Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: _____ Informatyka techniczna (ITE)
Specjalność: _____ Systemy informatyki w medycynie (IMT)

PRACA DYPLOMOWA Inżynierska

Implementacja systemu optycznego rozpoznawania znaków oraz opracowanie środowiska eksperymentalnego opartego o samodzielnie skonstruowany zbiór danych

Krzysztof Zalewa

Opiekun pracy
Dr inż., Paweł Zyblewski

Słowa kluczowe: 3-6 słów kluczowych

WROCŁAW (2025)

Streszczenie

Dodaj streszczenie pracy w języku polskim. Staraj się uwzględnić wymienione na stronie tytułowej słowa kluczowe. Uwaga przedstawiony rekomendowany szablon dotyczy pracy dyplomowej pisanej w języku angielskim. W przeciwnym wypadku, student powinien samodzielnie zmienić nazwy „Chapter" na „Rodział" itp stosując odpowiednie pakiety systemu LaTeXoraz ustawienia w pliku *latex-settings.tex*.

Abstract

Streszczenie w języku angielskim.

Spis treści

1. Wstęp

1.1. Opis problemu

1.2. Cel pracy

Test

Celem pracy jest porównanie, w oparciu o eksperymenty komputerowe, działania wybranych algorytmów optycznego rozpoznawania znaków. Kluczowym elementem pracy jest pozyskanie i opracowanie autorskiego zbioru danych, składającego się z treści dostępnych za pośrednictwem publicznie dostępnego API serwisu wolnelektury.pl, który pozwoli na rzetelną ewaluację znanych z literatury algorytmów OSR. Opracowany zbiór zawierał będzie dokumenty o zróżnicowanej charakterystyce, uwzględniając m.in. różne kroje i stopnie pisma, a także modyfikacje utrudniające poprawne odczytanie treści.

2. Przegląd literatury

2.1. Narzędzia OCR

2.1.1. Tesseract

Tesseract OCR to najstarszy z wybranych algorytmów optycznego rozpoznawania znaków. Został on stworzony przez firmę HP w latach 1984 - 1994. Algorytm ten działa w kilku fazach:

1. Analiza komponentów, gdzie zarys tych komponentów jest przechowywany. Takie podejście mimo że nakłada dodatkowe koszty obliczeniowe pozwala na łatwiejsze rozpoznawanie tekstu w odwróconych kolorach (biały tekst na czarnym tle) [3].

2. Wyszukiwanie lini w komponentach. Celem tego kroku była eliminacja potrzeby korekty przekrzywienia.

3. Podział linii na słowa.

4. Pierwsza iteracja rozpoznawania. Zaczynając na górze strony algorytm próbuje rozpoznać każde kolejne słowo. Jeżeli jest duże prawdopodobieństwo że słowo jest poprawne jest ono wykorzystywane do douczenia klasyfikatora. W ten sposób z każdym kolejnym słowem celność klasyfikatora powinna rosnąć.

5. Druga iteracja rozpoznawania. Po wykonaniu pierwszej iteracji jest duże prawdopodobieństow że klasyfikator uzyskałby lepsze wyniki. Więc po raz drugi algorytm próbuje rozpoznać tekst na stronie i aktualizuje słowa które były minej celnie rozpoznane.

