УДК 004.932

WEB-ПРИЛОЖЕНИЕ ДЛЯ ОБРАБОТКИ ЦИФРОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ ИИ

П. А. Легушева

Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, polina.leg2003@gmail.com

Разработано WEB-приложение для очистки цифровых документов от загрязнений и шумов, а также для распознавания текста с помощью фреймворка *Flask* и искусственного интеллекта. Был изучен материал о современных подходах распознавания текста и «деноизинга» *(denoising)* изображений, реализованы модели с помощью сверточных и рекуррентных нейронных сетей, и оценены благодаря метрикам качества. Полученные модели были встроены в приложение. Также была реализована авторизация и запись в базу данных имя, электронной почты, пароля и количества выполненных вычислений с документами пользователя.

Ключевые слова: WEB-приложение; фреймворк Flask; «деноизинг» изображений; распознавание текста; сверточные нейронные сети; автокодировщик; рекуррентные нейронные сети; метрики качества.

Современные методы распознавания текста включают в себя использование глубокого обучения и нейронных сетей. Большая часть подходов строится на:

- Сверточных нейронных сетях (CNN): этот метод используется для изучения признаков текста на изображениях и его структуры.
- Рекуррентных нейронных сетях (RNN): они учитывают контекст и последовательность символов или слов при распознавании текста.
- Сочетании различных алгоритмов: некоторые современные методы объединяют в себе CNN и RNN для повышения точности распознавания текста.

Ниже приведены примеры популярных и высококачественных инструментов для распознавания текста.

Tesseract – это библиотека оптического распознавания символов *(OCR)*, разработанная *Google*. Она предоставляет возможности распознавания текста на изображениях с использованием методов машинного обучения. *Tesseract* способен обрабатывать текст на различных языках и работать с различными типами шрифтов, что делает его универсальным инструментом для распознавания текста на изображениях и отсканированных документах.

EasyOCR – еще одна библиотека *OCR*, предоставляющая простой и понятный *API* для распознавания текста на изображениях. Она обеспечивает возможность работы с различными языками и типами текста, а также поддерживает распознавание текста на многоязычных изображениях. *EasyOCR* удобен для быстрой и эффективной обработки текста на изображениях, что делает его популярным среди разработчиков и исследователей.

Было создано приложение, в котором были разработаны следующие этапы:

- Создание маршрутов и шаблонов для управления навигацией пользователей по веб-приложению.
- Создание форм регистрации и входа для создания учетных записей пользователей и входа в систему.
- Настройка базы данных для хранения информации о пользователях и вывода сообщений об ошибках.
- Защита страницы от доступа неавторизованных пользователей и перенаправления на страницу входа.

• Загрузка загрязненных изображений, их очистки с использованием встроенных моделей и распознавания текста.

• Отображение информац+ии о пользователе на странице профиля, включая имя и количество проведенных вычислений.

Далее представлены эксперименты и результаты технологии *Tesseract* для распознавания текста очищенного и зашумленного изображений:

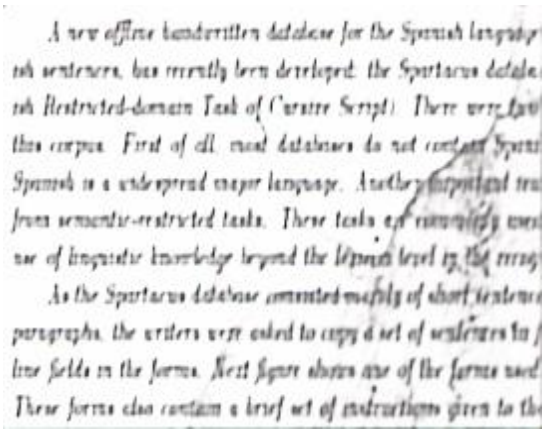

Рис. 1. Очищенное моделью изображение

```
A sev offre bestertien detcene Jor th Spruah Lrepaty
wh aeatescre ben rrraty een detrlopr the Spurtaren dete
1 Retr ever Del of Cre Sng) Drew
len corpue Fart of ll cal dealnacy ds et 1
Spcaed ow enderpread cnepr language, Aneta
frees emvendwcrrtrvted tasks There teal
tae of baste rele lepeed the Up elt 04 rg

Anh Stare dor sand ly of a at
ongraghe the eras vem ed toa dito wars
lee foe th force. est fn aaen ane of he ore sed

Theme fren cao vont to th

et of tudratign
```

Рис. 2. Результат распознавания текста на очищенном изображении с помощью технологии Tesseract

Хорошо показано, что есть неточности в распознавании из-за размытости изображения.

