Trabajos, Comunicaciones y Conferencias

Actas del Workshop Iberoamericano de Estudios métricos de la actividad científica orientada a temas locales/regionales

Sandra Miguel
(coordinadora)

FaHCE
FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN
UNIVERSIDAD NACIONAL DE LA PLATA

ACTAS DEL WORKSHOP IBEROAMERICANO DE ESTUDIOS MÉTRICOS DE LA ACTIVIDAD CIENTÍFICA ORIENTADA A TEMAS LOCALES/REGIONALES

La Plata, 21 y 22 de agosto de 2018

Sandra Miguel
(coordinadora)

FaHCE
FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN
UNIVERSIDAD NACIONAL DE LA PLATA

2019

Diseño: D.C.V. Celeste Marzetti
Tapa: D.G.P. Daniela Nuesch
Editora por Prosecretaría de Gestión Editorial y Difusión: Leslie Bava

Queda hecho el depósito que marca la Ley 11.723

ISBN: 978-950-34-1742-3
Colección: Trabajos, comunicaciones y conferencias, 37

Cita sugerida: Miguel, S. (Coord.). (2019). *Actas del Workshop Iberoamericano de estudios métricos de la actividad científica orientada a temas locales/regionales* (2018 : La Plata). La Plata : Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la Educación. (Trabajos, comunicaciones y conferencias ; 37). Recuperado de https://www.libros.fahce.unlp.edu.ar/index.php/libros/catalog/book/130

Índice

Prólogo

Este libro de actas reúne las intervenciones presentadas en el **Workshop Iberoamericano de estudios métricos de la actividad científica orientada a temas locales/regionales**, realizado en la ciudad de La Plata, Argentina, el 21 y 22 de agosto de 2018.

El evento estuvo organizado por el Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCS- UNLP/CONICET) en el marco del proyecto PICT 2015-2144 "La producción científica sobre los países latinoamericanos. Aproximación a su estudio desde una perspectiva bibliométrica y relación con indicadores del contexto económico y social", acreditado por la Agencia Nacional de Promoción Científica y Tecnológica del Ministerio de Educación, Cultura y Ciencia y Tecnología de la República Argentina. Para la realización de la reunión se contó con el financiamiento de este proyecto y del subsidio para reuniones científicas RC 2017-0323 otorgado por la Agencia Nacional de Promoción Científica y Tecnológica.

El Workshop reunió investigadores con diferentes perfiles formativos y trayectorias, cuyas intervenciones permitieron generar un espacio compartido de debate y un intercambio de perspectivas teóricas y metodológicas relacionadas con las políticas científicas y con la obtención de métricas y visualizaciones derivadas de la producción y su relación con indicadores del contexto económico y social. Cabe destacar que aunque el interés en los estudios sobre América Latina y España no es nueva, sí es novedoso que desde las regiones periféricas se indague con una perspectiva bibliométrica la producción científica que se realiza sobre los países de la región, sea ésta producida en el propio país o llevada a cabo desde el extranjero.

Las exposiciones se desarrollaron en los siguientes ejes temáticos:

- Presencia de temas orientados a lo local / regional en las políticas y agendas de investigación
- Aproximaciones metodológicas para el abordaje cienciométrico de la producción en temas locales / regionales
- Estudios métricos sobre la producción científica en temas locales de países iberoamericanos

Finalmente agradecer a todos los expositores y asistentes que hicieron posible el intercambio de conocimientos y experiencias en los temas abordados, y la concreción de esta publicación que recoge los principales resultados del encuentro.

Sandra Miguel
La Plata, 2019

Aproximación metodológica para la extracción de temas de un corpus bibliográfico referencial a partir del lenguaje natural

Sebastián Varela[1] *y Claudia M. González*[2]

Introducción

Se presenta un ejercicio exploratorio de text mining en el que se busca explorar y comparar la producción científica de de Argentina y México en el área de las ciencias ambientales. Para ello se utilizaron un conjunto de registros bibliográficos extraídos de la base de datos Scopus en el periodo 2007-2016 en cuyos títulos, resúmenes o palabras claves aparece mencionado el nombre del país.

