Revista Estudios
Número especial 2024
DOI: 10.15517/re.v0iEspecial.62955

ISSN: 1659-3316
Conferencia

Sobre la explotación y visualización de grandes datos para la investigación

About the mining and visualization of big data for research

Universidad de Costa Rica, 30 de agosto de 2023

Recibido: 1-10-2024

Aprobado: 8-11-2024

Gimena del Rio Riande
Instituto de Investigaciones Bibliográficas y Crítica Textual,
Consejo Nacional de Investigaciones Científicas y Técnicas
Buenos Aires, Argentina
gdelrio@conicet.gov.ar
ORCID: 0000-0002-8997-5415

Las humanidades digitales son un campo de estudio que se interesa por la explotación de los datos, o sea, se interesa en ver los textos a través de los datos y buscar distintas formas de abordar el conocimiento. Algo interesante en las humanidades digitales es que explotan muchísimas formas de lectura, pensando en que no solamente hoy en día es posible analizar un texto literario, por ejemplo, sino cualquier texto, incluso los mensajes de TikTok, Facebook o Twitter. Todo es un texto capaz de ser *minable*, explotable y visualizable en distintas formas. A eso se dedican las humanidades digitales.

Hoy me interesa presentar qué son los datos en las humanidades digitales, porque se habla de datos, se habla de textos y se habla de formas de entender los datos. Hoy en día también hay que pensar qué es todo esto para una computadora, porque en el fondo lo que estamos haciendo es interactuar con las computadoras. Recién los colegas nos hablaban de análisis del discurso y también de análisis automatizado del texto. Los humanos hacemos análisis del discurso, pero las máquinas no tienen ni idea de qué es el discurso, "ellas" entienden en términos de ceros y unos, hacen lo que pueden, y nosotros después, somos capaces de hacer ese análisis del discurso de una forma cuantitativa, acompañados por los ordenadores, pero de una forma cualitativa también. MAXQDA, uno de los softwares que mostraron los colegas, hace un análisis cuantitativo, ATLAS.ti también, pero somos nosotros los que estamos interactuando para sumar lo cualitativo en ese análisis del discurso.

Esto nos lleva a pensar ¿de qué hablamos cuando hablamos de datos en las humanidades digitales y qué estamos haciendo? A mí me parece interesante empezar por ahí porque hoy en día estamos todos inmersos en esta cuestión de la ciencia de datos y la importancia de la ciencia de datos en este nuevo paradigma que tenemos de lo digital. Hay carreras acerca de las ciencias de datos, que es una disciplina emergente, que se basan en el conocimiento de la metodología estadística y las ciencias de la computación; está muy interesada en esta cuestión de la creación de predicciones y clasificaciones, y se aplica en una gran cantidad de campos, no solamente en las ciencias digamos más duras, sino que hoy en día la ciencia de datos se ha movido mucho, por ejemplo, en el análisis de las ciencias sociales y en las humanidades tradicionales, de una forma si se quiere más cercana a la de las humanidades digitales que se interesan mucho más por esta cuestión de la *minería de datos*.

