Projekt zaliczeniowy

Proces ETL

Grupa projektowa:28

Imię	**Nazwisko**	**Numer albumu**	**Grupa dziekańska**	**Wkład w prace nad projektem**[1]	**Udział procentowy,**
Tomasz	Reczyński	194496	WZISS2-1111	- mockupy aplikacji - skonfigurowanie buttlera - stworzenie scrapera html - stworzenie procesów E, T, L, ETL - stworzenie metod E T L - dane statystyczne - poprawki błędów w scraperze - wstępne wersje podstron: opinionsPage, runETL, runE, runT, runL	**34%**
Adam	Pacholak	194221	WZISS2-1111	- mockupy aplikacji - konsultacje projektowe - strona domowa - README.md na gicie - dokumentacja techniczna - instrukcja obsługi - organizacja potrzebnej bazy wiedzy (linków)	**27%**
Daniel	Słowik	194704	WZISS2-1111	- mockupy aplikacji, - konsultacje projektowe, - struktura projektu, - podstawowy wygląd	**39%**

[1] proszę wymienić konkretne zadania

				strony, - podstrona z produktami, - podstrona z detalami produktu, - podstrona z opiniami, - podstrona do przeprowadzania procesu ETL, - podstrona procesu extract, - podstrona procesu transform, - podstrona procesu load, - możliwość usuwania rekordów z bazy danych, - model bazy danych, - export do CSV, - blokada wielokrotnego wywoływania procesu ETL, - CSS, - README.md, - dokumentacja techniczna	

__/70 pkt

Spis treści

Dokumentacja techniczna

Nazwy i wersje użytych technologii (języki programowania, system zarządzania bazą danych itp.)

1. Python v.3.7.1
    a. Manager pakietów: pip v.18.1
    b. Framework: Django v.2.1.3
    c. Baza danych: SQLite v.3.26.0
    d. Baza danych: SQLite v.3.26.0
2. Biblioteki
    a. beautifulsoup4 v.4.6.3
    b. django-request v.1.5.4
    c. equests v.2.20.1
    d. csv v.3.7
3. HTML5
4. CSS3
    a. Bootstrap4
5. JS ES6

Informacje na temat środowiska (minimalne wymagania sprzętowe, biblioteki itp.) potrzebne do instalacji aplikacji.

Minimalne wymagania sprzętowe*:

- Processors: Intel Atom® processor or Intel® Core™ i3 processor
- 256MB DDR3
- Disk space: 1 GB
- Operating systems: Windows* 7 or later, macOS, and Linux

* brak informacji dotyczącej wymagań na oficjalnej stronie.

Linki do oprogramowania, które to środowisko tworzą.

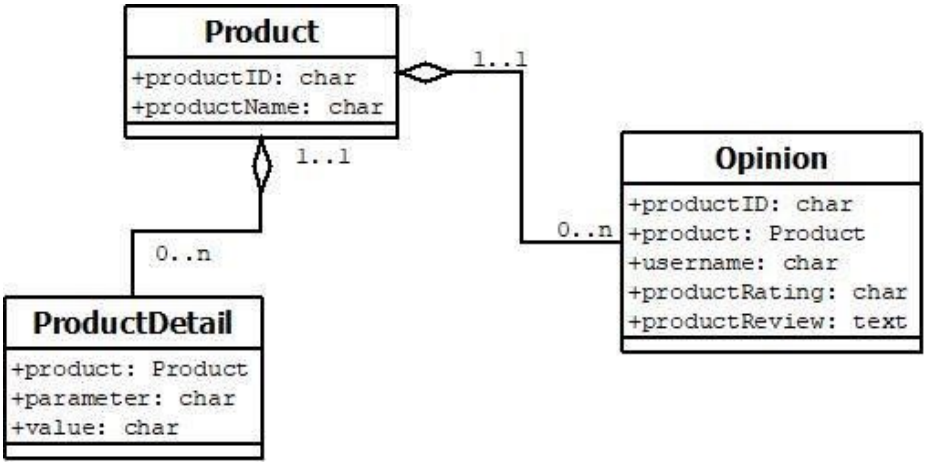
1. https://virtualenv.pypa.io/en/latest/
2. https://www.python.org/
2. https://www.djangoproject.com/
3. https://www.sqlite.org/index.html
4. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
5. https://getbootstrap.com/

