launion.com.mx @uniondemorelos

CIENCIA

launion.com.mx
@uniondemorelos

Explorando textos sin leerlos

R. María del Río Chanona y J. Antonio del Río Portilla

Rita María del Rio-Chanona es becaria posdoctoral de la JSMF (Fundación James S. McDonnell) en Com-plexity Science Hub Viena desde junio de 2021 y afi-liada al Growth Lab en el Centro para el Desarrollo In-ternacional (CID) de la Universidad de Harvard. María tiene un doctorado en matemáticas de la Universidad de Oxford, donde formó parte del grupo de economía de la complejidad del Instituto para el Nuevo Pensa-miento Económico, Oxford Martin School. Ha trabaja-do con organizaciones políticas internacionales, inclui-do el Fondo Monetario Internacional y la Organización Internacional del Trabajo. Realizó sus estudios de li-cenciatura en física en la Universidad Nacional Autó-noma de México UNAM. Su investigación se enfoca en la ciencia de redes, el procesamiento del lenguaje natural y el modelado con agentes y se centra en tópi-cos de futuro del trabajo, la transición ecológica, la Gran Resignación y el impacto económico de la pan-demia de Covid-19.

Jesús Antonio del Río Portilla es físico y doctor en ciencias por la Facultad de Ciencias de la UNAM. Dis-tinguido con el Premio Weizman por su tesis doctoral, Premio Efraín Hernández Xolocotzin por la Universi-dad de Chapingo, Medalla de Honor en Ciencia y Tec-nología otorgada por el Congreso del Estado de More-los, Medalla VASE y el Reconocimiento al Mérito Es-tatal en Investigación REMEI 2021 por las contribu-ciones a la Divulgación y Vinculación. Director funda-dor el Centro Morelense de Innovación y Transferen-cia Tecnológica (2007-2008) y primer director del Insti-tuto de Energías Renovables de la UNAM (2013-2021). Es miem-bro de las academias Mexicana de Ciencias, de Ingeniería de México y de Ciencias de Morelos.

Esta publicación fue revisada por el comité editorial de la Academia de Ciencias de Morelos.

Sobre la pandemia de COVID-19

Durante la pandemia de COVID-19, en Estados Unidos (EUA) se observó un fenómeno que inquietó a los economistas: el reporte del mayor núme-ro de renuncias a empleos en la historia. A partir de febrero del 2020 se incrementaron las renuncias a los trabajos. Este fenómeno pudo haberse presentado en otras regiones, pero no fue tan observado como en EUA. A esta renuncia masiva se le denominó la “Great Resignation” (Gran Resig-nación, GR). Seguramente también te preguntarás, como muchas otras per-sonas: ¿cuáles son las razones que la pandemia despierta en la gente para renunciar a sus trabajos? En época de pandemia las incertidumbres sobre el futuro crecen y si además le aumentamos las renuncias, esta incertidumbre se incrementa.

En el ámbito científico se han estudiado algunos fenómenos similares. Las formas clásicas de estudiar y entender este tipo de fenómenos sociales eran mediante encuestas y trabajo de campo que realizaban personal especializado en áreas de ciencias sociales y humanidades. Sin embargo, precisa-mente una de las consecuencias de la pandemia fue el aislamiento social. De esta manera, el aislamiento dificultaba la realización de estudios estándares de este tipo de problemas. Adicionalmente, la forma de proceder en el pasado mediante encuestas era muy personal y consumía mucho tiempo, por lo que podría haber sesgos al hacer preguntas específicas y descuidar aspectos que no se habían considerado.

El uso de redes

A diferencia del siglo pasado, en la actualidad se ha extendido el uso de redes sociales como FaceBook o Twitter. En este tipo de redes es común que las personas viertan sus opiniones sobre temas de política, deportes o de su situación personal. Por lo cual, es probable que, de poderse analizar estas redes, se puedan encontrar en ellas las explicaciones de fenómenos socioeconómicos como la GR. Esta inquietud fue una de las que generó un estudio en la red Reddit que buscó respuestas sobre las motivaciones que llevaron a renunciar a las personas masivamente durante la pandemia [1]. Precisamente el uso de una red de este tipo permitió, mediante el uso de técnicas de procesamiento de lenguaje natural, obtener algunas respuestas a las inquietudes que despertó la GR. En ese trabajo, se analizaron 198081 posts en Reddit de los cuales 88092 fueron antes del COVID-19 (desde 2018 a 2021), para tener un antes y un después significativo.

