Cita sugerida para este artículo
González, C. M., Varela, S. y Miguel, S. (2017). Aplicación de algoritmos no supervisados para la detección de tópicos de investigación. Presentado en *V Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología.* Universidad Nacional de La Plata, Facultad de Humanidades y Ciencias de la Educación, La Plata. Recuperado de

Aplicación de algoritmos no supervisados para la detección de tópicos de investigación

***Claudia M. González*[1,3], *Sebastián Varela*[2,3], *Sandra Miguel*[1,3]**
[1] Universidad Nacional de La Plata. Facultad de Humanidades y Cs de la Educación. Dpto de Bibliotecología, Argentina.
[2] Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la Educación. Dpto de Sociología, Argentina.
[3] Instituto de Investigaciones en Humanidades y Cs Sociales- IdIHCS- (CONICET/UNLP)

Resumen

En el trabajo se realiza un estudio exploratorio sobre la aplicación de algoritmos no supervisados basados en agrupamientos (*clustering)* y probabilidades para la detección de tópicos de investigación. El estudio se realiza usando un corpus bibliográfico sobre Ciencias Sociales de la base de datos Scopus para el periodo 2010-2015. Se muestran los resultados obtenidos aplicando la técnica de *clustering* basado en *k-means* y el modelado de tópicos usando *Latent Dirichlet Allocation* (LDA).

Palabras claves

Bibliometría - Modelado de tópicos - Clustering – Latent Drichlet Allocation (LDA)

Introducción

A los fines del tratamiento automático de la información, ciertas tareas como la clasificación automática, la detección de novedades, la generación automática de resúmenes y la elaboración de juicios de similitud y relevancia en la recuperación de información son desde hace muchos años foco de numerosas investigaciones. En todas ellas el interés está puesto en desarrollar algoritmos que logren elaborar descripciones breves de cada ítem de información miembro de una colección, que lo hagan de manera eficiente en colecciones grandes, y además que sean no supervisados, es decir que actúen sin intervención humana.

ISSN 1853-5631- web: http://jornadabibliotecologia.fahce.unlp.edu.ar

Estos desarrollos se basan principalmente en el modelado matemático de las colecciones de textos, y en muchos casos suelen complementarse con pre-procesamientos lingüísticos. Dentro de los enfoques utilizados, el más tradicional se basa en la estadística de frecuencias, aunque hace algunos años ya se viene trabajando en modelados basados en probabilidades. De cualquier manera, más allá de cualquier técnica particular, a todas ellas las rige el principio de conservar las relaciones estadísticas/probabilísticas esenciales de los documentos (intra-documental) en el contexto de la colección a la que pertenecen (inter-documental).

En el caso particular de la detección de temas/tópicos, las aplicaciones más importantes tienen que ver con la indización automática, la generación de estructuras de navegación de información y el análisis de tendencias, entre otras. Así, en este trabajo se propone su uso para el análisis de tendencias en temas de investigación. Se considera que dado que un tema o tópico se define como un conjunto coherente de contenido semánticamente relacionado que se refiere a un solo argumento, si se tiene un conjunto de productos bibliográficos científicos, identificar en dicho corpus los agrupamientos semánticos que ocurren con frecuencia sirve para caracterizar ese aspecto de la actividad científica que tiene que ver con el "qué" se investiga.

Objetivo

El objetivo de este estudio es explorar los resultados que arroja la aplicación de dos técnicas específicas, una estadística y otra probabilística, en la detección de temas en corpus bibliográficos referenciales. Esto implica conformar un corpus textual experimental, estudiar las técnicas, identificar las herramientas que posibilitan la aplicación de las mismas de manera no supervisada y evaluar los resultados obtenidos a la luz de un objetivo cienciométrico.

Las técnicas seleccionadas son dos: el agrupamiento (clustering) basado en k-means y Latent Dirichllet Allocation (LDA). La primera es una técnica estadística que se basa en el cómputo de las frecuencias de las palabras, la vectorización documental en base a ellas y el cálculo de medidas de similaridad vectorial para realizar los agrupamientos. La segunda, corresponde estrictamente a los denominados *topics models*, que son modelos matemáticos estocásticos debido a la existencia de incertidumbre al momento de formular respuestas o salidas de dichos modelos, es decir, esto implica que los resultados o salidas son probabilidades.