Metodología

En la base de datos Scopus se encontraron en dicho periodos 5.943 artículos mexicanos y 3.190 argentinos. Los registros fueron descargados en formato BibTexT y convertidos en *data frames* del lenguaje de programación R utilizando el paquete *bibliometrix*.

La unidad de análisis es el resumen o abstract de artículo. Se procedió a desmenuzar el texto en tokens[3] individuales -términos en este caso- y a transformarlo en una matriz ordenada de datos (Silge y Robinson, 2017).

[1] varela.sebastian@gmail.com

[2] Instituto de Investigaciones en Humanidades y Cs Sociales (UNLP/CONICET), La Plata, Argentina. cgonzalez@fahce.unlp.edu.ar

[3] Un token es una unidad significativa de texto, y el proceso de tokenización es el proceso de división de un corpus textual en tokens. En este trabajo usamos como tokens en primer lugar terminos, y posteriormente bigramas (grupos de dos palabras).

La matriz de datos resultante tiene 1.931.639 filas (términos), de los cuales 1.233.227 provienen de abstracts mexicanos y 698.412 de argentinos. Posteriormente se aplican diccionarios de stopwords[4]: los diccionarios estándar del paquete *tidytext* y uno propio para eliminar números y otros caracteres inapropiados. Como resultado quedan un total de 73.666 términos, 44.240 procedentes de resúmenes mexicanos y 29.426 de resúmenes argentinos.

Resultados y discusión

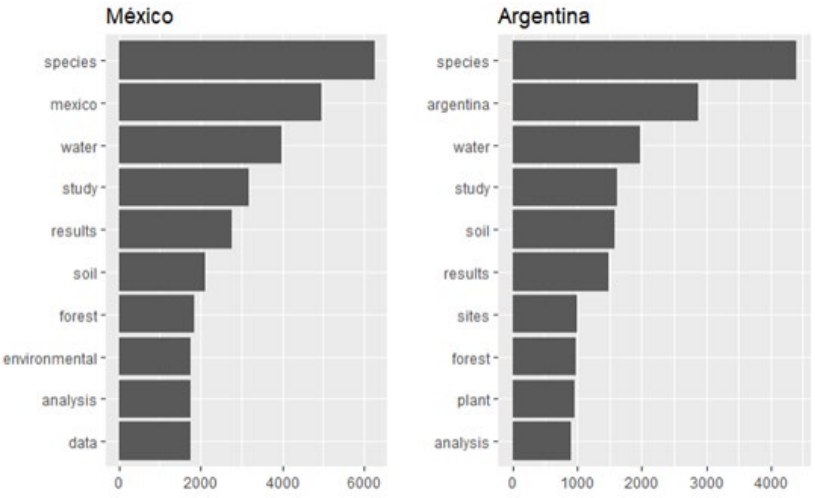
En la Figura 1 podemos observar con gráficos de barras el top ten de los términos más frecuentemente utilizados en la producción académica de ambos países. Llama la atención que no aparezca el término *grassland* (pastizal, pradera), lo que invita a pensar que hay un descuido en la agenda de investigación en relación a ese ecosistema, habida cuenta de que es el ecosistema predominante en Argentina, y muy relevante en México (aunque en este país los bosques y los ambientes áridos tienen mayor proporción)[5]. Sería interesante indagar en qué medida los artículos con las palabras *soil* y *plant* en sus abstracts refieren a estudios sobre *grasslands*.

Los términos *results*, *study*, *data*, *analysis* son palabras típicas del lenguaje científico que no aportan desde el punto de vista disciplinar, y se podrían eliminar del análisis. El examen global de estos gráficos permite inferir que no hay diferencias importantes entre las agendas de investigación de ambos países. Se encuentran los siguientes términos en común: *species*, *water*, *soil*, *forest*. En cambio, el término *environmental* aparece sólo en el corpus mexicano, mientras que *plant* entra en el top ten de Argentina y no en el de México.

[4] Las stopwords son palabras comunes que son inútiles para el objetivo analitico. Algunos ejemplos en inglés son: *“the”, “of”, “to”, “from”* etc.

