



UNIWERSYTET MERITO

PRACA INŻYNIERSKA

# Analiza i Wizualizacja Danych na Przykładzie Web Scrapingu Rynku Obuwniczego

*Michał Rudawski*

Nr albumu: 80765

Przedstawiona do  
Wydział Informatyki  
UNIWERSYTET MERITO

Wrocław, 2024

# Spis treści

<b>1</b>	<b>Wprowadzenie</b>	<b>3</b>
1.1	Przegląd Projektu . . . . .	3
1.2	Cele i Zakres . . . . .	3
1.3	Struktura Pracy . . . . .	3
<b>2</b>	<b>Tło i Przegląd Literatury</b>	<b>4</b>
2.1	Przegląd Web Scrapingu . . . . .	4
2.2	Znaczenie Analizy Danych w E-commerce . . . . .	4
2.3	Ewolucja Narzędzi Web Scrapingu . . . . .	5
2.4	Wyzwania Związane z Web Scrapingiem . . . . .	5
<b>3</b>	<b>Projekt Systemu i Architektura</b>	<b>5</b>
3.1	Ogólna Architektura Systemu . . . . .	5
3.2	Opis Modułów . . . . .	6
3.2.1	main.py: Punkt Wejścia Systemu i Orchestrator . . . . .	6
3.2.2	logger_module.py: Mechanizm Logowania . . . . .	6
3.2.3	_base_scraper.py: Klasa Bazowa dla Scrapingu . . . . .	6
3.2.4	Indywidualne Scrapery: eobuwie.py, nike.py, stockx.py, adidas.py . . . . .	6
3.2.5	shoes_purchase_analyzer.py: Moduł Analizy Danych . . . . .	7
3.3	Przepływ Pracy Systemu i Diagram Przepływu Danych . . . . .	7
<b>4</b>	<b>Szczegóły Implementacji</b>	<b>7</b>
4.1	Mechanizm Scrapingu . . . . .	7
4.2	Zbieranie i Przechowywanie Danych . . . . .	7
4.3	Techniki Analizy . . . . .	8
<b>5</b>	<b>Testowanie i Walidacja</b>	<b>8</b>
5.1	Metodologie Testowania . . . . .	8
5.2	Walidacja Dokładności Scrapingu . . . . .	9
5.3	Walidacja Analizy Danych . . . . .	9
<b>6</b>	<b>Wyniki i Dyskusja</b>	<b>10</b>
6.1	Zebrane Dane . . . . .	10
6.2	Wnioski z Analizy . . . . .	10
6.3	Porównanie z Oczekiwanymi Wynikami . . . . .	10
<b>7</b>	<b>Wyzwania i Ograniczenia</b>	<b>10</b>
7.1	Wyzwania Techniczne . . . . .	10
7.2	Ograniczenia Obecnego Systemu . . . . .	11
7.3	Etyczne Rozważania Web Scrapingu . . . . .	11

<b>8</b>	<b>Praca w Przyszłości i Usprawnienia</b>	<b>11</b>
8.1	Proponowane Usprawnienia . . . . .	11
8.2	Skalowalność i Poprawa Wydajności . . . . .	12
8.3	Potencjalne Dodatkowe Funkcje . . . . .	12
<b>9</b>	<b>Zakończenie</b>	<b>12</b>
9.1	Podsumowanie Wyników . . . . .	12
9.2	Wpływ Projektu . . . . .	12
9.3	Końcowe Uwagi . . . . .	13
<b>10</b>	<b>Bibliografia</b>	<b>13</b>
<b>11</b>	<b>Załączniki</b>	<b>13</b>
11.1	Listingi Kodu . . . . .	13
11.2	Dodatkowe Wykresy i Tabele z Analizy . . . . .	13

# 1 Wprowadzenie

## 1.1 Przegląd Projektu

Projekt ten stanowi kompleksowe podejście do analizy i wizualizacji danych, uzyskanych za pomocą technik web scrapingu. Skupia się na rynku obuwniczym, który jest dynamicznie rozwijającym się sektorem e-commerce. Celem jest stworzenie zaawansowanego narzędzia, które umożliwia efektywne zbieranie, przetwarzanie i analizowanie danych o produktach obuwniczych z różnych sklepów internetowych. Wykorzystując język programowania Python oraz szereg narzędzi do analizy danych, projekt ten ma na celu nie tylko gromadzenie informacji, ale również ich analizę w celu identyfikacji trendów rynkowych i preferencji konsumentów. Projekt ten integruje różne metody i narzędzia, począwszy od skryptów scrapujących, aż po zaawansowane techniki wizualizacji danych, aby zapewnić bogate i użyteczne wglądy w branżę obuwniczą.