Las técnicas

Revisamos en este apartado los componentes, y en algún caso los antecedentes de las técnicas propuestas. Dada la complejidad que implica la explicación exhaustiva de cada una de ellas, hemos optado por un hacer un resumen abreviado con los rasgos principales, que por otra parte es lo que nos permite la extensión permitida para este trabajo. Hablaremos del *clustering* basado en *k-means,* de la reducción de dimensiones *tf-idf*, de dos antecedentes que nos permiten entender la transición hacia la técnica LDA que son las técnicas *Latent Semantic Indexing* (LSI) y *Probabilistic Latent Semantic Analisys* (pLSA). Por último introducimos la *Latent Diritchlet Allocation.*

Clustering basado en k-means

El clustering es una técnica de exploración de datos utilizada para descubrir grupos o patrones en un conjunto de datos. Existen dos estrategias estándar para la generación de los agrupamientos (clusters): el método de particiones y el método jerárquico. Particularmente, el algoritmo de *k-means* (MacQueen, 1967) pertenece al método de particiones, en el que cada cluster está representado por un centro (centroide) que es la media de los puntos de datos del cluster. La idea es hacer una clasificación en la que los objetos dentro del mismo cluster sean lo más similares posibles (alta cohesión intraclase), a la vez que los objetos en *clústeres* diferentes sean lo más disímiles posibles (baja interrelación entre clases).

Para utilizar esta técnica, el primer paso es indicar el número de clusters (*k*) que se generarán. El algoritmo comienza seleccionando aleatoriamente *k* objetos del conjunto de datos para que sirvan como centros iniciales para los conglomerados. Los objetos seleccionados también se conocen como centroides. A continuación, cada uno de los objetos restantes se asigna a su centroide más cercano y esto se hace utilizando la distancia euclidiana entre el objeto y la media del centro del grupo. Este paso se llama "paso de asignación de clúster". Después del paso de asignación, el algoritmo calcula el nuevo valor medio de cada grupo, diremos que es el paso de "actualización del centroide". Con los centros recalculados, cada observación se verifica nuevamente para ver si puede estar más cerca de un clúster diferente. Todos los objetos se vuelven a asignar usando las medias de clúster actualizadas. Los pasos de actualización de centroides y asignación de clúster se repiten de forma iterativa hasta que las asignaciones de clúster dejan de cambiar, es decir, hasta que se logra la convergencia (los clusters formados en la iteración actual son los mismos que los obtenidos en la anterior iteración).

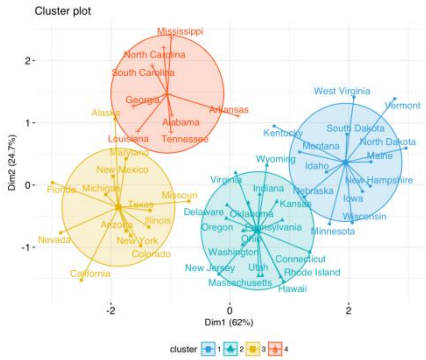


Fig 1: Ejemplo de clustering k-means

Reducción tf-idf

Está técnica es muy importante porque fue la primera en presentar la vectorización documental. Se basa en reducir cada documento de la colección a un vector de números reales, cada uno de los cuales representa proporciones de recuentos de palabras. En el conocido esquema propuesto por Salton y McGill (1983) se genera un vocabulario con las palabras significativas de los documentos de una colección, llamados términos, y para cada uno de los documentos se realiza un conteo de las frecuencias de las palabras del vocabulario en el documento. Luego de una normalización adecuada, estas frecuencias a nivel del documento (*tf*) se comparan con las frecuencias de esos términos a nivel de la colección (*idf*) y se expresan en escala logarítmica y normalizada. El resultado final es una matriz término-documento en cuya intersección se encuentran los valores *tf-idf*. Esta técnica realiza una reducción del documento de longitud arbitraria a una lista de números (lo que llamamos vector) de longitud fija.

La crítica que se le hace a esta técnica es que la reducción lograda en la descripción es relativamente pequeña, además de revelar poco de la estructura estadística inter e intra documental.

Titles:
c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Terms	Documents							
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Fig 2: Ejemplo de matriz término-documento

Latent Semantic Indexing (LSI)

Esta técnica propuesta por Deerwester y otros (1988) se basa en el principio de que las palabras que se utilizan en los mismos contextos tienden a tener significados similares. Una característica clave de LSI es su capacidad para extraer el contenido conceptual de un corpus mediante el establecimiento de asociaciones entre los términos que aparecen en contextos similares. Básicamente la técnica tiene la capacidad de detectar las correlaciones más fuertes entre los términos, y lo hace proponiendo una reducción de las dimensiones mediante una descomposición de la matriz término-documento en valores singulares (SVD). Esto permite identificar un subespacio lineal en el espacio *tf-idf* que captura la mayor parte de la varianza en la colección. Es una técnica basada en el álgebra lineal que trabaja descomponiendo matrices de co-ocurrencia.

