Projekt zaliczeniowy

Proces ETL

Grupa projektowa:...

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Imię** | **Nazwisko** | **Numer albumu** | **Grupa dziekańska** | **Wkład w prace nad projektem[[1]](#footnote-0)** | **Udział procentowy** |
| **Mateusz** | Otręba | 194209 | WZISS2-1111 | Budowa aplikacji i bazy danych | 33 |
| **Jarosław** | Starmach | 194798 | WZISS2-1111 | Budowa aplikacji i bazy danych | 33 |
| **Maciej** | Dominik | 192616 | WZISS2-1111 | Budowa aplikacji i bazy danych | 33 |

**WEB SCRAPPER STRONY CENEO.PL - DOKUMENTACJA TECHNICZNA**

**Do zaprojektowania i zaprogramowania aplikacji użyto następujących technologii:**

1. Python w wersji 3.7.1 - Całość aplikacji została zaprogramowana w tym języku. Dodatkowo wykorzystano do niego takie biblioteki jak (wszystkie biblioteki bez podanej wersji, mają wersje języka tjk. 3.7.1):
   1. Json - biblitoeka wykorzystana do eksportowania pobranych danych do formatu json
   2. Urllib.request - bibliopteka do pobierania strony internetowej z wykorzytaniem jej adresu url
   3. Tkinter - jest to standardowe GUI pythona i zostało wykorzystane do zbudowania interfejsu użytkownika
   4. "The Fish-Footman began by producing from under his arm a great letter, nearly as large as himself."
   5. Beautiful Soup w wersji 4.4.0 - biblioteka Pythona wykorzystana do wyciągania danych z pliku html
   6. MySQLdb w wersji 1.2.4 - bilibliotek wykorzystywana do ustanowienia połączenia i komunikacji z bazą danych
2. Pip w wersji 18.1 - narzędzie do zarządzania pakietami i bibliotekami pythona, wykorzytane do instalowania większossci wyżej wymienionych bibliotek
3. Xampp - pakiet narzędzi bazodanowych z których do aplikacji wykorzytane zostały:
   1. Apache w wersji 2.4.37 - otwarty serwer HTTP wykorzystany do przechowania bazy danych
   2. MySQL w wersji 8.0.12 - wykorzystany do zbudowania bazy danych
   3. phpMyAdmin w wersji 4.8.4 – narzędzie służące do łatwego zarządzania bazą danych MySQL
4. Kartka i ołowek do zaprojektowania działania

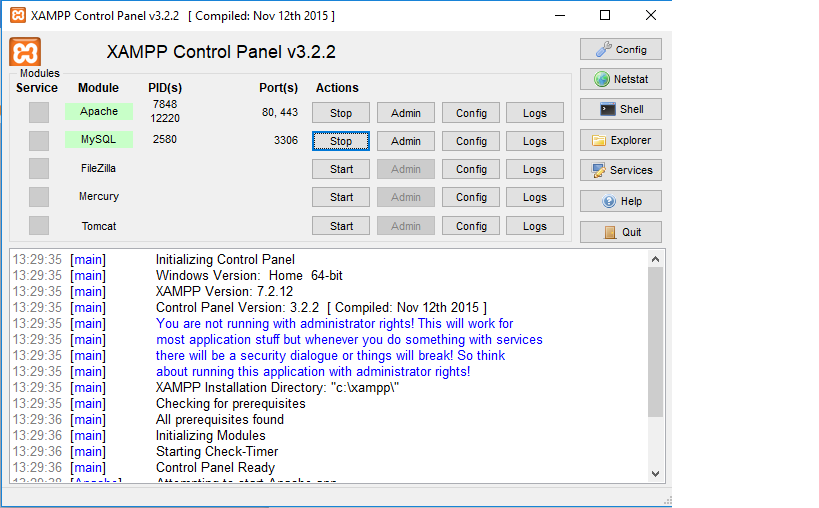
**Do instalacji i popranego działania aplikacji niezbędne jest:**

