Zastosowanie nowoczesnych technologii informatycznych w korzystaniu z alternatywnych źródeł danych

Program NCBiR GOSPOSTRATEG
Projekt INSTATCENY

Zespół Podstaw Sztucznej Inteligencji IPI PAN

13 czerwca 2022

Plan prezentacji

IPI PAN

Instytut Podstaw Informatyki Polskiej Akademii Nauk zajmuje się

- prowadzeniem badań podstawowych z informatyki - matematycznych lub bliskich matematyce (m.in. logika, teoria grafów, kryptografia, związki z teorią informacji, analizą modeli regresyjnych i innymi działami probabilistyki),
- z naciskiem na tworzenie oryginalnych rozwiązań o ewidentnych walorach aplikacyjnych (np. w lingwistyce komputerowej czy systemach odkrywania wiedzy z danych).
- poczynając od logicznych i teorio-informacyjnych podstaw informatyki, poprzez prace dotyczące systemów rozproszonych, kryptografii, maszynowego uczenia się oraz inżynierii lingwistycznej i ekstrakcji informacji z wielkich kolekcji dokumentów.

Zespół Podstaw Sztucznej Inteligencji IPI PAN

prowadzi prace naukowo-badawcze w dziedzinie systemów inteligentnych w następujących kierunkach:

- inżynieria pozyskiwania wiedzy
-
  - wyszukiwarki internetowe
  - pozyskiwanie wiedzy z danych, tekstu i hipertekstu
- metaheurystyki w optymalizacji
  - optymalizacja w środowiskach dynamicznych
  - inspirowane naturą metody optymalizacji trendów

Zespół opracował i wdrożył masowo-równoległą wyszukiwarkę internetową NEKST.PL, która gromadzi zasoby całego polskiego Internetu.
Naszą specjalnością jest systematyzowanie zasobów internetowych oraz ich udostępnianie użytkownikowi. Systematyzacja oznacza automatyczny podział zasobów internetowych na grupy tematyczne, wyróżnianie kanałów tematycznych w serwisach internetowych, oraz etykietowanie i kategoryzowanie dokumentów i ich grup.

GOSPOSTRATEG / INSTATCENY

GOSPOSTRATEG

strategiczny program badań naukowych i prac rozwojowych NCBR „Społeczny i gospodarczy rozwój Polski w warunkach globalizujących się rynków”

INSTATCENY

Projekt ”Budowa zintegrowanego systemu statystyki cen detalicznych” realizowany przez GUS, SGH i IPI PAN w latach 2019-2022 w ramach programu GOSPOSTRATEG

Zadania realizowane przez IPI PAN

1. Identyfikacja obecnie stosowanych rozwiązań informatycznych w procesie zbierania, przetwarzania i analizy danych o cenach detalicznych.
2. Ocena głównych kierunków prac będących podstawą do stworzenia innowacyjnego systemu informatycznego zbierającego dane o cenach i służącego do ich analizy.
3. Ocena na podstawie rekomendowanych metod zbierania i pomiaru cen możliwości i ograniczeń do zbudowania zintegrowanego systemu informatycznego uwzględniającego uwarunkowania infrastrukturalne i technologiczne.
4. Opracowanie funkcjonalności systemu automatycznego pozyskiwania danych z Internetu.

Wyniki

- Oprogramowanie do wykorzystania alternatywnych źródeł danych w analizie cen
  - `IC.Scraper`- pająk/scraper stworzony w ramach projektu `InstatCeny`, powstał jako alternatywa dla ankieterów fizycznie odwiedzających sklepy celem notowania informacji o cenach produktów znajdujących się w koszyku inflacyjnym GUS.
  - `IC.Classifier`- Klasyfikator produktów do kategorii COICOP
  - `IC.Matcher`- Moduł dopasowania produktów / śledzenia między okresami czasu
  - `IC.PriceIndexer`- wrapper opracowanego na SGH modułu obliczania wskaźników zmienności cen
- Badania nad skutecznością w/w metod

Sposób implementacji modułów

- Funkcjonalność oprogramowania stworzonego w ramach projektu `InstatCeny` nie jest wystawiana do użytku publicznego
- Mimo to jest zrealizowana w konwencji usług (serwisów) intranetowych, w wewnętrznej sieci lub w pseudosieci na pojedynczym komputerze.
- Dostęp do usług w tym kontekście uzyskuje się przez tzw. *endpointy*, czyli kanały wymiany informacji, opisane (rozbudowanymi) adresami URL usług.

Implementacja IC.Scrapera

zadanie IC.Scrapera

odwiedzanie stron internetowych sieci handlowych, a następnie parsowanie pobranego kodu źródłowego strony celem ekstrakcji istotnych cech produktu.

Technika pracy

zespół wzajemnie współpracujących mikroserwisów i w taki sposób pozwala na tworzenie ciągów przetwarzania danych (komunikacja przez zapytania w protokole HTTP, czyli możliwość pracy rozproszonej w sieci WWW).