Instrukcje instalacji aplikacji

Proces instalacji:

1) Należy rozpocząć od zainstalowania Pythona z oficjalnej strony
2) Następnie wykonać polecenie “pip install virtualenv” w celu zainstalowania wirtualnego środowiska.
3) Kolejnym krokiem jest utworzenie wirtualnego środowiska poleceniem “ . virtualenv”
4) Uruchomienie środowiska odbywa się poprzez “. Scripts/activate”
5) Instalujemy Django poleceniem “pip install django”
6) Instalujemy requests poleceniem “pip install requests”
7) Instalujemy Beautiful Soup poleceniem “pip install beautifulsoup4”
8) Przechodzimy do folderu z aplikacją - Application i tworzymy swoje konto administratora komendą “python manage.py createsuperuser”

Model danych użyty w projekcie



Klasy i funkcje

Plik forms.py

W tym pliku znajduje się tylko klasa ProductForm, która jako parametr przyjmuje zawartość formularza o nazwie productID w formie modelu ProductID

Plik models.py

Klasa **Product** reprezentująca podstawowe dane o produkcie:
productID - ID produktu (to z Ceneo)
productName - nazwa produktu

Klasa **ProductDetail** reprezentująca szczegółowe dane na temat konkretnego produktu:
product = models.ForeignKey(Product, on_delete=models.CASCADE)
parameter - parametr
value - wartość

Klasa **ProductID** z atrybutem productID wykorzystywana podczas pobierania danych z formularza.

Klasa **Opinion** przedstawiająca opinie dotyczące konkretnego produktu:
product = models.ForeignKey(Product, on_delete=models.CASCADE)
z następującymi atrybutami:
productID - ID produktu
username - nazwa użytkownika wystawiającego opinię
productRating - ocena produktu w skali 0-5
productReview - opinia na temat produktu

Plik opinionETL.py

Funkcja **getOpinionCount**() zwracająca ilość opinii

Funkcja **opinionRunETL**(prodID, product), która przyjmuje jako parametry ID produktu oraz sam produkt, a następnie wywołuje funkcje generateOpinionLinkList, extractOpinions, transformOpinions, loadOpinions

Funkcja **opinionRunE**(prodID), która jako argument przyjmuje ID produktu, a zwraca zeskrapowane dane w postaci html'a. Wywołując funkcje generateOpinionLinkList oraz extractOpinions

Funkcja **opinionRunT**(extractedData), która jako argument przyjmuje wynik funkcji opinionRunE, a zwraca listę wyników, wywołując funkcję transformOpinions.

Funkcja **opinionRunL**(transformedData, product) dostaje na wejściu wynik funkcji opinionRunT, czyli rezultaty, oraz produkt, którego te opinie dotyczą. W wyniku działania tej funkcji za pośrednictwem funkcji loadOpinions dane zostają umieszczone w bazie danych.

Funkcja **generateOpinionLinkList**(prodID) z ID produktu generuje tablice stron linków z opiniami.

Funkcja **extractOpinions**(listOfOpinionLinks) wydobywa z wykorzystaniem tablicy ston z linkami do opinii zeskrapowany HTML

Funkcja **transformOpinions**(scrapedHTML) przetwarza HTML do modelu opinii wymaganego przez bazę danych. Na wyjściu zwraca tabelę z opiniami.

Funkcja **loadOpinions**(opinions, prod) z wykorzystaniem atrybutów takich jak tabela z opiniami oraz produkt, któtego te opinie dotyczą zapisuje dane w bazie danych.

Plik productETL.py

Funkcja **productRunETL**(prodID) z atrybutem jakim jest ID produktu wywołuje w swoim wnętrzu kolejno funkcje extractProduct, transformProduct, loadProduct.

Funkcja **productRunE**(prodID) wywołuje z ID produktu funkcję extractProduct

Funkcja **productRunT**(extractedData) z atrybutem jakim jest tablica z nazwą i danymi produktu wywołuje funkcję transformProduct

Funkcja **productRunL**(transformedData) z atrybutem postaci wyniku transformacji wywołuje funkcję loadProduct. Wynkiem czego są wpisy w bazie danych.