## 1.2 Cele i Zakres

Głównym celem tej pracy jest zaprojektowanie i implementacja złożonego systemu do web scrapingu oraz analizy danych, który pozwoli na dogłębne zrozumienie rynku obuwniczego. W szczególności, praca koncentruje się na następujących aspektach:

- Rozwój skutecznych metod scrapingu danych, które mogą radzić sobie z dynamicznymi treściami stron internetowych sklepów obuwniczych.
- Przetwarzanie i czyszczenie zebranych danych w celu zapewnienia ich użyteczności i wiarygodności.
- Zastosowanie metod analizy statystycznej i technik wizualizacji danych w celu identyfikacji kluczowych trendów i wzorców w danych.
- Ocena wydajności i efektywności różnych technik scrapingu w kontekście ciągle zmieniającego się środowiska e-commerce.

Zakres pracy obejmuje więc nie tylko techniczne aspekty scrapingu i analizy danych, ale również interpretację i zrozumienie wyników w kontekście rynkowym.

## 1.3 Struktura Pracy

Praca została zorganizowana w sposób, który ułatwia czytelnikowi zrozumienie zarówno teoretycznych podstaw projektu, jak i praktycznych aspektów jego realizacji. Struktura pracy przedstawia się następująco:

1. **Tło i Przegląd Literatury:** Ten rozdział zawiera przegląd istniejących badań i literatury związanej z web scrapingiem i analizą danych, dając czytelnikowi solidne tło teoretyczne.

2. **Projekt Systemu i Architektura:** W tym rozdziale opisano szczegółowo projektowane systemy, ich architekturę oraz indywidualne moduły.
3. **Szczegóły Implementacji:** Tutaj przedstawiono konkretne techniki i metody użyte w realizacji projektu.
4. **Testowanie i Walidacja:** Rozdział ten koncentruje się na metodach testowania systemu oraz na walidacji zebranych danych i analiz.
5. **Wyniki i Dyskusja:** Ten rozdział prezentuje wyniki projektu, wraz z analizą i interpretacją tych wyników.
6. **Wyzwania i Ograniczenia:** Omówienie wyzwań napotkanych podczas realizacji projektu i ograniczeń systemu.
7. **Praca w Przyszłości i Usprawnienia:** Propozycje na przyszłe usprawnienia i kierunki rozwoju projektu.
8. **Zakończenie:** Podsumowanie pracy, wnioski końcowe i refleksje nad jej wpływem i znaczeniem.

Każdy rozdział został zaprojektowany tak, aby stopniowo prowadzić czytelnika przez wszystkie etapy projektu - od koncepcji, przez realizację, aż po analizę wyników i wnioski.

## 2 Tło i Przegląd Literatury

### 2.1 Przegląd Web Scrapingu

Web scraping, będący procesem ekstrakcji danych ze stron internetowych, zyskał znaczące znaczenie w wielu dziedzinach, włączając w to e-commerce, badania rynkowe, i analizę konkurencji. W kontekście dynamicznie zmieniającego się rynku obuwniczego, web scraping jest wykorzystywany do gromadzenia danych na temat cen, dostępności produktów, trendów mody, i preferencji klientów. Techniki te obejmują zarówno proste metody parsowania HTML, jak i bardziej złożone rozwiązania wymagające interakcji z dynamicznie generowanymi treściami stron. W obliczu coraz częstszego wykorzystania JavaScript i AJAX na stronach e-commerce, narzędzia takie jak Selenium, Scrapy, czy BeautifulSoup stają się kluczowe w efektywnym zbieraniu danych. Również kwestie etyczne i prawne związane z web scrapingiem, takie jak przestrzeganie warunków użytkowania stron internetowych i praw autorskich, są istotnym aspektem tej praktyki.