Algorytmika

Skupienie się wyłącznie na witrynach internetowych sieci handlowych, które umożliwiają zakupy online, a tym samym udostępniają w pełni istotne, z punktu badania inflacji, informacje o produktach - tj. np cenę produktu.

Założenia o sieci handlowej

- Dostępne w ofercie produkty sieci handlowych są pogrupowane w odpowiednie kategorie (np. nabiał, kawy, owoce itp.).
- Parsowanie kodów źródłowych stron internetowych kategorii produktów pozwala na wyłuskanie adresów URL pojedynczych produktów.
- Parsowanie kodów źródłowych stron pojedynczych produktów pozwala wyłuskać informacje dotyczące konkretnego produktu.

Praca IC.Scrapera

- Logika IC.Scrapera uruchamiana co 24h, licząc od momentu rozpoczęcia procesu (czyli startu całej aplikacji).
- Po uruchomieniu - pobieranie z odpowiedniej tabeli w bazie danych konfiguracji sieci handlowych do przetworzenia. Konfiguracja każdej z sieci handlowych to szereg adresów URL definiujących kategorie produktów.
- Adresy te są odwiedzane i strony WWW są ściągane.
- Adresy produktów są ekstrahowane z pobranych stron kategorii, a następnie ekstrahuje się z nich informację o produktach

Nowa sieć handlowa - opis bazodanowy

- Sieć handlowa jest opisana w ramach niniejszego systemu przez dwa komponenty: (1) uzupełnienie zawartości bazy danych `IC.Scrapera` oraz (2) uzupełnienie kodu źródłowego `IC.Scrapera` o metody obsługi nowej sieci handlowej.
- Po wybraniu odpowiedniej sieci handlowej jako źródło danych należy w pierwszej kolejności dodać tą sieć jako jedno ze źródeł `IC.Scrapera`. Służy do tego odpowiednie zapytanie szerzej opisane w dokumentacji technicznej.
- Należy określić kategorie pobieranych produktów podając adresy internetowe (URL) stron WWW sieci handlowej, z tymi kategoriami produktów. Adresy URL muszą zostać podane wprost, z racji mnogości dostępnych produktów oraz skorelowanego z tym czasem działania `IC.Scrapera`. Adresy te należy dodać do konfiguracji sieci handlowej poprzez API `IC.Scrapera`.

