НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ

Факультет кібербезпеки, комп’ютерної та програмної інженерії

кафедра Комп’ютеризованих систем захисту інформації

**ЗВІТ**

**З ЛАБОРАТОРНОЇ РОБОТИ №\_1**

**З ДИСЦИПЛІНИ «Основи автоматизованої обробки інформації»**

**OCR-системи. Сканування інформації. Етапи сканування текстової та табличної інформації та графічної інформації.**

Роботу виконав студент:

Білецький Дмитро Олександрович

1\_курс, група \_142

  захищено з оцінкою\_\_\_\_\_

             Викладач:

Бурбела О. О.

                                          \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

                                                                  (підпис)

                    “\_\_”\_\_\_\_\_\_ 2019р.

КИЇВ 2019

**Мета роботи:** Отримати навички сканування і розпізнавання документів за допомогою програми ABBYY FineReader.

**Основні теоретичні відомості**

Системи оптичного розпізнавання символів (Optіcal Character Recognіtіon - OCR) призначені для автоматичного введення друкованих документів у комп'ютер. FіneReader - омніфонтова система оптичного розпізнавання текстів. Це означає, що вона дозволяє розпізнавати тексти, набрані практично будь-якими шрифтами, без попереднього навчання. Особливістю програми FіneReader є висока точність розпізнавання і мала чутливість до дефектів друку, що досягається завдяки застосуванню технології "цілісного цілеспрямованого адаптивного розпізнавання". Комп'ютер призначений для роботи з документами, що мають електронну форму. В той же час, нам часто доводиться мати справу з паперовими виданнями і документами: журналами, книгами, листами, службовими записками і так далі. Щоб в роботі з інформацією такого роду теж можна було використовувати комп'ютер, необхідні засоби перетворення паперових документів в електронну форму. Якщо передбачається, що документ містить в основному текстову інформацію, то можна виділити наступні основні етапи такого перетворення: •в ході сканування за допомогою пристроїв оцифрування зображення проводиться створення електронного образу (зображення) документу; •процес розпізнавання дозволяє перетворити електронне зображення в текстові дані (із збереженням елементів форматування оригіналу або без них). Всі операції, необхідні в ході перетворення паперового документу в електронну форму, можуть бути виконані за допомогою програми FineReader (рис. 2.1.). Ця програма здатна виконувати сканування і розпізнавання текстів на різних мовах, у тому числі і змішаних двомовних текстів. З її допомогою можна виконувати пакетну обробку багатосторінкових документів, а також настроювати режим розпізнавання для поліпшення відповідності електронного документа паперовому оригіналу при поганій якості останнього або використанні в нім шрифтів, далеких від стандартних. Основні операції обробки паперового документу в програмі FineReader виконуються за допомогою панелі інструментів Scan&Read. З погляду цієї програми, процес обробки документі складається з п'яти етапів: •сканування документу (кнопка Сканировать); •сегментація документу (кнопка Сегментировать); •розпізнавання документу (кнопка Распознать); •редагування і перевірка результату (кнопка Проверить); •збереження документу (кнопка Сохранить).

**2.1. Сканування документів**

Процес створення електронного зображення паперового документу нагадує його фотографування і вимагає застосування відповідного пристрою. Сьогодні в якості такого пристрою виступає сканер. Такі пристрої, як цифрові камери, поки не можуть забезпечити для документів стандартного формату якість зображення, яке гарантувало б їх надійне розпізнавання. Різні моделі сканерів розуміють різні команди керування. Щоб уникнути різнобою, був прийнятий універсальний стандарт взаємодії сканера і додатків. Цей стандарт називається TWAIN. Додаток посилає команди драйверу TWAIN який перетворить їх в інструкції, розпізнавані конкретним сканером. Таким чином, для додатків конкретна модель сканера перестає мати значення. Операційна система Windows ХР підтримує

інтерфейс TWAIN, а всі сучасні сканери сумісні з ним і надають необхідні драйвери нижнього рівня.

Сканування за допомогою інтерфейсу TWAIN здійснюється таким чином. Спочатку слід включити сканер. Команда сканування знаходиться в додатку в меню Файл (наприклад, в програмі Рaint відповідний пункт називається З сканера або камери). Після вибору цієї команди відкривається діалогове вікно драйвера TWAIN вигляд якого залежить від моделі сканера. У цьому вікні задають параметри сканування: чорно-білий або кольоровий режим, дозвіл, корекція яскравості і контрастності. Більшість сканерів дозволяють також провести попереднє чорнове сканування з низьким дозволом і по його результатах точно задати область сканування — частину сторінки документа. Після налаштування всіх параметрів слід натиснути на кнопці Сканировать (напис на кнопці може бути іншій). Процес сканування відбувається автоматично, і зображення передається в додаток. Діалогове вікно драйвера TWAIN автоматично не закривається, так що, наприклад, в багатовіконних графічних редакторах (таких, як Adobe PhotoShop) можна відразу провести сканування декількох зображень.

