Sharik Stefanny Guzmán Mera[2]
Jaime Andrés Valencia Gaviria[3]

Para citar este capítulo:

Guzmán Mera, S. y Valencia Gaviria, J. (2022). Modelo Rutas: un aporte de la ingeniería de datos en la investigación de las ciencias sociales y de la comunicación. En P. Rendón-Cardona (Comp.),

Rutas: narraciones de paz en Colombia desde el periodismo universitario, 2000-2021 (pp. 16-28).
Universidad Católica de Pereira.
DOI: https://doi.org/10.31908/eucp.69.c664

Modelo Rutas: un aporte de la ingeniería de datos en la investigación de las ciencias sociales y de la comunicación

Sharik Stefanny Guzmán Mera [2]
sharick.guzman@ucp.edu.co

Jaime Andrés Valencia Gaviria [3]
jgaimer42@gmail.com

Resumen

La automatización en algunas etapas de los procesos de investigación se ha convertido en un método eficaz para realizar de manera más rápida y precisa el análisis de grandes volúmenes de información. A pesar de esto, algunas herramientas digitales continúan siendo un reto en el campo de las ciencias sociales, pues aún en el siglo XXI muchos investigadores desconfían de las herramientas digitales y prefieren utilizar métodos tradicionales para recopilar y analizar la información. Sin embargo, gran parte de los investigadores han puesto a prueba el uso de métodos de automatización en sus procesos de recopilación. Gracias a esto, han obtenido resultados mucho más precisos en la interpretación y han abarcado incluso mayores volúmenes de información de una manera más rápida. En este capítulo se da a conocer con detalle qué es la minería de datos, para qué sirve y cuál es su función dentro de la investigación cualitativa, mediante la descripción del proceso en un caso de aplicación en el proyecto "Rutas: relatos universitarios de paz en Colombia".

Palabras clave: Rutas, investigaciones sociales, *big data*, automatización, *web scraping*, *web crawling*.

[2] Estudiante de doble titulación de la Maestría en Comunicación y Máster en Comunicación Corporativa, Comunicadora social-periodista egresada de la Universidad Católica de Pereira. Joven investigadora de MinCiencias en el proyecto CaPAZ. Enlace CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001764012

Introducción

En este capítulo se podrá conocer de cerca cómo es el proceso de extracción, transformación, limpieza y almacenamiento de los datos dentro de una investigación cualitativa, teniendo en cuenta el aporte de los métodos de automatización de la información y minería de datos para el alcance de los objetivos y obtención de resultados, con mayor precisión y tecnología frente al análisis de grandes volúmenes de información. Es este un valor agregado al estudio de las ciencias humanas, aplicado a los datos recopilados durante la primera fase de investigación del proyecto RUTAS, realizado por el Centro analítico de producciones culturales universitarias en el marco del conflicto (CAPAZ) de la Universidad Católica de Pereira. En este, mediante el análisis informatizado de textos, se habló sobre cómo los estudiantes de periodismo universitario han narrado el conflicto en Colombia.

¿Los datos también hablan de la sociedad?

Dentro de las necesidades del campo de investigación de las ciencias sociales, se encuentra el análisis de datos en grandes volúmenes, una fuente de valor sujeta hasta hace algunos años a la interpretación humana. La evolución de los recursos digitales y tecnológicos han permitido que muchos de ellos comiencen a ser aplicados en los procesos de investigación como herramientas que facilitan la extracción, jerarquización y comprensión de la información en un rango de tiempo menor al habitual, y con un margen de error reducido al momento de extraer y transformar los datos.

En este punto se comienza a hablar de la minería de datos o *data mining* (DM por sus siglas en inglés), un método de análisis matemático que expone datos, patrones o tendencias, pero de manera controlada por el científico de datos. Barzanallana establece por ello que:

> Extraer conocimiento útil y comprensible previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, es un campo de las ciencias de la computación. Las herramientas de *minería de datos* predicen tendencias futuras y comportamientos, permitiendo por ejemplo la toma de decisiones en los negocios (2019 p.1)

Para el investigador social, es fundamental analizar el uso del lenguaje al momento de hablar de un objeto de estudio, y de manera más específica, comprender los patrones de comportamiento, expresión o base histórica de un grupo poblacional

[3] Estudiante de Ingeniería de Sistemas y Computación, Universidad Tecnológica de Pereira. https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001847019