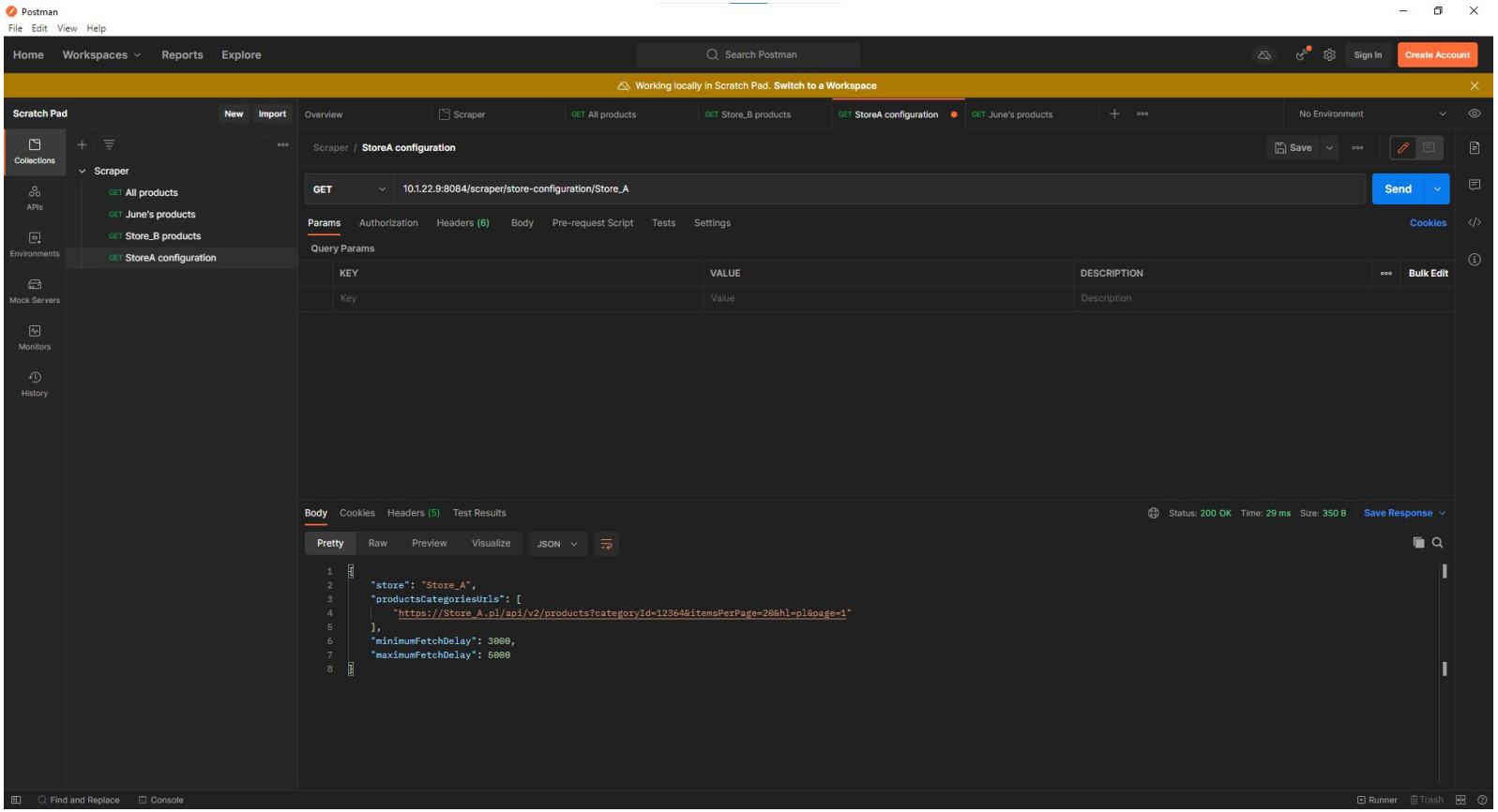
Nowa sieć handlowa - oprogramowanie

- Należy utworzyć klasę w kodzie źródłowym `IC.Scrapera`[1], której nazwa musi być tożsama z nazwą sieci handlowej, wpisaną do bazy danych. Klasę tę definiuje się jako rozszerzenie klasy abstrakcyjnej „`Store`", przy czym wymagane jest zaimplementowanie poniższych metod.
  - `parseProductsLinks` - ekstrakcja linków do produktów ze strony kategorii produktów
  - `parseProduct` - ekstrakcja własności produktu (opis, cena itd.)
- Nazwę stworzonej klasy należy dodać do bazy danych `IC.Scrapera`. Kluczem głównym rekordu w bazie danych przechowującym konfigurację sieci handlowej jest wprost nazwa zaimplementowanej klasy w kodzie źródłowym scrapera.

[1] `IC.Scraper` jest napisany w języku programowania `Java`

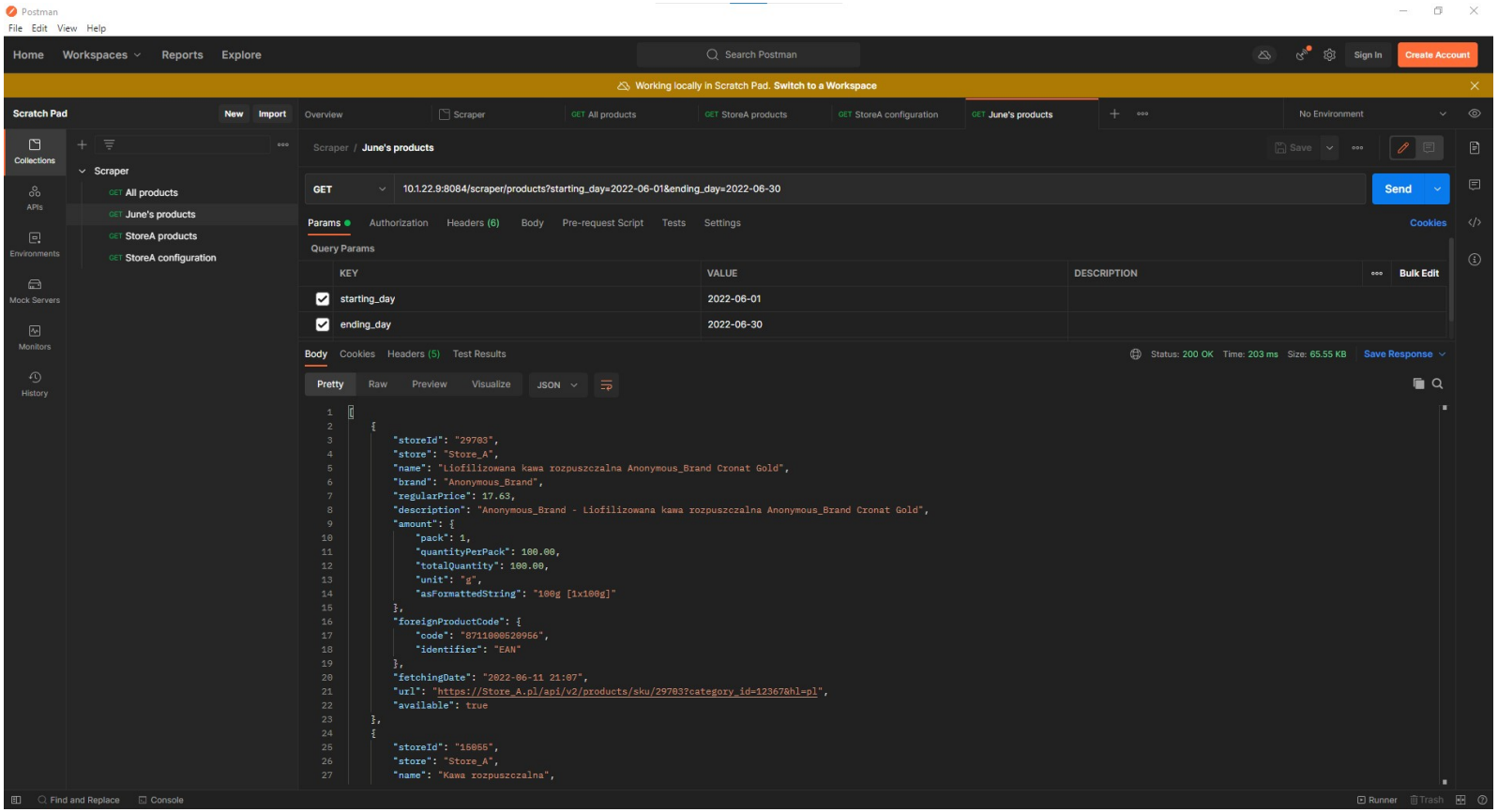
Implementacja IC.Scrapera - konfigurowanie

