КОМАНДА “МАССИВ ПРИМАТОВ”

РАЗРАБОТКА ВСТРАИВАЕМОГО ИТ РЕШЕНИЯ ДЛЯ ПРОВЕРКИ ТЕКСТА НА НАЛИЧИЕ ССЫЛОК И ИНЫХ ДЕСКРИПТОРОВ (В ТОМ ЧИСЛЕ ИХ ВАРИАЦИЙ) ИЗ ЧЕРНОГО СПИСКА. (ФАКТЧЕКИНГ)

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

Москва, 2024

ПРОЦЕСС ИСПОЛЬЗОВАНИЯ РЕШЕНИЯ

В имеющемся датасете обнаруживаются папки, в которых хранились файлы json, в которых содержатся статьи.

Совершается проход по всем папкам в датасете и использовала endswith для правильного считывания json-файлов.

В файле blacklist.txt, содержащем черный список, находится своеобразная запись запрещенных ссылок.

Переходя к этапу разметки имеющегося датасета для дальнейшего обучения по нему модели, программа проходится по всем статьям и размечает их на содержащие в себе запрещенные ссылки и не содержащие.

С использованием правильно считаных запрещенных ссылок производится разметка датасета для дальнейшего обучения по нему модели. Используя возможности языка программирования Python, а именно метода search, используется функция проверки наличия ссылок, находящихся в статье, в черном списке, и соответствующее помечание статей.

Создается словарь, для хранения данных меток. Статья помечается нулем, если не содержит запрещенных ссылок, и единицей, если такие ссылки обнаружены в статье.

Программа использует метод опорных векторов, который хорошо подходит для работы с текстами и обладает достаточной производительностью для обработки такого объема данных. Таким образом, разместив датасет статей с учетом черного списка, обучается модель, которая может предсказывать содержит ли поданная на рассмотрение статья запрещенные ссылки. Обученная модель, которая имеет точность 0.97.

По итогам работы программы в датасете ищется количество статей с запрещенными ссылками. И предсказывает наличие в тестовой статье запрещенных ссылок.