Zadaniem funkcji **extractProduct**(productID) z atrybutem ID produktu jest zeskrapowanie danych dotyczących produktu ze strony Ceneo oraz zwrócenie ich w postaci nazwy oraz detali.

Zadaniem tej funkcji - **transformProduct**(scrapedProductHTML) jest przetworzenie danych do modelu wymaganego przez bazę danych oraz przekazanie ich na wyjściu.

Zadaniem funkcji **loadProduct**(productParameters) jest załadowanie produktu i jego parametrów do bazy danych.

Plik views.py

home(request) funkcja renderująca plik home.html

runETL(request) funkcja, która w przypadku POST'a pobiera dane z formularza, sprawdza czy ID jest poprawne i wywołuje metodę productRunETL, ponadto zlicza liczbę opinii.

extract(request) funkcja, która w przypadku POST'a pobiera ID z formularza sprawdza jego poprawność i wywołuje metodę opinionRunE oraz productRunE. Oraz renderuje stronę run-etl.html

transform(request) funkcja, która w przypadku POST'a wywołuje metody opinionRunT oraz productRunT, renderuje transform.html przekazując ilość opinii i parametrów na wyjściu.

load(request) funkcja, która w przypadku POST'a wywołuje funkcję productRunL, oraz renderuje stronę load.html

products(request) funkcja, która renderuje stronę products.html z parametrem zawierającym wszystkie produkty z bazy danych.

deleteProduct(request, product_id) funkcja, której argumentem jest ID produktu, co umożliwia jego usunięcie z bazy danych oraz przekierowanie na stronę z produktami.

deleteProducts(request) funkcja, która umożliwia usunięcie zawartości całej bazy danych.

productDetails(request, product_id) ta funkcja ma za zadanie wyrenderowanie konkretnych parametrów produktu jako strony product-details.html przekazując na wyjściu szczegóły produktu.

productCSV(request, product_id) funkcja zwraca plik CSV z danymi produktu.

productsCSV(request) funkcja zwraca plik CSV z danymi wszystkich produktów.

opinions(request, product_id) funkcja renderuje stronę opinions.html z opiniami dotyczącymi konkretnego produktu.

opinionsCSV(request) funkcja zwraca plik CSV z wszystkimi opiniami o produkcie.

opinionCSV(request, opinion_id) funkcja zwraca plik CSV z wybraną opinią o produkcie.

sortNameAscending(request) funkcja sortuje rosnąco produkty po nazwie

sortNameDescending(request) funkcja sortuje malejąco produkty po nazwie

Instrukcja obsługi

instrukcja uruchomienia aplikacji

Aby uruchomić aplikację należy przede wszystkim wykonać kroki zawarte w dokumentacji technicznej, a dokładniej podpunkcie dotyczącym instalacji środowiska aplikacji. Następnie należy uruchomić wirtualne środowisko, przejść do folderu z aplikacją i uruchomić serwer. Domyślnie aplikacja znajduje się pod adresem localhost:8000/

1. *. Scripts/activate*
2. *cd ../Application*
3. *python manage.py runserver*

Opis funkcjonalności aplikacji

W głównym założeniu aplikacja ma za zadanie poprawne przeprowadzenie procesów ETL. Aplikacja pozyskuje dane ze strony Ceneo.pl, następnie przetwarza je do utworzonego wzorca, po czym załadowuje je do bazy danych. Efektem tego procesu są czytelnie zaprezentowane dane o produktach. W zakładce z bazą danych istnieje możliwość przeprowadzania takich czynności jak: sortowanie danych, wyświetlanie parametrów produktów, oraz opinii na ich temat, usunięcie danych z bazy oraz eksport opinii do plików w formacie *csv*. Usuwanie oraz eksport odbywa się zarówno dla pojedynczych rekordów jak i całych modeli.