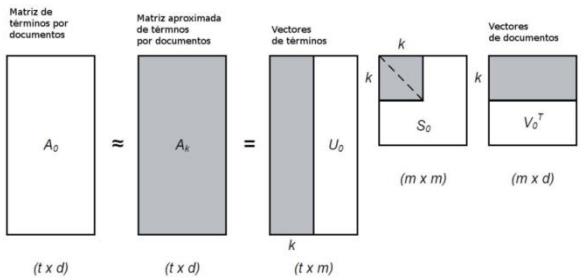


Fig 2: Interpretación de la descomposición de una matriz en valores singulares (SVD)

Si bien esta técnica logró una mayor reducción, muy útil para el caso de colecciones grandes, investigaciones posteriores creen encontrar una mejor resolución apoyándose en los métodos bayesianos (probabilidad condicionada).

Probabilistic Latent Semantic Indexing (pLSI)

Este avance sobre el modelado LSI fue propuesto por Hofmann (1999) y se encuadra dentro de los modelos estadísticos denominados “modelado de aspecto”. Es un modelo de variable latente de co-ocurrencia de datos que asocia una variable no-observada, la variable latente, con cada observación. En él, cada palabra en el documento es vista como componente de un compuesto mixto mayor que son variables aleatorias multinomiales que pueden ser vistas como representaciones de tópicos. Cada palabra es generada por un solo tópico, y diferentes palabras en el documento pueden ser generadas por tópicos diferentes. Cada documento es representado como una lista de proporciones de estos componentes mixtos y por lo tanto reducido a una distribución de probabilidad en un conjunto fijo de temas. Esta distribución es la descripción resumida

que se asocia a cada documento.

Suponiendo que *d* es un documento, *z* es un tópico, *w* es una palabra y N_d es el número de palabras en el documento *d*, *P(z|d)* denota la probabilidad del tópico *z* en el documento *d,* y la *P(w|z)* como la probabilidad de la palabra *w* en el tópico *z.* Para el PLSA el procedimiento de generación de cada palabra en el documento es *P*(*w*|*z*). Es decir que

Para cada documento $d \in \{1, ..., N\}$

Para cada palabra *w* en el documento *d*

Se genera aleatoriamente un tópico *z* extraído de la distribución de tópicos *P(z|d)*

Se selecciona aleatoriamente una palabra *w* de la distribución de palabras (vocabulario) *P*(*w*|*z*)

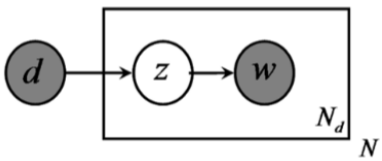


Fig 3 modelo Probabilistic Latent Semantic Analysis

Si bien esta propuesta fue un avance hacia el modelado probabilístico de textos, la crítica que se le hace es que es incompleta ya que no provee un modelado probabilístico a nivel de los documentos. Cada documento es representado como una lista de números que reflejan las proporciones de tópicos, pero la técnica en sí no ofrece un modelo probabilístico generativo para esos números.

Latent Dirichlet Allocation (LDA)

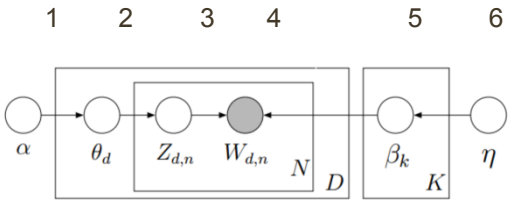
Es uno de los métodos de modelado de tópicos más usados, el cual integra la clase de modelos que se denominan generativos. Dentro de la familia es la técnica más simple y puede pensarse que es como el *Latent Semantic Analysis* pero implementada con métodos Bayesianos. Parte de considerar que hay temas latentes presentes en todos los documentos, y cada palabra en el documento contribuye con el tema o tópico, el cual nos termina dando una aproximación sobre lo que trata el documento o la colección. Así cada documento es una amalgama de múltiples tópicos en el contexto del corpus, y cada tópico es un surtido de miles de palabras, mientras que cada palabra es una entidad que contribuye con el tema del documento.

El modelado de tópicos presenta una estrategia en tres frentes para atacar la complejidad: en primer lugar cada palabra en cada documento es asignada a un tópico.