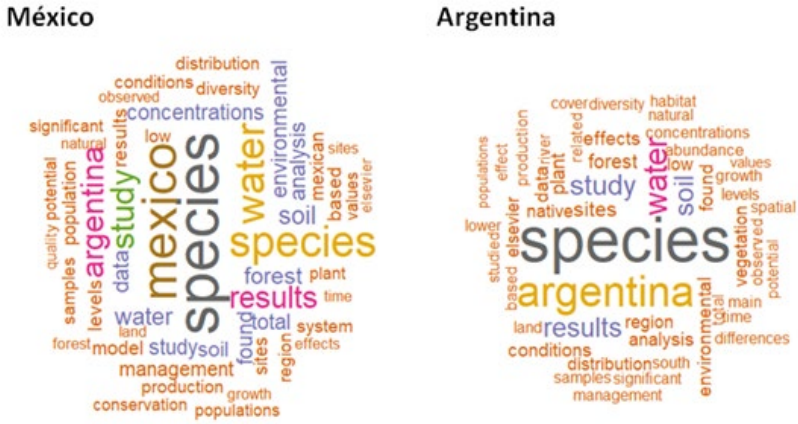
[5] En Argentina *grassland/grassland* -singular y plural sumados- tiene 517 menciones, y en México 193.

Figura 1. Términos más frecuentes



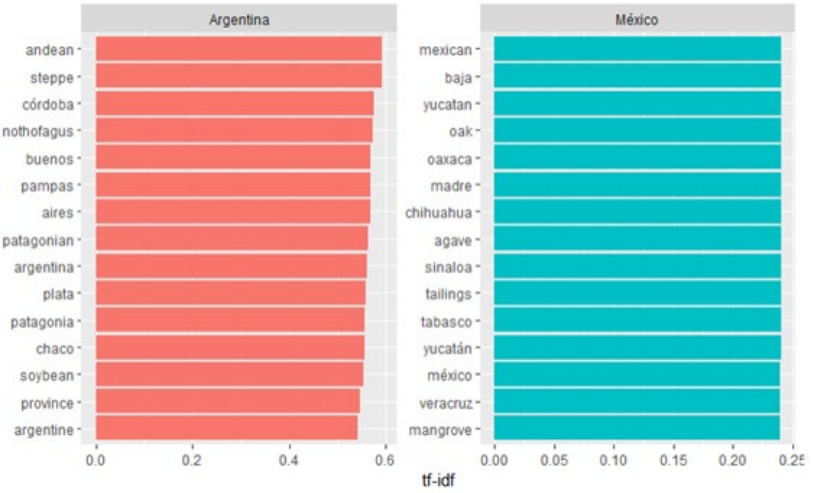
A continuación, en la Figura 2, están las word clouds (nubes de palabras) para cada país. Las nubes de palabras son listados en los que la importancia de los términos se pondera según el tamaño y color de letra. Respecto de estas nubes, cabe decir algo similar a lo dicho respecto de la Figura 1: no se observan mediante diferencias relevantes entre ambos corpus.

Figura 2. Nubes de palabras



Para observar comparativamente las características de cada colección, el paso siguiente sería utilizar el estadístico *tf-idf* (frecuencia de término – frecuencia inversa de documento), un estadístico usado para identificar términos que son especialmente relevantes para un documento (o colección de documentos) particular[6]. Se observan en la Figura 3 los bigramas más frecuentes para cada corpus: en el caso de Argentina aparecen las referencias geográficas más relevantes para la producción de conocimiento medioambiental[7]: *andean*, *cordoba*, *buenos*, *aires*, *patagonian*, *patagonia*, *plata* (referencia al río o cuenca de La Plata), *pampas*, *chaco*. Por otro lado hay referencias a un ecosistema: *steppe*(referencia mayormente a la Patagonia); y finalmente a especies vegetales: *nothofagus* (planta patagónica) y *soybean* (soja), este último término relacionado a la actividad agrícola.

Figura 3. TF-IDF



En el caso de México, las referencias geográficas son *baja* (California), *yucatan* (que aparece con y sin acento), *oaxaca*, *chihuahua*, *sinaloa*, *madre* (Sierra), *tabasco*, y *veracruz*. Por otro lado hay referencias a especies

[6] Para más detalles sobre este estadístico véase Silge y Robinson (2017).

