**Técnicas para el análisis de datos digitales**

**Análisis de redes sociales online y minería de texto para las ciencias sociales**

Camilo Cristancho

camilo.cristancho@uab.cat

En la última década el creciente uso de las tecnologías de información y comunicación ha proporcionado un volumen sin precedentes de información sobre el comportamiento humano. El seguimiento en tiempo real de las interacciones sociales proporciona información sobre la estructura y el contenido de las relaciones. El precepto de una ciencia social computacional que *aprovecha la capacidad de recopilar y analizar los datos con una amplitud y profundidad y escala sin precedentes[[1]](#footnote-1)* implica enormes oportunidades para la investigación en ciencias sociales en la medida en que posibilita la explicación de una gran diversidad de fenómenos colectivos que difícilmente podrían ser estudiados de otro modo a un nivel agregado. Sin embargo, la disponibilidad de grandes volúmenes de información también conlleva desafíos epistemológicos y metodológicos en las posibilidades de gestionar e interpretar los datos.

Este módulo es un punto de partida para la investigación con métodos digitales y una oportunidad de reflexionar sobre sus implicaciones teóricas y prácticas. El curso tiene como propósito desarrollar habilidades prácticas para el uso básico de herramientas de análisis de redes sociales y de minería de texto. Se parte de una introducción general a la ciencia de datos y a teorías para el análisis social basadas en huellas digitales. Mediante ejercicios prácticos se busca el desarrollo de habilidades técnicas para la obtención de información estructurada y no estructurada de múltiples formatos y fuentes tales como las redes sociales y páginas de Internet. Finalmente, se discuten las consideraciones éticas y las limitaciones de este tipo de información y las posibilidades de interpretación de los resultados.

* Retos de la extracción de diferentes tipos de datos (Datos estructurados y no estructurados)
* Técnicas y aplicaciones para la obtención de datos (Scrapping y APIs)
* Formato y limpieza, de datos
* Análisis de la información (Análisis de redes y minería de texto)
* Visualización de los resultados

Los participantes deben instalar R, RStudio, NodeXL (template para Excel) y Gephi.

**Programa**

**Día 1 – 3 horas**

## **Introducción a la ciencia social computacional**

* 1. Los datos como artefactos digitales de la vida social y política
	2. Interpretación y relevancia del contexto
	3. Los desafíos del muestreo y la representatividad
	4. Ciencia de datos y ciencias sociales

## **Obtención de datos estructurados**

* 1. Consideraciones teóricas
	2. Fuentes e Interfaces de programación de aplicaciones (APIs)
	3. NodeXL - Excel add-ons (interfaces desde MS Office) y Google Docs TAGS

**Día 2 – 3 horas**

* 1. Paquetes de R para Twitter
	2. Scraping
	3. Redes de hipervínculos

## **Procesamiento y estructuración de datos**

* 1. JSON parser
	2. Extracción de texto

**Día 3 – 3 horas**

## **Análisis de redes sociales**

* 1. Consideraciones teóricas
		1. Descripción
		2. Actores más prominentes y prestigiosos
		3. Patrones de difusión y niveles de influencia
		4. Análisis dinámico
	2. Estudio de caso - SNA – NodeXL
		1. Tweets
		2. Hipervínculos

**Día 4 – 3 horas**

* 1. Reducción de complejidad – Clusters
	2. Visualización – Gephi

## **Minería de texto**

* 1. Consideraciones teóricas
	2. Representación y descripción: convirtiendo el texto en datos
	3. Interpretación y exploración

**Día 5 – 3 horas**

* 1. Visualización y análisis
		1. Nubes de términos, tags,
		2. Redes de términos y clusters
	2. Análisis de sentimiento – clasificación por diccionario
	3. Visualización

**Destinatarios**

El taller está especialmente diseñado para estudiantes de ciencias sociales con interés en el análisis de datos obtenidos a través de Internet y medios sociales.

**Metodología**

Los datos obtenidos y los métodos presentados se contextualizan en diferentes teorías de las ciencias sociales (ciencia social computacional, teorías de redes sociales, economía política de los medios de comunicación, etc. ). Se analizará un caso basado en Twitter y otras fuentes de datos. Se presentarán las herramientas de recolección de datos, análisis de redes y análisis de texto.

**Requerimientos**

Conocimiento básico en el uso de Software de escritorio, estadística básica y familiaridad con la red social Twitter. Es necesario tener un perfil de Twitter activo y verificado (con un email y número de móvil válido) para usar las aplicaciones de recolección de datos. Las herramientas utilizadas exigen algunas habilidades en el uso de R a nivel de usuario[[2]](#footnote-2).

*Camilo Cristancho*

Doctor en Ciencia Política por la Universidad Autónoma de Barcelona e Ingeniero por la Universidad de los Andes. Investigador en comportamiento político, actitudes y protesta. Trabaja principalmente con enfoques metodológicos cuantitativos con datos de encuesta, minería de texto y métodos digitales. Miembro del equipo de investigación Democracia, Elecciones y Ciudadanía del Departamento de Políticas de la Universitat Autònoma de Barcelona y parte de equipos de investigación a nivel español y europeo. Tutor en diversos talleres de introducción al manejo de datos digitales y big data para científicos sociales (UOC, UAB, UAM, CPB).

1. Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915), 721. [↑](#footnote-ref-1)
2. Se recomienda ver un tutorial introductorio sobre el uso de RStudio. Varias posibilidades online (i.e. https://www.youtube.com/watch?v=IEPKW05lSrw) [↑](#footnote-ref-2)