En este proceso se estima la distribución de probabilidad conjunta para todas las variables. Se calcula el peso probable de las palabras, y se crean los tópicos basados en el peso de cada palabra, cada tópico asignará diferentes pesos a diferentes palabras. Para este modelo, el orden de las palabras no es un problema, porque cada documento es tratado como una "valija de palabras", tampoco lo es el orden de los documentos. El tópico puede asumirse que es una distribución de probabilidad a través de una multitud de palabras, por lo cual el modelado de tópicos no es más que una relación probabilística entre tópicos no observados y variables lingüísticas observadas. Luego, la proporción de cada tópico es estimada para cada documento. Para ello hace un cálculo de la probabilidad de los documentos dados los tópicos (prior) para asignar los tópicos a los documentos de la colección. Luego hace un cálculo de la probabilidad de los tópicos dados los documentos (posterior) para crear los tópicos de la colección. Finalmente se explora la distribución de tópicos en todo el corpus. LDA modeliza las probabilidades prior/posterior como distribuciones Dirichlet, β y θ con hiperparámetros η y . Esta es un tipo de distribución probabilística multivariada.

Para encontrar los tópicos, LDA explora dos distribuciones de probabilidad:
$\alpha = P(k|d)$, la probabilidad del tópico *k* en el documento *d;*
$\beta = P(w|k)$, la probabilidad de la palabra *w* en el tópico *k*.

Inicialmente, α y β se inician aleatoriamente como sigue: cada palabra en el documento se generó al elegir aleatoriamente un tema (de la distribución de temas en documentos) y luego al azar se escoge una palabra (de la distribución de palabras en los temas). Sucesivas iteraciones del algoritmo cuentan las implicaciones de un muestreo *prior*, a la vez que incrementalmente actualiza α y β.



1- Parámetro de proporciones
2- Proporciones de los tópicos por documento
3- Asignación de tópicos por palabra

4- Palabra observada
5- Tópicos
6- Parámetro de Tópico

1) Generar cada tópico $\varphi_i \sim Dir(\beta)$ para $i = 1, \ldots, k$

2) Para cada documento:

primero generar las proporciones de tópicos $\theta_d \sim Dir(\alpha)$

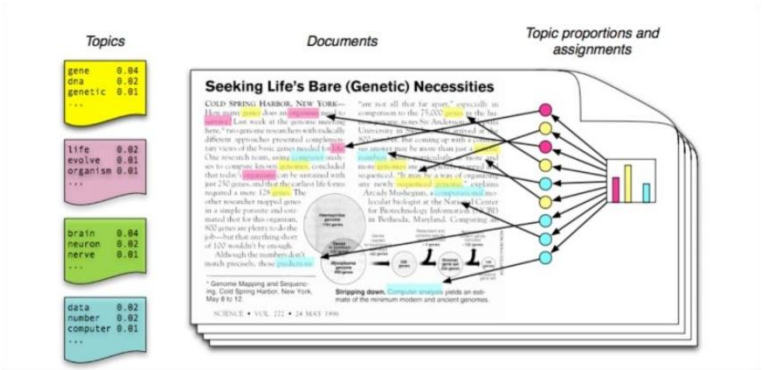
por cada palabra dentro del documento:

a) Generar $z_{d,n} \sim Multi(\theta_d)$

b) Generar $w_{d,n} \sim Multi(\varphi_{z\,d,n})$

La distribución conjunta define la probabilidad posterior $(\varphi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})$.

Resumiendo, LDA es un modelo que genera aleatoriamente valores de datos observables basándose en algunos parámetros ocultos y sigue un proceso generativo. Mientras los documentos son examinables, la estructura de tópicos, la distribución de tópicos por documento y la asignación de tópicos a los documentos, son estructuras ocultas, que es a lo que estos modelos llaman "latente". Para una colección de documentos se tiene que inferir 1) la asignación del tópico por palabra $z_{d,n}$, 2) la proporción del tópico en el documento θ_d *3)* la distribución del tópico en el corpus φ_k



Corpus de análisis

Para el desarrollo del presente trabajo se tomó un corpus de registros bibliográficos descargados de la base de datos *Scopus.* El subconjunto seleccionado 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. Esto obedece a un interés particular que estamos llevando adelante en otra investigación que tiene que ver con 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.

La razón de seleccionar Scopus como fuente obedece a que se trató de evitar sumar en esta etapa exploratoria inicial una complejidad derivada de la selección e ingesta de datos. En esta base de datos, implementar estrategias de búsqueda de cierta complejidad y descargar registros con un nivel de normalización aceptable se hace con facilidad. Por el contrario, la elección del área de Ciencias Sociales & Humanas tiene que ver con que *a priori* se considera que posee un tipo de discurso académico más ambigüo, con la dificultad que ello conlleva. Se debe tener en cuenta que este trabajo tiene como finalidad hacer una exploración metodológica, no obtener resultados concluyentes sobre la actividad científica del área.

