СЕРВИС ДЛЯ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕКСТА

Лапшина Вероника Вячеславовна (Россия, Санкт-Петербург, школа №564, класс: 10м)

Руководитель: Штукенберг Дмитрий Григорьевич, преподаватель программирования

В современном мире информацию удобно хранить в электронном виде, однако большое количество знаний хранится на бумажных носителях – книги, документы и т.д. Поэтому инструменты для оптическое распознавание текста (OCR) имеют большое значение. Существующие программы распознают текст, допуская при этом большое количество ошибок, и нуждаются в редактировании распознанного текста человеком. Поэтому, задача повышения качества распознавания является крайне актуальной.

Процесс OCR состоит из 3 этапов: препроцессинг, распознавание, постпроцессинг. Препроцессинг – подготовка текста к распознавание, выделение в изображение блоков, содержащих текст. Распознавание – преобразование изображений в текст. Постпроцессинг – корректирование текста после распознавания. По мнению автора, слабое место в OCR это постпроцессинг, т.к. качественное исправление ошибок требует некоторого понимания текста, а это невозможно без обширной статистики. Используемые же программы работают у каждого пользователя по отдельности, что не позволяет объединять и обобщать опыт распознавания текста.

Автором разработан онлайн сервис, позволяющий конвертировать изображения в текст и использовать для постпроцессинга опыт предыдущих распознаваний, что позволяет обмениваться знанием об ошибках в распознанных текстах между пользователями. Программа составляет статистику по встреченным словам и их исправлениям и использует ее при следующих распознаваниях.

Пользователю предлагается веб-страница с возможностью загрузки файла. Полученный файл обрабатывается программой tesseract , преобразующей изображение в текст. Для каждого слова из текста осуществляется поиск по базе данных. При нахождении слова, выбирается оптимальный вариант исправления из существующих. Для выбора используется статистика исправлений – выбирается чаще всего встречаемый вариант. При существовании вариантов, с примерно одинаковой частотой исправлений, для выбора используется наивный байесовский классификатор. После пользователь получает возможность редактировать текст. Результата редактирования сравнивается с текстом, полученным в результате обработки, и учитываются в базе данных.

Язык программирования – python. Система управления базами данных – MySQL.

Литература:
1.David MacKenzie, Paul Eggert, and Richard Stallman. «Comparing and Merging Files with GNU Diff and Patch»
2. E. Myers (1986). «An O(ND) Difference Algorithm and Its Variations»