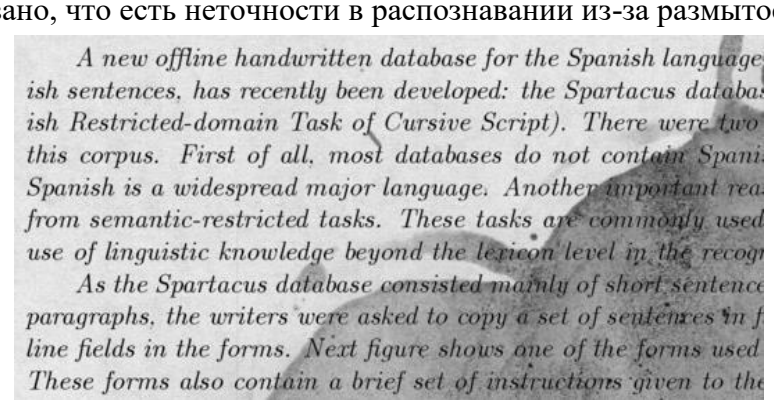

Рис. 3. Зашумленное изображение

```
A new offline handwritten database for the Spanish lang
ish sentences, has recently been developed: the Spartacus data
ish Restricted-domain Task of Cursive Script). There weng
this corpus. First of all, mos
Spanish is a widespread major language. Anothep
from semantic-restricted tasks.
use of linguistic knowledge beyond the

As the Spartacus ddbabase. consis
paragraphs, the writers ‘we
line fields in the forms.
These forms also conti
```

Рис. 4. Результат распознавания текста на зашумленном изображении с помощью технологии Tesseract

Видно, что меньше распознается текст там, где есть шум.

```
['v5 r letrll Udskls"u', '"Jav', 'mnlh', 'enb;n', 'Siine', 'Hestratc LAnur Tal
of (\xa0 fr 677l', 'Ikr', 'Fru 4 dl', "Aa'", 'mmnlm hnu', 'Aud', 'mrrhl Wul', '
Iku ul,', 'Ka', 'kcly nlk Vnnabuln', 'Quluu', 'rmbd mufhy4[*4431', '7rthi',
'{m', 'dnlum;', "K Hunilk7'", 'Jvn #m', 'Il kr', 'Uvl']
```

Рис.5. Результат технологии EasyOCR на очищенном изображении

```
['new', 'offline handwritten database for the Spanish language', 'sentences_',
'recently been developed: the Spartacus databaz', 'ish Restricted-domain Task o
f Cursive Script) .', 'There', 'were two', 'this', 'corpus_', 'First of all,',
'databases do', 'contam', 'Spani=', 'Spanish is', 'widespread major language:',
'Another important rea-', 'from semantic-restricted tasks_', 'These tasks', 'co
mmonly used', 'use', 'of linguistic knowledge beyond the lexicon level in-the',
'recog', 'As the Spartacus database consisted mainly of short sentence', 'parag
raphs.', 'the writers were asked to copy', 'of sentenres in', 'line fields in t
he forms', 'Next figure shows', 'one of the forms', 'These forms also', 'contam
', 'brief set of instructions given to the', 'ish', 'has', 'most', 'not', 'set'
, 'used']
```

Рис.6. Результат технологии EasyOCR на зашумленном изображении

В работе были использованы следующие метрики качества:

- *Word Error Rate (WER)* – частота ошибок в символах.
- *Character Error Rate (CER)* – частота ошибок в словах.
- *Levenshtein distance* – метрика сходства между двумя строковыми последовательностями. Чем больше расстояние, тем более различны строки.ё

В данном исследовании были использованы пять загрязненных изображений из тестового датасета и пять таких же, но уже очищенных ранее моделью, изображений.

Полученные результаты оценки модели показаны в следующей таблице:

Значения метрик на очищенных и зашумленных изображениях

	Mean CER	Mean WER	Levenshtein distance
Tesseract clean image	45,11%	85,312%	55,011%
Tesseract dirty image	49,826%	69,664%	50,235%
EasyOCR clean image	96,281%	100%	96,287%
EasyOCR dirty image	99,594%	98,847%	99,594%

Из имеющихся результатов метрик лучшей оказалась технология *Tesseract* на очищенном изображении. Данная технология получилась более устойчива к шумам и размытым изображениям. Для улучшения результата следует усовершенствовать качество модели по очистке шумов на изображении.

Заключение

В данной статье было представлено веб-приложение, в котором встроены модели для очистки изображений от шума и нежелательных элементов, а также для распознавания текста, что позволяет удобно обрабатывать старые документы, книги, учебники и извлекать информацию из них. Оно будет полезно для образовательных учреждений, преподавателей, исследователей, археологов, историков и др. Также сравнивались две технологии *Tesseract* и *Easy-OCR*. В данном эксперименте *Tesseract* по качеству распознавания оказался лучше.

Библиографические ссылки

1. 1.Андрианов С. Н., Веремей Е. И. О ходе подготовки бакалавров по направлению «Информационные технологии» на факультете ПМ-ПУ СПбГУ // Тр. Первой международной научно-практической конференции «Современные информационные технологии и ИТ-образование». М.: МАКС Пресс, 2005. С. 92-98.

2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. 2022.

3. Шолле Ф. Глубокое обучение на Python. 2022.