Konfigurowanie sieci handlowej



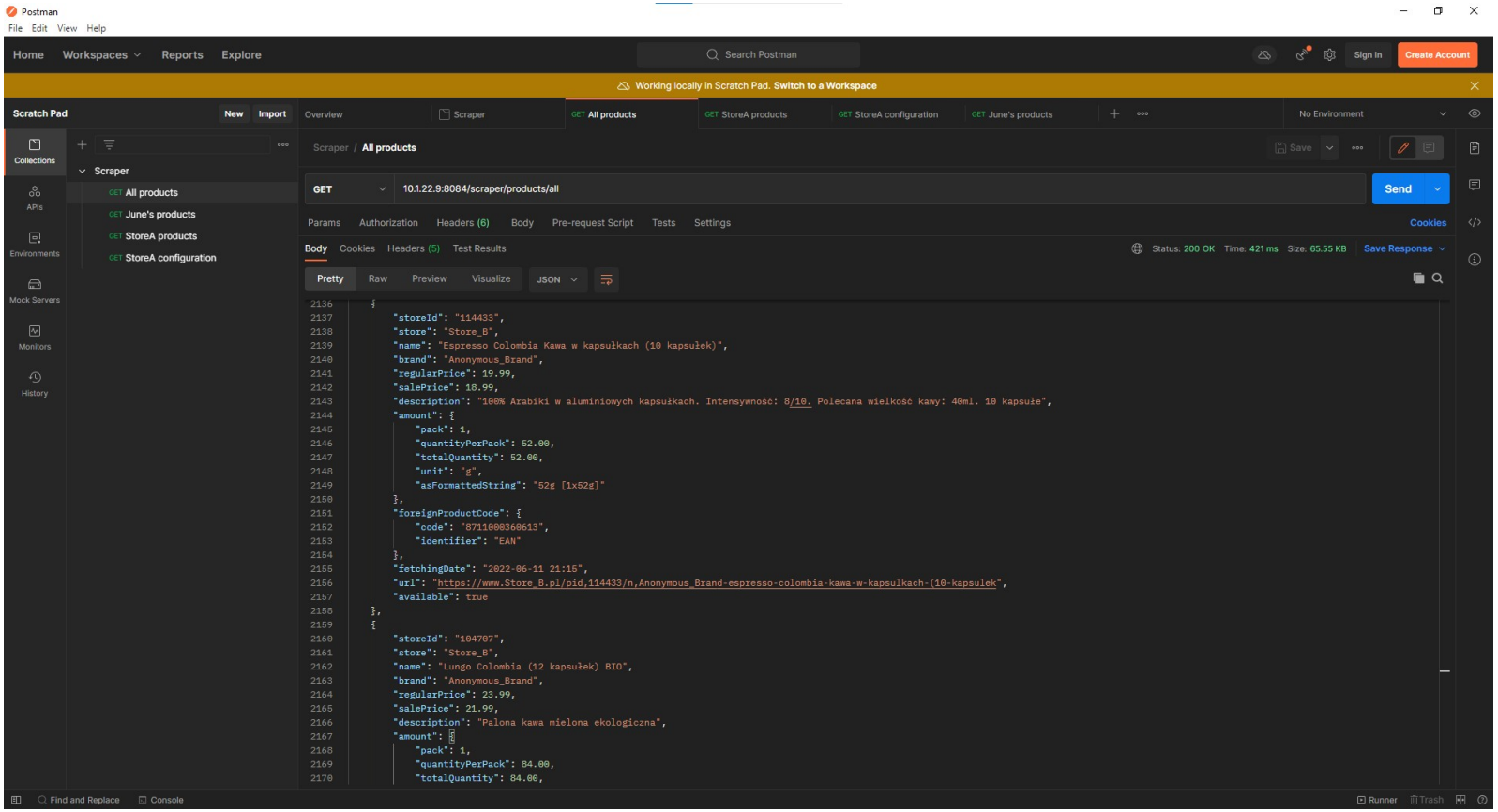
Implementacja IC.Scrapera - zapytania do bazy

