Web Scraping con R.

Fundamentación.

En las últimas décadas, a través de los avances tecnológicos, la humanidad ha generado un volumen de datos enorme que crece exponencialmente a cada año. Si bien esta producción inmensa de datos permite analizar cuestiones y/o temas novedosos, también implica la necesidad de utilizar nuevas herramientas más adecuadas para manejar grandes volúmenes de datos.

El lenguaje de programación R surge en 1993 en la Universidad de Auckland con la característica de poseer una gran potencia a la hora de procesar grandes volúmenes de datos, generar gráficos y realizar cálculos estadísticos. Este lenguaje tiene la característica de ser gratuito y de código abierto. Esto implica en primer lugar una gran accesibilidad al estar exento de pago por licencia. Por otro lado, el código abierto ha permitido el desarrollo del lenguaje a través de una comunidad de usuarios que ha ido creando una innumerable cantidad de librerías que han facilitado su utilización y expandido sus posibilidades.

En relación al análisis de datos, el web scraping permite extraer datos de sitios web de manera eficiente para satisfacer distintas necesidades para transformar datos en formato HTML (web) en datos estructurados como puede ser una tabla Excel, csv o JSON que permiten una manipulación y análisis más factible.

Objetivos.

A lo largo del curso se buscará que los asistentes sean capaces de descargar datos de su interés desde la web, estructurarlo y realizar las operaciones correspondientes para poder extraer información de valor a través de código en lenguaje R. Dentro de las capacidades que se espera que puedan realizar los estudiantes se encuentra:

-Conocer los principales elementos de HTML

-Identificar elementos en páginas web estáticas

-Descargar información de manera estructurada

-Automatizar proceso de descarga de información

-Transformar la descarga en una función

-Descargar sitios de noticias

-Obtener palabras más frecuentes de un corpus de palabras

-Obtener herramientas de análisis de texto como conteo de palabras, TF-IDF, modelado de tópicos.

Requisitos.

En relación a los requisitos, se requieren conocimientos previos básicos en R (principalmente comandos básicos de tidyverse) y tener instalado R y RStudio. Se requiere tener computadora propia con buen funcionamiento, internet de buena conexión y un espacio libre de distracciones para poder realizar la cursada.

Modalidad de cursada.

La cursada está planificada para 4 encuentros de 2 horas reloj cada uno en modalidad a distancia a través de google meet.

Las clases quedarán grabadas y a disposición de los asistentes.

Cada clase atravesará distintas instancias:

-Presentación de los temas a trabajar en la clase

-Una sección expositiva en la que los alumnos irán escribiendo código a medida que el profesor explica las funciones y presenta el código

-Espacio de dudas y consultas sobre los temas trabajados durante la clase o en encuentros previos

-Un desafío semanal a resolver fuera del horario de clase

Cada encuentro tendrá material propio enviado a los inscriptos:

-Material explicativo de los temas de la clase

-Código ejecutado durante la clase

-Archivo de desafío semanal

Contenidos.

Unidad 1.

Introducción a elementos básicos de HTML e introducción al scraping de texto plano.

-Elementos y etiquetas HTML

-Descarga de página HTML

-Identificación de elementos de la página

-Descarga de distintos elementos de interés de la página

-Funciones principales de rvest (librería)

Unidad 2.

Descarga datos de un portal de noticias identificando posibilidades e inconvenientes.

-Descarga de noticias en formato de tabla

-Creación de una función para automatizar la descarga

-Identificación de patrones de url

-Función de frecuencia de palabras

-Term Frequency – Inverse Document Frequency

Unidad 3.

Introducción a RSelenium para trabajar con páginas dinámicas, redes sociales y web crawling.

-RSelenium

-Scrolleo automático

-Scrapeo de redes sociales

-Web Crawling

Unidad 4.

Modelado de tópicos con LDA (Latent Dirichlet Allocation) con artículos de revistas.

-Acercamiento teórico a LDA

-Exploración de datos

-Construcción de tópicos emergentes

-Interpretación de tópicos

-Visualización de datos

Docente.

El curso está a cargo de Diego Pacheco. Lic. En Sociología (UBA), doctorando en Ciencias Sociales (UBA) y docente del seminario de investigación "Explorando la periferia" de la carrera de Sociología (UBA).

Trabajó como analista de datos tanto en el ámbito público como privado.

Ámbito público:

-Ministerio de Desarrollo Social de la Nación (Programa POTENCIAR TRABAJO)

-Ministerio de Seguridad de la Nación (Secretaría de políticas criminales)

-Gobierno de la Ciudad de Buenos Aires (Dirección General de Sistemas de Información Sanitaria)

Ámbito privado:

-Data engineer en proyectos de machine learning para YPF

Becas:

UBACyT (Universidad de Buenos Aires Ciencia y Técnica): Actualmente su proyecto de tesis doctoral incluye trabajo con información cuantitativa de barrios populares en el conurbano e información georreferenciada.

Ganador beca FUNDATOS II (fundación Fundar). Proyecto sobre detección de déficit habitacional en el conurbano bonaerense a partir de algoritmos de aprendizaje automático.