2.1.2. Easy OCR

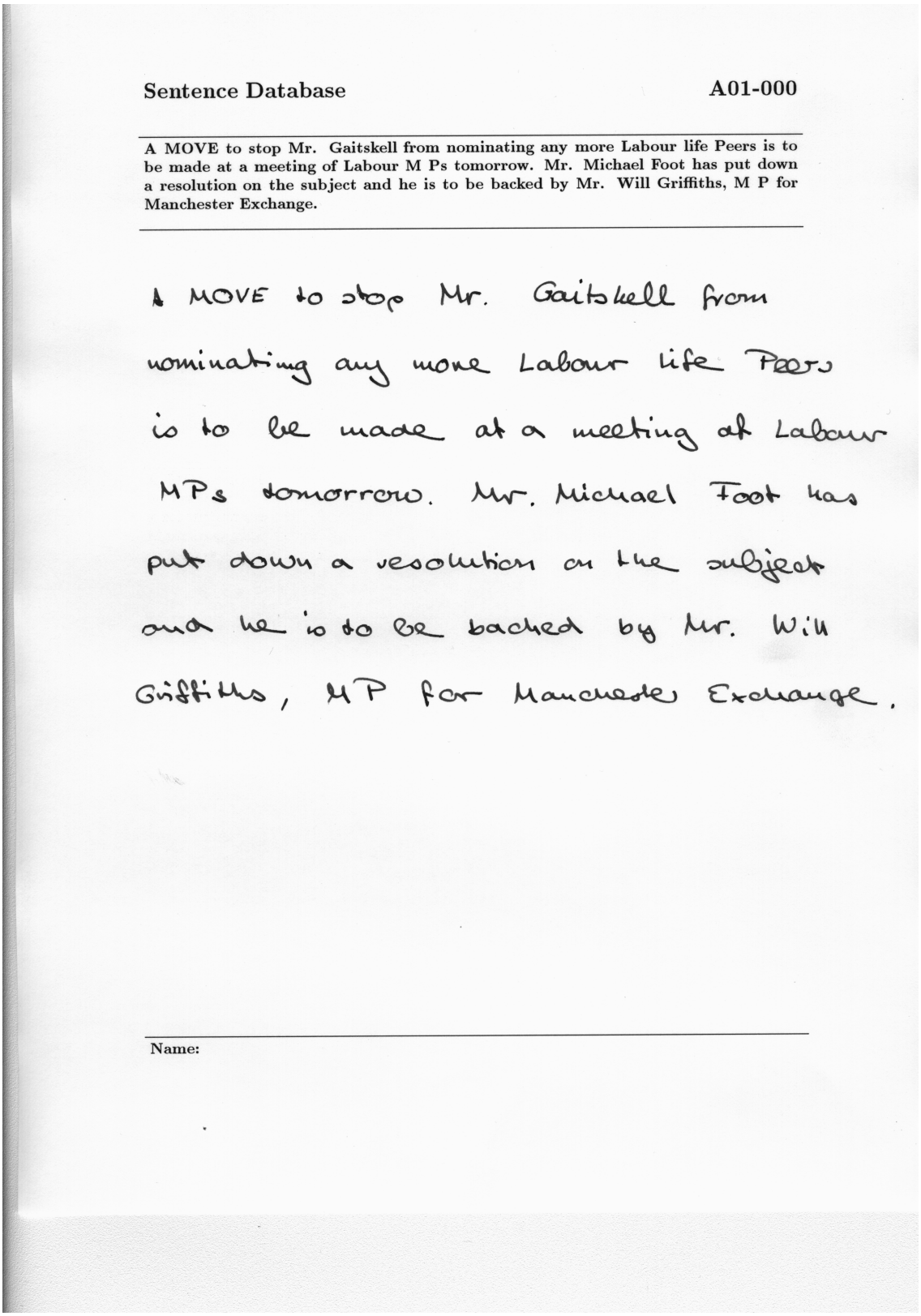
2.1.3. DocTR OCR

2.1.4. Paddle OCR

2.2. Zbiory danych

2.2.1. IAM

IAM to zbiór ręcznie zapisanych tekstów w języku angielskim. Wykonany przez Instytut matematyki i informatyki na Uniwersytecie Breńskim. [2] Zbiór zawiera obrazy w rozdzielczości 300dpi zapisane w formacie PNG w 256 odcieniach szarości. Każdy pod katalog zawiera teksty zapisane przez jedną osobę.



Sentence Database **A01-000**

A MOVE to stop Mr. Gaitskell from nominating any more Labour life Peers is to be made at a meeting of Labour M Ps tomorrow. Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, M P for Manchester Exchange.