Щоб почати сканування в програмі FineReader, необхідно включити сканер і натиснути на кнопці Сканировать на панелі інструментів Scan&Read. Cканування може проводитися як через драйвер TWAIN, так і в обхід нього. Перший спосіб використовують, коли потрібне точне налаштування параметрів сканування, коли документ включає кольорові ілюстрації, які необхідно зберегти, а також коли різні сторінки 24

багатосторінкового документу сильно розрізняються за якістю. Другий варіант забезпечує максимальну швидкість і зручність сканування. Вибір варіанту здійснюється за допомогою прапорця Показывать діалог TWAIN-драйвера сканера (Сервіс • Опції • Сканування).

Після закінчення сканування значки всіх оброблених сторінок відображаються у вікні Пакет. В основній частині робочої області з'являється вікно Изображение, що містить зображення поточної сторінки. Додавати сторінки в пакет можна не лише шляхом сканування, але і шляхом відкриття файлів із зображеннями, що є на комп'ютері.

**2.2. Розпізнавання документів.**

Етап розпізнавання документу полягає в перетворенні електронного зображення (фактично набору кольорових або чорно-білих крапок) в текстовий документ. Раніше для опису цього процесу зазвичай використовувався термін OCR (Optical Character Recognition), який відповідає одному з використовуваних методів. При такому підході відбувається «порівняння» елементу зображення з еталонними варіантами зображення символів, після чого вибирається найбільш відповідний символ. Цей підхід вимагає використання спеціального комплекту шрифтів, але дає найкращі результати. Сучасні алгоритми розпізнавання не прив'язані до конкретного зображення символів: програма, як і людина, здатна дізнаватися букви при будь-яких зображеннях (а іноді і при значних спотвореннях).

Під час розпізнавання спочатку в зображенні виділяються великі елементи тексту: колонки, абзаци, окремі текстові блоки (наприклад, підмалюнкові підписи), комірки таблиць. Цей етап називають сегментацією, він може виконуватися автоматично або

вручну. Після цього виконується автоматичний етап розпізнавання: блоки розбиваються на рядки, рядки — на окремі символи, кожен з яких розпізнається незалежно і поміщається в підсумковий текстовий документ. Всередині кожного блоку, виділеного на етапі сегментації, зберігається природний порядок розпізнавання. Блоки нумеруються, виходячи з порядку включення їх в документ. Програма FineReader спроможна, виконати сегментацію (визначити межі блоків) автоматично навіть для вельми складних сторінок. При цьому враховуються поля документу, просвіти між колонками, рамки. Якщо структура сторінки дуже складна або містить явні дефекти, яких не завжди вдається уникнути при скануванні, зручніше використовувати ручну сегментацію або ручне редагування результатів автоматичної сегментації. Блоки відображаються у вигляді кольорових прямокутників з номером в лівому верхньому кутку. Новий блок створюють перетягуванням миші по діагоналі прямокутника. Поточний блок позначається виділеною лінією, а його кути – прямокутними маркерами. За допомогою цих маркерів можна змінити розмір або положення блоку. Команди редагування блоків виведені на панель Изображение. Вони дозволяють: •об'єднати два блоки в один (Додати частину до блоку); •видалити фрагмент блоку (Видалити частину блоку); •змінити порядок нумерації блоків (Перенумерувати блоки); •змінити розбиття таблиці на комірки (Додати вертикаль, Додати горизонталь. Видалити лінії). Різні типи блоків обробляються програмою по-різному. Щоб змінити тип блоку, потрібно натиснути правою кнопкою миші в його межах і призначити новий тип за допомогою меню Тип блока в контекстному меню. Програма FineReader підтримує наступні типи блоків: •автоматичний (Зона розпізнавання) — на етапі розпізнавання аналізується і розбивається на блоки автоматично; •текстовий (Текст) — на етапі розпізнавання перетвориться в текст; •табличний (Таблиця) — є набір комірок, кожна з яких перетвориться в текст окремо; •зображення (Картинка) — включається в документ без змін як графічна ілюстрація, якщо формат збереження перетвореного документа допускає вставні об'єкти. Процес розпізнавання тексту після сегментації повністю автоматизований. Кнопка Распознать на панелі Scan&Read дозволяє автоматично виконати сегментацію і розпізнавання, хоча в ході сегментації враховуються обмеження, встановлені вручну. Під час процесу відображається діалогове вікно Распознать, що дозволяє перервати процес. Крім того, в цьому вікні відображаються повідомлення, які вказують на наявність проблем при розпізнаванні. Проблеми зазвичай викликаються невірними налаштуваннями або поганою якістю зображення, що розпізнається.