> Hoy en día la información que nos rodea lo hace en su gran mayoría en forma de texto. El volumen de información no estructurada crece continuamente de tal manera que resulta necesario separar por medio de técnicas de procesamiento de texto lo esencial de lo que no lo es así, como distinguir proposiciones subjetivas de las objetivas. (Lanzarini, 2019 p.4)

Sin embargo, al ser un proceso completamente humano, existe gran influencia en los métodos de crianza, experiencias y conocimientos adquiridos previamente por el investigador al momento de manipular la información. Esto hace que el nivel de precisión, veracidad de los resultados y la capacidad de abarcar la extracción, jerarquización e interpretación de un gran volumen de información de manera manual en medio de los procesos de investigación cualitativa continúe siendo para la comunidad científica un tema de gran discusión.

> El *data mining* permite encontrar información escondida en los datos que no siempre resulta aparente, ya que, dado el gigantesco volumen de datos existentes, gran parte de ese volumen nunca será analizado. La minería de datos es un proceso de identificación de información relevante extraída de grandes volúmenes de datos, con el objetivo de descubrir patrones y tendencias estructurando la información obtenida de un modo comprensible para su posterior utilización. (ESIC, 2018 p.1)

Por esta razón, los métodos de extracción de datos y análisis de información automatizados dejaron de ser útiles únicamente en el campo de las ciencias exactas, como la ingeniería, para pasar a ser parte de los métodos de innovación y estrategia digital en el interior de los proyectos de investigación en el área de las ciencias sociales. En estas, los *software* para el análisis de textos, transcripción de recursos multimedia y jerarquización de datos se han convertido en un soporte que ayudan a orientar y fortalecer estos procesos humanos propios de la investigación cualitativa.

> El avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos, tales como imagen, video, texto, y otros datos numéricos, en una base de datos sencilla (...) allí se pueden aprovechar dos cosas: la ingente cantidad de datos que se almacenaban en áreas como el comercio, la banca o la sanidad, y la potencia de los nuevos ordenadores para realizar operaciones de análisis sobre esos datos. (ESIC, 2018 p.1)

No hay que olvidar que la automatización de estos procesos se convierte entonces en un complemento para el desarrollo de la investigación, pues al momento de interpretar los datos habrá mayor claridad frente a las agrupaciones, categorías o subtemas en los que se clasifica la

información. Sin embargo, pese al avance tecnológico, continúa siendo un recurso administrado y ejecutado bajo criterios humanos, y complementado por aquellos aspectos narrativos y estéticos que se escapan del análisis realizado por medio de computadoras. En línea con Tagnin (2019) con la DM los aprendizajes automatizados, análisis de clústeres y demás procesos asociados al *big data* generan una multiplicación de las entidades observables en el mundo social, a escalas nunca antes vistas.

Por este motivo la DM entra en esta discusión implementando modelos y simulaciones computarizadas que pueden predecir trayectorias fundamentales en la toma de decisión en la etapa de análisis multivariado. De manera simple la ingeniería de datos es el conjunto de técnicas de computación, matemáticas y dominio sobre un tema que permiten entender de manera más clara los datos y las fuentes que se van a analizar, mediante tres pasos: la extracción, la transformación o modificación del dato, y la carga o almacenamiento. Estas fases serán desarrolladas en este artículo mediante en estudio de caso del proyecto ''Rutas: Relatos Universitarios de Paz''.

¿Qué es la ingeniería de datos y cómo funciona?

La automatización en los procesos investigativos es una técnica usada actualmente en varios tipos de investigaciones con dos objetivos: el primero es facilitar determinados procesos que son repetitivos, y el segundo, permitir el manejo, análisis e interpretación de grandes volúmenes de datos o información.

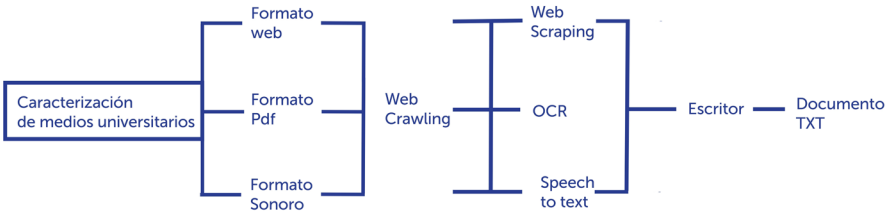
El proyecto RUTAS se centró en entender cómo se ha narrado el conflicto armado en Colombia desde la mirada del periodismo universitario. Se trató de un primer proyecto realizado por el Centro analítico de producciones culturales universitarias en el marco del conflicto (CAPAZ), y se presentó por la necesidad de generar estadísticas textuales sobre más de 7000 archivos de noticias recopiladas en 24 medios pertenecientes a la Red Colombiana de Periodismo Universitario.