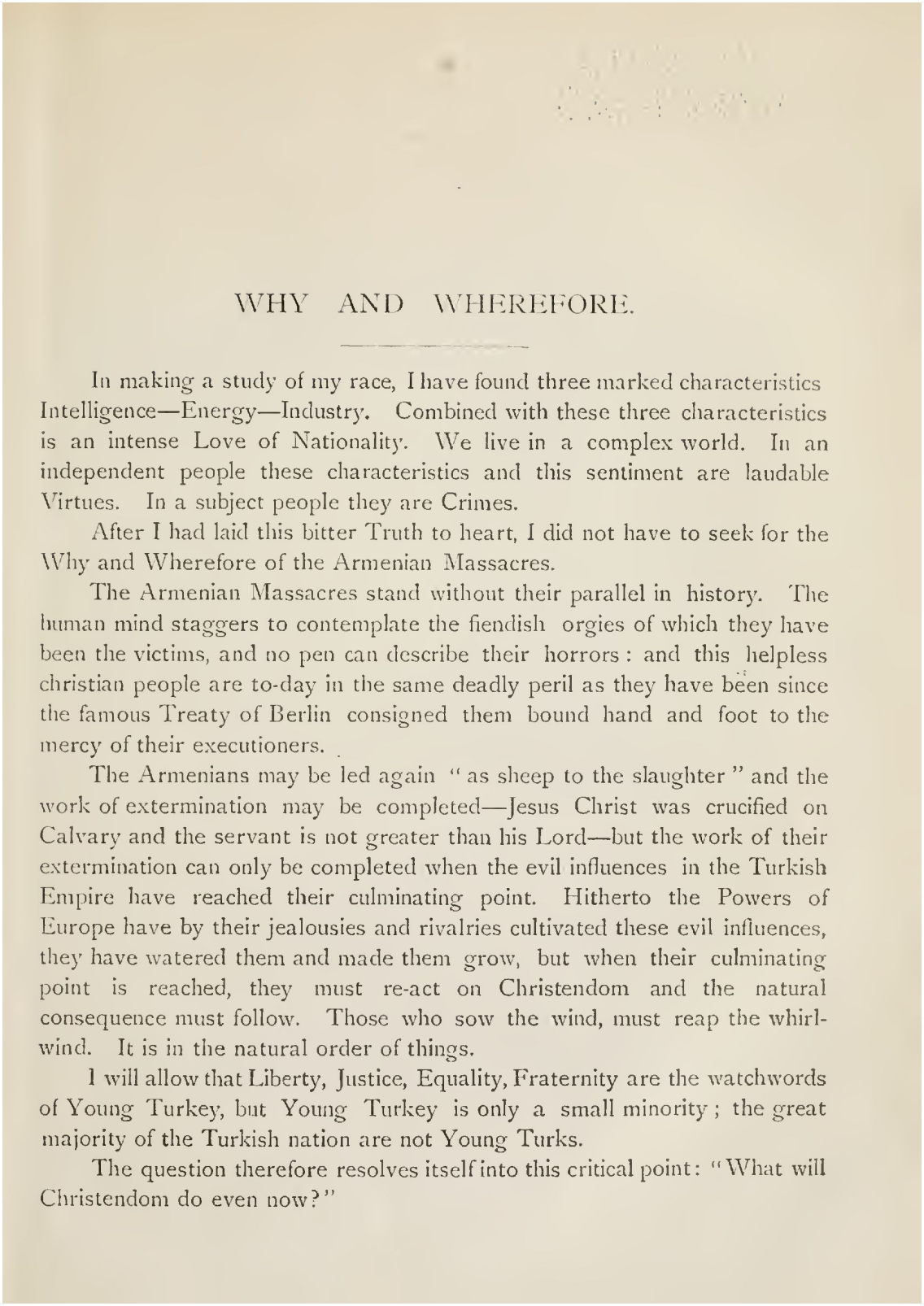
A MOVE to stop Mr. Gaitskell from nominating any more Labour life Peers is to be made at a meeting of Labour MPs tomorrow. Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, MP for Manchester Exchange.

Name:

Figure 2.1: Przykładowy obraz ze zbioru danych IAM

2.2.2. oldbooksdataset

Zbiór udostępniony na platformie git hub zawierający skany książek w języku angielski. Książki zapisane są w formacie .tiff w rozdzielczości 300dpi oraz 500dpi [1].



WHY AND WHEREFORE.

In making a study of my race, I have found three marked characteristics Intelligence—Energy—Industry. Combined with these three characteristics is an intense Love of Nationality. We live in a complex world. In an independent people these characteristics and this sentiment are laudable Virtues. In a subject people they are Crimes.

After I had laid this bitter Truth to heart, I did not have to seek for the Why and Wherefore of the Armenian Massacres.

The Armenian Massacres stand without their parallel in history. The human mind staggers to contemplate the fiendish orgies of which they have been the victims, and no pen can describe their horrors: and this helpless christian people are to-day in the same deadly peril as they have been since the famous Treaty of Berlin consigned them bound hand and foot to the mercy of their executioners.

The Armenians may be led again "as sheep to the slaughter" and the work of extermination may be completed—Jesus Christ was crucified on Calvary and the servant is not greater than his Lord—but the work of their extermination can only be completed when the evil influences in the Turkish Empire have reached their culminating point. Hitherto the Powers of Europe have by their jealousies and rivalries cultivated these evil influences, they have watered them and made them grow, but when their culminating point is reached, they must re-act on Christendom and the natural consequence must follow. Those who sow the wind, must reap the whirlwind. It is in the natural order of things.

