*Projekt zaliczeniowy*

*Proces ETL na przykładzie strony pyszne.pl*

***Grupa projektowa 28:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Imię** | **Nazwisko** | **Numer albumu** | **Grupa dziekańska** | **Wkład w prace nad projektem** | **Udział procentowy** |
| **Anna** | Łuszczkiewicz | 217025 | WZISN2-1112 | Python:  proces Transform,  refaktoryzacja kodu - wydzielenie procesów  Dokumentacja Techniczna  Instrukcja Obsługi Aplikacji | 25% |
| **Karolina** | Pytlak | 216850 | WZISN2-1112 | PHP: index.php  JavaScript: index.js  Dokumentacja Techniczna: tabela zmiennych, tabela funkcji | 25% |
| **Sebastian** | Krawczyk | 193492 | WZISN2-1112 | Python: proces Extract, proces Load  PHP: delete.php, result.php  JavaScript: CSV oraz Datatables w result.php  SQL database  HTML, CSS, Bootstrap w index.php i result.php | 25% |
| **Tomasz** | Krawczyk | 217006 | WZISN2-1112 | - Dodanie przycisków do formularza służących do uruchamiania poszczególnych etapów.  - Przerobienie formularza w taki sposób, żeby uruchomił kod PHP po jego zatwierdzeniu.  - Poprawienie skryptu python tak by można było go wykonać z poziomu PHP w całości i każdy etap pojedynczo.  - Wykonanie komendy uruchamiającej kod python z poziomu skryptu PHP | 25% |

|  |
| --- |
| \_\_/70 pkt |

**DOKUMENTACJA TECHNICZNA PROJEKTU „PROCES ETL NA PRZYKŁADZIE SERWISU PYSZNE.PL”**

**1. Użyte technologie**

* Python v.3.7.6
* MySQL v. 10.1.38-MariaDB
* PHP v. 7.3.4
* Java Script v. 1.6

**2. Środowisko**

Minimalne wymagania sprzętowe:

* Aplikacja działa na urządzeniach z systemem operacyjnym Windows, minimalna wersja: Windows 7

Biblioteki:

* urlopen
* BeautifulSoap
* request
* mysql.conector
* intertools
* sys
* json
* os

**3. Linki do oprogramowania**

* Link do repozytorium aplikacji:

<https://github.com/JosephSodomized/pysznyScraping>

* Link do Python v.3.7.6:

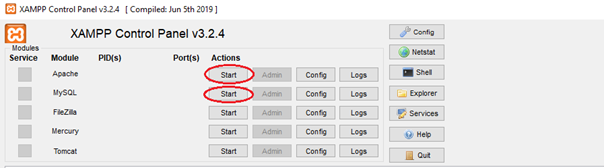
<https://www.python.org/downloads/release/python-376/>

* Link do XAMPP:

<https://www.apachefriends.org/index.html>

**4. Instalacja aplikacji**

Aby aplikacja działa poprawnie należy przed jej uruchomieniem zainstalować program XAMPP. W trakcie instalacji programu postępować zgodnie z instrukcją. Po instalacji programu należy uruchomić XAMPP control, następnie uruchomić Serwer Apache oraz bazę danych MYSQL wciskając odpowiednie przyciski “START”, co zostało przedstawione na rysunku 1.



*Rysunek 1 Zaznaczone odpowiednie przyciski w programie XAMPP*

Następnie po uruchomieniu serwera należy pobrać repozytorium projektu ze strony github.

Projekt należy rozpakować w ścieżce xampp/htdocs.

W celu uruchomienia aplikacji należy wprowadzić w pasku przeglądarki internetowej:

http://localhost /pysznyScraping/index.php

**5. Model danych**

Dane zostały zawarte w modelu relacyjnym składającym się z jednej tabeli, posiadającej dziewięć kolumn: ID, Nazwa restauracji, Rodzaj kuchni, Liczba recenzji, Czas dostawy, Koszt dostawy, Zamówienie minimalne, Ocena, Ostatnia recenzja.