Tales archivos base fueron recopilados de los medios teniendo como centro aquellos relacionados con el conflicto armado en Colombia, por medio de distintos formatos técnicos como piezas audiovisuales, sonoras y textuales. Un aspecto que permitió determinar de manera prioritaria el esquema Extract, Transforma and Load (ETL) como el método automatizando requerido para el manejo de grandes volúmenes de datos.

Para esto, se realizó un proceso específico acorde a cada producto periodístico identificado y etiquetado de manera manual cada producto; tras un proceso de extracción de la información llevada a archivos (txt),

se avanzó a un proceso fundamental de análisis lexicométrico desde el programa Iramuteq.

Figura 1.
Proceso de automatización de los datos



Como se mencionó anteriormente, la ingeniería de datos permite entender de manera más clara los datos y las fuentes que se van a analizar, mediante tres pasos: la extracción, la transformación o modificación del dato, y la carga o almacenamiento, las cuales dentro de esta investigación se llevaron a cabo de la siguiente manera.

Extracción

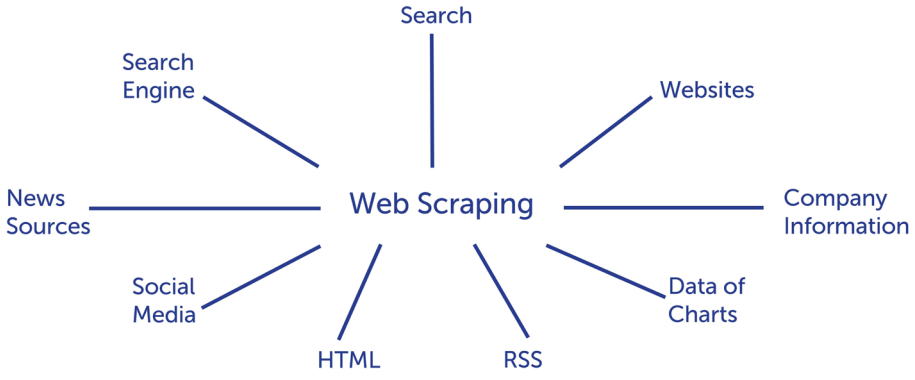
De manera previa, se definieron los elementos más relevantes en el proceso de investigación, como el objetivo de la investigación, las fuentes de información, qué definen y acotan, cuál es la información que va a ser analizada tanto en el tiempo como en el espacio, entendiendo el espacio como el lugar de origen de la información, y de igual forma se define el formato a analizar. Con su aplicación al proyecto Rutas, se determinaron las fuentes de datos, que en este caso particular fueron los portales de noticias pertenecientes a la Red Colombiana de Periodismo Universitario.

Es vital para este proceso comprender las razones por las que se escogieron dichas fuentes y el peso de su aporte a la investigación, así como algunos detalles técnicos que se relacionan a la naturalidad de las fuentes. Es decir, si se encuentran presentes como medios digitales o físicos, y en caso de ser digitales, conocer si están expuestos por medio de API (*application programing interface*, por sus siglas en inglés, es un método por el cual dos o más sistemas computacionales se comunican entre sí, por medio de protocolos de red o transferencia de datos). En el caso de las páginas web, se debe tener presente si hay que realizar un cambio de formato de los datos o si es posible acceder legalmente a ellos, según las tecnologías con las que están hechas dichas páginas.

Este fue un aspecto técnico que se evidenció particularmente con la Universidad Jorge Tadeo lozano, pues contaba con ciertos bloqueos en su página web que impedían de manera legal y técnica acceder a determinadas secciones e información de la página web. En este punto se dio solución a la problemática mediante la gestión con las personas encargadas de manejar la página de esta universidad.

Por otro lado, las fuentes de información, las variables a analizar y el diseño de datos permiten tener claridad acerca de cuáles son los elementos específicos de las fuentes que se deben tener en cuenta, como su distribución temporal, distribución geográfica, tipos de clasificación o agrupación, en general, los datos específicos se van a tomar en cuenta para la investigación. Para esto es fundamental definir también la clasificación de aquellos datos relevantes que posiblemente no contienen la información suficiente para situarse en la distribución escogida.