[7] Esto puede deberse en parte a que en dichas regiones hay instituciones de investigación importantes dedicadas a la producción de conocimiento en ciencias ambientales.

vegetales: *oak, mangrove* y *agave*. Finalmente aparece el término *tailings*, referido a desechos tóxicos de la actividad minera.

Algo llamativo cuando se analizan estos gráficos es que no aparecen las disciplinas del campo ambiental: *physiology, ecology, restoration, conservation*. Se requiere mayor elucidación sobre este punto aunque se puede conjeturar que no aparecen porque en los resúmenes se tiende a mencionar las diversas subdisciplinas y especialidades del campo ambiental.

Continuando el proceso de exploración, se cambia el proceso de tokenización de palabra a bigrama (conjunto de dos palabras), lo cual puede ser más interesante para la detección de temas de investigación. Para facilitar la interpretación se utilizan gráficos de redes de palabras (generados con los paquetes de R *widyr* y *ggraph*). Estas redes tienen la ventaja de permitir visualizar las relaciones entre las palabras más frecuentes de manera simultánea. A partir de la interpretación de los nodos resultantes se pueden inferir temáticas. El orden de aparición de las palabras -o direccionalidad- se indica con una flecha.

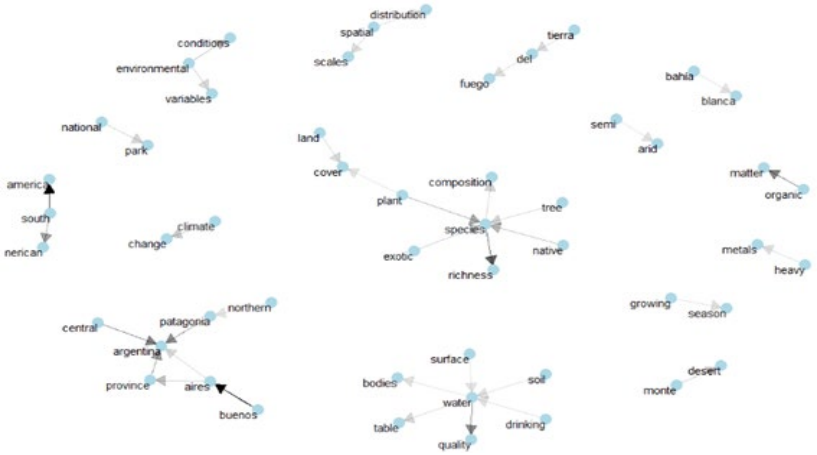
En primer lugar cabe mencionar referencias geográficas: de izquierda a derecha aparece una pequeña red de bigramas con referencia geográfica: *south → america/south → american*, y luego un cluster de bigramas probablemente explicado por la producción de grupos de investigación con referencia geográfica en la provincia de Buenos Aires y la Patagonia norte, aunque esta conjetura requiere una indagación más profunda[8]. Otras referencias geográficas son Tierra del Fuego, que a pesar de ser una provincia pequeña aparece como una provincia relevante para las ciencias ambientales[9]; y Bahía Blanca que también aparece como una referencia geográfica destacada[10].

[8] Cabe señalar que una parte de dichos trabajos se deben a la producción deI instituto de Investigaciones Fisiológicas y Ecológicas vinculadas a la Agricultura (IFEVA), de la Facultad de Agronomía (UBA).

[9] En dicha provincia se encuentra el Centro Austral de Investigaciones Científicas (CADIC-CONICET).

[10] En dicha ciudad está la Universidad Nacional del Sur e importantes institutos de investigación del CONICET orientados investigación ambiental y oceánica.