### 2.2 Znaczenie Analizy Danych w E-commerce

W erze cyfrowej, analiza danych stała się nieodzownym elementem strategii e-commerce. Wykorzystanie danych w celu zrozumienia zachowań klientów i optymalizacji strategii sprzedażowych jest kluczowe dla sukcesu w konkurencyjnym środowisku online. W branży obuwniczej, analiza danych pozwala na identyfikację

wzorców zakupowych, zrozumienie wpływu trendów mody, oraz przewidywanie przyszłych trendów rynkowych. Techniki takie jak uczenie maszynowe, analiza predykcyjna, i zaawansowana wizualizacja danych, umożliwiają firmom nie tylko przewidywanie, ale i kształtowanie przyszłych trendów rynkowych. Wzrost danych strukturalnych i nieskukturalnych z różnych źródeł, włączając w to media społecznościowe i strony internetowe, stanowi wyzwanie w zakresie zarządzania i analizy danych, wymagając skutecznego wykorzystania narzędzi do big data.

## **2.3 Ewolucja Narzędzi Web Scrapingu**

W ostatnich latach narzędzia web scrapingu znacząco się rozwinęły, od prostych skryptów parsujących HTML do złożonych systemów, które mogą naśladować interakcje użytkownika i dynamicznie reagować na zmiany na stronach internetowych. Rozwój technologii takich jak headless browsing (przeglądanie bez interfejsu użytkownika) i automatyzacja przeglądarek umożliwił efektywniejsze zbieranie danych z dynamicznych stron internetowych, które używają ciężkich skryptów JavaScript. Te postępy umożliwiły firmom dostęp do szerszego zakresu danych, włączając w to te, które były wcześniej niedostępne z powodu ograniczeń technologicznych.

## **2.4 Wyzwania Związane z Web Scrapingiem**

Mimo swojej użyteczności, web scraping niesie ze sobą wyzwania techniczne i etyczne. Jednym z głównych wyzwań technicznych jest utrzymanie aktualności scraperów w obliczu ciągłych zmian w strukturach stron internetowych. Firmy muszą regularnie aktualizować swoje narzędzia, aby zapewnić ciągłość i dokładność zbieranych danych. Z punktu widzenia etycznego, ważne jest przestrzeganie praw autorskich i zasad prywatności, co wymaga od firm zrozumienia i przestrzegania prawnych i moralnych granic scrapingu danych.

# **3 Projekt Systemu i Architektura**

## **3.1 Ogólna Architektura Systemu**

Projektowany system charakteryzuje się modularną architekturą, która umożliwia efektywne zbieranie, przetwarzanie i analizowanie danych z różnorodnych źródeł internetowych. Bazując na modelu klient-serwer, system dzieli funkcje między serwer - który zarządza przetwarzaniem danych i logiką biznesową, a klientów - którzy zajmują się bezpośrednim zbieraniem danych ze stron internetowych. Serwer centralny działa jako węzeł koordynujący, zapewniając optymalizację zadań i równowagę obciążenia między różnymi modułami. Wykorzystanie architektury opartej na mikrousługach znacząco zwiększa skalowalność systemu, umożliwiając jednoczesne scrapowanie danych z wielu stron, co jest kluczowe w kontekście dużych wolumenów danych. Dodatkowo, system jest wyposażony w mechanizmy do zarządzania sesjami i pamięcią podręczną, co zapewnia wysoką wydajność i niezawodność w środowiskach produkcyjnych.

## 3.2 Opis Modułów

System składa się z kilku dobrze zdefiniowanych modułów, które są ze sobą ściśle zintegrowane, aby zapewnić kompleksowe funkcjonowanie:

### 3.2.1 main.py: Punkt Wejścia Systemu i Orchestrator

Moduł 'main.py' pełni rolę głównego koordynatora systemu, zarządzając uruchamianiem i interakcjami między innymi modułami. Odpowiada za rozpoczęcie procesu scrapingu, koordynację przepływu danych oraz zarządzanie zadaniami analizy. Ten moduł również implementuje logikę decyzyjną dotyczącą wyboru źródeł danych, określania priorytetów zadań i obsługi wyjątków na poziomie systemowym. Zapewnia on elastyczne API do interakcji z użytkownikiem końcowym oraz innymi systemami zewnętrznymi, co umożliwia łatwą integrację i rozbudowę systemu.