Opis scenariuszy użycia aplikacji

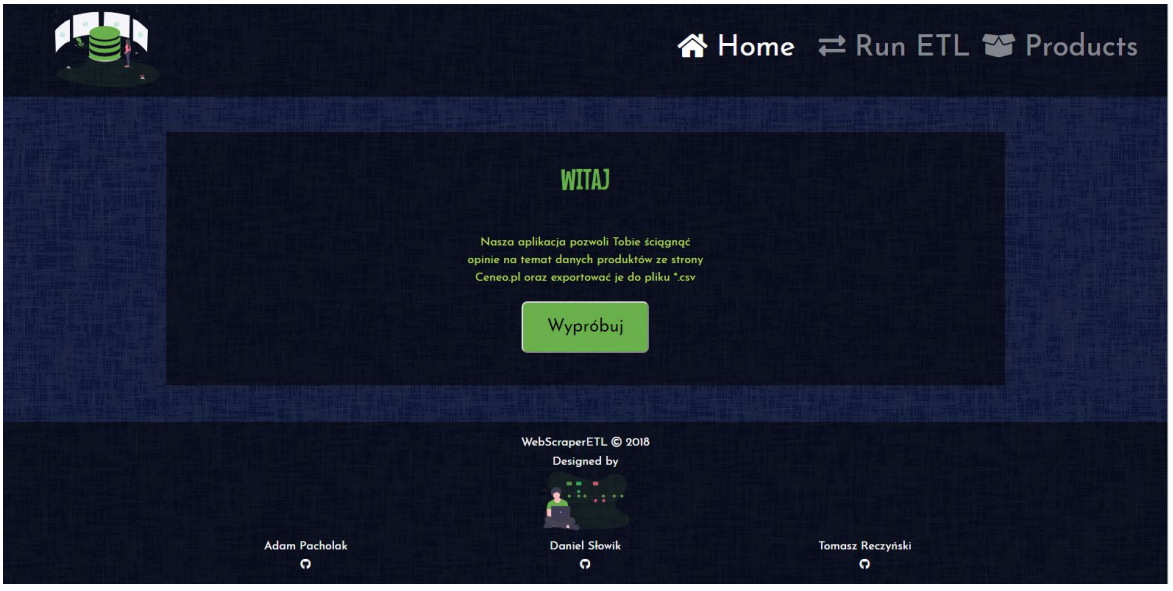
Jako użytkownik:

1. Mam możliwość przeprowadzenia procesów ETL.
2. Mam możliwość przeprowadzenia osobno jeden po drugim procesów E, T oraz L.
3. Mam wgląd do bazy danych produktów.
4. Mogę wyświetlić parametry produktów z bazy danych.
5. Mam możliwość wyświetlenia opinii na temat interesującego mnie produktu.

6. Mogę sortować produkty w bazie danych w porządku alfabetycznym, bądź odwrotnie do alfabetycznego.
7. Mam możliwość usunięcia rekordu z bazy danych.
8. Mam możliwość wyczyszczenia całej bazy danych.
9. Mam możliwość wygenerowania produktów z detalami lub opinii do pliku w formacie *.csv .