Herramientas de software utilizadas

Se decidió utilizar el lenguaje de programación orientado a la estadística R (2017) por considerar que es una herramienta muy potente para el análisis y la visualización de datos. Posee una sintaxis intuitiva, a la vez que permite implementar nuestras propias funciones y rutinas a medida que crecen nuestras necesidades. Se utiliza el aplicativo RStudio, lo que permite contar con un entorno interactivo que nos brinda acceso al editor de código, la consola de ejecución y el visor de gráficos de manera ágil. Pero por sobre todas las cosas es abierto y su uso está extendido ampliamente en el medio científico, por lo cual se tiene acceso a librerías, llamadas comúnmente paquetes, que han sido desarrollados para fines específicos. En este trabajo se utilizaron los siguientes 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 alguna técnicas de procesamiento del lenguaje natural dentro de las cuales se encuentra la detección de n-gramas.

Procedimiento y resultados

En primer término, se realiza la búsquedas en Scopus respondiendo a la estrategia de búsqueda:

(TITLE-ABS-KEY(Argentina OR argentino OR argentinos OR Argentine OR argentinian OR argentinians) AND AFFILCOUNTRY(Argentina OR Argentine)) AND SUBJAREA(MULT OR ARTS OR BUSI OR DECI OR ECON OR PSYC OR SOCI) AND PUBYEAR > 2009 AND PUBYEAR < 2016

Se descargan los resultados en formato *bibtex* y se realiza una exploración bibliométrica general con la finalidad de obtener una primera aproximación a la estructura temática del corpus. Dicha estructura se hará perceptible desde algunos títulos de revistas y desde las palabras claves utilizadas en los registros bibliográficos.

Producción por año	
Año public.	Cant. trab.
2010	389
2011	448
2012	546
2013	535
2014	660
2015	811
TOTAL	3389

Tabla 1: Producción por año

Indicadores bibliométricos	
Tasa de crecimiento porcentual anual	15,8
Índice de colaboración	2,66
Promedio de citas por artículo	2,429
Total de fuentes distintas (revistas, libros, etc.)	1172
Keywords Indexadas diferentes(ID)	5042
Keywords de Autor (DE)	8145

Tabla 2: Indicadores bibliométricos generales

Revistas más productivas

	Fuentes	Art.
1	INTERSECCIONES EN ANTROPOLOGIA	164
2	MUNDO AGRARIO	61
3	MAGALLANIA	57
4	CHUNGARA	41
5	REVISTA ESPANOLA DE ANTROPOLOGIA AMERICANA	41
6	ARQUEOLOGIA	40
7	INTERDISCIPLINARIA	34
8	JOURNAL OF ARCHAEOLOGICAL SCIENCE	30
9	INTERCIENCIA	28
10	DESARROLLO ECONOMICO	27
11	IZQUIERDAS	26
12	ESTUDIOS MIGRATORIOS LATINOAMERICANOS	25
13	PROBLEMAS DEL DESARROLLO	25
14	ANTIPODA	24
15	ARCHAEOFAUNA	24
16	EDUCATION POLICY ANALYSIS ARCHIVES	24
17	ESTUDIOS ATACAMENOS	23
18	HISTORIA CIENCIAS SAUDE - MANGUINHOS	22
19	REVISTA ESTUDOS FEMINISTAS	22
20	ECOLOGICAL INDICATORS	19

Tabla 3: Revistas más productivas

Palabras claves más usadas

	Keywords de Autor (DE)	Artí.	Index Keywords (ID)	Art.
1	ARGENTINA	623	ARGENTINA	389
2	PATAGONIA	67	FEMALE	250
3	LATIN AMERICA	56	MALE	249
4	HUNTER-GATHERERS	54	HUMAN	182
5	LATE HOLOCENE	42	HUMANS	164
6	GENDER	36	ARTICLE	152
7	EDUCATION	35	ADULT	122
8	BUENOS AIRES	32	MIDDLE AGED	108
9	PERONISM	27	AGED	65
10	STATE	27	CHILD	62
11	ZOOARCHAEOLOGY	27	BUENOS AIRES [AR]	60
12	CHILDREN	26	YOUNG ADULT	48
13	POVERTY	25	PRIORITY JOURNAL	39
14	BRAZIL	24	ADOLESCENT	38
15	POLITICS	24	HOLOCENE	33
16	SOUTH AMERICA	24	MAJOR CLIN STUDY	32
17	ARGENTINE	23	COMPARAT STUDY	31
18	ARCHAEOLOGY	22	QUESTIONNAIRE	30
19	DEVELOPMENT	21	CONTROLLED STUDY	29
20	LITHIC TECHNOLOGY	20	PATAGONIA	29

Tabla 4: Palabras claves de autor y del sistema más usadas (unigramas)