Figura 4. Bigramas corpus argentino



Hay luego pequeños clusters de bigramas que dan cuenta de referencias técnicas: análisis espacial y de mapas (*spatial, distribution, scales*) y *environmental → variables*. Finalmente aparecen los agrupamientos relacionados con *temáticas específicas* de investigación: en primer lugar dos nodos numerosos, uno referido a estudios sobre ecología de comunidades vegetales (*species, richness, composition, plant, exotic*, etc.); y otro sobre limnología (*water, quality, surface, drinking*, etc.). También una serie de bigramas de fácil interpretación: contaminación con metales pesados (*heavy → metals*); cambio climático (*climate → change*); estudios de ecosistemas semi áridos (*semi → arid*), estudios del crecimiento de las plantas (*growing → season*); estudios sobre montes desérticos (*monte → desert*); estudios sobre materia orgánica (*organic → matter*); cambio climático (*climate → change*), y estudios sobre los parques nacionales (*national → park).*

A continuación (Figura 5) se examina la red de bigramas de frecuencias absolutas resultante de los abstracts mexicanos:

Figura 5. Bigramas corpus mexicano

Encontramos en este caso algunos bigramas que sugieren temáticas en común con la producción argentina: *heavy → metal; organic → matter; environmental → conditions; climate → change*. Algunas redes de bigramas también son similares sugiriendo temáticas comunes: ecología de comunidades vegetales (*species, richness, plant*, etc.), y limnología (*water, quality, drinking* -aunque en este caso se agrega el bigrama *air → quality*).

Por otro lado las siguientes son temáticas distintivas de México: diversidad genética (*genetic → diversity*); servicios de ecosistemas (*ecosystem → services*); cobertura del suelo (*land → cover*); bosques de neblina o altura (*cloud → forest*) y biosfera de reserva (*biosphere → reserve*)

Finalmente las referencias geográficas más frecuentes son *sierra → madre*; *yucatan → peninsula;* y el cluster que menciona Baja California, México central, Veracruz México y México City. Se ve a continuación una comparación de los corpus interpretando los bigramas.

Figura 6. Comparación temática de los corpus

Conclusión

El ejercicio propuesto permite explorar la producción científica de ambos países una disciplina específica, permitiendo identificar similitudes y diferencias.

El uso de bigramas (y eventualmente otros n-gramas) resulta de potencial utilidad para la identificación de temas de investigación.

Luego de este primer ejercicio exploratorio, resulta de interés propiciar estudios entre la producción científica Argentina y la de países más cercanos (Brasil o Chile), focalizando sobre tópicos menos generales y más acotados, restringiendo el foco a ecosistemas específicos (por ejemplo ambiente acuático o estepa), o bien en temáticas particulares como la contaminación.

La identificación de temas o tópicos en grandes volúmenes de datos sigue siendo un desafío para las ciencias de la información y de la computación. El uso de algoritmos y técnicas estadísticas cada vez más sofisticados parece ser un camino posible.

Referencias bibliográficas

Aria, M. y Cuccurullo, C. (2017) *bibliometrix: An R-tool for comprehensive science mapping analysis*, Journal of Informetrics, 11(4), pp 959-975, Elsevier.

Pedersen, T. L. (2017). *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. Recuperado de https://cran.r-project.org/package=ggraph.

Silge J, R. D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). Recuperado de http://dx.doi.org/10.21105/joss.00037.

Silge, J. and Robinson, D. (2017). *Text mining with R.* New York: O'Reilly.

En un contexto de creciente incorporación del acontecer local y nacional en contextos globales, el impulso de políticas que orientan la investigación hacia temas locales y a la resolución de problemas sociales, productivos y medioambientales, plantea nuevos desafíos en la valoración de los resultados e impacto que esa investigación produce, así como también de la transferencia e innovación. Este libro de actas reúne las ponencias presentadas en el Workshop Iberoamericano de estudios métricos de la actividad científica y tecnológica en temas locales/regionales, con aportes de autores de diferentes perfiles formativos y trayectorias, en un intento por contribuir a los debates teóricos y metodológicos para la obtención de métricas y visualizaciones derivadas de la producción científica de países iberoamericanos y la relación con indicadores del contexto económico y social. Los contenidos están organizados en tres ejes temáticos. El primero se enfoca en cuestiones relativas a las políticas y agendas de investigación; el segundo presenta diferentes aproximaciones metodológicas para el abordaje cienciométrico de la producción científica en temas locales/regionales, y el tercero recoge estudios de caso de un grupo de países de la región.

Trabajos, Comunicaciones y Conferencias, 37

ISBN 978-950-34-1742-3