1. Istalacja python w wersji conajmniej 3.0 (zalecane Python 3.7.1). W tym celu należy pobrac instalator pythona ze strony: <https://www.python.org/downloads/windows/>, uruchomić go i przejsc przez proces istalacji. UWAGA! Należy się upewnić że w instalatorze jest zaznaczona opcja „Add Python 3.x to PATH”. Po dokonaniu instalacji można się upewnić, czy process przebiegł pomyślnie poprzez uruchomienie terminala cmd i wpisanie komendy „python -v” w przypadku poprawnej instalacji, powinna zostać wyświetlona wersja języka.
2. Instalacja bilioteki Beautiful Soup, w tym celu konieczne jest uruchomienie teminala i wpisanie komendy „pip3 install beautifulsoup4” która przy pomocy menadżera pakietów „pip” (zainstalowanego razem z pythonem) pobierze i zainstaluje pakiet.
3. Instalacja biliboteki MySQLdb - W tym celu wpisujemy w terminalu komende „pip3 install mysqlclient” która przeprowadzi process pobierania i instalacji.
4. Instalacja biblioteki Request - komenda „pip3 install requests” w terminalu
5. Instalacja pakietu bazodanowego Xampp - W tym celu pobieramy instalator z https://www.apachefriends.org/pl/download.html i uruchamiamy go UWAGA! Przed rozpoczęciem instalacji zalecane jest wyłączenie oprogramowania antywirusowego.

Przechodzimy przez kolejne etalpy instalacji wybierając scieżkę docelową inną niż C:Program Files (x86) (problemy z uprawnieniami). Po instalacji przechodzimy do folderu docelowego, odnajdujemy tam plik „config.inc.php” a w nim wyrażenie „$cfg['Servers'][$i]['auth\_type']” po czym zmieniamy jego wartość na „cookie”, tak żeby docelowo wyglądało tak: „$cfg['Servers'][$i]['auth\_type'] = 'cookie';”

1. Uruchamiamy XAMPP Control Panel a następnie klikamy start przy opcji „Apache”

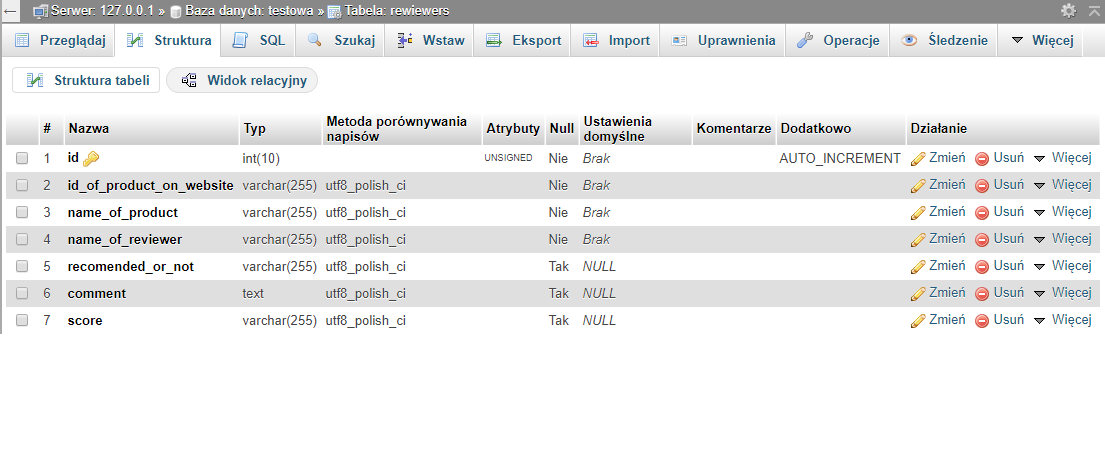
oraz „MySQL”, przy poprawny połączeniu panel powinien wyglądać jak na załączonym niżej zrzucie