Produkty pobrane w czerwcu - zrzut bazy danych



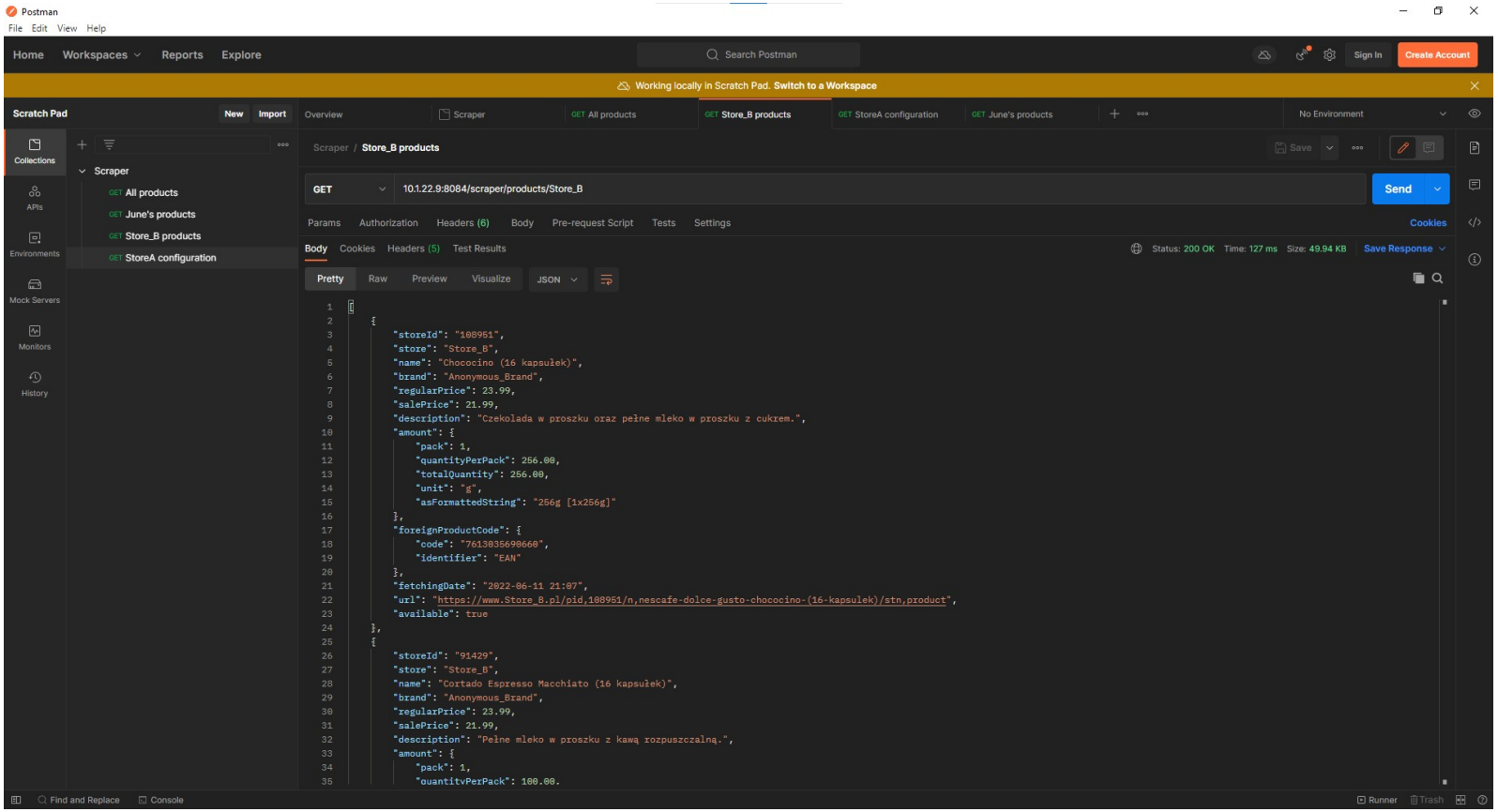
Implementacja IC.Scrapera - zapytania do bazy

Wszystkie pobrane produkty - zrzut bazy danych



Implementacja IC.Scrapera - zapytania do bazy

Produkty pobrane z konkretnej sieci handlowej - zrzut bazy danych



Implementacja IC.Scrapera - wykonanie (filmik)

- uruchomienie scrapera z linii komend,
- wczytanie z bazy danych informacji na temat konfiguracji sklepów (m. in. linki do kategorii, czy wielkość zwłoki pomiędzy kolejnymi zapytaniami do serwerów sieci),
- odpytywanie serwerów sieci zgodnie z adresami do kategorii parsowanie otrzymanej odpowiedzi celem ekstrakcji adresów URL do poszczególnych produktów,
  - odpytywanie serwerów sieci
  - zgodnie z adresami do poszczególnych produktów
  - parsowanie otrzymanej odpowiedzi celem ekstrakcji informacji o produkcie
  - zapis sparsowanego produktu do bazy
- zakończenie działania scrapera