### 3.2.2 logger\_module.py: Mechanizm Logowania

Moduł 'logger\_module.py' jest zaawansowanym systemem logowania, który rejestruje wszystkie istotne zdarzenia w systemie. Ten moduł jest kluczowy dla monitorowania stanu systemu, diagnostyki problemów oraz analizy wydajności. Oferuje on różne poziomy logowania (np. debug, info, warning, error), co pozwala na szczegółowe śledzenie procesów i szybką reakcję na potencjalne problemy. Logi są przechowywane w bezpieczny sposób, z możliwością łatwego dostępu i analizy, co jest niezbędne do utrzymania wysokiej jakości usług.

### 3.2.3 \_base\_scraper.py: Klasa Bazowa dla Scrapingu

Klasa '\_base\_scraper.py' działa jako fundament dla wszystkich specjalistycznych scraperów. Definiuje wspólne metody i struktury danych, które są wykorzystywane przez wszystkie scrapery, co zapewnia spójność i ponowne wykorzystanie kodu. Ta klasa bazowa zawiera funkcje do zarządzania połączeniami sieciowymi, przetwarzania danych HTML, obsługi błędów i ograniczeń stron, oraz przechowywania danych pośrednich. Dzięki temu indywidualne scrapery mogą skupić się na specyfikach konkretnych stron internetowych, jednocześnie korzystając z solidnej, wspólnej bazy funkcjonalności.

### 3.2.4 Indywidualne Scrapery: eobuwie.py, nike.py, stockx.py, adidas.py

Każdy z tych dedykowanych scraperów jest zaprojektowany do pracy z konkretną stroną internetową, taką jak Eobuwie, Nike, StockX czy Adidas. Są one zbudowane na bazie klasy '\_base\_scraper.py' i dostosowane do specyficznych wyzwań związanych ze strukturą i dynamiką danych na poszczególnych stronach. Te indywidualne scrapery zarządzają ekstrakcją danych, takich jak informacje o produktach, ceny, dostępność i opinie klientów, z zachowaniem wysokiej dokładności i aktualności danych.

### 3.2.5 shoes\_purchase\_analyzer.py: Moduł Analizy Danych

Moduł 'shoes\_purchase\_analyzer.py' jest sercem analizy danych w systemie. Wykorzystuje zaawansowane techniki statystyczne, uczenie maszynowe i analizę danych do wydobywania wartościowych wglądów z zebranych danych. Ten moduł jest odpowiedzialny za przeprowadzanie kompleksowych analiz, takich jak identyfikacja trendów zakupowych, analiza segmentacji rynku, i prognozowanie popytu. Oferuje on również narzędzia do wizualizacji danych, które umożliwiają intuicyjne przedstawienie złożonych analiz i trendów, co jest kluczowe dla podejmowania decyzji biznesowych.

## 3.3 Przepływ Pracy Systemu i Diagram Przepływu Danych

Opis przepływu pracy systemu prezentuje, jak dane są zbierane, przetwarzane i analizowane. Rozpoczyna się od scrapingu danych przez indywidualne scrapery, po czym dane są przekazywane do centralnego systemu przetwarzania. Tutaj są one czyszczone, normalizowane i przechowywane w bazie danych. Następnie, moduł analizy danych przetwarza te informacje, wykorzystując różne metody statystyczne i algorytmy uczenia maszynowego, aby wydobyć użyteczne wglądy i wzorce. Cały proces jest monitorowany i rejestrowany przez moduł logowania, zapewniając przejrzystość i umożliwiając optymalizację systemu.

## 4 Szczegóły Implementacji

### 4.1 Mechanizm Scrapingu

Mechanizm scrapingu w projekcie został zaprojektowany, aby efektywnie radzić sobie z różnymi wyzwaniami związanymi z ekstrakcją danych ze stron internetowych. Centralnym elementem jest zastosowanie hybrydowego podejścia do scrapingu, które łączy zarówno metody bazujące na analizie statycznych dokumentów HTML, jak i dynamicznego renderowania stron za pomocą przeglądarek internetowych. Do parsowania HTML wykorzystano biblioteki takie jak BeautifulSoup, które umożliwiają szybką i efektywną ekstrakcję danych. W przypadkach, gdy zawartość strony jest generowana dynamicznie za pomocą JavaScript, użyto narzędzi takich jak Selenium, które umożliwiają kontrolę przeglądarki i interakcję ze stroną w sposób imitujący użytkownika. Mechanizm ten jest również wyposażony w zaawansowane strategie omijania ograniczeń (np. CAPTCHA, rate limiting), co pozwala na niezawodne zbieranie danych z różnorodnych źródeł internetowych.