El principal hallazgo es que la pandemia exacerbó las preocupaciones sobre la salud mental de las personas. En el trabajo se encuentra, a través de las palabras usadas, que dichas preocupaciones se hicieron desproporcionadamente presentes en los textos de Reddit relacionados con renuncias de empleo. Si bien el aumento de las vacantes laborales y el cambio de trabajo fueron factores presentes en períodos de recuperación anteriores, la pandemia de COVID-19 desató fuerzas que llevaron a comportamientos de abandono donde se enfatizaban problemas de salud mental, que no estaban presentes en anteriores casos de abandono. Estos factores adicionales podrían ayudar a explicar las tasas inusualmente altas de renuncias a los empleos en 2021.

Aquí seguramente la pregunta está latente en ustedes: ¿cómo en este trabajo se encontraron las causas?, ¿se leyeron los 198081 textos? o ¿qué método se utilizó? En particular, en ese trabajo se utilizó la técnica de modelado de tópicos (topic modeling) que tiene como objetivo descubrir temáticas latentes en los textos sin imponer categorías predefinidas [2]. Es decir sin prejuiciar al descubrimiento. Abundando sobre el método, podemos decir que clasifica de manera no supervisada documentos. Cuando decimos no supervisada quiere decir que sin la intervención de una persona se clasifican los textos por similitudes entre ellos. Esta técnica encuentra algunos tópicos incluso cuando no se está seguro de lo que estamos buscando [3]. Además, el agrupamiento puede ser suave, cuando un documento pertenece a más de un tópico o puede pertenecer a uno solo. Analicemos un ejemplo: supongamos que tenemos muchos documentos (Doc) compuestos por muchas palabras; aquí solo ponemos las palabras que aparecen con más frecuencia en tres de ellos:

Doc1: Perro, correa, caminar, croquetas, entretenimiento...

Doc 2: Libro, novela, entretenimiento, comedia, historia...

Doc 3: Correr, caminar, gimnasio, futbol, nadar...

El algoritmo encuentra que las palabras que se mencionan pueden agruparse, es decir cuando se menciona “perro” siempre cerca está “correa” y “caminar”, digamos esas palabras frecuentemente son separadas por dos o tres otras palabras. Así se infiere que este grupo de palabras conforman un tópico. Se procede con la conformación de todos los posibles cúmulos de palabras para definir los tópicos, luego se analizan los tópicos y se puede hacer una tabla. En la tabla 1 presentamos una porción de la tabla de tres cúmulos A, B, C que llamaremos tópicos.

Tópico	Perro	Correa	Caminar	Libro	Novela	Entretenimiento	Correr
A	0.62	0.43	0.2	0.01	0.001	0.27	0.1
B	0.02	0.01	0.05	0.47	0.57	0.35	0.02
C	0.03	0.1	0.3	0.01	0.001	0.19	0.62

Tabla1

Primero notemos que en las listas Doc 1, Doc 2 y Doc 3 solamente hemos presentado algunas de las palabras más frecuentes en los documentos. La Tabla 1 también está abreviada ya que sería muy grande y no cabría en el periódico. En la tabla completa se presenta la probabilidad de que una palabra pertenezca a un tópico, es decir, del total de palabras en el tópico cuántas veces esa palabra aparece en el grupo A o en el B o en el C de cada documento. En nuestra ilustración solo presentamos un fragmento de la tabla y vemos que faltan algunas palabras que aparecen en alguno de las listas, pero con este fragmento de tabla podemos ilustrar el procedimiento.

Observemos que en el tópico A: “perro” y “correa” pueden ser parte importante de este tópico. En particular, el tópico parece ser: “sacar a caminar a un perro”. En cambio, esas palabras, aunque aparecen, no son importantes para el tópico B, que más bien parece referirse a la lectura. Para el tópico C, que podríamos asociar con hacer ejercicio, las palabras caminar y correr son relevantes, pero no lo son “Perro”, “Correa”, “Libro” o “Novela”. En cambio, el “entretenimiento” pudiera tener alguna relevancia. Observa que en este tercer tópico las palabras “correr” y “caminar” también tienen alguna relevancia. ¿Qué nombre le pondrías al tercer tópico? Parece que conformar los tópicos es una tarea interesante.

