SEMINARIOS
(SEMINARS)

MÁSTER UNIVERSITARIO EN SISTEMAS DE INGENIERÍA CIVIL

-CURSO 2019/2020-

PROPUESTA DE SEMINARIO
(SEMINAR PROPOSAL)

UNIVERSIDAD POLITÉCNICA DE MADRID
ETSI CAMINOS, CANALES Y PUERTOS

Título (*Title*)

Captación automatizada de datos de la web mediante técnicas de WebScraping en R.

Ponente (*Lecturer*)

Borja Moya Gómez (Investigador Posdoctoral Juan de la Cierva – Formación, en TRANSyT – UPM)

Resumen (*Abstract*)

Las páginas web son una gran fuente de datos publicados accesibles desde cualquier dispositivo con conexión a internet. Sin embargo, y pese a que algunos de estos datos están disponibles en conjuntos de datos que pueden ser fácilmente usados/trabajados/analizados on-line o mediante su descarga en formatos ampliamente usados, otra gran cantidad de datos no están disponibles, o los conjuntos de datos disponibles no presentan el mismo nivel de desagregación publicada en la web. Este hecho requiriere establecer algunas estrategias para capturar los datos de la web y poder usarlos.

Afortunadamente, la web tiene una estructura semipermanente, tanto en sus relaciones con los servidores que le proporcionan los datos y servicios a usar, como en cómo deben estar dispuestos los datos para su visualización. Entender la estructura de cada web de interés permite poder "enseñar" al ordenador a encontrar los datos que se necesitan para el estudio que se esté llevando a cabo y poder tratarlos de la mejor manera para almacenarlos o trabajar.

En este seminario, se van a introducir conceptos básicos útiles sobre la web para hacer "raspados" de la web, en la medida de lo posible, mediante técnicas de WebScraping. Estas técnicas se aplicarán en diversas páginas web mediante la creación de scripts (archivos de órdenes al ordenador) en el lenguaje de programación enfocado al análisis estadístico R (libre, gratuito y ampliamente usado por analistas de datos) y el entorno gráfico R-Studio.

Para asistir a este seminario no es necesario tener conocimientos previos de R. La primera parte del seminario va a proporcionar los conocimientos necesarios básicos para capturar y almacenar datos.

Programa (*Agenda*)

1. Motivación del seminario. *(15 min)*
2. Presentación de R y R-Studio. Instalación y primeros pasos. *(120 min)*
3. ¿Cómo funciona la web? Conceptos básicos para el WebScraping: HTML y REST. *(45 min)*
4. WebScraping con R (I). Consultas a API. *(60 min)*
5. WebScraping con R (II). Entendiendo la web. *(120 min)*
6. WebScraping con R (y III). De tabla a tabla. *(45 min)*
7. Guía para los trabajos de evaluación del seminario. *(75 min)*

Tiempos estimados

Evaluación (*Evaluation*)

40% Asistencia a clase
60% Trabajo por parte de los alumnos equivalente a 32 h de dedicación. Obtención de datos por WebScraping y visualización simple de los datos.