Eksploracja danych w Internecie

(Kod przedmiotu: 02 64 5054 00)

Laboratorium 03

Cel: Nabranie umiejętności posługiwania się narzędziem *Selenium*.

Zadanie 1: Poprzednio utworzony projekt miał problemy (np. część zawartości nie była dostępna) z wieloma stronami internetowymi, które korzystały z interfejsu (frontend'u) napisanego w popularnych obecnie silnikach *JavaScript*, takich jak np. *Angular.js, React.js, Vue.js*. Jest to spowodowane tym, że zawartość strony jest dynamicznie generowana za pośrednictwem metod uruchamianych po stronie klienta. W skrócie, w pierwszej kolejności pobierana z serwera jest strona internetowa zawierająca podstawowe komponenty a dopiero po ich pobraniu budowana jest zawartość strony.

W związku z powyższym narzędzie zaimplementowane na poprzednich laboratoriach nie zadziała (chyba, że strona korzysta z *Server Site Rendering*). Problem ten można rozwiązać na wiele sposobów i jednym z nich jest zastosowanie narzędzia *Selenium*.

Twoim zadaniem jest utworzenie Scrapera, który będzie pobierać dane (adresy e-mail) korzystając z narzędzia *Selenium*. Wybierz taką stronę, która pobiera dane za pomocą skryptów *javascript*. Porównaj wyniki z poprzednio zaimplementowanym rozwiązaniem.

Do zadania dodany jest notatnik *lab07-09-selenium-przyklad.ipynb* z przykładem użycia Selenium.

Zadanie 2: Wykorzystaj narzędzie *Selenium* w celu wypełnienia formularza na dołączonej do zadania stronie internetowej (napisana we flasku – instrukcja uruchomienia w pliku *app.py*). Po wypełnieniu i wysłaniu formularza wyświetli się stosowny komunikat.

Materiały:

- Dokumentacja Selenium
  https://www.selenium.dev/documentation
- Pillow (do przetwarzania plików graficznych)
  https://python-pillow.org/
- OpenCV (do przetwarzania plików graficznych)
  https://docs.opencv.org/master/index.html
- Matplotlib (do wizualizacji)
  https://matplotlib.org/
- Metryki
  https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics
- Tablica pomyłek
  https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html