Con esta forma de clasificar los grupos, se pueden definir las palabras relevantes en cada tópico y definir su contexto. Cuando una persona revisa el tópico puede interpretar lo que se está diciendo y cómo se está diciendo sin necesidad de leer todos los textos. Este algoritmo no solo usa las palabras también usa lo que se conoce como “lexicón” que son conjuntos de palabras sinónimos o con significados equivalentes, pero esos son elementos más avanzados que no discutiremos por el momento.

Algoritmo de nube de palabras

Hace algunos años, explicamos en los artículos de esta sección, la Ciencia desde Morelos para el mundo, cómo podríamos utilizar análisis de textos usando el desorden en ellos, es decir la entropía de los textos y sonaba un trabajo complicado [4]. Las técnicas que hoy están disponibles en el mundo de la computación utilizan herramientas de cómputo y análisis de lenguaje natural sofisticado como el usado para analizar los textos en Reddit, pero también existen herramientas más sencillas al alcance de cualquier estudiante de bachillerato con algunas habilidades computacionales. En lo que sigue explicaremos el algoritmo de Nube de palabras (WordCould, WC en corto) que es uno de los más simples. Este algoritmo lo que hace es contar las veces que una palabra aparece en un texto y considerarla relevante si aparece muchas veces. Sin embargo, en cualquier idioma, hay palabras que se usan mucho, de hecho, los textos suelen tenerlas miles de veces, y no tienen un significado por sí mismas, como “el”, “la”, “se” “con”, “en”, etc. Por eso antes de contar, el algoritmo de WC elimina este tipo de palabras. Este conjunto de palabras a eliminar se conoce en el lenguaje computacional como “stopwords”. En las implementaciones computacionales de estos algoritmos de WC ya se tienen definidos esos conjuntos de palabras para diferentes idiomas. Así una vez eliminadas las stopwords, WC cuenta el número de veces que aparece una palabra en el texto y las ordena por número de aparición o por frecuencia. Una vez que se tiene esta lista se hace con ellas una gráfica donde el tamaño de cada palabra es proporcional a la frecuencia con la que aparece. La conformación de nubes de palabras es un ejercicio donde se usan herramientas sencillas de análisis de texto.

Con el programa de la liga [5] se puede hacer el análisis de cualquier texto en formato simple, de preferencia txt. En el programa de la liga está un primer ejemplo donde analizamos un texto de Wikipedia de Historia de México con 5500 caracteres y el resultado se muestra en la figura 1. Como vemos, tenemos algo de información, sabemos que en la Historia de México los metales como el cobre, la plata y el oro han sido importantes y que la región es Mesoamérica. Quizá podemos decir algo más, pero tenemos muy poca información, solo 5500 caracteres que podemos leer rápidamente y entender el documento completo. En cambio, ¿qué pasa si analizamos libros donde tenemos muchos más caracteres?

Figura 1. Análisis de un Texto de Wikipedia [5].

A continuación, presentamos dos nubes de palabras de los textos de dos libros: “La casa dorada” [6] y “Las nanoaventuras del maestro Fonseca” [7] (con más

de 80 mil caracteres) y (con más de 200 mil). Ambos libros son de divulgación y escritos mayoritariamente por miembros de la Academia de Ciencias de Morelos. Lo que hicimos fue tomar los textos de esos libros y utilizar el mismo programa de la liga [5], pero en lugar del texto de Wikipedia poner los textos de los libros. Antes de continuar con el análisis, tomémonos un tiempo para revisar las dos figuras donde se muestran las nubes de palabras figuras 2 y 3.

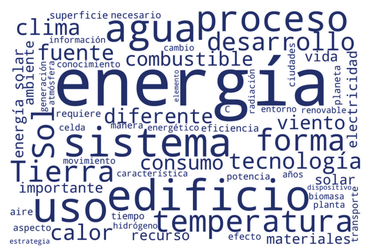

Figura 2. Análisis de “La Casa dorada” [6].

Figura 3. Análisis de “Las nanoaventuras del maestro Fonseca” [7].

