Uso de R en bibliometría. Exploración de técnicas para la detección de temas locales de investigación

Claudia M. González[1,3], Sebastián Varela[2,3], Sandra E. Miguel [1,3]

1 Universidad Nacional de La Plata. Facultad de Humanidades y Cs de la Educación. Dpto de Bibliotecología, Argentina
cgonzalez@fahce.unlp.edu.ar
2 Universidad Nacional de La Plata. Facultad de Humanidades y Cs de la Educación. Dpto de Sociología, Argentina
svarela@fahce.unlp.edu.ar
3 IdIHCS - Instituto de Investigaciones en Humanidades y Cs Sociales (CONICET/UNLP), Argentina
smiguel@fahce.unlp.edu.ar

Palabras Claves: Cienciometría - Clustering - Modelado de tópicos - Latent Drichlet Allocation (LDA)

1 Introducción

La investigación de aspectos de la ciencia a partir de las publicaciones puede hacerse desde una perspectiva bibliométrica, que implica el análisis estadístico sobre la información descriptiva, de contribuciones y de citación de las producciones científicas de carácter bibliográfico. O también desde una perspectiva temática, que permite detectar los frentes de investigación a partir de elicitar los tópicos allí tratados de manera que se pueda obtener información resumida útil para establecer relaciones analíticas. Esta identificación de temas puede hacerse utilizando el método de análisis de contenido tradicional, generalmente usado en estudios cualitativos sobre poca cantidad de documentos, o utilizando modelado estadístico, lo que permite extraer y analizar información de grandes colecciones de texto, aplicando algoritmos no supervisados. Diversas técnicas se han utilizado para la realización de representaciones temáticas de corpus textuales. El análisis de co-palabras es la técnica básica, la cual ha sido trabajada complementariamente con técnicas estadísticas de reducción del volumen de datos sin mayores pérdidas de información, tales como el análisis de *clusters*, el escalamiento multidimensional (MDS), el análisis factorial (FA), el *blockmodeling*, el análisis de redes sociales (ARS) y dentro de esta, técnicas específicas como las redes *Pathfinder*. En el último tiempo, la técnica *Latent Dirichlet Allocation* (LDA) ha sido una de las más usadas (Blei et al, 2003), y puede utilizarse además para mostrar la evolución de tópicos y asociaciones entre temas y autores. Es una forma de minería textual que funciona identificando patrones de co-ocurrencia de palabras y produciendo agrupamientos por similitud que denominamos tópicos. Cuando las técnicas basadas en co-ocurrencias de palabras se aplican sobre las bases de datos bibliográficas comprensivas, explotando los títulos y los resúmenes de los documentos como corpus textual, se las toma como técnicas bibliométricas.

2 Materiales y Métodos

En este trabajo se muestran resultados preliminares obtenidos al aplicar la técnica de *clustering* basado en *k-means* y un modelado de tópicos usando *Latent Dirichlet*

Allocation (LDA) sobre un corpus de registros de la base de datos Scopus utilizando paquetes del lenguaje R. El objetivo general es detectar aquellas áreas que permitan estimar el esfuerzo que realizan los recursos humanos de investigación de determinado lugar geográfico para abordar los problemas que son propios de ese territorio y sus habitantes. Por ello, el corpus responde a una estrategia de búsqueda que comprende la producción del gran área Ciencias Sociales & Humanas en el periodo 2010-2015, restringida a aquellos trabajos que tuvieran algún autor con afiliación argentina, además de contener Argentina (o alguna de sus variaciones explicitadas en la estrategia de búsqueda) en los campos título, resumen y palabras clave. Para el procesamiento se utilizaron los paquetes *bibliometrix (2017),* que sirve para realizar análisis bibliométricos y de co-citación; el paquete *topicsmodels* (2017) que permite implementar LDA y CTM (*Correlated Topics Models*); el paquete *tidytext* (2017) que permite aplicar algunas técnicas de procesamiento del lenguaje natural dentro de las cuales se encuentra la detección de n-gramas. En este trabajo, se procedió a sacar bigramas y se los interpretó de manera cualitativa, detectando 7 áreas (ver referencia de colores en figura 1). Se generaron los clusters mediante la técnica de *K-medias* y se procedió a realizar un análisis de los clusters obtenidos a la luz de las categorías que se derivaron de los bigramas. Luego se realizó lo mismo aplicando modelado de tópicos con LDA.

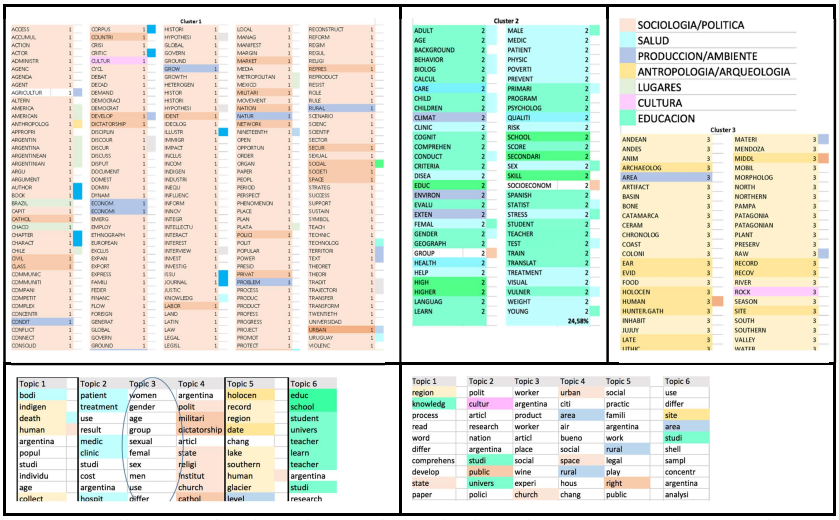

Fig. 1. Fila 1: Selección de términos de los clusters generados con k-means coloreados según áreas detectadas a partir del análisis manual de bigramas. Fila 2: Tópicos generados con LDA aplicando 2 métodos diferentes de ajuste: el *Variational Expectation Maximization* (LDA-VEM) y *Correlated Topic Model* (CTM-VEM). Interpretación de pertenencia a áreas.

Referencias

1. Blei, D.M., Ng, A.Y. y Jordan, M.I. (2003). Latent dirichlet allocation. *J. Mach. Learn*. p. 993-1022.
2. Contador Pachón, S. (2015). *Clasificación de textos científicos con R*. Trabajo presentado en la *VII Jornadas de Usuarios de R.* Universidad Complutense de Madrid, Madrid.
3. Silge, J. and Robinson, D. (2017). *Text mining with R.* New York: O'Reilly.