### 4.2 Zbieranie i Przechowywanie Danych

Zbieranie danych jest zautomatyzowane i odbywa się zgodnie z zaplanowanym harmonogramem, co zapewnia aktualność i ciągłość informacji. Zebrane dane są następnie przetwarzane, co obejmuje ich czyszczenie, normalizację i walidację, aby upewnić się, że są one kompletna i spójne. Po przetworzeniu, dane są przechowywane w scentralizowanej bazie danych, co ułatwia ich dalszą analizę i przetwa-



rzanie. Baza danych została zaprojektowana w taki sposób, aby była skalowalna i mogła efektywnie radzić sobie z dużymi wolumenami informacji. Stosuje się także techniki zapewniania bezpieczeństwa danych, takie jak szyfrowanie i backupy, co jest kluczowe w kontekście ochrony danych osobowych i zapewnienia zgodności z przepisami prawnymi, takimi jak RODO.

## 4.3 Techniki Analizy

Moduł analizy danych wykorzystuje szeroki zakres technik statystycznych i algorytmów uczenia maszynowego do przetwarzania i interpretacji zebranych danych. Techniki te obejmują między innymi analizę eksploracyjną danych, analizę korelacji, segmentację rynku, oraz predykcyjne modele cenowe. Kluczowe jest tutaj zastosowanie zaawansowanych metod wizualizacji, takich jak heat mapy, wykresy punktowe i diagramy sankey, które umożliwiają intuicyjne przedstawienie złożonych wzorców i trendów. Moduł analizy jest również wyposażony w dashboardy i interaktywne panele kontrolne, które umożliwiają użytkownikom szybką i łatwą eksplorację danych oraz generowanie spersonalizowanych raportów. Całość jest zintegrowana w taki sposób, aby zapewnić płynne przejście od zbierania danych, poprzez ich przetwarzanie, aż po wizualizację i interpretację wyników.

# 5 Testowanie i Walidacja

## 5.1 Metodologie Testowania

Testowanie systemu jest kluczowym etapem zapewnienia jego niezawodności i efektywności. W projekcie zastosowano kompleksowe podejście do testowania, obejmujące zarówno testy jednostkowe, integracyjne, jak i testy akceptacyjne użytkownika.

- **Testy Jednostkowe:** Skupiają się na indywidualnych komponentach systemu, takich jak funkcje scrapingu czy algorytmy analizy danych. Każdy moduł jest testowany izolacyjnie, co pozwala na identyfikację i naprawę błędów na wczesnym etapie rozwoju.
- **Testy Integracyjne:** Sprawdzają współpracę między różnymi modułami systemu, aby upewnić się, że dane są poprawnie przekazywane i przetwarzane przez cały system. Testy te są kluczowe dla weryfikacji, czy system działa jako spójna całość.
- **Testy Akceptacyjne Użytkownika:** Służą do oceny systemu z perspektywy końcowego użytkownika. Zapewniają, że wszystkie funkcje spełniają wymagania i są intuicyjne w obsłudze.
- **Testy Wydajnościowe:** Pozwalają na ocenę, jak system radzi sobie z dużymi wolumenami danych i wieloma równoległymi zadaniami, co jest istotne dla zapewnienia wysokiej wydajności i skalowalności systemu.

- **Testy Bezpieczeństwa:** Sprawdzają, czy system jest odporny na różne wektory ataków, co jest kluczowe z punktu widzenia ochrony danych i zapewnienia zgodności z przepisami prawnymi.

## 5.2 Walidacja Dokładności Scrapingu

Walidacja dokładności scrapingu jest niezbędna do zapewnienia, że dane zbierane przez system są kompletne, aktualne i wiarygodne. Proces walidacji obejmuje:

- **Porównanie ze Źródłami:** Dane zebrane przez scrapery są porównywane z danymi dostępnymi bezpośrednio na stronach internetowych, aby upewnić się, że proces scrapingu nie wprowadza błędów.
- **Weryfikacja Integralności Danych:** Sprawdzenie, czy wszystkie wymagane informacje są zbierane, i czy dane są spójne i kompleksne.
- **Testy na Różnych Urządzeniach i Przeglądarkach:** Wykonanie scrapingu na różnych urządzeniach i przeglądarkach, aby upewnić się, że system jest odporny na zmienne warunki.
- **Automatyczne Wykrywanie Anomalii:** Implementacja systemów do automatycznego wykrywania anomalii w zebranych danych, co pozwala na szybką interwencję w przypadku wykrycia błędów.