Implementacja IC.Scrapera - wykonanie (filmik)

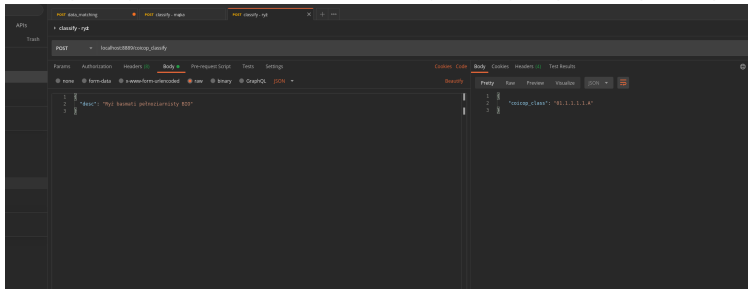
Kategorie pobieranych produktów zawężono do jednej kategorii zawierającej kawy.

Okrojono liczbę produktów do scrapowania (Auchan [Store_A] - 6 sztuk, Frisco [Store_B] 5 sztuk).

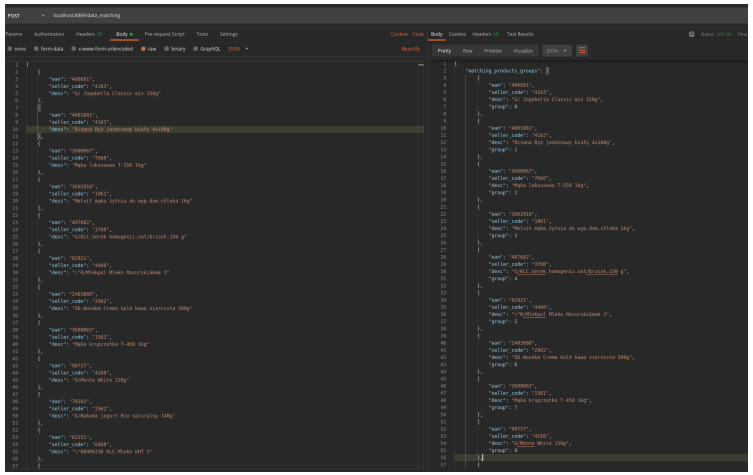
Dla niektórych produktów parser dodatkowo ekstrahuje informację o cenie promocyjnej, czy produkt jest dostępny oraz informację o ilości (istotne np. dla zgrzewki wody 6x1.5l).

Implementacja IC.Classifiera i IC.Matchera

IC.Classifier uruchamianie



IC.Matcher uruchamianie



zaprezentuje dr Piotr Borkowski

Implementacja `IC.PriceIndexera`

- Aplikacja `IC.PriceIndexer` udostępnia funkcje z projektu *PriceIndices* utworzonego przez prof. Jacka Białka.
- Aplikacja wystawia wybrane funkcje z projektu jako *endpointy* RESTowe
- Wystawione funkcje pozwalają na
  - przygotowanie danych skanowanych do obróbki pod kątem liczenia wskaźników inflacji
  - wybór danych
  - klasyfikację na bazike cech numerycznych
  - filtrację
  - obliczania samych wskaźników inflacji.
- Formaty danych są takie, jak dla pozostałych zaimplementowanych modułów
- szczegółowe informacje - prof. Jacek Białek

Badania zaimplementowanych modułów - Zakres

- Unikatowość kodów i etykiet produktów
- Efektywność klasyfikacji do kategorii COPICOP
- Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP
- badanie klasyfikatora K1 na bazie zidentyfikowanych słów pozytywnych i negatywnych związanych z klasami COICOP
- Śledzenie produktów
  - Klasyfikacja produktów miesiąc po miesiącu
  - Dopasowanie produktów miesiąc po miesiącu
- Wsparcie informatyczne przy wyborze reprezentantów dla kategorii COICOP