Figura 2
WebScraping



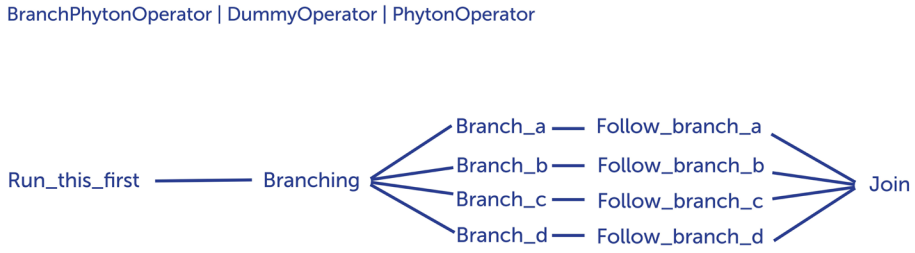
Nota. Elaboración propia a partir de GRID Digital Solutions: https://www.grid.cl/blog/el-web-scraping-que-es-aplicaciones-y-consecuencias/

Una vez estos elementos están claros, se puede iniciar la fase de extracción, la cual es una etapa en su mayoría técnica en donde se usan diferentes técnicas de computación como *web scrapping*, *optical character recognition* (ocr), o *speech to text*. Estos permiten extraer, cambiar el formato y normalizar los datos, como es el caso de la extracción de los contenidos periodísticos de la Red Colombiana de Periodismo Universitario, en los que se utilizó *speech to text* para convertir los archivos en formato sonoro y audiovisual a formatos de texto plano (TXT) para el análisis lexicométrico.

Transformación

Una vez extraídos los datos, se continúa a un proceso de limpieza y enriquecimiento de los datos, que en algunos casos también aplica en pequeñas "adaptaciones" del dato para uso de programas. Para la limpieza de los datos se tomó en cuenta cuáles son aquellos que pueden presentar errores dependiendo del dominio. Un dato contaminado puede ser cuando en una lista de eventos temporales hay un evento cuya fecha está más en el futuro, o muy en el pasado, lo cual no concuerda con los demás datos. Este dato se debería revisar o eliminar según un criterio humano.

Figura 3
Automatización de procesos GAP



Nota. Elaboración propia a partir de programa Apache Airflow.

Por ejemplo, el medio universitario *Datéate al minuto* no contaba con el criterio de fecha de creación o publicación de la totalidad de sus contenidos periodísticos que por la temática abordada fueron pertinentes para la investigación, mas no podían clasificarse fácilmente, al no contar con los datos exactos de distribución. Para este propósito, se decidió entonces dar solución a través de una representación general para los datos que presentaban inconsistencia en su fecha de creación, mediante una agrupación llamada 00-00-00. La representatividad de algunos datos es especialmente importante cuando los datos se van a usar para promedios, pues estos datos pueden aumentar o disminuir el valor del promedio por sí solos.

En la limpieza de datos se deben realizar agrupaciones de estos acordes a la necesidad de la investigación, para iniciar un proceso de enriquecimiento en el cual se agregan elementos de otras fuentes que permitan correlacionar las fuentes de información y de esta manera ver cómo interactúan ambos datos juntos, y así obtener un primer análisis de cuáles tienen mayor relación. Cabe resaltar que actualmente las técnicas de inteligencia artificial permiten hacer este proceso de manera automática, por medio de técnicas de análisis de imágenes

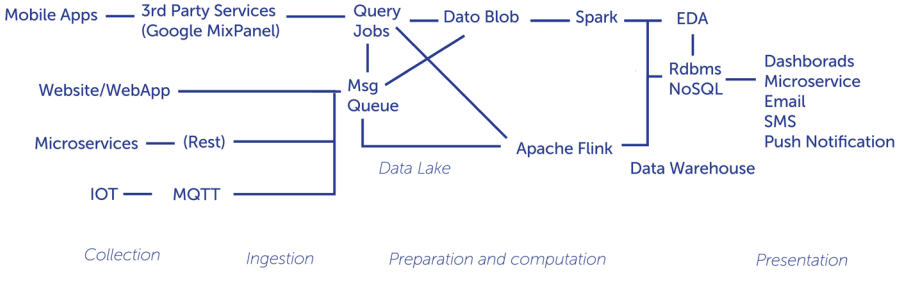
o procesamiento de lenguaje natural (NLP) y el análisis de texto, de manera simple, lo que permite el uso de estas técnicas simulando el criterio humano en tareas como clasificación, agrupación o etiquetado, pero de manera superficial.

