Politechnika
Śląska

PRACA MAGISTERSKA

Moduł do obsługi importów danych finansowych z plików PDF

Mateusz MARCZEWSKI
Nr albumu: 282700

Kierunek: Informatyka
Specjalność: internet i technologie sieciowe

PROWADZĄCY PRACĘ
Dr hab. inż. Arkadiusz Biernacki
KATEDRA Katedra Sieci i Systemów Komputerowych
Wydział Automatyki, Elektroniki i Informatyki

Gliwice 2023

Tytuł pracy

Moduł do obsługi importów danych finansowych z plików PDF

Streszczenie

(Streszczenie pracy – odpowiednie pole w systemie APD powinno zawierać kopię tego streszczenia.)

Słowa kluczowe

(2-5 slow (fraz) kluczowych, oddzielonych przecinkami)

Thesis title

Module for importing inventory data from PDF files

Abstract

(Thesis abstract – to be copied into an appropriate field during an electronic submission – in English.)

Key words

(2-5 keywords, separated by commas)

Spis treści

1 Wstęp **1**

1.1 wprowadzenie w zagadnienie 1

1.2 cel pracy 1

1.3 charakterystyka rozdziałów 2

2 Definiowanie wymagań **5**

2.1 Formaty plików występujące w pracy 5

2.1.1 Pliki PDF 5

2.1.2 Pliki CSV 6

2.2 Przegląd typów danych do importowania 7

3 Badanie istniejących rozwiązań **9**

3.1 Przegląd istniejących modułów i bibliotek do importowania danych z PDF 9

3.2 Funkcje istniejących rozwiązań 9

3.3 Oceny użytkowników i opinie na temat istniejących rozwiązań 9

4 Ocena dostępnych opcji **11**

4.1 Porównanie istniejących rozwiązań na podstawie zdefiniowanych wymagań 11

5 Rozwój istniejących rozwiązań **13**

5.1 Rozważania dotyczące budowy niestandardowego modułu 13

5.2 Zasoby i umiejętności wymagane do rozwoju i utrzymania modułu 13

6 Rozwój wybranego rozwiązania **15**

6.1 Testowanie wybranego rozwiązania 15

6.2 Udoskonalenie wybranego rozwiązania na podstawie napotkanych problemów 15

7 Dokumentacja i udostępnianie wybranego rozwiązania **17**

7.1 Dokumentacja rozwiązania 17

7.2 Udostępnianie rozwiązania jako projektu open-source 17

8 Podsumowanie **19**
8.1 Podsumowanie badania 19
8.2 Implikacje dla przyszłej pracy 19

Bibliografia **21**

Spis skrótów i symboli **25**

Lista dodatkowych plików, uzupełniających tekst pracy (jeżeli dotyczy) **27**

Spis rysunków **29**

Spis tabel **31**

Rozdział 1

Wstęp

1.1 wprowadzenie w zagadnienie

Zarządzanie dostępnymi zasobami jest niezbędnym procesem dla dużych biznesów. Proces ten składa się z czynności takich jak, śledzenie ilości, lokalizacji i wartości towarów posiadanych przez firmę. Jest to niezbędna czynność dla utrzymania wydajności finansowej firmy oraz maksymalizacji zysków. Zważając na opisane operacje można wyszczególnić duże wyzwanie stojące przed firmami chcącymi optymalnie zarządzać swoimi zasobami. Jest nim importowanie danych z różnych źródeł, takich jak faktury zamówień, potwierdzenia sprzedaży, dane o załadunkach i tym podobne, a następnie przechowywanie ich w bazie danych, dzięki czemu można łatwiej nimi zarządzać. Zazwyczaj takie dane będą dostępne w formie plików PDF, lub będzie możliwe ich do tej postaci zeskanowanie, co ułatwia ich przechowywanie i podgląd jednak utrudnia ekstrakcje danych z takich plików.

W celu poradzenia sobie z tym problemem, wiele firm wykorzystuje specjalnie przygotowane aplikacje, które pozwalają na automatyzację procesu importowania danych z plików PDF. Takie aplikacje mogą importować dane w postaci plików PDF, a następnie znajdywać w nich najważniejsze dane i eksportować je do żądanego formatu, obsługiwanego przez bazę danych lub inne aplikacje służące do zarządzania biznesem.