Wnioski z badań

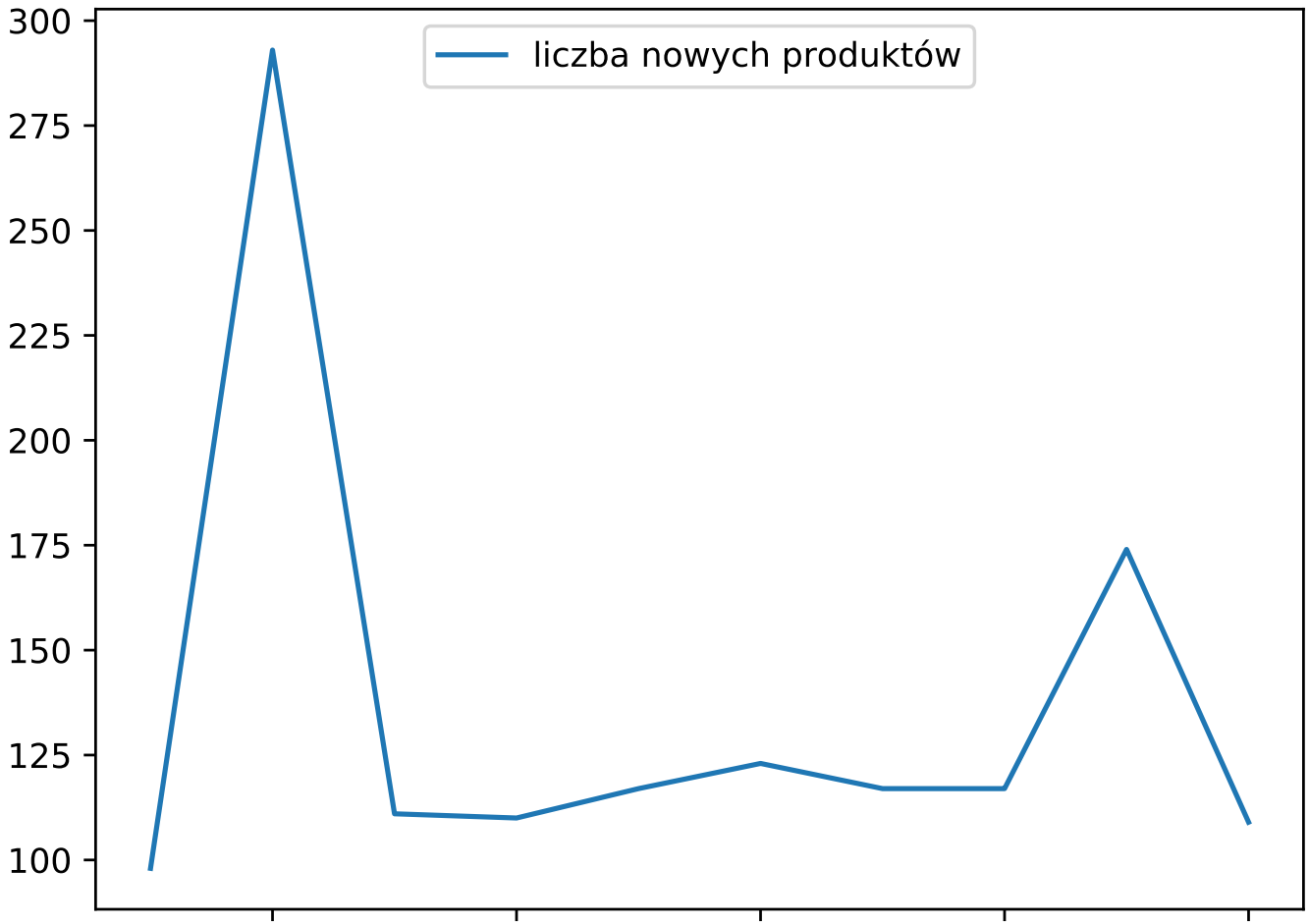
- **Unikatowość kodów i etykiet produktów**
  - Kody produktów nie są unikatowe w relacji do GITN, ani w drugą stronę.
  - Jest to poważny problem merytoryczny w ramach pobierania danych od gestorów - żaden z tych kodów nie może być wykorzystywany do stwierdzenia pojawiania się nowych produktów ani zanikania sprzedaży już istniejących
  - Gestorzy przywiązują większą wagę do nadania własnego kodu produktu niż do norm międzynarodowych.
  - Kod produktu od gestora nie daje pewności ciągłości / nieciągłości sprzedaży tego samego produktu z punktu widzenia analizy cen, gdyż zmiana tego kodu może być związana z ruchem cen, a nie wprowadzeniem nowego produktu do obrotu
  - Konieczne jest więc wykorzystanie modułu `IC.Matcher` do wskazania sytuacji wątpliwych.

Liczba kodów „gtin" odpowiadająca pojedynczemu „id"	liczba przypadków
1	1353
2	5800
3	1026
4	916
5	273
6	181
7	59
8	39
9	27
10	30
11	6
12	11
13	2
14	4
15	3
16	1
19	2

Wnioski z badań cd.

- **Zmienność produktów**
  - Zmienność oferty produktów jest dość wysoka - z miesiąca na miesiąc wynosi zwykle powyżej 5%, ale może sięgać nawet 40% oferowanych produktów.
  - W tym kontekście zdolność do automatycznej klasyfikacji do COICOP nowych produktów przybiera na znaczeniu.

Liczba nowych towarów wg cechy „id” w 2018 - `GestorAlfa()`



Wnioski z badań cd.