## 5.3 Walidacja Analizy Danych

Walidacja analizy danych jest kluczowa dla zapewnienia, że wnioski wyciągnięte z danych są wiarygodne i mogą być wykorzystane do podejmowania decyzji biznesowych. Proces ten obejmuje:

- **Weryfikacja Modeli Statystycznych:** Ocena, czy zastosowane modele statystyczne są odpowiednie dla rodzaju i charakteru danych.
- **Walidacja Krzyżowa:** Użycie technik walidacji krzyżowej do oceny skuteczności i wiarygodności modeli predykcyjnych.
- **Porównanie z Wcześniejszymi Wynikami:** Porównanie wyników analiz z historią danych i wcześniejszymi analizami, co pozwala na ocenę spójności i ewolucji trendów.
- **Ekspertskie Recenzje i Konsultacje:** Konsultacje z ekspertami w dziedzinie e-commerce, aby upewnić się, że analizy są zgodne z realiami rynkowymi i mogą być użyteczne w praktycznym zastosowaniu.

## 6 Wyniki i Dyskusja

### 6.1 Zebrane Dane

W ramach projektu udało się zebrać szeroki zakres danych z różnych sklepów internetowych, w tym Eobuwie, Nike, StockX i Adidas, co stanowiło podstawę do dalszej analizy. Zebrane dane obejmowały informacje o produktach, takie jak ceny, dostępność rozmiarów, kolory, opinie użytkowników oraz dane dotyczące trendów zakupowych. System, wykorzystując moduł 'shoes\_purchase\_analyzer.py', efektywnie przetwarzał i analizował te informacje, co pozwoliło na generowanie użytecznych wglądów w aktualne trendy rynkowe.

### 6.2 Wnioski z Analizy

Analiza danych wykazała kilka kluczowych trendów. Przede wszystkim, zidentyfikowano sezonowe wzorce w preferencjach zakupowych, co ma bezpośredni wpływ na strategie cenowe i promocyjne sklepów. Ponadto, analiza danych pozwoliła na zrozumienie, które marki i style są obecnie najbardziej popularne wśród różnych grup demograficznych. System był również w stanie przewidywać nadchodzące trendy na podstawie analizy wzorców zakupowych i opinii użytkowników, co stanowi cenne informacje dla decydentów biznesowych w branży obuwniczej.

### 6.3 Porównanie z Oczekiwanymi Wynikami

Porównanie wyników uzyskanych przez system z oczekiwanymi wynikami biznesowymi wykazało wysoką zgodność. System efektywnie identyfikował zmieniające się trendy rynkowe, co pokrywało się z niezależnymi badaniami rynkowymi i raportami branżowymi. Dzięki temu, możliwe było potwierdzenie wiarygodności przetwarzania danych przez system oraz skuteczności zastosowanych metod analizy statystycznej i uczenia maszynowego. Wyniki te świadczą o wysokiej skuteczności i wiarygodności implementacji systemu, odzwierciedlając precyzję i dokładność zarówno procesu scrapingu danych, jak i ich późniejszej analizy.

## 7 Wyzwania i Ograniczenia

### 7.1 Wyzwania Techniczne

W trakcie projektowania i implementacji systemu napotkano szereg wyzwań technicznych. Jednym z najważniejszych było zapewnienie efektywności i dokładności procesu scrapingu danych w obliczu dynamicznie zmieniających się treści stron internetowych. Sklepy e-commerce często aktualizują swoje strony, co wymagało ciągłego dostosowywania scraperów do nowych struktur HTML i skryptów JavaScript. Kolejnym wyzwaniem było zarządzanie dużymi wolumenami zebranych danych, co wymagało implementacji wydajnych rozwiązań bazodanowych i algorytmów przetwarzania danych. Dodatkowo, zapewnienie wysokiej dostępności i niezawodności systemu w warunkach produkcyjnych, szczególnie w sytuacjach

dużego obciążenia, stanowiło istotne wyzwanie. Wyzwania te wymagały zastosowania zaawansowanych technik programistycznych, w tym asynchronicznego przetwarzania danych, optymalizacji zapytań do bazy danych i skalowania poziomego infrastruktury.

