ОСОБЛИВОСТІ ФУНКЦІОНУВАННЯ СУЧАСНИХ ПРОГРАМ РОЗПІЗНАВАННЯ ТЕКСТУ

Вороніна Є.С., Кудрявцева К.С., студентки
Національний авіаційний університет, м.Київ
Науковий керівник – Денисенко С.М., к.п.н., доцент кафедри КММТ

Програми розпізнавання тексту не можуть повноцінно та правильно розпізнавати масиви тексту згідно вимог сучасного користувача. На даний момент такі програми не є інновацією та мають досить широкий спектр використання, однак якість їх роботи сумнівна. В умовах тотального переходу до використання електронних баз даних, електронних бібліотек, хмарних сховищ даних тощо програми розпізнавання тексту набувають небаченої актуальності, тому їх функціонал потребує суттєвого доопрацювання.

Новизна та наукові здобутки авторів виявляються в аналізі можливостей сучасних найпопулярніших програм розпізнавання тексту, а також наданні рекомендацій необхідних нових можливостей таких програм для комфортного та зручного їх використання сучасним користувачем.

Програми розпізнавання тексту дозволяють сканувати друкований текст і переводити його в цифровий формат. Ці програми підходять студентам, робітникам, що працюють з документами, перекладачам текстів, будь-якій людині, якій потрібно оцифрувати текст.

На даний момент програми розпізнавання тексту мають досить простий функціонал. Переважна більшість таких програм розпізнає лише друкований текст виконаний однією мовою та відформатований стандартним способом (одна колонка, жодних таблиць).

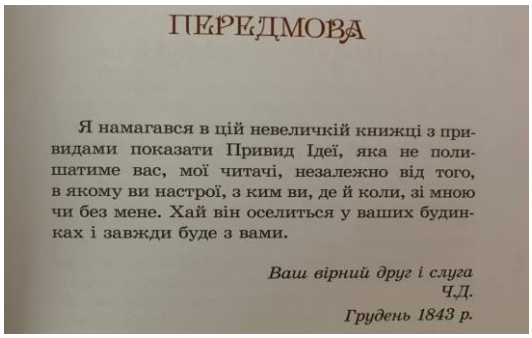
Програми сканування тексту використовують *OCR*-зчитувач — технологію оптичного розпізнавання символів. Процес розпізнавання включає вивчення тексту і переклад символів в код, який можна використовувати для обробки даних. Отримані електронні копії користувачі можуть редагувати, форматувати за допомогою звичайних редакторів тексту. Процес перекладу в цифровий формат працює за таким принципом: фотографія розбивається на безліч фрагментів, для кожного з них додаток створює кілька варіантів. Символи перевіряються і порівнюються між собою, знаходяться збіги. Так програма ідентифікує символ і виводить його в поле вбудованого текстового редактора.

OCR діє за об'єднаною технологією апаратного і програмного забезпечення. Апаратне забезпечення (оптичний сканер або спеціалізована монтажна плата) слугує для копіювання або читання тексту, в той час як програмне забезпечення відповідає за розширену обробку. *OCR* працює досконаліше з системою інтелектуального розпізнавання (*ICR*). *ICR* дозволяє ідентифікувати різні мови та стилі рукописних текстів. Але через сканування тексту через фотографії можуть виникнути деякі проблеми, наприклад, спотворення перспективи, засвічення від фотоспалаху, вигини рядків. При роботі з більшістю додатків такі дефекти можуть істотно ускладнити процес розпізнавання. Для усунення цих проблем останні версії програм сканування містять технології попередньої обробки зображення, що пришвидшить та поліпшить процес розпізнавання.

Зазвичай найкращі програми є платними, тому недоступними для простого користувача. Дана робота базується на доступних безкоштовних програмах розпізнавання тексту.

Aspose OCR — це безкоштовний онлайн інструмент, що дозволяє розпізнати текст на фотографіях і переводити його в електронний варіант. Відсканований текст доступний для подальшого редагування у середовищах *Word, Pdf, Excel* і багато та багатьох інших. Є функціонування перекладу тексту, без смислових втрат, втрати змісту, також є розпізнавання тексту на різних мовах з фото формату *JPG, BMP, TIFF, PNG* та інші.

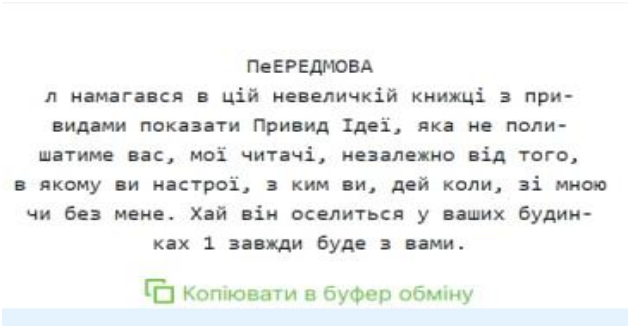
Для дослідження була обрана передмова книги «Різдвяна історія» Чарльза Діккенса (рис. 1а). Виявлені наступні дефекти: помилкове розпізнавання великої літери «Я», тексту, написаного курсивом (рис. 1а-б).