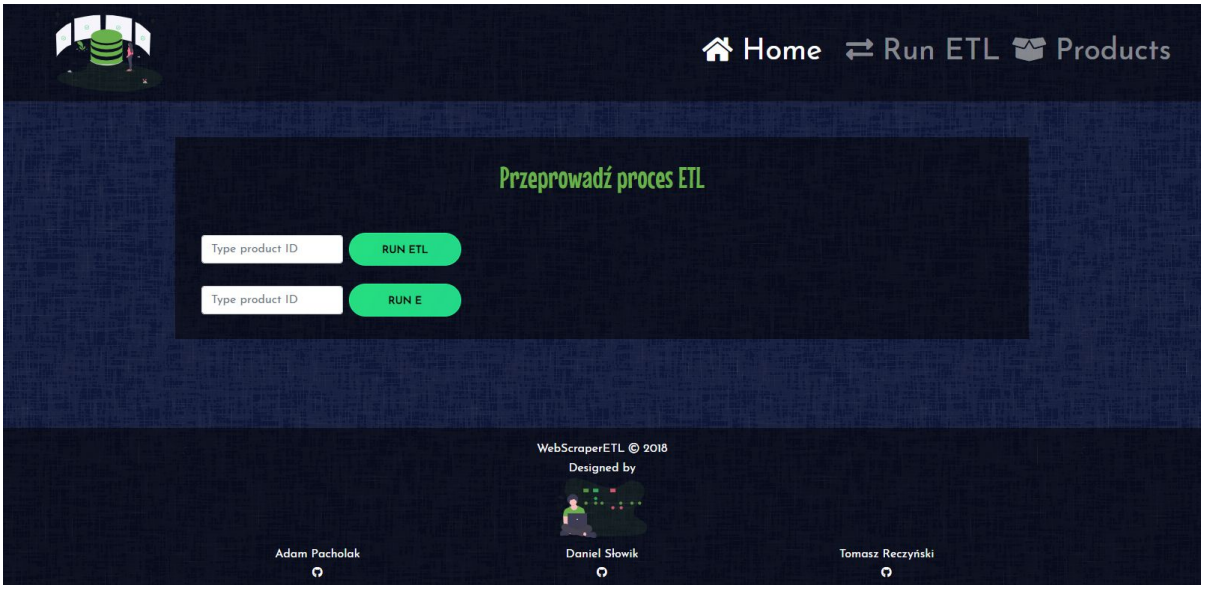
opis menu i widoków okna aplikacji

HOME

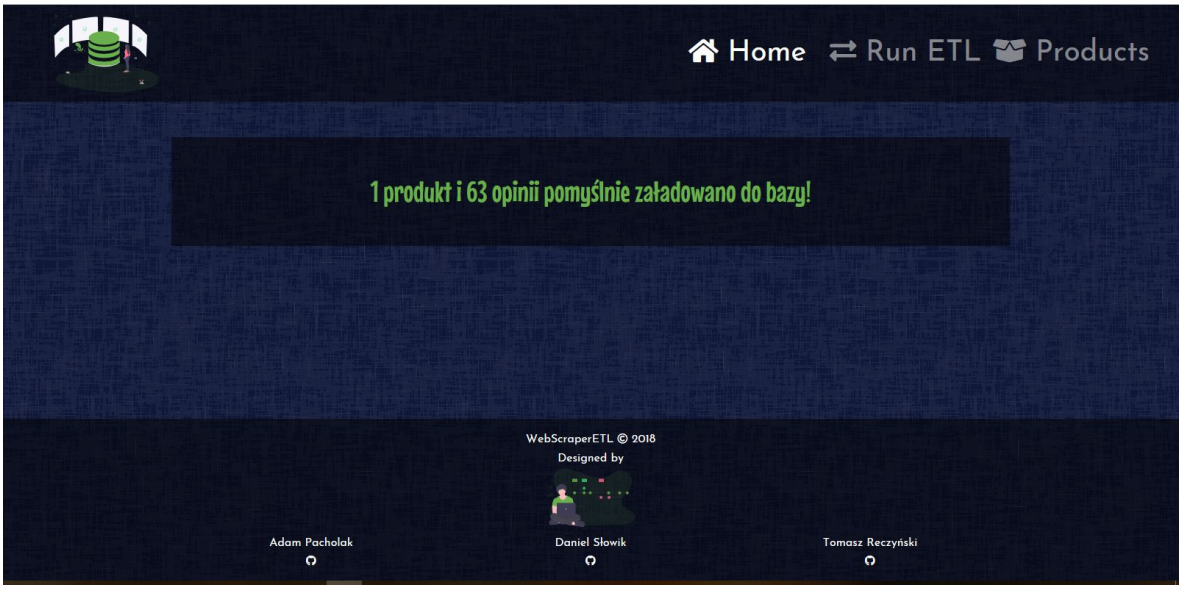


Strona domowa projektu składa się z powitania użytkownika oraz krótkiego opisu działania aplikacji, jak i jej przeznaczenia. Przycisk *Wypróbuj* przekierowuje użytkownika do zakładki RUN ETL.