Finalmente, el último paso en el proceso de la transformación de datos es la adaptación en términos de forma para su uso en programas externos. Un caso muy específico que no se desarrollará en este capítulo, pero que, llevado a un ejemplo real, se evidencia en el proyecto CAPAZ con el programa Iramuteq, en donde para poder hacer uso de los análisis textuales, fue necesario codificar y formatear la información extraída. Esta consiste en una serie caracteres especiales, (*, -, #, que indican cada variable a analizar, bajo un estándar y forma dada por el mismo programa.

Carga o almacenamiento

El proceso de almacenamiento de los datos es el paso final del proceso de ingeniería de datos, en donde se define cómo se va a almacenar la información para optimizar su uso posterior y con el menor costo posible en términos de peso de almacenamiento. Normalmente en este punto se suelen usar herramientas técnicas, como lo son Hadoop o Apache Spark, programas conocidos como *data lakes*. Estos son un tipo de base de datos basadas en archivos que optimizan de manera significativa el peso de la información, y suelen tener lenguajes de consulta que permiten relacionar la información de manera rápida. Sin embargo, esta no es la única alternativa; se pueden usar bases de datos SQL o NOSQL tradicionales, una decisión más de corte técnico que depende de cada caso particular, pues la persistencia y uso de los datos en el tiempo varía según las necesidades de la investigación.

Figura 4
Canalización de datos



Nota. Elaboración propia a partir de ichi.pro, arquitectura para la canalización de macrodatos.

La analítica de datos aplicada al campo social

Una vez terminado este proceso de ingeniería de datos se puede decir que se tiene una información confiable para ser analizada e interpretada, pero aquí surge un interrogante: ¿cómo acceder a dicha información de manera clara y eficiente? En muchas ocasiones el volumen y cantidad de datos suele ser tan grande que no se logra interpretar de manera clara o sencilla. Es por eso que existen programas o herramientas de visualización de datos cruciales para la investigación, pues permiten generar correctamente el análisis, evaluación de indicadores e interpretación de los datos, los cuales hasta la actualidad son habilidades sujetas completamente a la naturaleza humana.

Se debe tener en cuenta que el aspecto más importante de este proceso de investigación es conservar el factor humano sujeto a la interpretación de los datos, el cual, sin perder su esencia cualitativa inicial, es fortalecido y complementado con los procesos de automatización de la información. En un caso de aplicación analizar cómo se narró el conflicto en Colombia desde la mirada de los periodistas universitarios requiere extraer una cantidad considerable de información de distintas fuentes de manera rápida, precisa y efectiva, pero también comprender desde lo social la intención narrativa y estética dada por los estudiantes de periodismo universitario en cada uno de los productos para generar intención, emociones y contexto a la información que está siendo narrada. Por ejemplo, la música, los paisajes sonoros, la colonización de la imagen y los recursos gráficos que terminan de entregar un sentido a la narración. "Se trata de la capacidad de retener la atención de quién escucha una historia para que la interprete como queramos que sea interpretada" (Platzi, 2021).

A partir de esto, por ejemplo, se puede tener una imagen mental acerca de los diferentes sucesos vividos dentro del proyecto RUTAS, en el cual la ingeniería de datos modificó algunos elementos que aún no estaban completamente en la primera etapa de la investigación. El primer elemento modificado por la ingeniería de datos fue el nivel de alcance del proyecto, pues al poder automatizar los procesos de extracción y procesamiento de datos se dio paso a un gran número de posibilidades para el análisis del formato, pues RUTAS inicialmente estaba concebido para analizar únicamente la información existente en formatos escritos (web). Por esta razón, es en esta etapa del proyecto en que la ingeniería de datos permite agregar nuevos formatos para el análisis, como los audiovisuales, sonoros y periódicos, lo que permite tener una mayor perspectiva y representatividad de algunas universidades.

