



# PROMiDAT

## IBEROAMERICANO

Programa Iberoamericano de  
Formación en Minería de Datos

### **WEB Mining con R** **(Minería de la WEB con R)**



(506) 2268.8823 - (506) 8708.9091



[info@promidat.com](mailto:info@promidat.com)



[facebook.com/oldemarrodriguez](https://facebook.com/oldemarrodriguez)



[www.promidat.com](http://www.promidat.com)

**Tutor:** El curso será impartido por Dr. Oldemar Rodríguez graduado de la Universidad de París IX y con un postdoctorado en Minería de Datos de la Universidad de Stanford.

**Duración:** Cuatro semanas.

**Descripción:**



En este curso se presentarán los fundamentos de la minería de datos aplicados con datos de la web. Se estudian los principales formatos de los documentos web como son HTML (Hyper Text Markup Language), XML (eXtensible Markup Language) y JSON (JavaScript Object Notation). Se estudiará el uso de expresiones regulares y su aplicación para la minería de texto así como la conexión con redes sociales y APIs de terceros para la extracción de datos geoespaciales y análisis de sentimientos.

**Objetivo:**

En este curso el estudiante será capaz de:

1. Entender la estructura básica los principales formatos de los documentos WEB como son HTML (Hyper Text Markup Language), XML (eXtensible Markup Language) y JSON (JavaScript Object Notation).
2. Entender la sintaxis de una expresión regular.
3. Aplicar expresiones regulares para la limpieza y extracción de datos.
4. Estudiar los principales métodos de procesamiento estadístico de textos.
  - a. Usar paquetes en R para crear Nubes de palabras (wordclouds).
5. Utilizar técnicas especiales para minar datos en redes sociales con APIs publicos como Twitter.
6. Estudiar la conexión con APIs de google para datos geoespaciales y análisis de texto.

## Metodología:

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas.

- Una vídeo conferencia semanal, las cuales quedarán grabadas en Webex, para que los alumnos la puedan acceder en cualquier momento.
- Trabajos prácticos semanales.
- Foros para plantear dudas al tutor y compañeros.
- Aula virtual en Moodle.

## Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren extracción intensa de datos desde la red de internet.

## Contenido:

### 1- Almacenamiento de documentos y datos en WEB

- a. Documentos HTML (Hyper Text Markup Language).
- b. Documentos XML (eXtensible Markup Language).
- c. Documentos JSON (JavaScript Object Notation).
- d. Xpat como lenguaje de consultas para documentos WEB (a query language for web documents).
- e. Fundamentos de HTTP.

### 2- Minería de texto (Text Mining)

- a. Expresiones regulares
  - i. Sintaxis básica
  - ii. Tipos de datos
  - iii. Operadores de agrupación y repetición
  - iv. Vista hacia adelante y hacia atrás
- b. Manipulación y limpieza de textos
  - i. Detección de patrones
  - ii. Reemplazo de patrones
  - iii. Extracción de patrones

### 3- Extracción de datos de la web (Web Scraping)

- a. Principios de protocolo HTTP
- b. Tipos de peticiones HTTP

- c. Uso del paquete httr para peticiones GET y POST desde R
- d. Extracción de datos
- e. Proceso de limpieza de datos web

#### **4- Análisis de datos en formato texto**

- a. Procesamiento estadístico de textos.
- b. Métodos no supervisados.
- c. Métodos supervisados.
- d. Nubes de palabras (wordclouds).
- e. TextPlot.
- f. Wordlayout.
- g. Análisis de casos reales.

#### **5- Minería sobre las Redes Sociales**

- a. Minería de opiniones, exploración de tendencias y más con Twitter.
- b. Creando un “app” en la plataforma de Twitter.
- c. Visualización de actividad en Twitter.

#### **6- Datos geoespaciales y consulta de direcciones con google maps API.**

#### **7- Análisis de sentimientos con Google Sentiment Analysis.**

### **Bibliografía:**

1. Bing Liu. “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)”. Springer, 2011.
2. Matthew A. Russell. “Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More”. O’Reilly, 2015.
3. Owen Jones, Robert Maillardet and Andrew Robinson. Introduction to Scientific Programming and Simulation using R. Chapman & Hall/CRC Taylor & Francis Group, FL. 2009.
4. R Development Core Team. “R: A Programming Environment for Data Analysis and Graphics”. The R Project for Statistical Computing, 2010. <http://www.r-project.org/>
5. R Development Core Team. “Writing R Extensions”. The R Project for Statistical Computing, 2010. <http://www.r-project.org/>
6. Sharan Kumar Ravindran Vikram Garg. “Mastering Social Media Mining with R”. Packt Publishing, 2015.

7. Simon Munzert, Christian Rubba, Peter Meiner y Dominic Nyhuis. "Automated Data Collection with R". Wiley, 2015.
8. Soumen Chakrabarti. "Mining the Web: Discovering Knowledge from Hypertext Data". Morgan y Kaufmann. 2013.
9. Williams, G. "Data Mining with Rattle and R", Springer, 2011.