Reporte 05

Nombre: Martínez López Andrés
Fecha: 28/05/2021

Referencia bibliográfica	**APA** **Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.,Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson,B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., et al. (2020). CORD-19: The Covid-19 Open Research Dataset.** **IEEE** **Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B.,et al. "CORD-19: The Covid-19 Open Research Dataset" ArXiv, Abril de 2020.**
Autor (es)	Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D., Weld, D.,Etzion, O. y Kohlmeier,S.
Título	CORD-19: The Covid-19 Open Research Dataset
Año	2021
Tipo de publicación	Artículo de investigación
Nombre de la revista, conferencia, Editorial u otro	arXiv
Número de páginas	11 páginas
Problema abordado	La forma en que se adquiere o extrae la información relacionada con la Covid-19 ya sea con respecto a la enfermedad en sí o al desarrollo de la vacuna provenientes de artículos científicos que con el paso del tiempo aumenta de forma considerable.
Objetivo	Creación de un sistema computacional capaz de extraer toda la información posible vertida de los artículos científicos que aborden la Covid-19 y almacenarlos de manera efectiva en una colección de la misma (dataset), así como el lenguaje utilizado y sus palabras para que todos los especialistas médicos puedan encontrar información relevante de manera más rápida y concisa.

Justificación	La existencia de un gran número de artículos científicos que abordan el tema de la Covid-19 por lo que la creación de una nueva tecnología o herramienta se vuelve fundamental para la recaudación de datos.
Marco teórico	COVID-19: Es una enfermedad infecciosa causada por el virus recientemente descubierto. Datasets: Conjunto de datos correspondientes a los contenidos de una única tabla de base de datos o una única matriz de datos de estadística, donde cada columna de la tabla representa una variable en particular, y cada fila representa a un miembro determinado del conjunto de datos que se esté tratando Clúster: Grupo de empresas interrelacionadas que trabajan en un mismo sector industrial y que colaboran estratégicamente para obtener beneficios comunes.
Método utilizado	Los artículos cientificos que abordan la Covid-19, así como los derivados de los virus del coronavirus crece a un paso agigantado, ya que originalmente se estipulaba considerar aproximadamente 28 mil artículos, sin embargo, al paso de poco tiempo ya se tenían dentro del sector médico más de 50 mil. Dicho lo anterior se llevo a cabo una recolección de datos mediante un preprocesamiento en el cual se utilizaban palabras clave (como covid, coronavirus, vacuna, etc) para destacar los artículos que enfatizaran esta temática. Una vez extraída la información de manera concisa se convierte XML y posteriormente se convierte a formato JSON, creando así identificadores que ayuden a no insertar dentro del dataset artículos con información que ya se había abordad con anterioridad. Adecuando a su vez un pipeline o cadena de procesos a seguir que se encargase de actualizar de manera periódica el dataset con los artículos publicados en ese intervalo de tiempo.
Fuentes de investigación utilizada	Artículos científicos y documentación de herramientas similares al objetivo.
Herramientas utilizadas	Archivos en formato XML y JSON, las cuales son fundamentales para la creación de bases de datos con un gran volumen de datos.
Resultados alcanzados	Se implementó un sistema computacional que, a través de la minería de datos, creo una colección de datos (dataset) en relación a la pandemia de Covid-19, llamado “CORD-19”. Teniendo como problemática algunos lenguajes y la adición de imágenes o gráficos que ayuden a la obtención de datos. En el futuro se planea que el sistema pueda actualizarse de manera diaria y se logren incluir gráficos.

Aspectos de interés	Es de destacar como se aplica la minería de datos para la extracción de información valiosa en donde la rapidez es de suma importancia. Así mismo la inclusión de una propuesta por estandarizar la forma en que se presentan los metadatos para mejorar la calidad de las colecciones que se lleven a cabo así como la integración de nuevas tecnologías y herramientas que mejoren las anteriores.