

Lab 08

WebScrapping – wyszukiwanie artykułu na Wikipedii

W tym zadaniu postanowiłem wykorzystać zdobytą wiedzę z poprzednich laboratoriów o WebScrappingu do pobrania artykułu z serwisu **Wikipedia** wskazanego przez użytkownika. W ramach tego zadania będę pobierał jedynie informacje tekstowe takie jak nazwa opisywanego artykułu, jego treść oraz ewentualne linki zewnętrzne.

```
C:\Java\jdk1.8.0_191\bin\java.exe ...
Enter a name of an article to find on Wikipedia(use underscore(_) as space)
Screen_scraping
Scrapping from http://pl.wikipedia.org/wiki/Screen_scraping
Article name:Screen scraping
Article text:
Z Wikipedii, wolnej encyklopedii Przejdź do nawigacji Przejdź do wyszukiwania Screen scraping (dosł.
Screen scraping (dosł. 'zeskrobywanie z ekranu') to technika, za pomocą której program komputerowy wy
Istnieje wiele synonimów screen scrapingu: data scraping, data extraction, web scraping, page scrapin
Zazwyczaj transfer danych między programami odbywa się za pośrednictwem struktur danych przystosowan
Pierwotnie termin ten odnosił się do czytania danych z pamięci ekranu terminala komputerowego. Analog
Screen scraping jest uznawany za nieelegancką technikę, używaną tylko jako ostateczność, kiedy żaden
External links:1
http://www.perl.com/pub/a/2003/01/22/mechanize.html<- Screen-scraping with WWW::Mechanize
Process finished with exit code 0
```

W powyższym przykładzie wykorzystałem artykuł odnośnie Screen Scrapingu. Niestety nie udało mi się wydzielić czystego tekstu artykułu(bez dopisów o źródle czy tekstów linków skokowych), ponieważ zmieniając parametry pobieranych informacji wraz ze zbędnymi informacjami utraciłbym część informacji z artykułu (na przykład informacje z artykułu o zmienionej kursywie -przykład poniżej).

```
Article name:Screen scraping
Article text:
(dosł. 'zeskrobywanie z ekranu') to technika, za pomocą której wydobywa dane z wyjścia innego programu. Służący do
Istnieje wiele synonimów screen scrapingu: , , , oraz (przy czym ostatnie trzy odnoszą się do ).
Zazwyczaj transfer danych między programami odbywa się za pośrednictwem struktur danych przystosowanych do maszyn,
Pierwotnie termin ten odnosił się do czytania danych z ekranu . Analogicznie screen scraping oznacza także skomput
Screen scraping jest uznawany za nieelegancką technikę, używaną tylko jako ostateczność, kiedy żaden inny mechaniz
External links:1
http://www.perl.com/pub/a/2003/01/22/mechanize.html<- Screen-scraping with WWW::Mechanize
```

Oczywiście program działa również na innych artykułach oraz jest zabezpieczony przed wprowadzeniem nieistniejącego artykułu. (w moim przypadku ograniczyłem się do polskiej wersji językowej portalu)

```
Enter a name of an article to find on wikipedia(use underscore(_) as space)
Program komputerowy
Scrapping from http://pl.wikipedia.org/wiki/Program\_komputerowy
Article name:Program komputerowy
Article text:
Z Wikipedii, wolnej encyklopedii Przejdź do nawigacji Przejdź do wyszukiwania Spis treści 1 Terminol
Program komputerowy (ang. computer program) - sekwencja symboli opisująca realizowanie obliczeń zgod
Formalne wyrażenie metody obliczeniowej w postaci języka zrozumiałego dla człowieka nazywane jest ko
Programy komputerowe można zaklasyfikować według ich zastosowań. Wyróżnia się zatem aplikacje użytko
Określenie „program komputerowy” jest często używane zamiennie z określeniami „oprogramowanie” (ang.
W najprostszym modelu wykonanie programu (zapisanego w postaci zrozumiałej dla maszyny) polega na um
Program komputerowy będący w trakcie wykonania nazywany jest procesem lub zadaniem.
Specjalnym rodzajem programu komputerowego jest system operacyjny (występujący często jako zbiór pro
Aby jakikolwiek program mógł zostać uruchomiony, sprzęt musi zostać do tego przygotowany. Odpowiedzi
Jeżeli program nie jest zapisany w postaci binarnej, ale w postaci kodu źródłowego czytelnego dla cz
Program można podzielić na dwie części (obszary):
O ile operowanie na danych przez program jest zupełnie normalne i oczekiwane, o tyle operowanie prog
Programowanie jest procesem tworzenia programów (względnie: oprogramowania). Jest to proces cykliczn
```

```
Enter a name of an article to find on Wikipedia(use underscore(_) as space)
CoreThreading
Couldn't find article with that name
Exception in thread "main" java.lang.NullPointerException
    at wikiScraper.main(wikiScraper.java:26)
Process finished with exit code 1
```