

Web scraping para características de empresas

Iparraguirre García, Ander
EUSTAT
a-iparraguirregarcia@eustat.eus

Ledesma Sereno, Matias
EUSTAT
m-ledesma@eustat.eus

PALABRAS CLAVE

Web scraping, Sociedad de la información, carga de respuesta

RESUMEN

Inspirado en los objetivos de la ESSnet Big Data I de *Análisis de las metodologías para el uso de internet para la obtención de datos de la sociedad de la información* el siguiente proyecto presenta la aplicación práctica de la recolección de variables estadísticas vía web scraping. Las variables estadísticas objetivo son el uso de medios de comunicación sociales, los idiomas utilizados en la página web, recepción de pedidos o reservas on-line y publicidad de ofertas de empleo o recepción de solicitudes de trabajo. La obtención de características de empresas parte de un input de urls validado. Partiendo de estas se ha elaborado software propio en lenguaje Python para la recolección de datos utilizando las librerías más características para web scraping como Requests, Beautiful Soup o Selenium.

La segunda fase del proyecto evalúa los resultados obtenidos utilizando la herramienta R Studio para realizar una validación de los resultados obtenidos en la primera fase contrastándolos con los resultados de la actual *Encuesta sobre la sociedad de la información* elaborada por Eustat. El objetivo del proyecto es la eliminación algunos ítems para evitar la carga de respuesta y agilizar el proceso para el encuestado.