En segundo término, se decide cual es la información relevante a los fines del objetivo perseguido. Para realizar una caracterización temática lo habitual es trabajar con los campos: título (TI), keywords de autor (DE), keywords del sistema (ID), resumen (AB) y título de la fuente (SO). Por un lado, los campos AB, TI, DE e ID aportan un volumen importante de vocabulario portador de semántica. En los tres primeros casos se trata de expresiones en lenguaje natural, y en el caso de ID son expresiones en lenguaje controlado asignadas por la propia base de datos. Por su parte, el campo SO provee información que de manera contextual puede utilizarse para

determinar temas de colecciones de documentos, aunque esto de manera derivada. Para este trabajo se decidió tomar solo el campo AB. La decisión se fundamenta en que la caracterización temática utilizando los títulos de las revistas es una forma demasiado indirecta, ya que asignar a los documentos individuales la categoría temática de la revista fuerza a una excesiva generalización. Por otro lado, el procesamiento del corpus y las técnicas que deben emplearse para trabajar con las revistas son diferentes a las que aquí se plantean. El campo de palabras claves del sistema tampoco será tomado en cuenta dado que estaríamos incorporando al corpus de análisis un lenguaje que no responde al lenguaje natural del autor y que finalmente puede aportar ruido por estar pensado con una lógica diferente. Si bien el campo título se considera adecuado, al hacer una primera revisión de los datos, se detectó en los primeros 1000 registros que un 6% de estos carecen de título en inglés. Si bien un alto porcentaje de registros trae el título en inglés y luego el título entre corchetes en su idioma original, luego de hacer un borrado global de los títulos entre corchetes se observa que hay registros que solo contienen su título en el idioma original. Esto nos pone en situación de eliminar esos títulos, dado que el tratamiento de corpus multi-idioma se vuelve bastante más complejo por el uso de ciertas herramientas lingüísticas como el stemming, pero al mismo tiempo, la decisión de eliminar esos títulos produciría una diferencia de esos registros respecto al resto ya que algunos contarían con un refuerzo de vocabulario provisto por el título y otros no. Este último aspecto también se da en las palabras claves del autor, donde un 19,5 % de los registros no las poseen. Por lo tanto, de los 3389 registros originales, nos quedamos con 3153 resúmenes para conformar el corpus de análisis (los restantes registros no tienen contenido en el campo AB).

En tercer término, para la aplicación de la técnica de agrupamiento basada en *k-means* se realiza la conversión de los registros al formato de hojas de datos (*data.frame*) de R, se seleccionan las porciones de texto para conformar el corpus de análisis restringido, se eliminan los datos faltantes y se aplica una stopword para limpiar el vocabulario no significativo. Se utilizan las funciones *termExtraction* y *conceptualStructure* que integran el paquete *bibliometrix.* Se realizan diferentes pruebas de agrupamiento variando dos parámetros: el umbral mínimo de frecuencias a considerar (grado) y el uso/no uso de stemming. El número máximo de agrupamientos utilizado es 8, que es el máximo que posibilita la función *conceptualStructure.* Se realizan 22 combinaciones distintas y se elaboran gráficos para observar la tendencia de los clusters. Se revisa la medida TSS/BSS, donde TSS es la suma total de los cuadrados

de las distancias entre cada punto y la media global, es decir, es una medida de la variación total de los datos, mientras que BSS es la suma de los cuadrados de las distancias de cada centroide de cluster con la media global. Cuanto más lejos están las medias de los grupos de la media global se supone que más discernibles son los clusters, por lo tanto, cuanto más alto sea el valor que nos da la proporción BSS/TSS mejor será.