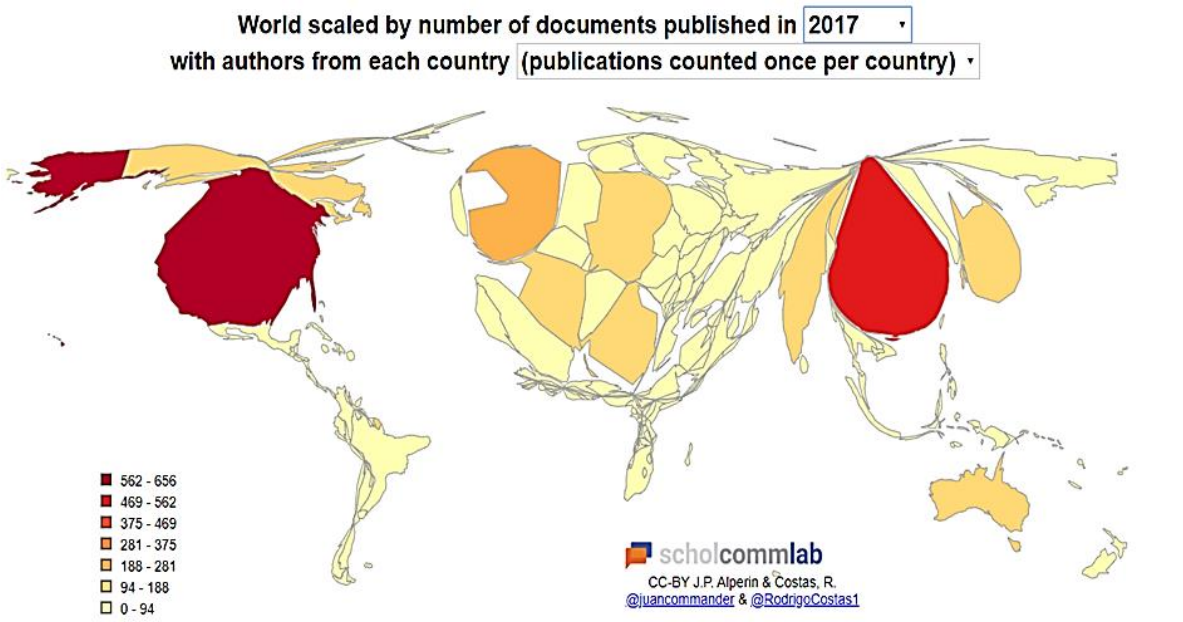
Nosotros, cuando estamos haciendo humanidades digitales, estamos haciendo mucho trabajo de *minería de datos*, un trabajo con grandes cantidades de datos donde nos interesa extraer patrones de distinto tipo. De hecho, en la conferencia inaugural, yo mostraba algunos ejemplos que habíamos trabajado sobre algunas crónicas históricas y literarias donde explotamos patrones narrativos o temas entre distintos textos. A medida que se va sumando un *corpus* a mayores, para el humano es más difícil entender esos patrones, se nos empiezan a escapar esos detalles. A las máquinas no. Las máquinas no entienden la semántica, el significado, pero sí son capaces de tomar esas estructuras y ayudarnos a hacer otro tipo de análisis.

Donde converge, a lo mejor, la discusión hoy en día entre lo que es el campo de las humanidades y las ciencias de datos es ¿cómo hacemos para que las máquinas aprendan? ¿Automáticamente? ¿De una forma semiautomatizada? Aquí entra otro tema del que estuvimos hablando el día de ayer sobre la inteligencia artificial (IA) que a mí me parece muy interesante. Por el momento, su uso es bastante limitado en el campo de las humanidades digitales por los costos que tienen este tipo de desarrollos y también por toda una cuestión ética. Desconocemos los textos que se están utilizando en muchos de los softwares de IA para alimentar esos modelos de datos. Sí lo que estamos seguros es que, cuando los probamos, en muchos casos el Sur Global no aparece, está completamente borrado, así que tenemos una idea de donde se están tomando gran parte de esos datos y de dónde no.

Otras cuestiones éticas que en este momento se discuten es que los modelos de inteligencia artificial también se están alimentando con el uso de trabajo bastante "esclavo". La última mejora que se hizo al ChatGPT fue a través de un etiquetado masivo del material y básicamente eso se hizo con trabajadores escasamente remunerados, como en Kenia, África (Perrigo, 2019), pero también en otros países de América Latina como Venezuela se contrató de esta forma a la mano de obra. Lo último que podemos agregar sobre las cuestiones éticas de la inteligencia artificial es la cantidad de recursos que consume. En esta actividad académica también se va a hablar de humanidades y ecología, un tema que es fundamental, pensar en si estamos trabajando con las herramientas de forma responsable. Creo que esto es parte también del uso de la tecnología, no sé si es parte del desarrollo de las humanidades digitales, pero es parte de nuestro día a día de todos como usuarios de tecnología. Deberíamos

