис. 9).

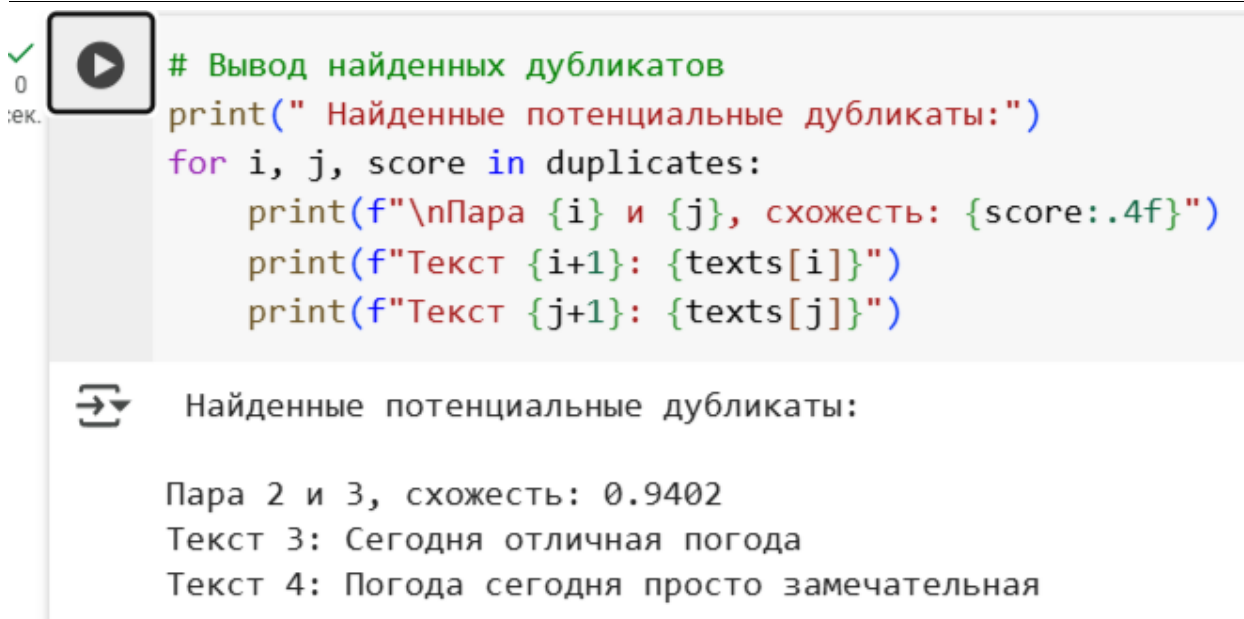


Рисунок 9. Дубликаты

4. Выводы

В ходе исследования была реализована методика определения текстовых дубликатов с использованием семантических эмбедингов на основе модели distiluse-base-multilingual-cased-v2 из библиотеки sentence-transformers. Подход заключался в преобразовании текстов в векторы (эмбединги) и вычислении их косинусного сходства. Установленный порог сходства (0,8) позволил выявлять потенциально дублирующие тексты. Например, пара текстов: Текст 3: «Сегодня отличная погода» и Текст 4: Погода сегодня просто замечательная». Они имеют сходство 0,9402, что значительно превышает порог, и, следовательно, были классифицированы как потенциальные дубликаты.

Результаты подтверждают эффективность использования моделей на основе трансформеров для обнаружения семантических повторений даже при наличии перефразирования. Реализация проводилась в облачной среде Google Colab, что обеспечило доступ к GPU и ускорило вычисления. Полученный метод может быть применён в образовательных и практических целях, например, для обнаружения плагиата или фильтрации контента.

Библиографический список

1. Дашонок В. Л., Хомоненко А. Д., Боброва В. А. Подход к устранению избыточности семантически близкой текстовой информации в хранилищах и базах данных // Бюллетень результатов научных исследований. 2013. №3 (8). С. 14-22.
2. Вартанян Р. А. Подход к обработке текста на естественном языке // Наукосфера. 2020. №11-2. С. 199-203.
3. Abbas A. et al. Semdedup: Data-efficient learning at web-scale through semantic deduplication // arXiv preprint arXiv. 2023. №2303. С. 25-30.

4. Devarajan V., Subramanian R. Analyzing semantic similarity amongst textual documents to suggest near duplicates // Indonesian Journal of Electrical Engineering and Computer Science. 2022. №3. C. 1703-1711.
5. Nguena I. M. et al. Fast semantic duplicate detection techniques in databases // Journal of Software Engineering and Applications. 2017. №6. C. 529.