- **Klasyfikacja** Przeprowadzone eksperymenty prowadzą do następujących wniosków:
  - dla dostępnych danych najlepiej sprawdzały się metody: LinearSVC (dokładność 93% poziom 3 COICOP)), Naiwny Bayes (93%), Regresja logistyczna (92%)
  - klasyfikacja do wyższych poziomów hierarchii wypada nieco lepiej (93% dla poz.3, 91% dla poz.5 dla maszyn wektorów nośnych);
  - stosowane metody pozwalają na ekstrakcję słów z nazw produktów charakteryzujących klasy;
  - klasy z małą liczbą produktów są czynnikiem ryzyka;
  - wyzwaniami będą: dane z większą liczbą rekordów / z większą liczbą klas;

  Wspomnianym problemom mogą zaradzić:
  - zwiększenie liczby klasyfikatorów (komitety)
  - klasyfikatory hierarchiczne (klasyfikacja na kolejne poziomy hierarchii)
  - klasyfikacja nie kategorii lecz słów należących do nazwy kategorii

Skuteczność klasyfikacji

GestorBeta			GestorGama		
coicop lvl: 3	coicop lvl: 4	coicop lvl: 5	coicop lvl: 3	coicop lvl: 4	coicop lvl: 5
klasyfikator LinearSVC					
0.933 (0.049)	0.909 (0.046)	0.917 (0.026)	0.988 (0.001)	0.930 (0.007)	0.932 (0.006)
klasyfikator MultinomialNB					
0.936 (0.063)	0.887 (0.042)	0.892 (0.029)	0.987 (0.003)	0.928 (0.012)	0.921 (0.015)
klasyfikator Regresja logistyczna					
0.921 (0.063)	0.880 (0.058)	0.887 (0.027)	0.980 (0.005)	0.907 (0.009)	0.902 (0.007)
klasyfikator Random Forest					
0.911 (0.065)	0.869 (0.045)	0.883 (0.039)	0.977 (0.005)	0.928 (0.011)	0.917 (0.004)

Tabela: Średnie i odchylenia standardowe miar dokładności różnych metod klasyfikacji uzyskanych dla 5-krotnej kroswalidacji na zbiorach `GestorBeta` i `GestorGama` .

Wnioski z badań cd.

- **Ekstrakcja słów kluczowych i klasyfikacja na ich podstawie**
  - Najbardziej skuteczną a zarazem intuicyjną metodą ekstrakcji słów kluczowych jest wykorzystanie modelu stworzonego na bazie naiwnego klasyfikatora bayesowskiego.
  - Pozwala on na zidentyfikowanie zarówno pozytywnych jak i negatywnych słów kluczowych, charakteryzujących klasę COICOP.
  - Wykorzystanie tych słów kluczowych daje bardzo dobre wyniki przy identyfikacji klasy na podstawie opisu produktu.
  - Podejście teoriomnogościowe, oparte o wykrywanie słów różnicujących klasy, jest mniej efektywne, choć daje także dobre efekty, gdyż dla niektórych klas COICOP nie da się takich słów ustalić ze względu na współdzielenie z innymi klasami.
  - Tym nie mniej klasyfikator klasyfikujący wyłącznie na podstawie słów pozytywnych uzyskuje dobre wyniki klasyfikacji do COICOP (dokładność ponad 95% dla 9 wybranych klas).

Przykłady wybranych słów kluczowych

- **CP1181**: **(CUKIER)** drobny, cukier, goldpack, trzcinowy
- **CP1142**: **(niskotłuste mleko)** hula, międzybórz, śmietana, mlektar, bychawa, wieluń, jarocin, homog, myszków, 10, łask, krasulamlekopaster, 406237, kar, jędrzejów, łowickie, grudziądz, osm, spoż, spożywcze, głubczyckie, osowa, litr
- **CP1143**: **(mlerko konserwowane)** zagęszcz, niesł7, puszka411g, 4kg, proszku,
- **CP11122**: 997, owsiana, żytnia, żytniat, schaer, dom, proso, gryczana, it, vitacorn, 720, 2000, provena, ryżowa, uniwersaln, owies, kukurydzia, jaglana,
- **CP11142**: hula, międzybórz, śmietana, mlektar, bychawa, wieluń, jarocin, homog, myszków, 10, łask, krasulamlekopaster, 406237, kar, jędrzejów, łowickie, grudziądz, osm, spoż, spożywcze, głubczyckie, osowa, litr,
- **CP1111**: hom, kuchnia, sys, 4001561, pełnoziarn, 5x100g, biała, mal, parabolicz, ze, 2x100, 2x150g, amerykańsk, długoziarn, jaśminowy, dzikim, torebka, gosposi, czerwony, długi, risotto, dziki, delebiały, arroz, carnaroli, tajlandii, parboiled, kupiec, ryż, cenos,

Skuteczność klasyfikacji na podstawie pozytywnych słów kluczowych

cechy do klasyfikacji: „opis towaru”	
klasyfikator	
`GestorAlfa`	
K1	0.959 (0.005)

Tabela: Średnie i odchylenia standardowe miar dokładności uzyskanych dla klasyfikacji na bazie wyłącznie słów pozytywnych, ocenionych z zastosowaniem 10-krotnej kroswalidacji ($cv = 10$).

Dziękuję

Serdecznie dziękuję za uwagę

Chętnie odpowiem na pytania, wysłucham sugestii czy komentarzy.