estar al menos avisados de todas estas cuestiones relacionadas con un uso responsable de los dispositivos, porque o si no, en el fondo, seguimos perpetuando brechas. Por ejemplo, el mapa a escala del mundo según el número de documentos publicados por países (Scholarly Communication Lab, 2017), que es bastante conocido y se muestra en la figura 1, lo hicieron desde un laboratorio de comunicación científica donde se cruzaron los datos de las publicaciones científicas de la Web of Science con el producto interno bruto (PBI) de cada país y la cantidad de publicaciones de los distintos países, nos permite ver cómo los *nortes* siempre están destacados y los *sures* siempre aparecen sin destacar. Y esto es algo que también podemos pensar, cómo a través de un uso responsable de la tecnología, de un uso crítico de la tecnología, surjan unas humanidades digitales críticas también, donde podamos colaborar para hacer otro mapa, elaborar otros mapas del conocimiento que no solamente sirvan para una bonita *página web* sino también para nosotros como desarrolladores de conocimiento.

Figura 1. Mapa a escala del mundo según el número de documentos publicados por países en 2017



Fuente: Scholarly Communication Lab, 2017.

La definición de humanidades digitales debe ser bastante simple y nos debe llevar a pensar dos cosas: la primera es la explotación de los datos, ver cómo trabajamos con herramientas digitales en los datos y cómo generamos conocimiento; no solamente quedarnos en un ejercicio instrumental y de juego lúdico con los datos, sino de pensar cómo generamos conocimiento, cómo pensamos en nuevas formas de entender los textos, nuevas formas de entender las disciplinas para los profesores (no solamente para los estudiantes). El uso de las herramientas es un desafío para los profesores porque los lleva a poner todo esto en un contexto donde se entienda por qué estamos utilizando herramientas digitales, que es muy diferente a trabajar con una lectura lineal, con un libro y con un *set* de datos.

También nos lleva a pensar nuevos marcos teóricos para el conocimiento, y como les decía, en las humanidades digitales tenemos todos estos procesos: pensar los datos, estructurarlos, analizarlos, visualizarlos (esta dimensión visual, a la que me referiré más adelante, y que me parece muy importante porque es también otra forma de generar conocimiento). Y creo que aquí hay una zona final que es la del texto: contar una historia, una narrativa, no quedarnos en el dato. El análisis simple del dato, contar cuántas veces aparece un término, no tiene demasiado sentido para hacer una investigación que sea cualitativa, que busque nuevos sentidos, que busque nuevo conocimiento, eso tenemos que enmarcarlo para que haya una narrativa, es lo que nosotros llamamos contar historias. ¿Cómo podemos contar historias de otra forma?, ¿cómo podemos contar también la investigación de otra forma?

Me gustaría hacer la salvedad sobre con cuáles datos trabajamos cuando estamos haciendo humanidades digitales y la importancia de que los datos sean abiertos, es decir, que estén en código abierto. Si nosotros no tenemos datos que podamos reusar, tenemos datos que están bajo *copyright*, sería bastante difícil poder hacer un trabajo que después pueda estar mostrado en abierto. Pero no solamente esta cuestión del dato abierto es lo que interesa, pueden haber muchos datos en abierto, pero si esos datos no los podemos reusar porque no están en formatos que sean reusables, bueno, ahí también estamos en problemas. Cuando el colega que expuso anteriormente hablaba del uso de Voyant Tools decía: yo transformé todo esto a un archivo txt. Es decir, hubo un proceso de esos datos, hubo una posibilidad de reusar esos datos y eso es porque el PDF sobre el que estaba trabajando, o el HTML, estaba en esas condiciones, estaba en código abierto, pero también era reusable desde muchos lugares.

En términos generales esto se viene denominando como datos FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*). Yo no voy a entrar demasiado en esta cuestión porque llevaría toda una charla hablar sobre el acceso a los datos y sobre la calidad que necesitamos de esos datos para hacer investigación, pero tenemos que pensar que, para poder hacer un trabajo en humanidades digitales primero, si el material está en un formato analógico como es un libro, ¿que tendríamos que hacer con el libro? Digitalizarlo ¿no? Después de digitalizarlo deberíamos ver en qué condiciones está ese PDF, si tenemos que hacer algo de procesamiento, de curaduría, hacer un OCR, quitar lo que llamamos toda esa "suciedad" que nos queda sobre los PDFS y luego ver qué hacemos con ese material.

