

PAD 06

Praca Domowa 6 – Web Scraping

1. Przed rozpoczęciem zainstaluj bibliotekę selenium poleceniem:

pip install selenium lub *conda install selenium*

2. Sprawdź wersję Chrome, którą masz zainstalowaną na komputerze.
3. Ściągnij chromedriver.exe ze strony <https://chromedriver.chromium.org/downloads>, która jest zgodna ze sprawdzoną wersją Google Chrome
4. Umieść chromedriver.exe w tym samym katalogu, w którym będziesz pracował.

Zadanie 1 (1 pkt)

Wejdź na stronę <https://www.pap.pl/> i sprawdź czy pozwala ona na web scraping sprawdzając plik robots.txt

Zadanie 2 (8 pkt)

Stwórz obiekt driver, który połączy się ze stroną Polskiej Agencji Prasowej. A następnie:

- a. Zaakceptuje pliki cookies
- b. Zwiększy okno przeglądarki na cały ekran
- c. Zmieni język strony na angielski
- d. Wejdzie w sekcję Business
- e. Z sekcji business ściągnie wszystkie tytuły do listy *titles*
- f. Ściągnie wszystkie zdjęcia z tej sekcji na dysk lokalny
- g. Zjedzie na dół strony
- h. Przejdzie na ostatnią stronę i zwróci printem jej numer (atrybut text)

Zadanie 3 (1 pkt)

Wybierz stronę, z której będziesz pobierał dane do projektu i podaj url w ramach odpowiedzi w MSTEams. Zastanów się nad tematem, który Cię interesuje i połączeniem go z uczeniem maszynowym (czyli budową modelu, który może coś przewidywać np. prognoza cen/ przypisanie klasy do danej kategorii). Unikamy tematów NLP i Computer Vision, bo na razie na to za wcześnie, a są bardziej wymagające, i będą w 3 semestrze 😊