ПЕРЕДМОВА

Я намагався в цій невеличкій книжці з привидами показати Привид Ідеї, яка не полишатиме вас, мої читачі, незалежно від того, в якому ви настрої, з ким ви, де й коли, зі мною чи без мене. Хай він оселиться у ваших будинках і завжди буде з вами.

Ваш вірний друг і слуга
Ч.Д.
Грудень 1843 р.



ПеЕРЕДМОВА
л намагався в цій невеличкій книжці з привидами показати Привид Ідеї, яка не полишатиме вас, мої читачі, незалежно від того, в якому ви настрої, з ким ви, дей коли, зі мною чи без мене. Хай він оселиться у ваших будинках 1 завжди буде з вами.

Копіювати в буфер обміну

а) б)

Рис. 1. а) фотографія тексту; б) текст, розпізнаний *Aspose OCR*

Free Online OCR — безкоштовний веб-сервіс для перенесення тексту з фото на електронний ресурс, який підтримує 106 мов. Крім 10 форматів графічних зображень, обробляє документи *pdf*, *djvu*, *doxc*, *odt*, архіви *zip* і стислі файли *Unix* та може зберегти вихідні файли в одному з 3 форматів: *txt*, *doc* і *pdf*. Є доступна можливість відразу перевести текст на іншу мову, використовуючи *Google Translate* або *Bing Translator*. Суттєвим недоліком є обмежена безкоштовна обробка (лише 20 сторінок).

Pdf24.org — проект німецької компанії *Geek Software GmbH*. *PDF24* пропонує безкоштовні та прості у використанні рішення *PDF* для багатьох проблем, в тому числі і сканування тексту з фото.

При практичному користуванні також залишилися помилки в розпізнаванні символів і текст «Передмова» не був розпізнаний. Також недоліком можна вважати, що текст потрібно копіювати самостійно (рис. 2).

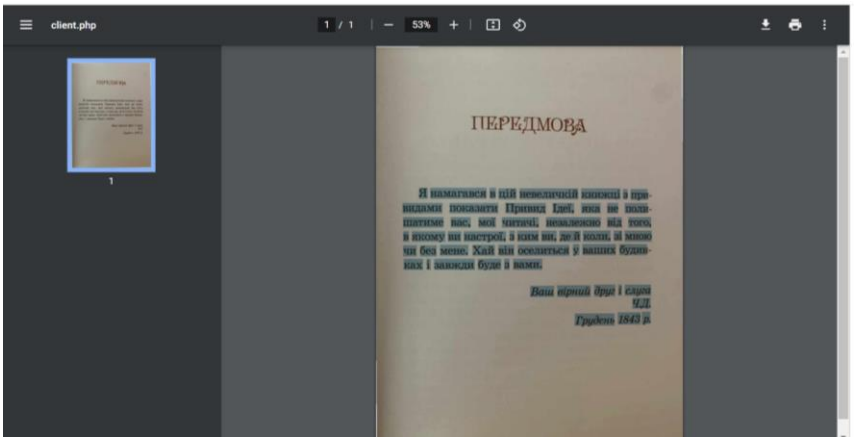


Рис. 2. Результат розпізнавання тексту *Pdf24.org*

Отже, програми розпізнавання тексту неякісно розпізнають текст найпростішого стандартного форматування. При опрацьовуванні двохколонкового тексту вихідний файл буде містити одну колонку зі змішаним з обох колонок вхідного файлу текстом; тобто програма не розпізнає форматування тексту декількома колонками. Аналогічна ситуація з таблицями: всі колонки таблиці зводяться в один рядок. Рукописний текст програми розпізнавання тексту визначають як зображення та не розпізнають. Текст, виконаний декількома

мовами, зазвичай, програми не ідентифікують. Переважно, програми розпізнають лише одну обрану мову, рідше — дві; текст у вхідному файлі, виконаний мовою, яка не була зазначена в програмі, у вихідному файли або не розпізнається взагалі та відображується як зображення, або відображується як нечитабельні символи.

Таким чином, на даний момент програми розпізнавання тексту не відповідають вимогам сучасного користувача та потребують істотних змін, а саме: можливості розпізнавання рукописного тексту, відформатованого тексту, тексту виконаного декількома мовами, табличного тексту.

Список використаних джерел

1. Зробити більше: Що таке оптичне розпізнавання символів (*OCR*)? – 2021 [Електронний ресурс] – Режим доступу до ресурсу: https://uk.go-travels.com/74733-optical-character-recognition-4158322-8335115.
2. Перетворити фото в текст онлайн [Електронний ресурс] – Режим доступу до ресурсу: https://products.aspose.app/ocr/uk/photo-ocr.
3. *Online ocr* [Електронний ресурс] – Режим доступу до ресурсу: https://www.onlineocr.net/.
4. Конвертувати зображення в *PDF* [Електронний ресурс] – Режим доступу до ресурсу: https://tools.pdf24.org/uk/images-to-pdf.