**2.3. Редагування документів.**

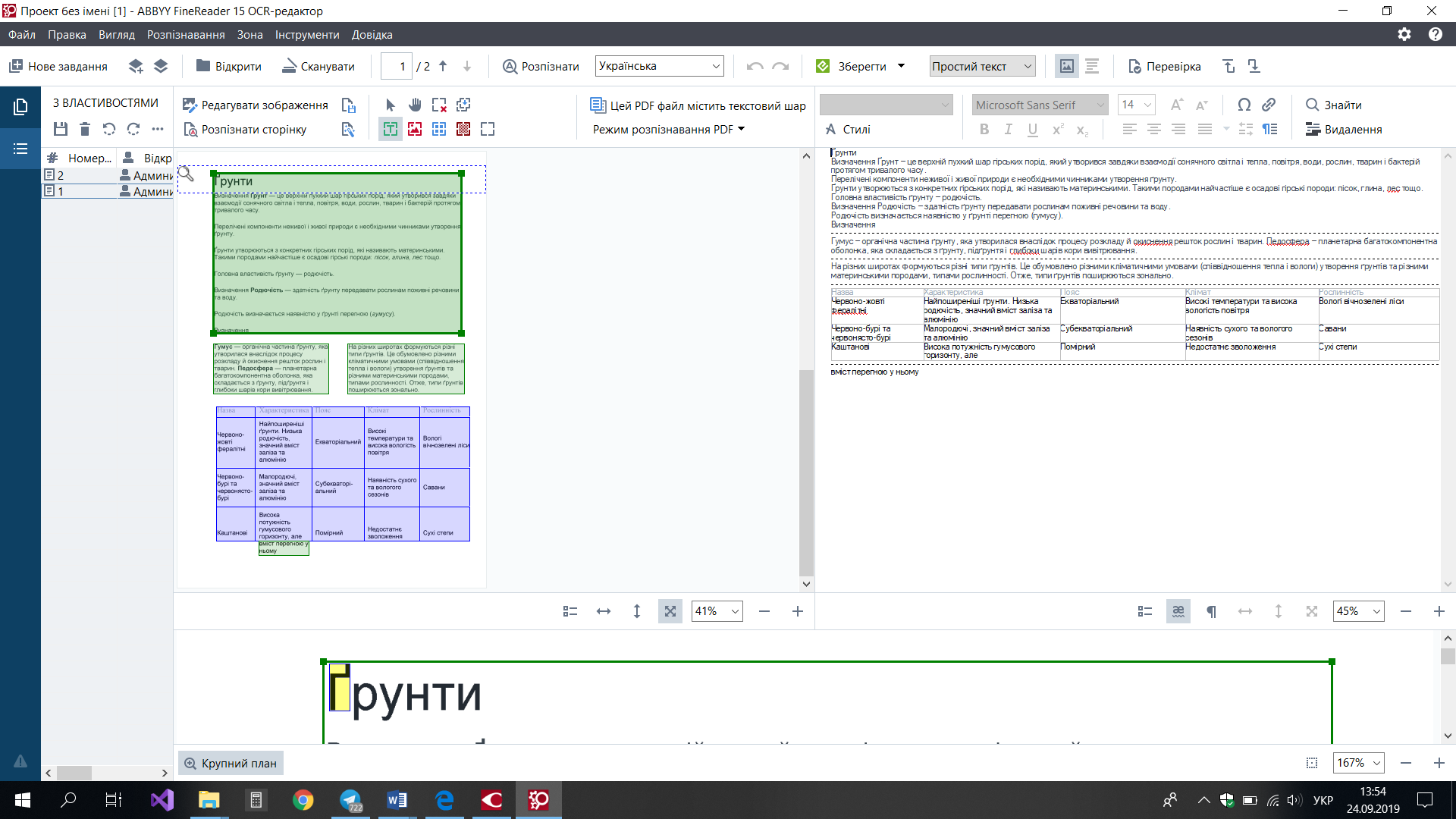
Коли розпізнавання даної сторінки завершується, текстовий документ відображається у вікні Текст. Завершальні етапи роботи дозволяють відредагувати отриманий текст за допомогою засобів, що нагадують текстовий редактор WordPad (панель для форматування відкривається за допомогою команди (Вид • Панель иструментов • Форматирование). Провести перевірку орфографії з урахуванням труднощів розпізнавання дозволяє кнопка Проверить на панелі інструментів Scan&Read. Інформація про можливі помилки і неоднозначності відображається в діалоговому вікні Проверка.