	Grado	Stemm	Gramas	Clusters	Tamaño por nro de cluster	tot.withinss (TSS)	Betweenss (BSS)	BSS/TSS %	Almacen
1	100	No	1	3	113(1), 215(2), 28(3)	710	437,04	61,6%	88 Mb
2	100	Si	1	3	48(1), 273(2),117(3)	874	322,96	63,0%	107,9 Mb
3	75	No	1	3	166(1), 58(2), 295(3)	1036	640,19	61,8%	127,6 Mb
4	75	Si	1	3	156(1),338(2),61(3)	1108	697,15	62,9%	136,3 Mb
5	50	No	1	3	113(1), 250(2), 460(3)	1644	1022,05	62,2%	201,4 Mb
6	50	Si	1	3	471(1),236(2),101(3)	1614	992,31	61,5%	197,8 Mb
7	30	No	1	3	324(1), 206(2), 868(3)	2794	1750,02	62,6%	341,2 Mb
8	30	Si	1	2	168(1),1023(2)	2380	1558,86	34,5%	290,8 Mb
9	25	No	1	3	1009(1),254(2),403(3)	3330	2083,84	62,6%	406,3 Mb
10	25	Si	1	2	196(1),1159(2)	2708	943,72	34,8%	330,7 Mb
11	20	No	1	3	478(1),1229(2),318(3)	4048	2538,78	62,7%	493,6 Mb
12	20	Si	1	2	256(1),1729(2)	3174	1111,14	35%	387,3 Mb
13	15	No	1	3	606(1),1527(2),437(3)	5138	3214,67	62,6%	626 Mb
14	15	Si	1	3	564(1),1059(2),303(3)	3850	2421,08	62,9%	469,5 Mb
15	10	No	1	3	803(1),663(2),2083(3)	7096	4488,41	63,3%	863,9 Mb
16	10	Si	1	3	703(1),432(2),1366(3)	5000	3154,33	63,1%	609,2 Mb
17	8	No	1	3	935(1),243182),795(3)	8320	5307,33	63,8%	1012,6 Mb
18	8	Si	1	2	568(1),2289(2)	5712	2054,03	36%	695,7 Mb
19	5	No	1	3	1205(1),3488(2),1195(3)	11774	17496,74	63,7%	1,4 Gb
20	5	Si	1	3	826(1),857(2),2262(3)	7888	4963,42	62,9%	960 Mb
21	2	No	1	3	11980(1),63(2),67(3)	24218	18272,83	75,5%	2,9 Gb
22	2	Si	1	3	2113(1),5804(2),59(3)	15950	11482,41	72%	1,9 Gb

Tabla 5: Indicadores obtenidos al generar clusterización k-means con diferentes combinaciones de parámetros (22 combinaciones)

En 18 casos de los 22, el resultado arroja 3 clusters. En los 4 casos que el resultado arroja 2, se observa un bajo valor en la relación BSS/TSS. Los gráficos correspondientes a los mejores resultados (casos 21 y 22) son los siguientes:

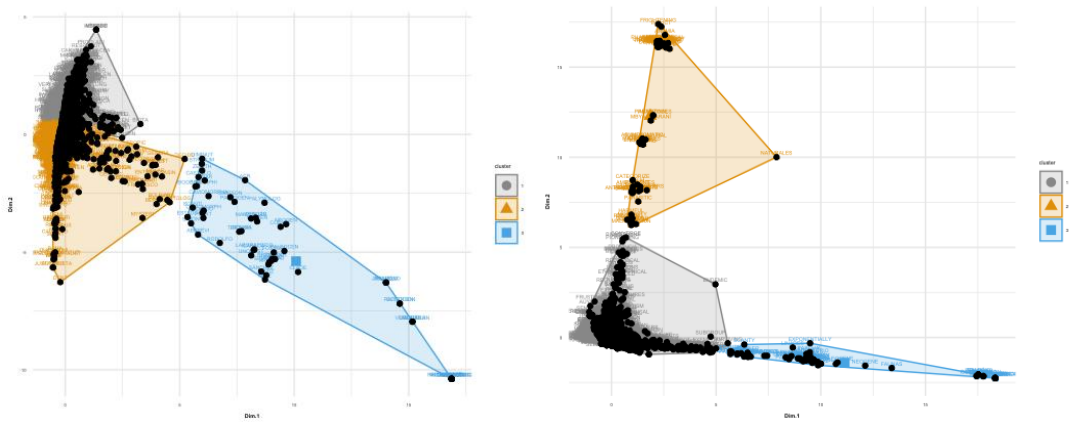


Figura 1: Clustering k-means, grado 2, con stemming

Figura 2: Clustering k-means, grado 2, sin stemming

Dado que es difícil realizar una interpretación del profuso vocabulario que queda agrupado en cada cluster, en cuarto término se exploró la riqueza que ofrece la generación de bigramas y se realizó una revisión manual del resultado que brindan los clusters. Para la generación de los n-gramas se utiliza la función *unnest_tokens* del paquete *tidytext* con apoyatura del código propuesto por Silger (2017). Se realiza la comparación de las cantidades de unigramas y bigramas que arroja el corpus según se aplique o no stemming. Asimismo se muestran las redes de co-ocurrencias de unigramas, las cuales muestran una diferencia a nivel de la densidad, aunque no así en su aspecto general.