1. W pgrzeglądarce internetowej wpisujemy localhost, następnie klikamy phpMyAdmin po czym wpisujemy pole użytkownik jako „root” i BEZ wpisywania hasła przechodzimy dalej. Wybieramy bazy danych tworzymy nową baze danych o nazwie „testowa” a jej kodowanie ustawiamy na utf8\_polish\_ci. Kolejno klikamy nową bazę danych i tworzymy w niej tabelę, według podanego niżej modelu

**Model bazy danych**

Baza danych została oparta na modelu MySQL z pakietu Xampp i jest „postawiona” na serwerze lokalnym Apache. Składa się z jednej tabeli o nazwie „rewiewers” która z kolei składa się z 7 kolumn, Nazwy kolumn i ich typy zostały zamieszczone na tabeli obrazie poniżej



**Składowe programu:**

**Klasa Rewiew:**

|  |  |
| --- | --- |
| Funkcje i atrybuty | Opis |
| Funkcja \_\_init\_\_() | Funkcja - konsrtuktor przyjmująca wszystkie atrybuty klasy i tworząca obiekt tej klasy w chwili jest wywolania |
| Atrybut Name | Atrybut przyjmujący nazwe recenzenta |
| Atrybut recomendOrNot | Atrybut mowiacy o tym czy rezenzent poleca produkt |
| Atrybut comment | Atrybut przyjmujacy komentarz na temat produktu |
| Atrybut score | Atrybu przyjmuujacy ocene produkty |
| Atrybut nameOfProduct | Atrybut przyjujacy nazwe produktu |
| Atrybut IdOfProduct | Atrybut przyjmujacy id produktu na stronie ceneo |

**Funkcje**

|  |  |
| --- | --- |
| Funkcja | Opis |
| clearDataBase() | Funkcja wykująca czyszczenie tabeli rewiewers w bazie danych za pomoca operacji TRUNCATE |
| processETL() | Funkcja wykonujaca cały proces etl poprzez wykonanie nastepujacych po sobie funkcji: extractData(), transform() oraz saveDataToDataBase |
| extractData() | Funkcja pobierajaca strone w postaci html i wyciągająca z niej wszystkie opinie |
| Transform() | Funkcja transofromujaca opinie do postaci umożliwiające ich zapisanie do bazy danych |
| saveDataToDataBase() | Funkcja zapisujaca dane do bazy danych i usuwajaca wszystkie nieaktualne opinie |
| checkIFMoreExist() | Fukcja sprawdzajaca, czy zostało wiecej opini do „wyciagniecia” |
| exportDatatoJson() | Funkcja slużąca do eksportowania przetrasformowanych danych do postaci json |

**Zmienne**:

|  |  |
| --- | --- |
| Nazwa zmiennej | Opis |
| HOST | Zmienna do przechowywania hosta serwera, w tym przypadku jest to localhost |
| USERNAME | Do przechowywania użytkownika do połączenia z bazą danych |
| PASSWORD | Przechowuje hasło do połączenia z ba |
| DATABASE | Przechowuje nazwę bazy danych |
| db | Użyta do ustanownia połączenia z bazą danych |
| idOfProduct | Przechowuje globalnie id produktu na stronie ceneo |
| nameOfProduct | Przechowuje globalnie nazwę produktu |
| listOfOpinios | Przechowuje listę wszystkich przetransformowanych już opini |
| extractedData | Przechowuje dane w postaci umożliwiających ich eksportowanie do postacie json |
| user\_agent | Wtyczka umożliwiająca połączenie się z witryną |
| numberOfPage | Zmienna używana do budowanie url do konkretnej podstrony z opiniami |
| rewiews | Przechowuje nieprzetransformowane jeszcze opinie |
| browser\_window | Instancja interfejsu użytkownika |
| label | Napis wyswietlany w oknie |
| Entry | Pole tekstowe sluzace do wprowadzania id |
| Button .... | Przycicki odpowiadające za uruchamianie kolejnych skłądowych procesu etl |
| Text | Pole sluzace do wyswietlania opini o produkcie |
| infoText | Pole tekstowe sluzace do wyswietlania dodatkowych informacji o przebiegu procesu |

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

1. proszę wymienić konkretne zadania [↑](#footnote-ref-0)