**6. Tabela z listą i opisem zmiennych, tabela z listą i opisem funkcji z podziałem na pliki**

|  |  |
| --- | --- |
| **ZMIENNA** | **OPIS** |
| **webscraper.py** | |
| namesList | Tablica zawierająca nazwy wyszukanych restauracji |
| kitchensList | Tablica zawierająca rodzaje kuchni wyszukanych restauracji |
| reviewCount | Tablica zawierająca liczby recenzji wyszukanych restauracji |
| averageDeliveryTime | Tablica zawierająca czasy dostaw w wyszukanych restauracjach |
| deliveryCost | Tablica zawierająca koszty dostaw wyszukanych restauracji |
| minimumOrder | Tablica zawierająca minimalną cenę zamówienia w wyszukanych restauracjach |
| hrefLinks | Tablica zawierająca nazwy linków przekierowujących na strony wyszukanych restauracji |
| ratingNumbers | Tablica zawierająca oceny wyszukanych restauracji |
| lastWrittenReviews | Tablica zawierająca ostatnie opinie wyszukanych restauracji |
| postcode | Kod pocztowy wprowadzony przez użytkownika |
| **index.php** | |
| process | Zmienna typu string. Nazwa aktualnie wykonywanego procesu. |
| postcode | Zmienna typu string. Kod pocztowy wprowadzony przez użytkownika. |
| error | Zmienna typu string. Informacja o nieistniejącym kodzie pocztowym. |
| countOfExtractRecords | Zmienna typu intiger, zawiera liczbę rekordów, które zostały pobrane ze strony pyszne.pl |
| countOfLoadRecords | Zmienna typu intiger, zawiera liczbę rekordów, które zostały załadowane do bazy danych |
| **result.php** | |
| conn | Połączenie z bazą danych |
| sql | Rezultat zapytania do bazy |
| **index.js** | |
| form | Formularz, w którym zawarty został kliknięty przycisk |
| formData | Tablica z danymi formularza po procesie serializacji |
| data | Tablica zawierająca kod pocztowy nazwę procesu |
| url | Url na który zostanie wysłane zapytanie. |
| errorMesage | Informacja o nieistniejącym kodzie pocztowym. |
| messageClass | Nazwa klasy, w której pojawić ma się informacja o pobranych/załadowanych rekordach |
| countOfRecords | Liczba pobranych/załadowanych rekordów |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NAZWA** | **PRZYJMOWANE PARAMETRY** | **OPIS DZIAŁANIA** | | **WARTOŚĆ ZWRACANA** |
| **webscraper.py** | | | | |
| main | postcode | Wywoływana gdy jeszcze żaden z przycisków ETL nie zostanie wciśnięty | |  |
| procesExstract | postcode | Wywoływana poprzez przycisk Ekstract. Używając wprowadzonego przez użytkownika kodu pocztowego wysyła zapytanie na stronę pyszne.pl otrzymując w odpowiedzi html zawierający listę restauracji. Następnie wywoływana jest ekstrakcja potrzebnych danych na temat znalezionych restauracji i przypisanie ich do konkretnych tablic, po czym wywoływana jest funkcja saveFiles | | - |
| saveFiles | - | Wywołuje funkcję saveFile dla każdego rodzaju wyekstrahowanych danych | | - |
| saveFile | name, content | Zapisuje każdy rodzaj wyekstrahowanych danych do pliku | | - |
| procesTransform | - | Wywołuje funkcję loadFiles a następnie wykonuje transformacje danych, w celu przekształcenia ich do formy odpowiedniej do danych. Następnie wywołuje funkcję saveFiles | | - |
| loadFiles | - | Powoduje otwarcie plików zawierających wyekstrahowane dane na temat wyszukanych informacji | |  |
| procesLoad | - | Używając funkcji loadFiles otwiera pliki z zapisanymi danymi na temat wyszukanych restauracji, następnie tworzy połączenie do bazy danych, zapisuje informację do danej tabeli w bazie danych, wywołując na końcu funkcję usuwającą pliki z danymi dotyczącymi wyszukanych restauracji. | |  |
| deleteFiles | - | Usuwa pliki z danymi dotyczącymi wyszukanych restauracji. | |  |
| **index.js** | | | | |
| clearClass | Nazwa klasy | Usuwa daną klasę | - | |

**INSTRUKCJA OBSŁUGI**

1. **Krótki opis aplikacji**

Aplikacja służy do pobierania danych restauracji dowożących pod dany adres ze strony [www.pyszne.pl](http://www.pyszne.pl), oraz załadowania ich do bazy danych.

1. **Uruchomienie aplikacji**

Dokładny opis instalacji środowiska niezbędnego do uruchomienia aplikacji znajduje się w dokumentacji technicznej aplikacji.

W celu uruchomienia aplikacji należy wpisać w pasek przeglądarki adres <http://localhost/pysznyScraping/index.php>. Po wpisaniu adresu w pasek przeglądarki można zauważyć stronę startową aplikacji, co zostało przedstawione na rysunku 1.



*Rysunek 1 Widok strony startowej aplikacji*

1. **Opis funkcjonalności**

Główną funkcjonalnością aplikacji jest przeprowadzenie procesu ETL (ang. Extract, Transform and Load) na danych pobranych ze strony www.pyszne.pl.

Zadaniem procesu ETL jest pozyskanie danych ze strony, przekształcenie tych danych oraz zaimportowanie ich do bazy danych.

