

## PROGRAMA DEL CURSO

## “Introducción a los textos como datos”

|                                  |                                                                                                                                                                |
|----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Datos de contacto:</b>        | Extensióndocente.ecp@ucr.ac.cr                                                                                                                                 |
| <b>Duración total del curso:</b> | 14 horas / certificado de participación                                                                                                                        |
| <b>Modalidad:</b>                | Virtual                                                                                                                                                        |
| <b>Sesiones y materiales:</b>    | Zoom – plataforma “UCR global”                                                                                                                                 |
| <b>Requisitos:</b>               | Contar con conexión a internet y equipo tecnológico para estar presente en cada sesión con video activado, conocimientos básicos de investigación cuantitativa |

**PUBLICO META:** Dirigido a analistas, consultores y demás profesionales en ciencias sociales interesados en aplicar métodos cuantitativos para el análisis de textos políticos, que tengan familiaridad con investigación cuantitativa y estadística inferencial. El curso no asume que los participantes tienen experiencia en programación en Python, pero experiencia previa con este u otro lenguaje de programación para el análisis de datos (como R) sería ventajoso.

**RESUMEN CORTO:** Este curso es una introducción a las técnicas de procesamiento y análisis de textos como datos cuantitativos, enfocado en dos técnicas particulares, el análisis de diccionario y la descomposición en valores simples para la modelación de tópicos. El curso inicia explicando el análisis de diccionario, pasando luego al uso del lenguaje de programación Python para la preparación de textos, su representación en forma de matrices numéricas, el uso de álgebra lineal para determinar similitudes entre textos en forma vectorial y finalmente la identificación de temas o tópicos en los textos.

**DESCRIPCION:** Este curso se plantea como una introducción a las técnicas de procesamiento y análisis de textos como datos cuantitativos. Se enfoca en dos técnicas introductorias: El análisis de diccionario enfocándose en las herramientas relacionadas con el proyecto *Linguistic Inquiry Word Count* (Boyd et al. 2022) incluyendo el análisis de sentimientos; y una técnica de *machine learning* no supervisado para modelación de tópicos, la descomposición en valores singulares. Como temas necesarios para realizar estos análisis, se abordarán temas como formas de obtener grandes cantidades de datos textuales, el pre-procesamiento, la tokenización, y otros necesarios para preparar textos para su análisis, y una explicación teórica de como los textos pueden representarse como matrices numéricas del tipo término-documento, y como esta representación permite identificar similitudes entre los textos.

**OBJETIVO GENERAL:** Capacitar a los profesionales en el uso de dos técnicas básicas de análisis de textos como datos y brindar las bases para el entendimiento de la representación de textos como matrices numéricas.

**OBJETIVOS ESPECIFICOS**

1. Comprender la lógica matemática subyacente en el análisis de textos representados como matrices numéricas y su manipulación mediante álgebra lineal
2. Desarrollar competencias en herramientas informáticas como LIWC y Python para realizar tareas básicas de análisis de textos como datos, tales como el análisis de diccionario y la modelación de tópicos.
3. Aplicar las herramientas y técnicas aprendidas para el análisis de textos políticos reales.