1.2 cel pracy

Celem tej pracy jest ocena różnych bibliotek oraz próba utworzenia własnego udoskonalonego rozwiązania, które umożliwi eksportowanie danych z plików o rozszerzeniu PDF, do łatwiej obsługiwanych plików wyjściowych. Prowadząc te badania, przeprowadzone zostaną testy na paru rozwiązaniach w celu porównania ich możliwości. Będziemy analizować zarówno biblioteki dostępne na rynku, jak i moduły open-source, które oferują funkcje eksportu danych z plików PDF. Przy ocenie tych rozwiązań będziemy kierować

się kryteriami takimi jak dokładność ekstrakcji danych, obsługiwane formaty plików wyjściowych, łatwość integracji, wydajność i wsparcie dla rozszerzeń PDF.

Końcowym celem tej pracy jest znalezienie najlepszej dla naszych wymagań biblioteki, która pozwoli stworzyć narzędzie, umożliwiające skuteczne poradzenie sobie z problemem importowania danych z plików PDF przez faktyczne firmy.

1.3 charakterystyka rozdziałów

W niniejszej pracy naukowej przedstawiamy szereg rozdziałów, które składają się na kompleksową analizę modułu do importowania danych inwentaryzacyjnych z plików PDF. Każdy z tych rozdziałów ma swoje unikalne znaczenie i koncentruje się na konkretnych aspektach badania. Poniżej przedstawiamy krótką charakterystykę poszczególnych rozdziałów:

Rozdział 2: Definiowanie wymagań

W tym rozdziale skupiamy się na ustaleniu konkretnych wymagań dla modułu lub biblioteki służącej do importowania danych z plików PDF. Analizujemy różne typy danych, które mają zostać zaimportowane, oraz identyfikujemy pola w plikach PDF, które będą przechwytywane. Przeanalizujemy również różne formaty plików, z którymi moduł będzie pracował, aby zapewnić wszechstronność i elastyczność rozwiązania.

Rozdział 3: Badanie istniejących rozwiązań

W tym rozdziale skoncentrujemy się na dokładnym przejrzeniu istniejących modułów i bibliotek dostępnych na rynku do importowania danych z plików PDF. Przeanalizujemy funkcje, jakie oferują te rozwiązania oraz przetestujemy ich możliwości. Celem tego rozdziału jest zgłębienie istniejących rozwiązań i dostrzeżenie ich mocnych i słabych stron w celu stworzenia solidnej podstawy do dalszych badań.

Rozdział 4: Ocena dostępnych opcji

W tym rozdziale przeprowadzimy szczegółową ocenę dostępnych opcji na podstawie wcześniej zdefiniowanych wymagań. Porównamy różne rozwiązania pod kątem ich zgodności z określonymi wymaganiami, aby wybrać najlepszą opcję dla dalszych badań i implementacji. Porównane zostaną również nasza opinia oraz opinie i oceny dostępne w artykułach naukowych, w celu poprawnej analizy rozwiązania.

Rozdział 5: Rozwój istniejących rozwiązań

W tym rozdziale skupimy się na rozwoju istniejących rozwiązań, takich jak dostosowanie ich do specyficznych potrzeb projektu. Przeanalizujemy również wymagane zasoby i umiejętności potrzebne do kontynuacji rozwoju i utrzymania modułu opartego na danej bibliotece.

Rozdział 6: Rozwój własnego rozwiązania

Ten rozdział koncentruje się na własnym rozwiązaniu, uwzględniając wyniki poprzednich analiz i badań. Przeanalizujemy wyniki, które udało się osiągnąć przy wykorzystaniu

własnego rozwiązania i porównamy je z istniejącymi na rynku.

Rozdział 7: Dokumentacja i udostępnianie wybranego rozwiązania

W tym rozdziale omówimy proces tworzenia dokumentacji dla wybranego rozwiązania, aby umożliwić innym użytkownikom skorzystanie z modułu. Przeanalizujemy również możliwość udostępnienia rozwiązania jako projektu open-source dla społeczności programistycznej.

Rozdział 8: Podsumowanie

W ostatnim rozdziale dokonamy podsumowania całego badania, podkreślając najważniejsze wnioski i rezultaty. Przedstawimy również implikacje wynikające z naszej pracy badawczej oraz sugestie dotyczące przyszłych kierunków rozwoju i badań w tej dziedzinie.

Każdy z tych rozdziałów przyczynia się do pełnego zrozumienia problemu i tworzy spójną strukturę pracy badawczej dotyczącej modułu do importowania danych inwentaryzacyjnych z plików PDF.