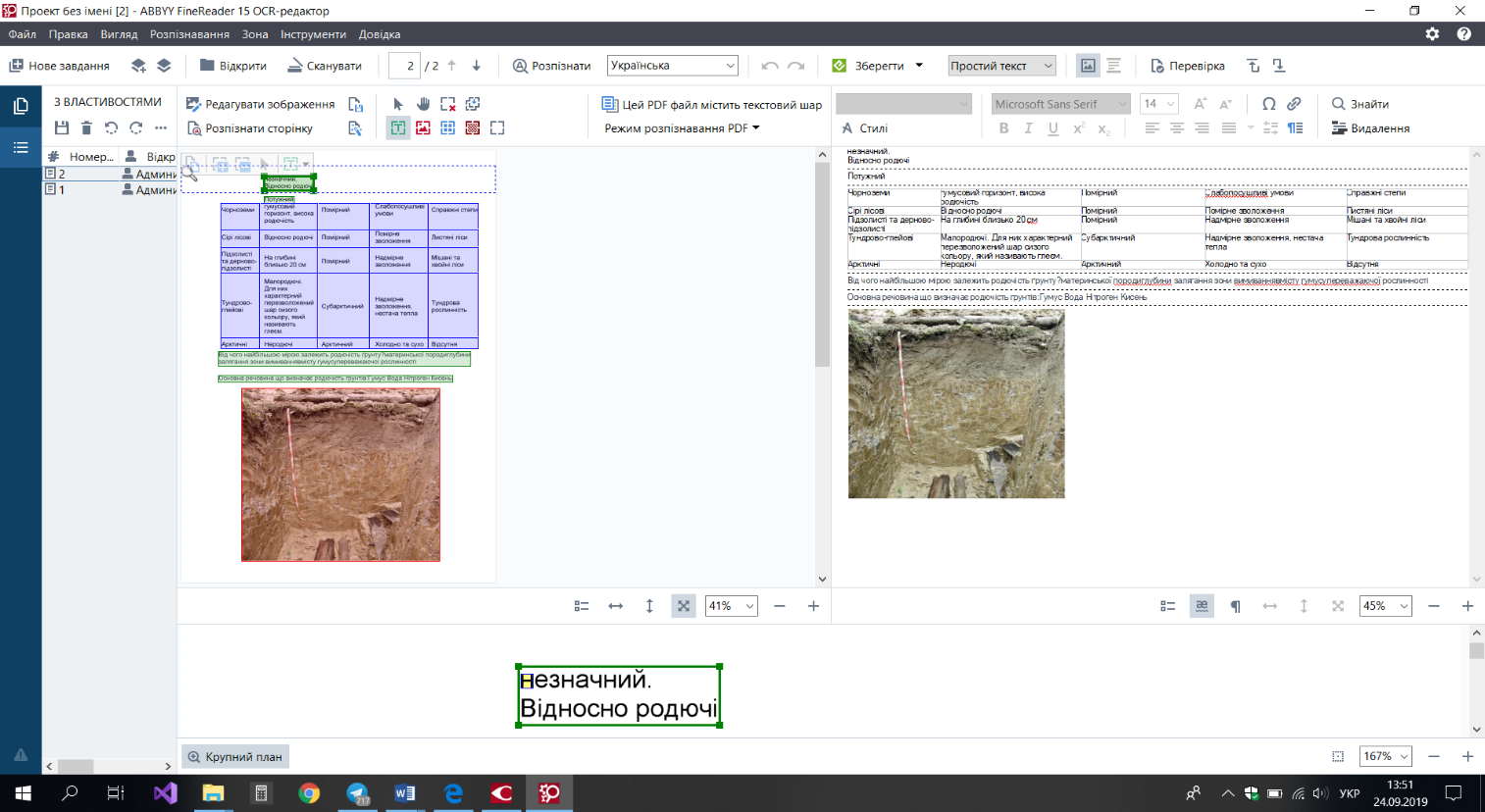
**2.4. Збереження документів.**

По натисненню на кнопці Сохранить на панелі інструментів Scan&Read запускається Мастер сохранения результатов. Він дозволяє зберегти розпізнаний текст або передати його в іншу програму (наприклад, в Microsoft Word) для подальшої обробки.

**Виконання лабораторної роботи**

1. Відсканував документ, який містить зображення і текст розбитий на

декілька колонок. Текст виділений зеленим, таблиці- фіолетовим, а зображеня- червоним.

2. Виконав функцію розпізнавання відсканованих сторінок.

3.1 Визначив правильні параметри розпізнавння.

3.2 Виконав сегментацію зображення сторінки.

3. Виконав функцію розпізнавання сторінок.

4. Перевірив правильність розпізнавання засобами ABBYY FineReader.

Та зробив висновки щодо dpi

* 100 dpi; 99%
* 200 dpi; 100%
* 300 dpi; 100%

5.Зберіг розпізнаний документ у файлі з форматом .doc, .pdf, .tiff.

**Висновок:** В результаті даної лабораторної роботи я навчився користуватись програмою ABBYY FineReader. Я отримав навички сканування і розпізнавання документів за допомогою даної програми.