Biblioteka BeautifulSoup

Czym jest?

BeautifulSoup w przeciwieństwie do biblioteki Requests (której również użyliśmy podczas naszego projektu) służącej do wykonywania żądań do stron internetowych, biblioteka bs4 służy jednak do parsowania kodu HTML i XML. Jest to bardzo ceniona biblioteka wśród programistów ze względu na swoją obszerną funkcjonalność.

Jednak samo pojęcie parsowania odnosi się do przetwarzania obszernego tekstu na mniejsze fragmenty, łatwiejsze do czytania lub analizowania. Parsowanie może dotyczyć dokumentów np w formacie XML czy HTML lub nawet po prostu obszernego tekstu.

Parsery biblioteki BeautifulSoup:

- html.parser-jest to domyślny parser tej biblioteki, lecz niestety nie jest najszybszy
- xml – parser ten jest przeznaczony do dokumentów XML
- lxml – najszybszy parser HTML. Właśnie z niego skorzystaliśmy
- html5lib – jest on najwolniejszym rozwiązaniem, jednak parsuje stronę identycznie do przeglądarek. Przydaje się jeśli inne parsery błędnie wykonują polecenie.

Dwa ostatnie parsery wymagają dodatkowej instalacji.

Bardzo pożyteczną funkcją jest wsparcie wykrywania kodowania, które może zapewnić lepsze wyniki dla stron HTML, które niepoprawnie deklarują swoje kodowanie.