Master Thesis

Michał Zaręba
196218

Automated Extraction and Categorization of Product Information from Receipts Using Tesseract OCR

Diploma thesis in the field of
Information Science

Thesis under the supervision of
dr hab. inż. Leszek Chmielewski, prof. SGGW
Institute of Information Technology
Department of Artificial Intelligence

Warsaw, year 2017

WARSAW
UNIVERSITY
OF LIFE SCIENCES

Faculty of Applied
Informatics and
Mathematics

Declaration of the Thesis Supervisor

I hereby declare that this thesis has been prepared under my supervision, and I confirm that it meets the conditions for presenting this work in the procedure for the award of a professional title.

Date ..... Supervisor's signature .....

Declaration of the Author of the Thesis

Aware of legal liability, including criminal liability for submitting a false declaration, I hereby declare that this diploma thesis was written by myself and did not contain the content obtained in a manner inconsistent with applicable law, in particular the Act of 4 February 1994 on copyright and related rights (Journal of Laws of 2019, item 1231, as amended).

I declare that the submitted work has not previously been the basis for any procedure related to awarding a diploma or obtaining a professional title.

I certify that this work version is identical to the attached electronic version.

I acknowledge that the diploma thesis will be subject to the procedure of the anti-plagiarism.

Date ..... Author's signature .....

Streszczenie

OCR + BERT

Tematem niniejszej pracy było zaimplementowanie klasy LaTeXowej pozwalającej na formatowanie tekstu zgodnie z wytycznymi nałożonymi przez uczelnię. Praca zawiera dwie główne części. Pierwsza z nich zawiera opis najważniejszych aspektów implementacji klasy. Natomiast druga część skupia się na sposobie użycia klasy przez osoby piszące prace dyplomowe.

Słowa kluczowe – OCR, BERT, Tesseract, thesis, implementation, SGGW, Warsaw University of Life Sciences

Summary

OCR + BERT

The subject of this thesis was to implement a LaTeX class that allows formatting text according to the guidelines imposed by the university. The thesis contains two main

Keywords – LaTeX, class, thesis, implementation, SGGW, Warsaw University of Life Sciences

Contents

1 Introduction

1.1 Motivation

Tracking of expenses and managing personal finances is an important aspect of modern life. With an increasing number of daily transactions and a vast variety of products available, individuals face significant challenges in effectively monitoring their spending and managing their budgets. Although receipts contain valuable details that could help consumers analyze and control their expenses, the majority of consumers either discard receipts shortly after purchase or find it too tedious and time-consuming to analyze them manually. Automating the extraction and categorization of product information from receipts could significantly simplify budget tracking and provide insights into spending habits, enabling consumers to understand precisely where their money goes. Simple and efficient way to track expenses is essential for individuals who wish to maintain a clear overview of their spending habits and make informed financial decisions.

1.2 Problem Statement

Most existing expense-tracking solutions focus primarily on invoices, bank statements, or require manual input. Large corporations and organizations typically possess the necessary budgets and technical resources to implement robust, automated systems for extracting and categorizing expense data from structured documents such as invoices or bank statements. For personal use, however, the most commonly available and practical source of spending information remains paper receipts. Current receipt-based solutions are often limited: many tools available today are either designed exclusively for commercial purposes, lack support for languages other than English, or are inadequately trained to accurately process Polish-language receipts. Thus, there is a clear gap and a significant need for a solution that effectively automates extraction and categorization of product details from receipts, specifically accommodating the complexity and linguistic characteristics of the Polish language.

1.3 Objectives of the Study

This study has two primary objectives, each directly addressing the challenges identified in the problem statement:

1. Develop a robust system capable of automatically extracting structured product information (such as product names, and prices) from Polish-language receipts using Optical Character Recognition (OCR).
2. Implement and evaluate a product categorization module based on the Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned specifically to classify extracted product information into predefined product categories relevant to personal expense management.

These objectives will be thoroughly addressed and analyzed in subsequent chapters. Given the complexity of Polish-language receipts and limited availability of labeled datasets, achieving optimal results will require careful integration and fine-tuning of multiple technologies. Critical aspects will include the effective integration of OCR and NLP components, targeted preprocessing strategies for receipts, fine-tuning Tesseract OCR for enhanced accuracy in Polish text recognition, and adapting the BERT model to handle the nuances of Polish-language product descriptions for precise categorization.

1.4 Scope and Limitations

The scope of this study is limited to the development of a system capable of automatically extracting product information from receipts and categorizing these products into predefined categories. The system specifically targets Polish-language receipts and will be evaluated primarily on its ability to accurately extract product costs and perform correct product categorization.

The limitations of this study include the following:

- The developed system will not include additional functionalities such as expense tracking over time, financial report generation, or integration with external personal financial management tools.
- The scarcity of comprehensive, labeled Polish-language receipt datasets restricts the potential accuracy and generalization capabilities of the models developed. Consequently, results may vary when encountering receipt formats or text variations not present in the training data.
- Tesseract OCR will be employed without utilizing spatial information or context regarding the positioning of text on receipts. Therefore, preprocessing steps such as image cropping, alignment, and noise reduction are necessary to ensure the OCR engine receives properly formatted and isolated textual inputs.

2 Literature Review

2.1 Optical Character Recognition (OCR) Technologies

[...]

2.2 Natural Language Processing (NLP) and BERT

[...]

2.3 Receipt Data Extraction Techniques

[...]

2.4 Integration of OCR and NLP

[...]

3 Methodology

3.1 System Architecture

[...]

3.2 Data Collection and Preprocessing

[...]

3.3 Optical Character Recognition with Tesseract

[...]

3.4 Text Parsing and Information Extraction

[...]

3.5 Product Categorization Using BERT

[...]

3.6 Data Grouping and Organization

[...]

4 Implementation

4.1 Development Environment

[...]

4.2 Integration of Tesseract OCR

[...]

4.3 BERT Model Fine-Tuning

[...]

4.4 System Workflow

[...]

5 Evaluation and Results

5.1 Evaluation Metrics

[...]

5.2 Experimental Setup

[...]

5.3 Results and Analysis

[...]

5.4 Comparison with Existing Methods

[...]

6 Discussion

6.1 Interpretation of Results

[...]

6.2 Challenges and Limitations

[...]

6.3 Recommendations for Future Work

[...]

7 Conclusion

7.1 Summary of Findings

[...]

7.2 Contributions to the Field

[...]

7.3 Final Remarks

[...]

8 References

[...]

9 Appendices

9.1 Sample Receipt Data

[...]

9.2 Code Snippets

[...]

9.3 Additional Figures and Tables

[...]

Bibliography

[1] *Zarządzenie nr 34 Rektora Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie z dnia 01 czerwca 2016 r. w sprawie wprowadzenia „Wytycznych dotyczących przygotowywania prac dyplomowych w Szkole Głównej Gospodarstwa Wiejskiego w Warszawie”*, Załączniki 1 i 2 `https://www.sggw.pl/dla-studentow/informacje-formalno-prawne/dokumenty-do-pobrania` → Praca dyplomowa (dostęp: 04.01.2017)

Wyrażam zgodę na udostępnienie mojej pracy w czytelniach Biblioteki SGGW w tym w Archiwum Prac Dyplomowych SGGW.

.....
(czytelny podpis autora pracy)