Ese material puede vivir como un dato, lo podemos subir a un repositorio y puede ser reusado por otros o, podemos seguir explotándolo y seguir trabajando con ese material. Por eso, este acrónimo FAIR que dice que los datos para reusarse tienen que ser encontrables, hallables en la *Web* para que sean *fair,* para que sean justos, tienen que ser accesibles de alguna forma. Creo que los repositorios institucionales en este momento son los lugares desde donde se puede acceder mejor al conocimiento y aquí hago, si se quiere, un paréntesis: pensemos siempre en los recursos que tienen las universidades antes de salir a buscar otro tipo de repositorios o lugares para ir a buscar datos, alimentemos también nuestros recursos dentro de la universidad.

Para las humanidades digitales los datos tienen que ser interoperables y deben ser, o pueden, deberían, queremos que sean, reusables por otros. La interoperabilidad en las humanidades digitales es muy importante, se refiere a que los sistemas que guardan y manejan los datos sean capaces de intercambiarlos de manera automática y segura. Si yo tengo un dato que solamente puede vivir en un formato y no lo puedo después reusar para hacer otros procesamientos, la verdad para poco me va a servir en los procesos que yo quiero hacer en este campo de las humanidades digitales.

Ahora me referiré a algunas ediciones que hemos hecho en el laboratorio de humanidades digitales. Yo soy filóloga, me gusta mucho trabajar sobre edición de textos y sobre formas de leer los textos. Habíamos trabajado en algunas crónicas del siglo XVII que definían el Río de la Plata, contaban sobre las visitas de extranjeros al Río de la Plata (del Rio Riande et al., 2018). Esto es un producto terminado, pero en humanidades digitales yo también tengo lo

que podemos llamar un *workflow* de trabajo que puede empezar desde cómo voy a buscar el *corpus* sobre el que voy a trabajar —esto ya se mencionó anteriormente, la importancia de hacer todo un proceso abierto en mi *corpus* para que facilite no solamente mi trabajo sino el trabajo de los otros, el trabajo de co-creación, el trabajo en comunidad.

Por ejemplo, lo que estaban mostrando los profesores antes, se puede usar para transformarlo en un trabajo en clase, entonces se podría disponer de *corpus* que estén en abierto, de software de código abierto, para pensar cómo se hace un trabajo colaborativo con los profesores pero también con los estudiantes y cómo se da crédito a todo ese trabajo en un campo como el de las humanidades digitales, donde estamos trabajando muy pocas veces individualmente y la mayoría de las veces en grupos de dos o muchas más personas. Es algo que merece que pongamos ahí nuestros ojos.

Insisto mucho, cuanto más interoperables los datos mejor. Textos que ya tengan un lenguaje marcado que nos pueda permitir reusarlos, textos y datos que estén en repositorios, que pueden ser tanto repositorios de software como repositorios de datos y de textos en general, pensar todo este proceso que hacemos con los datos desde el primer al último formato, es muy importante en las humanidades digitales. En las humanidades digitales estamos muy interesados en los procesos, no solamente en los resultados. A veces nosotros en las humanidades o en las ciencias sociales pensamos más en todo lo que queremos hacer, el resultado ¿no?, quiero hacer tal cosa. En humanidades digitales también es importante pensar cómo hacemos todo ese proceso, qué vamos haciendo en ese proceso y cómo vamos también trabajando los datos, cómo podemos abrir los datos en todo ese proceso. Entonces vale la pena no quedarnos solamente con los productos finales sino pensar esta dimensión tan importante de los datos en un *workflow*.

Creo yo, que para nosotros los humanistas digitales está muy ligado dato a *corpus*. Siempre insisto mucho con las definiciones, como buena filóloga soy un poco hincha con las definiciones, y creo que es muy importante porque nosotros no estamos trabajando o pensando en un análisis de la *Web* en general, no estamos haciendo un *scrapping* general de la *Web* sino que estamos de alguna forma pensando en cómo trabajamos un *dataset,* un *corpus,* una cantidad de texto. Esto para las humanidades digitales es crucial, siempre tengo

