Principios básicos de la minería de datos

La minería de datos (data mining), es un proceso para detectar información de conjuntos grandes de datos, de la manera más automática posible. Su objetivo es encontrar patrones, tendencias o reglas que expliquen el comportamiento de los datos en un contexto específico.

Encontrar la relación entre los datos es complicado, por diversas razones:

El volumen de datos suele ser muy grande.
Gestionar los datos es muy complejo, ya que estos proceden de diversas fuentes y tienen naturalezas distintas, pueden estar estructurados o no.
Los datos se actualizan constantemente, por lo que el proceso de búsqueda y almacenamiento de estos es continuo. Este procesamiento, además, debe llevarse a cabo en tiempo real, y aporta resultados al momento.

En general, requiere de automatización. Debe ser la propia tecnología la que explore los datos, encuentre patrones y los muestre a los profesionales al cargo. Esta tarea la realizan generalmente los desarrolladores, programadores o ingenieros en sistemas: estos cuadros técnicos se han incorporado en los últimos 10 años a la sala de noticias y trabajan junto a los periodistas, creando bases de datos que permitan hacer consultas sobre los ejes que fija el periodista.

En la Unidad de datos de Infobae, el equipo de datos trabaja junto a una experta en programación, que automatiza descargas, automatiza funciones de extracción de datos y nos ayuda a crear tablas dinámicas en Excel a partir del procesamiento de grandes conjuntos de datos.

Una lista de los trabajos que publicamos puede ser consultada aquí: https://www.infobae.com/infobae-data/

Nótese que al final de cada artículo, se relata cómo se procesó la información. Esto se corresponde con el principio de “Open Journalism” es decir de periodismo abierto, en el que reporteros contamos cómo desarrollamos el proceso, desde la fuente inicial hasta la hoja de cálculo que se comparte siempre y que puede ser descargada a elección del usuario (compartimos datos abiertos)

En minería básica de datos usamos Excel como herramienta.

Esta aplicación nos ofrece múltiples soluciones a la hora de gestionar grandes volúmenes de datos, porque nos hace posible organizar la información de la mejor manera que nos permita visualizar noticias dentro de un gran conjunto de números.

Aunque no es el objetivo de este curso el nivel avanzado de esta disciplina y nos concentramos en la lectura anterior de las técnicas básicas de scraping, aquí hay una lista de recursos que pueden explorar:

https://gijn.org/data-journalism/periodismo-de-datos/

Y en la clase siguiente, nos ocuparemos de las funciones básicas de Excel, como el primer gran paso que nos lleva a la creación de nuestro propio dataset o conjunto de datos.