Aplikacja posiada przycisk “ETL” który jest odpowiedzialny za przeprowadzenie procesu ETL, oraz przyciski odpowiedzialne za kolejne etapy procesu ETL.

Przycisk “E” odpowiedzialny jest za pozyskanie danych ze strony, przycisk “T” przekształca pobrane dane, przycisk “L” wczytuje przekształcone dane do bazy danych.

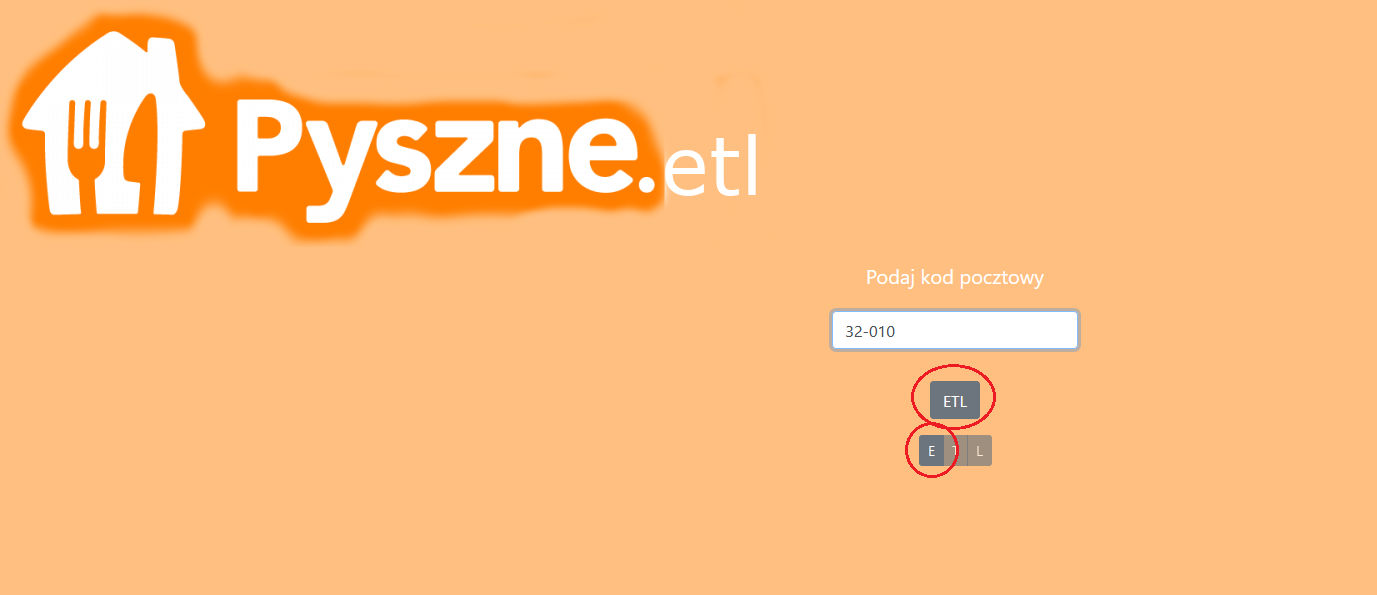
Po przeprowadzeniu pełnego procesu ETL dane wynikowe można eksportować do pliku o rozszerzeniu csv.

1. **Opis scenariuszy użycia**

Aplikacja posiada dwa scenariusze użycia.

Na początku po wprowadzeniu kod pocztowego aplikacja umożliwia przeprowadzenie procesu ETL dzięki przyciskowi „ETL”. Można również przeprowadzić procesy po kolei – jest to możliwe dzięki osobnym przyciskom „E”, „T” i „L”.

Przy uruchomieniu aplikacji możliwe jest uruchomienie procesu ETL lub pobrania danych, dzięki przyciskom “ETL” oraz “E”, co zostało przedstawione na rysunku 2.



*Rysunek 2 Procesy możliwe do przeprowadzenia prz starcie aplikacji*

**Scenariusz 1 - Uruchomienie pełnego procesu ETL przyciskiem “ETL”**

Po naciśnięciu przycisku “ETL” aplikacja przeprowadza proces ETL, oraz informuje użytkownika o liczbie załadowanych do bazy danych rekordów, co przedstawia rysunek 3.



*Rysunek 3 Wynik działania przycisku “ETL”*

**Scenariusz 2 - Przeprowadzenie procesu ETL poprzez przyciski “E”, “T”, “L”**

Po uruchomieniu procesu pobierania danych przyciskiem “E”, istnieje możliwość uruchomienia procesu transformacji danych, dzięki przyciskowi “T”, oraz ponownego uruchomienia pobierania danych. Ponad to aplikacja informuje użytkownika o liczbie pobranych rekordów, co zostało przedstawione na rysunku 4.