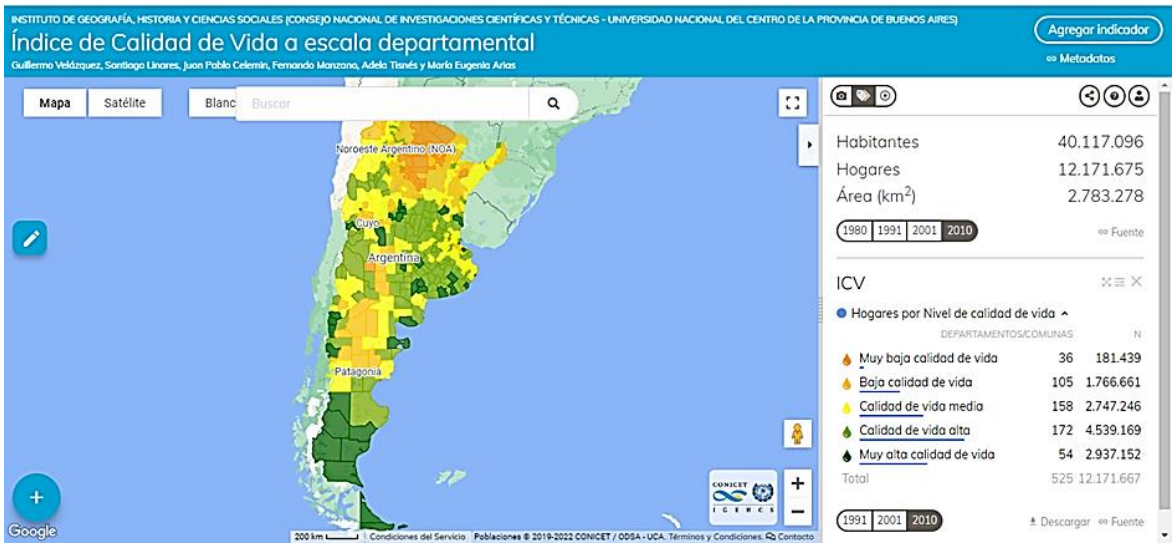
que recortar mi objeto de estudio, pensar un *corpus* y de ahí pensar, bueno, con qué datos trabajo.

Es fundamental que reflexionemos ¿qué son los datos? Decimos datos constantemente, pero a veces no vamos al diccionario y reflexionamos sobre lo que es un dato ¿no? Si bien el diccionario nos dice que los datos son algo dado, algo que se nos da. Este es un término que es bastante nuevo en nuestra lengua. Al hacer un trabajo de investigación sobre el *corpus* diacrónico de la lengua española, se encuentra que la palabra *datos* apenas la empezamos a documentar en el siglo XVIII. Cuando se va a Google Books a buscar en el Ngram el uso de *datos,* se descubre que es intensísimo en los últimos tiempos. Piensen cómo venían ganando terreno los datos en términos históricos como lo dado, lo que estaba antes. Pero desde 2000 en adelante, algo que venía siendo muy chiquitito que era *dato informático* después despega y ya nos hemos olvidado del sentido que tenía dato como testimonio, como antecedente ¿no? Y creo que esto es un lugar importante para pensar las humanidades.

Hay un libro muy interesante de Lisa Gitelman (2013) que se llama *Los "datos crudos" son un oximorón* que plantea que hablar de datos no tiene sentido, lo dado nunca es neutro. Ella dice que los datos se sirven siempre cocinados y nunca crudos. Cuando yo ya estoy haciendo o recabando datos de algún sitio, ese dato no es neutro, ese dato ya fue procesado por algo y por alguien, hay un humano detrás de la máquina, entonces nunca estamos dentro de esa neutralidad.

De hecho, para ponerles un ejemplo, cuando hicimos este mapa de la figura 2 sobre el índice de calidad de vida con unos colegas en Argentina (Instituto de Geografía, Historia y Ciencias Sociales Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Centro de la Provincia de Buenos Aires, 2021), se intentó hacer un mapa lo más neutro posible sobre cómo viven los argentinos, pero está basado en datos que nos habían sido provistos ya por las agencias de estadísticas de Argentina, que tal vez nunca muestran los datos con la crudeza que necesitamos.

Figura 2. Mapa del índice de calidad de vida en Argentina



Fuente: Instituto de Geografía, Historia y Ciencias Sociales Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Centro de la Provincia de Buenos Aires, 2021.

