

**CURSO ONLINE**  
**WEB SCRAPING Y TEXT MINING PARA ECONOMÍA Y NEGOCIOS**  
Grupo Lambda  
2020

---

**1. Profesor**

**Mg. Heber Julio Baldeón Paucar**

Coordinador de Asociaciones Público Privadas (APP), subsidios cruzados y métodos cuantitativos en la Gerencia de Regulación Tarifaria de la SUNASS. Economista por la UNMSM y Máster en Finanzas Cuantitativas en la Universidad de Alcalá, Madrid - España. Además, aprobó el 56 Curso de Extensión Universitaria de Economía Avanzada del BCRP y cuenta con especializaciones en el FMI, Banco Mundial, ESAN, UP, PUCP y UNMSM. Su ámbito de interés es la Econometría Aplicada, Machine Learning, Finanzas Cuantitativas y Gestión de Riesgos.

**2. Objetivos del curso**

- Proporcionar al estudiante aplicaciones de Social Media & Web Analytics y Text Mining en los negocios y economía.
- Brindar al estudiante las herramientas de programación en el software Python y R aplicado Webscraping y Text mining.
- Uso intensivo de información de sitios web de entidades del gobierno, sitios web de empresas privadas y de las redes sociales.

**3. Descripción del curso**

El web scraping (escarbar una web) es una técnica para transformar la información de sitios web y redes sociales que se encuentran sin estructura y transformarla en datos estructurados, de forma automatizada. Así, con los métodos de Text mining (minería de textos) realizar la extracción, la limpieza de textos y generación de insights mediante técnicas de Machine Learning y la lingüística computacional.

En el curso se desarrollarán los procedimientos y herramientas del web scraping y text mining aplicados al scraping de precios, el scraping de contenido, scraping news, análisis de sentimiento, detección de fraudes, Gestión de Riesgos y Business Intelligence.

Para el desarrollo de las aplicaciones de Web Scraping y Text Mining se utilizará, principalmente, Python y R.

**4. Requisitos**

Conocimiento básico de R o Python.

**5. Metodología**

- El curso consta de 16 horas lectivas. Las sesiones consistirán en exposiciones del Docente y resolución de casos aplicados.
- Se otorgará un certificado a nombre de la institución por la participación en el curso. Además, de un trabajo integrador opcional para la emisión de un certificado de aprobación a nombre de la institución.

## 6. Contenido

### 6.1 Social Media & Web Analytics (6 horas)

- Inteligencia de negocios vs Analítica de negocios.
- Identificación de fuentes web.
- Métodos de extracción de datos web estáticos.
- Métodos de extracción de datos web dinámicos.
- Uso de APIs sin autenticación y con autenticación.
- Análisis de datos web.
- Aplicaciones a Negocios y Economía.

### 6.2 Natural Language Processing (6 horas)

- Procesamiento de Lenguaje Natural (o NLP).
- Segmentación y Tokenización.
- POST, Parse Trees y Clasificación de textos.
- Detección de entidades y detección de relaciones.
- Análisis de sentimiento.
- Analítica de texto (Text Analytics o Text Mining).
- Aplicaciones a Negocios y Economía.

### 6.3 Visual Analytics (4 horas)

- Presentaciones de la información.
- Componentes de un Shiny y Arquitectura: server y ui.
- Construcción de dashboards con Shiny, Bokeh y Dash.
- Visualización de dashboards personalizados.
- Aplicaciones a Negocios y Economía.

Lima, mayo de 2020