ПРИЛОЖЕНИЕ ПО АНАЛИЗУ И ОБРАБОТКЕ ТЕКСТОВОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Охрименко В. Д., студент гр. 050503

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Перцев Д.Ю. – канд. техн. наук

В данной работе были исследованы основные технические решения и особенности создания приложения по анализу и обработке текстовой информации с использованием большой языковой модели BART, а также применение алгоритма выборочного реферирования текста – TextRank.

Приложения для создания кратких содержаний текстов нашли широкое применение во всем мире. Для их реализации можно выделить два метода: выборочный и абстрактный. Выборочный метод – это метод, при котором из исходного текста извлекаются наиболее важные и информативные фрагменты без изменения их первоначальной формулировки. Абстрактный метод, в свою очередь, включает переформулировку и обобщение основного содержания исходного текста, создавая новый текст, который отражает его ключевые идеи в сжатой форме [1]

Для реализации алгоритмов резюмирования текста были использованы комментарии с различных YouTube каналов (в основном англоязычных), чтобы обеспечить единообразие анализируемого материала, собранные с помощью YouTube Data API V3 [2]. Важной стадией подготовки данных к анализу был процесс предобработки. На этом этапе осуществлялась стандартизация текстовой информации, что включало в себя несколько ключевых процедур. Во-первых, все текстовые данные были приведены к нижнему регистру, что облегчало их последующую обработку и анализ, исключая влияние регистра символов на результаты обработки. Во-вторых, была проведена очистка текстов от дублирующих знаков препинания. Это не только улучшило читаемость комментариев, но и предотвратило возможные ошибки при разборе текста, обеспечивая более точное извлечение информации и анализ содержания. При работе с алгоритмами суммаризации стало очевидно, что наличие эмоджи в тексте комментариев вызывает проблемы при их обработке и анализе. Эмоджи не только усложнили задачу алгоритмам, но и исказили конечные результаты, делая их менее точными и релевантными. В связи с этим было принято решение о разработке и внедрении дополнительного этапа предобработки данных, который включал в себя удаление эмоджи из текстов комментариев. Этот шаг позволил повысить качество и точность алгоритмов суммаризации, устраняя источник потенциальных искажений.

В качестве основного инструмента для создания краткого содержания текста был выбран алгоритм TextRank [3] для выборочного резюмирования текста, который выбирает наиболее важные и информативные предложения из текста, формируя из них суммарный обзор. Для абстрактного метода использовалась предобученная модель BART [4], которая позволяет создавать связный и сжатый текст, извлекая ключевые идеи из исходного материала. Для достижения более высокого качества резюмирования комментариев использована комбинация этих двух алгоритмов. Этот подход был выбран для решения двух задач: необходимость сокращения объема данных и улучшение качества суммаризации. Сочетание TextRank для предварительного уменьшения размерности текстов и BART для генерации конечного краткого содержания позволило эффективно обрабатывать большие объемы данных. TextRank выступил в роли фильтра, уменьшая нагрузку на BART и позволяя ему сосредоточиться на генерации качественных резюме. Такой подход обеспечил не только управление большими массивами данных, но и повысил качество и точность резюмирования комментариев, делая результаты более полезными и информативными для последующего анализа и использования.

Таким образом, было создано приложение для анализа и обработки текстовой информации, которое использует большие языковые модели. Потенциал приложения к расширению и улучшению весьма значителен и включает в себя интеграцию с разнообразными источниками данных для обеспечения более широкого спектра анализа данных и создание расширенных настроек для анализа полученных данных.

Список использованных источников:

1. Суммаризация текста: подходы, алгоритмы, рекомендации и перспективы [Электронный ресурс]. – Электронные данные. – Режим доступа: *https://habr.com/ru/articles/514540.*

2. YouTube Data API v3 [Электронный ресурс]. – Электронные данные. – Режим доступа: *https://developers.google.com/youtube/v3.*

3. TextRank [Электронный ресурс]. – Электронные данные. – Режим доступа: *https://www.mdpi.com/2079-9292/12/2/372.*

4. BART [Электронный ресурс]. – Электронные данные. – Режим доступа: *https://huggingface.co/facebook/bart-large-cnn.*