Por eso hablar de *capta,* de datos capturados, tiene más sentido, porque el dato capturado ya tiene esa otra carga de esa otra semántica que en las humanidades nos es tan importante. Nosotros queremos pensar críticamente sobre los datos y no dar por sentado esa capa de neutralidad que evidentemente no existe. El libro *Algorithims of Oppression* de Safiya Noble (2018) es un libro que me parece interesantísimo y trata sobre esta cuestión de que nunca un dato es completamente neutral. Cuando uno hace una búsqueda de niña en Google, las primeras niñas que aparecen evidentemente son todas bonitas, pero también son todas blancas, son todas rubias, aparecen los problemas de los sesgos, pero también esto nos está hablando de la cuestión sobre *data* y *capta.* Trabajamos con lo capturado, lo previamente cocinado; no pensemos que los datos son neutrales.

En humanidades digitales deberíamos hablar, como dice Christof Schöch (2013), de datos inteligentes. Pensar en términos de *corpus,* recortar un *corpus, minar* ese *corpus,* es hablar no ya de la *big data,* que es enorme —vieron, que siempre están estas infografías de no sé cuántos millones de *tuits* en un segundo, todo eso es casi imposible de entender para el humano—, pero aquí estamos trabajando con un gran *corpus* que nosotros curamos, que le pasamos una capa para después poder hacer otros procesos. Por eso Schöch habla de s*mart data,* de datos inteligentes, y creo que esto es clave en las humanidades digitales, pensar en

que los datos los tenemos que limpiar, los tenemos que curar, para que después tengan sentido, para que después esos patrones, esas narrativas que queremos encontrar, tengan sentido.

Durante la pandemia un grupo de investigadores ideamos cómo pensar Twitter en términos de narrativa. Twitter ya no existe, ahora se llama X, ¿vieron cómo es esto de la volatilidad de las redes sociales? Es difícil pensar un *tuit* como una narrativa porque es tan breve, pero hay un diálogo, gente que propone algo, otra que responde, etcétera. Pero Twitter es muy difícil de *minar,* si lo hacemos a través de su API a veces se dificulta más, sin embargo, sí hay formas de curar esos datos, de transformarlos en una base de datos y después buscar sentidos. Eso fue lo que hicimos en este proyecto que se llamó *Narrativas digitales de La COVID-19* (Alles Torrent et al., 2020).

Como estábamos en casa, como queríamos ver un poco qué pasaba con las conversaciones en Twitter en los distintos países hispanohablantes, dijimos bueno, hagamos de Twitter una gran narrativa, una gran base de datos. Entonces hicimos un trabajo bastante larguito usando algunas notebooks de Python. Lo que hicimos fue una base de datos que tomara los *tuits* si se quiere "crudos"; pero después los limpiábamos, les quitábamos *emojis*, o sea, todas las cosas que a lo mejor después en un procesamiento textual podían resultar basura, suciedad, dejando el texto plano para poder procesarlo luego. Y lo que hicimos fue tratar de poner todos estos datos en abierto, en una base de datos que esté en abierto. En un momento tuvimos que cortar este *corpus* porque la base de datos seguiría creciendo al día de hoy y tomando datos que no fueran de la COVID. La base de datos trabajaba sobre unas búsquedas, unos parámetros que le dábamos nosotros sobre distintas terminologías relacionadas con coronavirus, COVID, etcétera, para que encontrara lo que estábamos buscando e hicimos nuestra propia base de datos con datos que fueran explorables.

¿Cómo hacemos para transformar un gran *corpus* que está en internet —aparentemente imposible de *minar* si no hacemos algo así como un *scrapping*— en una base de datos y en ella buscamos narrativas? Fue muy interesante, hicimos distintos trabajos, les voy a mostrar uno solo para terminar. Hicimos algunos trabajos con Voyant Tools —de hecho, analizamos un *corpus* con Voyant Tools— donde también exploramos otras formas de trabajo. Un tipo de trabajo bastante poco supervisado por el humano es el *topic modeling.* Era la tarea en

donde a nosotros como humanistas nos costaba un poco más tratarlo. Trabajamos muchísimo, aquí hemos dejado en el blog una máxima: preparar datos es 80% del trabajo de *data science*. Nos llevó un montón de tiempo hacer que todos estos datos estuvieran listos para hacer un buen trabajo de *data mining*. Algo que nos pareció muy interesante cuando estábamos haciendo todo este trabajo de estadística y de *data mining* en los distintos países es que, claro, tenemos que volver al contexto, el dato solo no nos decía nada.