Rozdział 2

Definiowanie wymagań

2.1 Formaty plików występujące w pracy

W tym punkcie skupiamy się na analizie formatów plików, z którymi będziemy pracować, są nimi formaty PDf oraz CSV. Plik PDF (Portable Document Format) jest zbudowany zgodnie z określonymi specyfikacjami opracowanymi przez firmę Adobe Systems. Format ten ma na celu zapewnienie niezależności od platformy i zachowanie spójności formatowania, niezależnie od urządzenia, na którym jest wyświetlany. Natomiast CSV (Comma-Separated Values) to prosty format przechowywania danych tabelarycznych w postaci tekstowej. Jest szeroko stosowany do wymiany danych między różnymi aplikacjami i systemami, ze względu na swoją łatwość w odczycie i zapisie.

2.1.1 Pliki PDF

Pliki PDF mogą różnić się wewnętrzną strukturą, wersjami formatu, a także zastosowanymi technologiami i funkcjonalnościami. W celu skutecznego importowania danych, konieczne jest zrozumienie tych formatów i ich cech.

W pierwszej kolejności warto zwrócić uwagę na różnice między starszymi a nowszymi wersjami formatu PDF. Pliki PDF są stale rozwijane i udoskonalane, dlatego ważne jest, aby zrozumieć, jakie funkcje i możliwości są dostępne w różnych wersjach. Niektóre starsze wersje mogą nie obsługiwać niektórych zaawansowanych funkcji, co należy wziąć pod uwagę przy tworzeniu modułu importującego.

Kolejnym aspektem jest analiza możliwości skryptowania i interaktywności w plikach PDF. Niektóre pliki PDF mogą zawierać skrypty JavaScript lub elementy interaktywne, takie jak formularze, przyciski, linki itp. Istotne jest zrozumienie tych elementów i określenie, czy i w jaki sposób można je obsłużyć podczas importowania danych.

Dodatkowo, badamy w tym typie dokumentów wykorzystuje się różne technologie i standardy związane z potrzebami użytkownika. Są nimi standardy takie jak PDF/A (standard archiwizacji), PDF/X (standard dla publikacji drukowanych) czy PDF/UA

(standard dostępności). W zależności od specyfiki aplikacji i wymagań postawionych przed rozwiązaniem, konieczne może być uwzględnienie tych standardów i zapewnienie zgodności modułu importującego.

Plik PDF składa się z różnych elementów i struktur, które razem tworzą dokument. Jego strukturę możemy podzielić na 4 części:

- Header - Jest to sekcja zawierająca informacje o strukturze i metadanych pliku PDF, takich jak wersja specyfikacji PDF, typ pliku, używane czcionki, rozmiar strony itp.
- Body - To główna część pliku PDF, która zawiera treść dokumentu, taką jak tekst, obrazy, grafiki, tabele itp. Ciało składa się z sekwencji obiektów PDF, które są odpowiedzialne za przechowywanie danych i struktury dokumentu.
- Xref table - Jest to sekcja pliku PDF, w której znajduje się spis wszystkich obiektów, używanych w dokumencie. Każdy obiekt ma unikalny numer identyfikacyjny i jest przechowywany w formacie klucz-wartość.
- Trailer - Ta sekcja zawiera informacje o lokalizacji i numerze generacji wszystkich obiektów w pliku PDF. Jest to wykorzystywane do odnalezienia i odzyskania obiektów podczas odczytu lub modyfikacji pliku.

Trudność przy obsłudze plików typu PDF jest zauważalna przy przeanalizowaniu tej struktury. Zazwyczaj pliki te odczytuje się od tyłu, wynika to z tego, że na końcu zawarte są informacje o wersji zapisu, oraz położeniu najważniejszych struktur, jednak plik ten przy otworzeniu go bez odpowiedniego narzędzia, takiego jak przeglądarki plików pdf firmy Adobe Systems, nie pozwoli nam uzyskać żadnyuch sensownych danych i informacji o ich położeniu. Z tego powodu aplikacje eksportujące te dane nie są trywialne, a ich zapotrzebowanie jest realnym problemem.

Analiza formatów plików PDF pozwala nam lepiej zrozumieć złożoność i różnorodność plików, które będziemy importować. Umożliwia nam to dokładne określenie wymagań naszego modułu i odpowiednie dostosowanie go do różnych formatów plików PDF, z którymi będziemy pracować.

2.1.2 Pliki CSV

W pliku CSV dane są przechowywane w formie tabeli, gdzie każdy wiersz reprezentuje rekord, a poszczególne pola w rekordzie są oddzielane separatorem, najczęściej przecinkiem, jak sugeruje sama nazwa tego formatu, jednak inne separatory, takie jak średnik, również są stosowane w poszczególnych przypadkach. Pliki CSV wyróżnia to, że mimo możliwości przechowywania złożonych tabel z danymi, jest on formatem tekstowym. Oznacza to, że dane są zapisywane jako tekst, bez złożonej struktury danych, a dopiero następnie można je wyświtlać w postaci tabel lub arkuszy. Można w nich przechowywać

dzięki temu różne rodzaje danych, takie jak tekst, liczby, daty itp. bez znacznego zwiększenia ich objętości pamięciowej. Wyżej wspomniane wariacje w postaci wykorzystanych separatorów, pozwalają na zawieranie dowolnych znaków specjalnych, takich jak przecinki w danych liczbowych.