RUN ETL

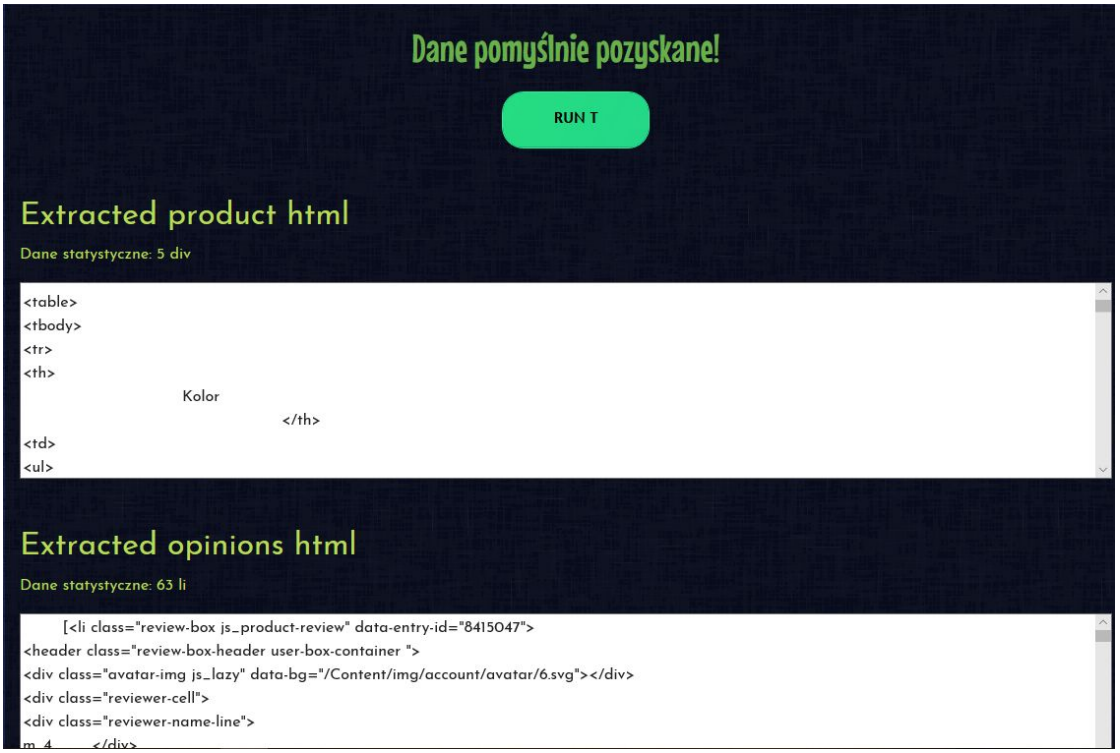


Zakładka RUN ETL składa się z pola do wpisywania ID produktu oraz przycisków uruchamiających proces ETL oraz E. Po wpisaniu ID produktu oraz wciśnięciu przycisku RUN ETL, wyszukany produkt zostaje poddany procesowi ETL, a informacja potwierdzająca zajście procesu wyświetlana jest na ekranie użytkownika.



Po wpisaniu id produktu oraz wciśnięciu przycisku RUN E, użytkownik jest przekierowywany do okna informującego o pozyskaniu danych o produkcie i

możliwości przeprowadzenia procesu transformacji danych, po naciśnięciu przycisku RUN T, a następnie procesu załadowania danych do bazy po przyciśnięciu przycisku RUN L. Po użyciu tego ostatniego wyświetlana jest informacja potwierdzająca dodanie produktu do bazy danych.



PRODUCTS

W zakładce PRODUCTS wyświetlane są wszystkie produkty dodane do bazy danych za pomocą procesów ETL. Są one przedstawione za pomocą tabeli składającej się z następujących kolumn:

a) **ID**, zawierającej ID produktu z platformy Ceneo.pl
b) **NAZWA**, zawierająca nazwę produktu. Kolumna ta umożliwia także sortowanie wyświetlanych produktów w porządku alfabetycznym, bądź odwrotnym do alfabetycznego
c) **PARAMETRY**, która pozwala użytkownikowi na wyświetlenie informacji o parametrach produktu po naciśnięciu ikony kół zębatych. Użytkownik zostaje przekierowany do zakładki Parametry produktu, gdzie wyświetlane są parametry produktu oraz ich wartość.

d) **OPINIE**, pozwalająca użytkownikowi na wyświetlenie opinii na temat produktu po naciśnięciu ikony trzech postaci. Użytkownik zostaje przekierowany do zakładki Opinie, gdzie wyświetlane są opinie na temat produktu, ocena w skali 5-stopniowej, nazwa użytkownika wydającego opinię oraz możliwość wygenerowania zarówno pojedynczej opinii jak i wszystkich do pliku *.csv.

e) **USUŃ**, umożliwiającą usunięcie produktu z bazy danych po naciśnięciu ikony kosza na śmieci oraz całkowite wyczyszczenie bazy danych po naciśnięciu napisu USUŃ

f) **CSV**, który eksportuje wszystkie jak i pojedyncze opinie na temat produktu do pliku w formacie csv.