2. Webscraping

Wszystkie zadania należy wykonać na następujące sposoby:

1. używając wyrażeń regularnych (regex),
2. ściągając dostępny plik csv (tam gdzie są),
3. używając parsera html,
4. wczytywania ramek Pandas (tam, gdzie to ma sens),
5. używając scrappy'ego

1.0 Napisz funkcję, która dla zadanej strony pobierze wszystkie linki na tej stronie. Funkcja powinna zwracać ramkę danych Pandas, która zawiera:

a) Adres strony internetowej, do której prowadzi link
b) Napis wyświetlany na stronie

1.1 Dla zadanej strony pobierz wszystkie e-maile na tej stronie. Przykładową stroną może być https://hunter.io/ .

1.2 Napisz funkcję Notowania(spolka), która dla danego waloru (np. MBK, CDR, FTE) zwróci nam ramkę danych Pandas, w której są następujące kolumny: data (razem z godziną), kurs, max, min, otwarcie, wolumen, obrót, transakcje. Ramka powinna zawierać 6 wierszy, oznaczających ostatnie 6 dni w których była czynna giełda. Należy użyć strony https://stooq.pl/q/?s=cdr , a także odnośników do ostatnich 6 dni z dołu strony. Bonus: narysuj wykres notowań spółki.

1.3 Podobnie jak poprzednio, napisz funkcję NotowaniaHistoryczne(spolka, dataStart, dataKoniec), która zwróci nam ramkę danych Pandas zawierającą dane pod adresem np. https://stooq.pl/q/d/?s=cdr , z przedziału dat zadanych jako argumenty. Być może trzeba będzie doczytywać kolejne strony. Pamiętaj, aby odpowiednio wczytać daty (i żeby to naprawdę były daty w ramce Pandas). Bonus: narysuj wykres notowań spółki.

1.4 Na wielu stronach wikipedii dotyczących poszczególnych miast na świecie można znaleźć tabelę z temperaturami/opadami dla poszczególnych miesięcy w danym mieście. Napisz funkcję pogoda(miasto), która pobierze taką tabelę z internetu i zwróci ją w formie ramki danych Pandas. Bonus: narysuj wykres temperatury w poszczególnych miesiącach. Może być także boxplot.

1.5a Napisz program, który dla zadanego tytułu filmu zwróci jego różne cechy na filmwebie (jako ramkę danych Pandas): ocenę, oryginalny tytuł, reżysera, scenariusz, gatunek, kraj produkcji, boxoffice. Pewnym problemem w tym zadaniu może być odnalezienie strony na filmwebie odpowiadającej danemu filmowi. Proponowane rozwiązanie jest następujące: wyszukaj film poprzez np. http://www.filmweb.pl/search?q=ciemniejsza+strona+greya , a następnie wejdź w pierwszy link (oczywiście programowo).

1.5b Na stronie http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm można znaleźć najlepiej zarabiające filmy w 2017 roku. Napisz funkcję najlepiejZarabiajaceFilmy(rok), która zwróci tabelę z tej strony w formie ramki danych Pandas. Bonus: narysuj boxplot zarobków w danym roku.

1.5c Napisz funkcję najlepiejZarabiajaceFilmyPlusFilmweb(rok), która zwróci tabelę, w której są najlepiej zarabiające filmy z danego roku, a także informacje o nich z Filmweba: ich ocena na filmwebie, reżyser, scenariusz, gatunek etc. Czy boxoffice z Filmweba zgadza się z tym z boxofficemojo? Dlaczego?

1.6 Na Wikipedii istnieją strony dotyczące danego roku, np. https://en.wikipedia.org/wiki/2010 . Na takiej stronie mamy różne wydarzenia, pogrupowane po miesiącach. Napisz kod, który wyciągnie i wrzuci w ramkę danych Pandas zdarzenia lub śmierci. Jest to zadanie

zainspirowane serwisem http://www.vizgr.org/historical-events/ , którego dotyczyło jedno z zadań w poprzedniej liście zadań.

1.7 Dla zadanej strony uczelni wyższej lub wydziału uczelni wyższej postaraj się pozyskać imiona, nazwiska i stopnie naukowe wszystkich jej pracowników (szukamy ciągów znaków typu “mgr inż. Imię Nazwisko”).

1.8 Dla strony http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm zwróć informację o najlepiej zarabiającym filmie w każdym roku. Po latach należy przechodzić poprzez linki u dołu strony (Previous year). Wersja trudniejsza: zwróć informację o średnich zarobkach 100 pierwszych filmów w każdym roku.