Nos pasó algo muy simpático, estábamos trabajando en *topic modeling* en datos de la COVID de México, Colombia y Argentina. Estábamos comparando estos datos y todo venía bastante similar en las estadísticas y en los tópicos como salud, pandemia, etcétera. En el caso de México y Colombia había mucha autorreferencialidad: "no tenemos vacunas en tal país". Pero cuando llegamos a trabajar esa semana en *topic modeling* en Argentina, nos aparecen cosas raras. Nadie en el equipo estaba mirando demasiado las noticias. Nos aparecía "hijo", "nombre"... Algo estamos haciendo mal, curamos todos los datos, hicimos la base de datos, no hicimos el *web scraping* total y ahora, ¡algo nos falla! ¡Nosotras estábamos fallando! Parece que esa semana, que fue la semana de inicio de la pandemia, había nacido un niño, el día que Alberto Fernández, nuestro presidente, dio el decreto de cierre de todo, unos padres decidieron ponerle al niño de segundo nombre COVID, por eso nos aparecía "nombre", "Ciro"... estábamos completamente desestructuradas… dijimos “tiremos las humanidades digitales ya, no hagamos nada más”.

En fin, este es un ejemplo para mostrarles cómo explotar y cómo visualizar datos, pero sobre todo para destacar que el dato solo, el dato sin un contexto no tiene sentido y que en las humanidades digitales lo más importante es, no pensar en este juego con las herramientas, sino pensar cómo hacemos para que todo este trabajo con los datos tenga un sentido en una narrativa. En humanidades digitales solemos llamar a estos marcos teóricos "lectura distante", "macroanálisis", "algoritmos críticos".

Cierro esta charla trayendo a colación la necesidad de pensar los datos casi siempre como *capta,* esta cuestión de que lo dado siempre tiene esta otra capa que, como investigadores, como estudiantes, como humanos, como humanistas, estamos llamados a analizar. También es preciso no quedarnos con lo cuantitativo, no quedarnos con el número, no quedarnos con la frecuencia, porque aparezca tantas veces un término no nos va a cambiar la vida, nos va a

cambiar la vida entender por qué pasa eso, darle un sentido y pensar cómo este proceso que estamos haciendo algorítmicamente con una máquina puede completar hipótesis que traemos de antemano. Hay muchas veces que las máquinas no nos dicen todo, nosotros ya teníamos esas hipótesis, pero podemos contrastarlas y podemos de algún modo también certificarlas, si se quiere, de esta forma más basada en los datos, pero sin perder de vista el texto, la narrativa, el contexto y el sentido que traen las humanidades.

Referencias

Alles Torrent, S., del Rio Riande, G., de León, R., Fila, M., Hernández, N., Bonnell, J. & Song, D. (2020). Narrativas digitales de la COVID-19 en Twitter: de los datos a la interpretación. *Publicaciones De La Asociación Argentina De Humanidades Digitales*, 1, e002. https://doi.org/10.24215/27187470e002

del Rio Riande, G. et al. (2018). Biblioteca Digital. https://hdlab.space/biblioteca-digital/

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron.* MIT Press.

Instituto de Geografía, Historia y Ciencias Sociales Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Centro de la Provincia de Buenos Aires (2021). Índice de Calidad de Vida a escala departamental (1991-2001-2010). *Poblaciones.org* https://poblaciones.org/2021/07/23/indice-de-calidad-de-vida-a-escala-departamental-1991-2001-2010/

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

Perrigo, B. (18 jan, 2019). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time.* https://time.com/6247678/openai-chatgpt-kenya-workers/

Schöch, C. (August 1, 2013) Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities, 2*(3). https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Scholarly Communication Lab. (2017). World scaled by number of documents published in 2017. https://scholcommlab.ca/cartogram/