Ćwiczenie 1 & 2

Celem biznesowym jest wyciąganie danych od konkurencji (sklepu books to scrape), aby dowiedzieć się więcej nt. Cen i dostępności tytułów. Dzięki wyciągniętym danym (do pliku .csv) będzie możliwe utworzenie raportów, dashboard'ów analizujących dane tytuły. Pomoże to w określeniu cen, które nasza firma może zaproponować za te same tytuły.

Strona, z której pobierane są dane o książkach konkurencji to https://books.toscrape.com/. Jest ona podzielona na podstrony, na każdej podstronie znajdują się informacje o 20 książkach. Podstron jest na moment tworzenia kodu 50. Liczba ta może się zmienić, gdy konkurencja doda/usunie książki. Kod jest na taki scenariusz odporny.

Najpierw wykonuje się zapytanie dla pierwszej podstrony, z której wyciągane są informacje o: tytule, cenie i dostępności. Te 3 zmienne reprezentują kolumny wynikowej ramki danych. Po wyciągnięciu danych z 1 strony pobierają się kolejne (z odstępem 1 sekundy pomiędzy każdą, aby uniknąć przeciążenia serwera). Każde wyciągnięcie danych z podstrony jest rejestrowane w terminalu:

```
PS C:\AGH - Business Intelligence> & "C:/AGH - Business Intelligence/BI/Scrip
-1.py"
Pobrano stronę 1
Pobrano stronę 2
Pobrano stronę 3
Pobrano stronę 4
Pobrano stronę 5
Pobrano stronę 6
Pobrano stronę 7
```

Po przejściu procesu wyciągania danych i zapisu ich do ramki danych, następuje eksport do pliku .csv. Poniżej zaprezentowano przykład tego pliku:

Ex-01 > export.csv

```
1   Tytuł,Cena,Dostępność
2   A Light in the Attic,51.77,In stock
3   Tipping the Velvet,53.74,In stock
4   Soumission,50.10,In stock
5   Sharp Objects,47.82,In stock
6   Sapiens: A Brief History of Humankind,54.23,In stock
7   The Requiem Red,22.65,In stock
8   The Dirty Little Secrets of Getting Your Dream Job,33.34,In stock
9   "The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria Woodhull",17.93,In stock
10  The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics,22.60,In stock
11  The Black Maria,52.15,In stock
12  "Starving Hearts (Triangular Trade Trilogy, #1)",13.99,In stock
13  Shakespeare's Sonnets,20.66,In stock
```

Dodano 2 przypadki obsługi błędów:

- W przypadku, gdy serwer nie odpowiada (błąd przy przetwarzaniu danej strony),

- W przypadku, gdy występuje problem z książką (błąd przy czytaniu wartości o danej książce).

Proponuję codzienne aktualizowanie danych nt. książek przy pomocy stworzonego przeze mnie algorytmu (np. przy pomocy darmowego oprogramowania Apache Airflow i języka Python).

Z racji, iż ciężko jest zweryfikować liczbę konkurencyjnych księgarni w okolicy, proponowałbym ograniczyć się tylko do sprzedaży internetowej. Argumentuję to również stale zmieniającym się rynkiem i malejącą ilością osób, które kupują książki w sklepach stacjonarnych.

Oczywiście, jest możliwe wyciąganie danych nt. pobliskich księgarni ze strony Głównego Urzędu Statystycznego, lecz opracowanie algorytmu odpowiedzialnego za ten proces mogłoby być zbyt skomplikowane. Argumentuję to faktem, iż wyciąganie danych z GUS-u przy użyciu API nie należy do najłatwiejszych. Ponadto GUS bardzo mocno ogranicza liczbę zapytań, które można za pomocą API wykonać (w celu nieprzeciążenia serwera).

Poniżej przedstawiam maksymalną liczbę żądań w danym okresie czasu. Proszę sugerować się czasami dla anonimowego użytkownika, gdyż z racji bezpieczeństwa firmy, w sposób anonimowy wysyłane byłyby żądania.