I will allow that Liberty, Justice, Equality, Fraternity are the watchwords of Young Turkey, but Young Turkey is only a small minority; the great majority of the Turkish nation are not Young Turks.

The question therefore resolves itself into this critical point: "What will Christendom do even now?"

Figure 2.2: Przykładowy obraz ze zbioru danych old-books-dataset

2.3. Metryki

Do oceny wyników zastosowano dwie główne metryki. CER i WER zasadniczą różnicą między nimi jest celność porównań. CER porównuje na poziomie pojednyńczych znaków natomiast, z kolei WER porównuje na poziomie poszczególnych słów. Z tego powodu CER jest przydaniejsze w kontektstach w których nawet pojedyńczy znak może zmienić znaczenie słowa (np. w medycynie). Natomiast WER jest lepsze do porównywania spójności zdań itp.

2.3.1. CER

CER (Character Error Rate z ang. częstotliwość błędnych znaków) to metryka dzięki której możliwa jest ocena różnic między tekstem wytworzonym poprzez model OCR a tekstem rzeczywistym. W tym przypadku CER obliczane jest poprzez zsumowanie operacji(wstawień, usunięć oraz zamian znaków) potrzebnych do uzyskania tekstu rzeczywistego.

$$CER = \frac{S + D + I}{N_c}$$

Gdzie:

- S - Liczba zamian znaków (ang. Substitutions)
- D - Liczba usunięć znaków (ang. Deletions)
- I - Liczba wstawień znaków (ang. Inserts)
- N_c - Liczba znaków w tekście (ang. Number of characters)

Na przykład

Tekst orginalny: Życiem wschód, śmiercia południe;
Tekst wygenerowany przez model: Zyciem wschod, siercia poudniex;

Aby przekształcić tekst wygenerowany do tekstu orginalnego należy wykonać 4 zamiany (Brakujące znaki polskie), 1 wstawienie (Brakujące 'm' w tekście wygenerowanym) oraz 1 usunięcie ('x' nie występuje w tekście orginalnym). Więc $CER = 6/28 = 0.2141 \approx 21.4\%$

2.3.2. WER

WER (Word Error Rate z ang. częstotliwość błędnych słów) podobnie jak CER jest to metryka dzięki której możliwa jest ocena różnic między tekstem wytworzonym poprzez model OCR a tekstem rzeczywistym. Jak sama nazwa wskazuje WER porównuje tekst na poziomie poszczególnych słów.

$$WER = \frac{S + D + I}{N_w}$$

Gdzie:

- S - Liczba zamian słów (ang. Substitutions), czyli słowa które występują w obu tekstach ale te w tekście są różne od tych w tekście orginalnym.
- D - Liczba usunięć słów (ang. Deletions), czyli słowa które występują w tekście orginalnym jednakże nie ma ich w tekście wygenerowanym.
- I - Liczba wstawień słów (ang. Inserts), czyli słowa nadmiarowe których nie ma w tekście orginalnym.
- N_w - Liczba słów w tekście (ang. Number of words)

Na przykład

Tekst orginalny: Życiem wschód, śmiercia południe;
Tekst wygenerowany przez model: Zyciem wschod, siercia poudniex;

Aby przekształcić tekst wygenerowany do tekstu orginalnego należy wykonać 4 zamiany(Słowa zbliżone do orginału ale nie takie same). Więc WER=4/4=1=100%

3. Aspekt inżynierski

Implementacja wybranych, znanych z literatury metod optycznego rozpoznawania znaków (ang. optical character recognition <OSR>). Opracowanie autorskiego zbioru danych – na podstawie treści samodzielnie pozyskanych z serwisu wolnelektury.pl z wykorzystaniem dostępnego publicznie API – na potrzeby ewaluacji zaimplementowanych algorytmów. Całość implementacji zostanie wykonana w języku Python z wykorzystaniem właściwie dobranych bibliotek programistycznych. Dodatkowo, na potrzeby ewaluacji, opracowane oraz zaimplementowane zostanie odpowiednie środowisko eksperymentalne.

3.1. Wybór algorytmów OCR

3.2. Akwizycja danych

4. Aspekt badawczy

4.1. Opis problemu

5. Podsumowanie

Bibliografia

[1] P. Barcha. Old books dataset. https://github.com/PedroBarcha/old-books-dataset, 2024. Accessed: 2024.

[2] U. Marti and H. Bunke. The iam-database: An english sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[3] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633, 2007.

Spis ilustracji

Spis tabel