## **7.2 Ograniczenia Obecnego Systemu**

Obecna implementacja systemu, pomimo swojej zaawansowanej funkcjonalności, posiada pewne ograniczenia. Głównym ograniczeniem jest zależność od zewnętrznych źródeł danych, co oznacza, że jakiegokolwiek zmiany na stronach internetowych mogą wpływać na efektywność scrapingu. Ponadto, system obecnie wspiera ograniczoną liczbę sklepów internetowych, co może wpływać na ogólny obraz rynku. W kontekście analizy danych, choć zastosowane metody są zaawansowane, istnieje potencjalna potrzeba dalszego rozwoju modeli predykcyjnych i algorytmów uczenia maszynowego, aby jeszcze lepiej zrozumieć złożone wzorce rynkowe. Na koniec, system wymaga regularnej konserwacji i aktualizacji, aby zapewnić jego ciągłą efektywność i zgodność z najnowszymi technologiami.

## **7.3 Etyczne Rozważania Web Scrapingu**

Web scraping, choć jest potężnym narzędziem do zbierania danych, wiąże się z szeregiem kwestii etycznych i prawnych. Najważniejszą z nich jest szacunek dla prywatności użytkowników i przestrzeganie przepisów prawnych dotyczących danych osobowych, w szczególności RODO. System został zaprojektowany z myślą o przestrzeganiu tych regulacji, jednak zawsze istnieje ryzyko, że zmiany w prawie lub w politykach prywatności stron internetowych mogą wymagać dostosowań w sposobie scrapingu. Ponadto, istotne jest, aby scraping danych był przeprowadzany w sposób, który nie obciąża nadmiernie serwerów stron internetowych, co wymaga odpowiedniego planowania i implementacji ograniczeń w zakresie częstotliwości i zakresu scrapingu. Te kwestie etyczne i prawne wymagają ciągłej uwagi i adaptacji, aby zapewnić, że działalność scrapingu pozostaje odpowiedzialna i zgodna z obowiązującymi normami.

# **8 Praca w Przyszłości i Usprawnienia**

## **8.1 Proponowane Usprawnienia**

W oparciu o doświadczenia zdobyte podczas obecnej fazy projektu, zidentyfikowano kilka kluczowych obszarów, które mogą zostać ulepszone w przyszłości. Po pierwsze, planuje się dalsze udoskonalanie algorytmów scrapingu, aby były one bardziej odporne na częste zmiany na stronach internetowych. Rozważa się także wprowadzenie zaawansowanych technik sztucznej inteligencji do automatycznego dostosowywania scraperów do zmieniających się schematów stron. Dodatkowo, planuje się rozbudowę modułu analizy danych o bardziej zaawansowane modele

statystyczne i algorytmy uczenia maszynowego, które pozwolą na jeszcze głębsze zrozumienie danych i generowanie bardziej precyzyjnych prognoz.

## **8.2 Skalowalność i Poprawa Wydajności**

Kolejnym ważnym aspektem przyszłych prac będzie skupienie się na skalowalności i wydajności systemu. Obecnie system radzi sobie efektywnie z obecnym wolumenem danych, jednak w przyszłości, z rosnącą ilością danych i źródeł, będzie konieczna dalsza optymalizacja. Planuje się wprowadzenie rozwiązań chmurowych i technik obliczeń rozproszonych, co pozwoli na bardziej elastyczne zarządzanie zasobami i zapewni lepszą skalowalność. Ponadto, rozważa się ulepszenia w zakresie zarządzania bazą danych, w tym zastosowanie nowoczesnych systemów bazodanych, które lepiej radzą sobie z dużymi zbiorami danych.