- Anonimowy użytkownik (niezalogowany) może wykonać do 5 żądań na sekundę, 100 żądań na 15 minut, 1000 żądań na 12 godzin, 10000 żądań na 7 dni.
- Zarejestrowany użytkownik może wykonać do 10 żądań na sekundę, 500 żądań na 15 minut, 5000 na 12 godzin, 50000 na 7 dni.

Poniżej prezentuję przykładowe wyciągnięcie danych. Tak jak wspominałem, dane są niezwykle chaotyczne i nieczytelne.


```
-3': 562, 'id-pozycja-3': 630880, 'id-okres': 251, 'id-sposob-prezentacji-miara': 5, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosc
i': 43, 'id-flaga': 36, 'wartosc': 118.4, 'precyzja': 1}, {'rownumber': 1180, 'id-zmienna': 305, 'id-przekroj': 736, 'id-wymiar-1': 2, 'id
-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 630880, 'id-okres': 251, 'id-sposob-p
rezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.4, 'precyzja': 1}, {'row
i': 43, 'id-flaga': 36, 'wartosc': 118.4, 'precyzja': 1}, {'rownumber': 1180, 'id-zmienna': 305, 'id-przekroj': 736, 'id-wymiar-1': 2, 'id
-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 630880, 'id-okres': 251, 'id-sposob-p
rezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.4, 'precyzja': 1}, {'row
-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 630880, 'id-okres': 251, 'id-sposob-p
rezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.4, 'precyzja': 1}, {'row
rezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.4, 'precyzja': 1}, {'row
number': 1181, 'id-zmienna': 305, 'id-przekroj': 736, 'id-wymiar-1': 2, 'id-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171
, 'id-wymiar-3': 562, 'id-pozycja-3': 630882, 'id-okres': 251, 'id-sposob-prezentacji-miara': 5, 'id-daty': 2023, 'id-brak-wartosci': 253,
 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 118.6, 'precyzja': 1}, {'rownumber': 1182, 'id-zmienna': 305, 'id-przekroj': 736, 'id-wymia
r-1': 2, 'id-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 630882, 'id-okres': 251,
'id-sposob-prezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.4, 'precyzja
': 1}, {'rownumber': 1183, 'id-zmienna': 305, 'id-przekroj': 736, 'id-wymiar-1': 2, 'id-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja
-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 631528, 'id-okres': 251, 'id-sposob-prezentacji-miara': 5, 'id-daty': 2023, 'id-brak-war
tosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 118.7, 'precyzja': 1}, {'rownumber': 1184, 'id-zmienna': 305, 'id-przekroj': 73
6, 'id-wymiar-1': 2, 'id-pozycja-1': 33617, 'id-wymiar-2': 563, 'id-pozycja-2': 6656171, 'id-wymiar-3': 562, 'id-pozycja-3': 631528, 'id-o
kres': 251, 'id-sposob-prezentacji-miara': 2, 'id-daty': 2023, 'id-brak-wartosci': 253, 'id-tajnosci': 43, 'id-flaga': 36, 'wartosc': 101.
4, 'precyzja': 1}]}
(BI) PS C:\AGH - Business Intelligence>
```

Podsumowując, z powodu zbyt skomplikowanych danych i nieczytelnych danych o konkurencji, a także małej liczby potencjalnych zapytań do strony GUS rekomenduję dwie ścieżki biznesowe:

1. Wycofanie się ze sprzedaży w sklepach stacjonarnych i ograniczenie się tylko i wyłącznie do sprzedaży internetowej
2. Zatrudnienie doświadczonego z danymi GUS-owymi specjalisty, w celu przygotowania, wdrożenia i utrzymywania/ulepszania algorytmu służącego do wyciągania i analizy danych o konkurencji ze strony GUS.