Pliki CSV są łatwe do tworzenia i edycji, nie wymagają zaawansowanych aplikacji, można je odczytać przy użyciu najprostszych edytorów tekstowych, jednak przy pracy na większej ilości danych wygodne jest użycie programu do obsługi arkuszy kalkulacyjnych. Ponieważ format CSV jest prostym i powszechnie wspieranym typem danych, pliki w tym formacie są szeroko używane do importu i eksportu danych, tworzenia raportów, analizy danych, integracji między systemami, itp. Z powodu na jego uniwersalność i prostotę, a mimo to duże możliwości, postanowiłem wykorzystać ten typ danych jako format wyjściowy z aplikacji importującej.

2.2 Przegląd typów danych do importowania

Plik PDF może zawierać różne formy zapisu informacji, takie jak tekst, tabele, obrazy, formularze itp. W celu efektywnego importowania danych z tych plików do innego formatu, konieczne jest zidentyfikowanie i zrozumienie charakterystyki tych typów danych.

Przede wszystkim, analizujemy sposoby ekstrakcji tekstu z plików PDF. Tekst może występować zarówno w postaci prostej, takiej jak akapity, nagłówki lub stopki. Jednak często występuje również w bardziej złożonej postaci np. listy, tabele. Ważne jest zrozumienie i rozróżnienie innych sposobów reprezentacji tekstu w pliku PDF, aby móc poprawnie go przechwycić i następnie przekształcić.

Kolejnym istotnym typem danych są tabele. Pliki PDF, które będziemy analizować zawierają tabele z danymi, które muszą być zaimportowane do innego formatu w taki sposób, aby zachować strukturę tabelaryczną. W tym punkcie analizujemy metody identyfikacji tabel oraz ekstrakcji danych z nich, uwzględniając zarówno prostsze tabele bez łączeń komórek, jak i bardziej zaawansowane struktury tabelaryczne. Różnią się one przede wszystkim, położeniem komórek, i danymi tekstowymi w nich zawartymi. Nie wszystkie komórki muszą zawierać opis jaki typ danych jest w nich zapisany, dodatkowo zapis ten może być niespójny jeżeli tabela byłaby uzupełniana przez innych pracowników.

W plikach PDF mogą również występować obrazy, jest to łatwy sposób reprezentacji danych Obrazy mogą mieć różne zastosowania, takie jak wykresy, diagramy, fotografie itp. W tym kontekście analizujemy techniki ekstrakcji tekstu i danych liczbowych z tabel, dlatego obrazy nie będą szczególnie obsługiwane, jednak wymagana jest świadomość istnienia takiego typu danych.

Wreszcie, rozważamy również importowanie danych z formularzy PDF. Pliki PDF często zawierają formularze, w których użytkownicy wprowadzają dane. W celu efektywnego importowania tych danych, badamy metody ekstrakcji danych z formularzy oraz możli-

wość przeniesienia ich do formatu obsługiwanego przez nasze rozwiązanie.

Przegląd typów danych do importowania umożliwia nam lepsze zrozumienie różnorodności informacji zawartych w plikach PDF oraz wyznaczenie kierunku dalszych badań i rozwoju modułu importującego.

Rozdział 3

Badanie istniejących rozwiązań

tekst

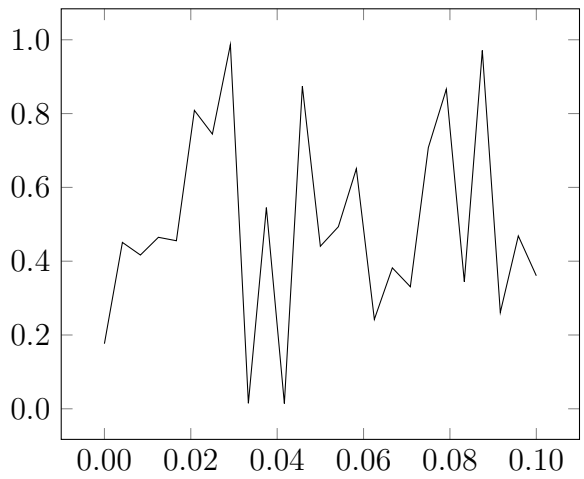
3.1 Przegląd istniejących modułów i bibliotek do importowania danych z PDF

3.2 Funkcje istniejących rozwiązań

3.3 Oceny użytkowników i opinie na temat istniejących rozwiązań

W całym dokumencie powinny znajdować się odniesienia do zawartych w nim ilustracji (rys. 3.1).