## **8.3 Potencjalne Dodatkowe Funkcje**

W przyszłości planuje się również rozbudowę systemu o dodatkowe funkcje, które zwiększą jego wartość i użyteczność. Jedną z propozycji jest wprowadzenie narzędzi do analizy sentymentu na podstawie danych z mediów społecznościowych, co pozwoli na lepsze zrozumienie opinii konsumentów i trendów rynkowych. Inną możliwą funkcją jest rozwój interfejsu użytkownika i dashboardów analitycznych, które będą jeszcze bardziej intuicyjne i dostosowane do potrzeb różnych użytkowników, od analityków rynku po menedżerów produktu. Dodatkowo, rozważa się integrację systemu z innymi platformami e-commerce i systemami zarządzania danymi, aby umożliwić łatwiejszy przepływ i wymianę danych.

# **9 Zakończenie**

## **9.1 Podsumowanie Wyników**

Projekt zakończył się sukcesem, demonstrując efektywność i użyteczność systemu do scrapingu i analizy danych w kontekście rynku obuwniczego. Kluczowe wyniki obejmują rozwój stabilnego i elastycznego systemu zdolnego do zbierania i analizowania dużych ilości danych z różnych źródeł internetowych. Analiza tych danych ujawniła istotne informacje o trendach rynkowych, preferencjach konsumentów i dynamice cen, co jest nieocenione dla decydentów biznesowych i analityków rynku. Wyniki te potwierdzają, że zastosowane metody scrapingu i analizy są skuteczne i mogą być z powodzeniem wykorzystane do uzyskiwania głębokich wglądów w zachowania rynkowe.

## **9.2 Wpływ Projektu**

Wpływ projektu jest wielowymiarowy. Po pierwsze, dostarcza on nowoczesnych narzędzi do analizy rynku obuwniczego, co ma istotne znaczenie dla firm działających w tej branży. Po drugie, projekt demonstruje praktyczne zastosowanie

technologii web scrapingu i analizy danych w realnych scenariuszach biznesowych. Wreszcie, przyczynia się on do akademickiej i praktycznej wiedzy na temat wykorzystania danych internetowych w celu lepszego zrozumienia dynamiki rynku i zachowań konsumentów.

## 9.3 Końcowe Uwagi

Podsumowując, projekt ten podkreśla znaczenie i potencjał analizy danych w kontekście e-commerce. Wskazuje na możliwości, jakie otwierają się dzięki zastosowaniu zaawansowanych technik scrapingu i analizy w celu zdobywania wiedzy rynkowej i podejmowania informowanych decyzji biznesowych. Jako kolejne kroki, rozważa się dalsze rozwijanie systemu, zarówno pod kątem technologicznym, jak i funkcjonalnym, aby jeszcze lepiej sprostać wymaganiom dynamicznie zmieniającego się rynku e-commerce. Projekt ten stanowi fundament do przyszłych badań i rozwoju w tej fascynującej dziedzinie.

## 10 Bibliografia

Bibliografia będzie zawierać wszystkie źródła, na które powoływano się w trakcie pracy nad projektem. Powinna obejmować zarówno literaturę akademicką, jak i inne istotne źródła, które pomogły w zrozumieniu problemów i wyzwań związanych z web scrapingiem, analizą danych i rynkiem obuwniczym. Każde źródło zostanie sformatowane zgodnie z przyjętymi standardami akademickimi, zapewniając czytelność i łatwość odnalezienia odwołań.

## 11 Załączniki

### 11.1 Listingi Kodu

W załącznikach zostaną dołączone listingi kodu, które są kluczowe dla zrozumienia i reprodukcji wyników projektu. Zostaną one przedstawione w czytelnej formie, z odpowiednimi komentarzami i opisami. Listingi będą obejmować kod kluczowych modułów systemu, w tym 'main.py', 'logger\_module.py', '\_base\_scraper.py', indywidualne scrapery, jak również moduły odpowiedzialne za analizę danych, takie jak 'shoes\_purchase\_analyzer.py'.

### 11.2 Dodatkowe Wykresy i Tabele z Analizy

W tej sekcji zostaną umieszczone dodatkowe wykresy i tabele, które pomogą w lepszym zrozumieniu wyników analizy danych. Wykresy będą obejmować m.in. analizę trendów cenowych, popularności poszczególnych marek i modeli, oraz wzorce zakupowe. Tabele będą zawierać szczegółowe dane liczbowe, takie jak statystyki sprzedaży, oceny produktów i inne kluczowe metryki. Te dodatkowe materiały wizualne i statystyczne zapewnią głębsze wglądy w dane i pomogą czytelnikowi w lepszym zrozumieniu wniosków wyciągniętych z projektu.