*Rysunek 4 Informacje zwrotne z procesu pobierania danych*

Następnie po przeprowadzeniu procesu transformacji danych, możliwe jest uruchomienie procesu importowania danych do bazy danych dzięki przyciskowi “L”, co zostało przedstawione na rysunku 5.



*Rysunek 5 Procesy możliwe do uruchomienia po przeprowadzeniu procesu transformacji danych*

Po przeprowadzeniu procesu wczytywania danych do bazy danych, użytkownik dostaje informację zwrotną dotyczącą liczby rekordów wczytanych do bazy danych oraz link “Zobacz wyniki”, gdzie może sprawdzić pobrane dane, widok został przedstawiony na rysunku 6.



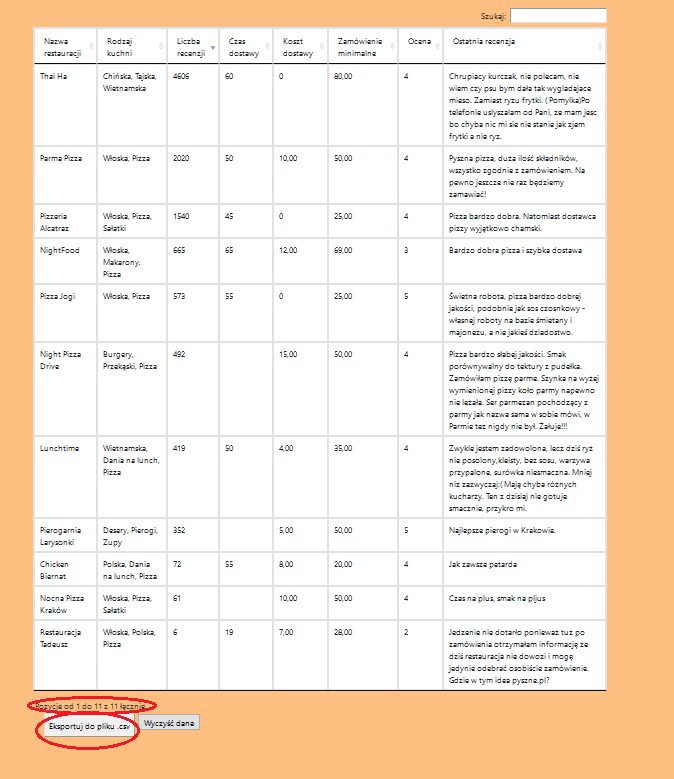
*Rysunek 6 Widok po przeprowadzeniu procesu ładowania danych do bazy danych*

Po wykonaniu pełnego procesu ETL wyniki można znaleźć pod adresem <http://localhost/pysznyScraping/result.php>, lub klikając w link “Zobacz wyniki”. Okno przedstawiające pobrane oraz przekształcone dane przedstawia rysunek 7.



*Rysunek 7 Wynik działania aplikacji*

Aplikacja wyświetla liczbę pobranych pozycji, oraz pozwala na eksport pobranych danych do pliku o formacie csv. Opisane funkcjonalności zostały zaznaczone na rysunku 8.



*Rysunek 8 Informacje dotyczące ilości pobranych pozycji*

1. **Opis menu i widoków okna aplikacji**

Aplikacja posiada dwa głowne okna aplikacji.

Na rysunku 9 zostało zaznaczone okno startowe aplikacji wraz z opisem.



*Rysunek 9 Okno startowe aplikacji wraz z zaznaczonymi polami aplikacji*

Legenda do rysunku 9:

1. Logo aplikacji

2. Pole do wpisania kodu pocztowego

3. Przycisk „ETL” odpowiedzialny za przeprowadzenie procesu ETL

4. Przycisk „E” odpowiedzialny za przeprowadzenie procesu pobierania danych

5. Przycisk „T” odpowiedzialny za transformację pobranych danych

6. Przycisk „L” odpowiedzialny za załadowanie danych do bazy danych

Drugim ważnym widokiem aplikacji jest widok z pobranymi już danymi, który można znaleźć pod adresem:<http://localhost/pysznyScraping/result.php>, został on przedstawiony na rysunku 10. Na samym środku okna można zauważyć pobrane dane.



*Rysunek 10 Widok przedstawiający wyniki działania aplikacji.*

Legenda do rysunku 10:

1. Wyszukiwarka pozwalająca wyszukać daną frazę w pobranych danych

2. Pole tekstowe wyświetlające liczbę wyświetlanych rekordów, a także zawiera informacje o wszystkich rekordach

3. Przycisk „Eksportuj do pliku csv” umożliwia pobranie danych i zapisanie ich do pliku o formacie csv

4. Przycisk „Wyczyść dane” usuwa informacje z bazy