Tekst dokumentu powinien również zawierać odniesienia do tabel (tab. 3.1).



Rysunek 3.1: Wykres przebiegu funkcji.

Tabela 3.1: Opis tabeli nad nią.

	metoda						
			alg. 3			alg. 4, $\gamma = 2$	
ζ	alg. 1	alg. 2	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\beta = 0.1$	$\beta = -0.1$
0	8.3250	1.45305	7.5791	14.8517	20.0028	1.16396	1.1365
5	0.6111	2.27126	6.9952	13.8560	18.6064	1.18659	1.1630
10	11.6126	2.69218	6.2520	12.5202	16.8278	1.23180	1.2045
15	0.5665	2.95046	5.7753	11.4588	15.4837	1.25131	1.2614
20	15.8728	3.07225	5.3071	10.3935	13.8738	1.25307	1.2217
25	0.9791	3.19034	5.4575	9.9533	13.0721	1.27104	1.2640
30	2.0228	3.27474	5.7461	9.7164	12.2637	1.33404	1.3209
35	13.4210	3.36086	6.6735	10.0442	12.0270	1.35385	1.3059
40	13.2226	3.36420	7.7248	10.4495	12.0379	1.34919	1.2768
45	12.8445	3.47436	8.5539	10.8552	12.2773	1.42303	1.4362
50	12.9245	3.58228	9.2702	11.2183	12.3990	1.40922	1.3724

Rozdział 4

Ocena dostępnych opcji

Odwołania do literatury: książek [4], artykułów w czasopismach [3], materiałów konferencyjnych [2] i stron www [1].

Równania powinny być numerowane

$$y = \frac{\partial x}{\partial t} \tag{4.1}$$

4.1 Porównanie istniejących rozwiązań na podstawie zdefiniowanych wymagań

Rozdział 5

Rozwój istniejących rozwiązań

5.1 Rozważania dotyczące budowy niestandardowego modułu

5.2 Zasoby i umiejętności wymagane do rozwoju i utrzymania modułu

Rozdział 6

Rozwój wybranego rozwiązania

6.1 Testowanie wybranego rozwiązania

6.2 Udoskonalenie wybranego rozwiązania na podstawie napotkanych problemów

Rozdział 7

Dokumentacja i udostępnianie wybranego rozwiązania

7.1 Dokumentacja rozwiązania

7.2 Udostępnianie rozwiązania jako projektu open-source

Rozdział 8

Podsumowanie

8.1 Podsumowanie badania

8.2 Implikacje dla przyszłej pracy

- syntetyczny opis wykonanych prac
- wnioski
- możliwość rozwoju, kontynuacji prac, potencjalne nowe kierunki
- Czy cel pracy zrealizowany?

Bibliografia

[1] Imię Nazwisko i Imię Nazwisko. *Tytuł strony internetowej*. 2021. URL: `http://gdzies/w/internecie/internet.html` (term. wiz. 30. 09. 2021).

[2] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. „Tytuł artykułu konferencyjnego". W: *Nazwa konferecji*. 2006, s. 5346–5349.

[3] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. „Tytuł artykułu w czasopiśmie". W: *Tytuł czasopisma* 157.8 (2016), s. 1092–1113.

[4] Imię Nazwisko, Imię Nazwisko i Imię Nazwisko. *Tytuł książki*. Warszawa: Wydawnictwo, 2017. ISBN: 83-204-3229-9-434.

Dodatki

Spis skrótów i symboli

DNA kwas deoksyrybonukleinowy (ang. *deoxyribonucleic acid*)

MVC model – widok – kontroler (ang. *model–view–controller*)

N liczebność zbioru danych

μ stopnień przyleżności do zbioru

$\mathbb{E}$ zbiór krawędzi grafu

$\mathcal{L}$ transformata Laplace'a

Lista dodatkowych plików, uzupełniających tekst pracy (jeżeli dotyczy)

W systemie do pracy dołączono dodatkowe pliki zawierające:

- źródła programu,
- zbiory danych użyte w eksperymentach,
- film pokazujący działanie opracowanego oprogramowania lub zaprojektowanego i wykonanego urządzenia,
- itp.

Spis rysunków

3.1 Wykres przebiegu funkcji. 9

Spis tabel

3.1 Opis tabeli nad nią. 10