Al observar las figuras, sin ver los pies de figura, podemos concluir a qué tema se refieren. Primeramente, la Figura 2 se relaciona con el libro “La casa dorada” donde se abordan temas de uso de energía desde una perspectiva de sistemas. Adicionalmente se analizan las formas de consumo de la energía en edificios. El cambio climático y desarrollo de tecnología deben ser puntos de vista que se discuten en el libro. Adicionalmente la energía del Sol y del viento son tópicos que seguramente se abordan desde una perspectiva del planeta. Esperamos que ustedes que leen y ven estas figuras coincidan con nuestra apreciación de que las nubes de palabras nos aportan información sintética, aunque no a detalle ni tampoco transmiten entornos emotivos. Continuemos y antes de seguir, revisa con cuidado la Figura 3.

Seguramente, en esta ocasión observamos que para hablar de nanotecnología o de nanoestructuras los conceptos luz y átomo son elementos fundamentales en la discusión. El maestro Fonseca seguramente se llama Eulogio Fonseca. Adicionalmente, se discuten diferentes efectos físicos y químicos en el texto. Parece ser que la forma y la estructura es muy importante en este mundo pequeño de la escala nanométrica donde las nanopartículas tendrán comportamientos ordenados o desordenados.

Siguiendo estos ejemplos, tenemos la seguridad que ustedes pueden hacer más conjeturas que podrían verificar con la lectura de estos libros.

Hasta aquí parecería que las nubes de palabras son una herramienta para, sin leer los textos, entenderlos; y aunque aportan información útil, tienen importantes limitaciones. Primero, la selección de las stopwords puede no ser adecuada para analizar todos los textos. Como verán en el ejemplo, tuvimos que añadir algunas palabras al conjunto de stopwords para español, pero quizá uno de los aspectos esenciales es que en algunos contextos las palabras en este conjunto, que se descartan a priori, pueden ser muy relevantes para entender el contexto. Este es uno de los aspectos que aborda el análisis de lenguaje natural que es una rama de la ciencia de datos que está muy activa en la actualidad.

Por otro lado, mientras en inglés existen ya algoritmos que incorporan estas y otras características, en español todavía hay trabajo por realizar.

Les invitamos a incursionar en el tópico de análisis de textos en español y seguramente se sorprenderán de sus hallazgos. La exploración con herramientas de minería de textos o de algoritmos de lenguaje natural puede ser muy útil y, si se usa con las precauciones adecuadas, nos aportará entendimiento de los textos, incluso sin leerlos.

Esta columna se prepara y edita semana con semana, en conjunto con investigadores morelenses convencidos del valor del conocimiento científico para el desarrollo social y económico de Morelos. Desde la Academia de Ciencias de Morelos externamos nuestra preocupación por el vacío que genera la extinción de la Secretaría de Innovación, Ciencia y Tecnología dentro del ecosistema de innovación estatal que se debilita sin la participación del Gobierno del Estado.

ACADEMIA DE CIENCIAS DE MORELOS, A.C.

ESTA PUBLICACIÓN FUE REVISADA POR EL COMITÉ EDITORIAL DE LA ACADEMIA DE CIENCIAS DE MORELOS

Para actividades recientes de la academia y artículos anteriores puede consultar: **www.acmor.org.mx**
¿Comentarios y sugerencias?, ¿Preguntas sobre temas científicos? CONTÁCTANOS: editorial@acmor.org.mx

REFERENCIAS

[1] R.M. del Río Chanona et al. https://arxiv.org/pdf/2208.07926.pdf
[2] R. Kulshrestha, https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2
[3] T. Lebryk https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383
[4] J. A. del Río et al. https://www.acmor.org/articulos-anteriores/f-sica-y-miner-a-de-textos-o-qu-investigan-los-cient-ficos-en-morelos
[5] https://colab.research.google.com/drive/17gE_9bJoV80HgfLw7HnpHmrnNm25Wlin?usp=sharing
[6] J.A. del Río, I. Marincic y J. Tagüeña “La casa dorada” (ADN, CONACULTA, México, 2013)
[7] J. A. del Río, J. Tagüeña y M. C. Vázquez “Las nanoaventuras del Maestro Fonseca”, (Adbo publicaciones y Academia de Ciencias de Morelos, México, 2012)