"La ingeniería de datos es una de las ramas de las ciencias de la computación que se ocupa del procesamiento de datos en grandes cantidades" (Python Colombia, 2022), por lo que otro aporte importante

de la ingeniería de datos al proyecto de RUTAS fue relacionada a la agrupación, clasificación y representación de la información. Al ser un volumen tan alto de datos, aún debía definirse de qué manera iban a ser narrados y representados los resultados correspondientes a cada una de las regiones y universidades escogidas como objeto de estudio en este proyecto. Una vez surgieron los primeros datos, sus consecuentes y representaciones gráficas, dejando en evidencia que existía cierta sobre representatividad de regiones y medios en particular, una situación que fue controlada y solucionada por medio de una agrupación de datos, acorde a sus características regionales, narrativas, formato y un grado de criterio personal dado por los miembros del equipo de trabajo.

Durante la investigación surgieron algunos momentos, como los mencionados anteriormente, en donde lo que más resalta es el conocimiento de los equipos de trabajo, pues los datos que existen no siempre pueden ser representados de manera clara, en función de la intencionalidad de la investigación o incluso del dominio sobre el cual se trabaja. A esto se suma la implementación y uso de modelos complementarios propios del dominio. En el caso particular de RUTAS, se usó como modelo complementario el modelo de Greimas (1987), lo que permitió conservar la intención humana en la narrativa existente en los textos de los estudiantes. Un aspecto en el que, dado el contexto de automatización de la extracción de la información, existía temor a perderlo; sin embargo, fue el equipo de trabajo quien adaptó e implementó dicho modelo acorde a las necesidades del proyecto y el estado de los datos.
Finalmente, el principal aporte de la ingeniería de datos es la reducción de tiempo, pues en este caso de aplicación, el proceso de extracción de datos en formato web estaba estimado en tres meses y se ejecutado en solo un mes, lo cual permitió al equipo centrarse en procesos de planeación, ejecución e implementación del proyecto.

Conclusiones

Se puede evidenciar que en los procesos de automatización por medio de las técnicas de ingeniería de datos en proyectos y procesos de investigación en el campo de las ciencias sociales se facilitan los procesos en la etapa de extracción de la información de manera automatizada. Así, se requiere únicamente un investigador o científico de datos con conocimientos técnicos en el área de la minería de datos para ejecutar esta tarea, lo que optimiza el tiempo y los recursos dentro de esta etapa de la investigación, que suele ser una de las más tardías y complejas de todas las etapas de un proyecto, si se tiene en cuenta también que estas técnicas de automatización permiten por su efectividad a los equipos de trabajo centrarse con mayor facilidad en los puntos más relevantes de la investigación, como la definición del análisis de la información y el objeto de estudio.

La ingeniería de datos aplicada en proyectos sociales permite agilizar los tiempos de extracción de los datos a analizar, pueden permitir identificar factores relacionados a las fuentes de datos que no estaban contemplados cuando se planteó el proyecto, y permite tener una mayor claridad sobre la investigación en general, pero en particular en temas relacionados a datos y posibles sesgos dentro de estos.

El modelo de investigación/trabajo implementado en este proyecto puede ser replicable de manera relativamente sencilla, e incluso puede ser adaptable a otros dominios o temas. Pero cabe resaltar que cuanto más complejo el dominio, mayor será el conocimiento técnico relacionado a la ingeniería de datos.

Este tipo de procesos de investigación se puede realizar de manera más eficiente si previo a la extracción de datos está más o menos claro qué modelos de análisis de dominio se van a usar. A su vez, este tipo de procesos pueden arrojar mejores resultados, si previo a la transformación de datos se implementan análisis exploratorios de los datos (EDA) obtenidos.

Referencias

Barzanallana, R. (2019). *Minería De Datos*. Informática aplicada a las ciencias sociales. https://guides.co/g/informatica-aplicada-a-las-ciencias-sociales/21926

ESIC. (2018). *Minería de datos: qué es, cómo es el proceso y a qué áreas se puede aplicar*. ESIC Business & Marketing School. https://www.esic.edu/rethink/tecnologia/mineria-datos-proceso-areas-se-puede-aplica

Lanzarini, W. H. (2019). Minería de Datos, Minería de Textos y Big Data. XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan, Argentina).

Platzi. (2021). ¿Cómo aprender a hacer storytelling? 5 consejos esenciales. Platzi. https://platzi.com/blog/aprender-storytelling-consejos/

Python Colombia. (2022). Introducción a la Ingeniería de Datos | Python Pereira [video]. YouTube. https://www.youtube.com/watch?v=P0-oaeHdFAY

Tagnin, J. (2019, 21 de junio). *Big data y ciencias sociales*. Bordes. Revista de política, derecho y sociedad. http://revistabordes.unpaz.edu.ar/big-data-y-ciencias-sociales/