Sin stemming		Con stemming	
Unigramas diferentes	17125	Unigramas diferentes	24672
Cant. de unigramas dif. con frecuencia >= 20	1980	Cant. de unigramas dif. con frecuencia >= 20	2463
Bigramas diferentes	188031	Bigramas diferentes	205054
Cant. de bigramas dif. con frecuencia >= 20	305	Cant. de unigramas dif. con frecuencia >= 20	254

Tabla 6: Valores totales obtenidos al contar frecuencias de n-gramas

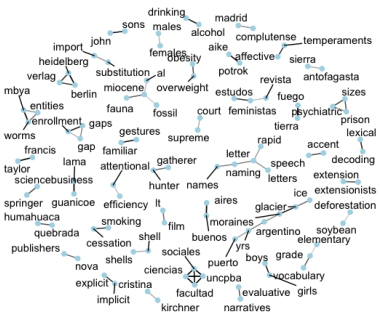


Figura3:
Red de bigramas sin stemming n=17 coeficiente Phi= 0,70

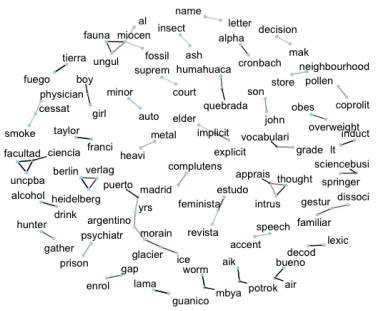


Figura4:
Red de bigramas con stemming n=17 coeficiente Phi= 0,70

Una interpretación cualitativa de los bigramas con frecuencia 20 o superior nos hace proponer los siguientes grupos:

LUGARES			ANTROPOLOGIA/ARQUEOLOGIA			SOCIOLOGIA/POLITICA			PRODUCCION/AMBIENTE		
1	buenos aires	505	11	late holocene	137	29	human rights	71	32	raw materials	70
3	latin american	198	21	archaeological sites	83	48	military dictatorship	57	83	rural areas	38
9	latin america	160	35	hunter gatherer	67	53	middle class	54	89	climate change	35
16	south america	104	42	hunter gatherers	61	59	public policies	47	97	natural resources	34
24	south american	80	45	archaeological site	58	62	public private	45	116	sea level	31
27	city buenos	72	49	archaeological record	56	101	social actors	34	126	economic growth	29
28	cordoba argentina	71	109	lithic raw	32	122	public policy	30	181	environmental conditions	23
37	argentina brazil	65	121	middle holocene	30	125	developing countries	29	187	rural development	23
38	santa cruz	65	163	early holocene	25	129	labor market	29	214	environmental problems	21
43	la plata	60	167	radiocarbon dates	25	158	social representations	26	224	regional development	21
51	american countries	54	195	human occupation	22	176	social groups	24	239	economic development	20
56	united states	52	232	archaeological evidence	20	177	socioeconomic status	24	248	rural extensionists	20
57	aires province	51	223	recovered archaeological	21	188	social security	23	TOTAL		365
64	santa fe	44	TOTAL		637	180	civil society	23	EDUCACION		
67	metropolitan area	43	CULTURA			211	catholic church	21	90	higher education	35
75	province buenos	40	76	rock art	40	225	social networks	21	93	social skills	35
79	argentina chile	38	193	cultural heritage	22	230	urban space	21	128	high school	29
85	region argentina	37	TOTAL		62	242	national identity	20	169	secondary school	25
86	countries argentina	36	SALUD			TOTAL		599	TOTAL		124
91	mendoza argentina	35	77	health care	39	GRUPOS			TIEMPO		
92	patagonia argentina	35	84	psychometric properties	37	88	young people	36	25	twentieth century	76
94	tierra del	35	133	public health	29	118	university students	31	40	nineteenth century	64
95	del fuego	34	155	mental health	26	142	sex age	28	50	pre hispanic	55
104	northwestern argentina	33	197	quality life	22	192	college students	22	213	early twentieth	21
113	aires city	31	TOTAL		153	235	children adolescents	20	TOTAL		216
120	cordoba argentina	30	METODOLOGIAS			203	spanish speaking	22	GENERICO		
123	urban areas	30	19	case study	89	235	children adolescents	20	6	ciencias sociales	176
124	del plata	29	73	case studies	40	241	general population	20	69	socio economic	43
141	plata argentina	28	96	depth interviews	34	TOTAL		199	102	economic social	33
146	mar del	27	114	comparative analysis	31				105	social political	33
156	pampean region	26	136	semi structured	29	30	sciencebusiness media	71	115	science technology	31
160	catamarca argentina	25	153	data collected	26	31	springer sciencebusiness	71	117	social sciences	31
162	de buenos	25	238	discourse analysis	20	150	science publishers	27	132	political social	29
165	los andes	25	253	structured interviews	20	174	publishers rights	24	151	social cultural	27
171	area buenos	24	TOTAL		289				157	political economic	26
172	de humahuaca	24							201	social economic	22
186	rio negro	23							202	socio cultural	22
194	entre rios	22							TOTAL		473
212	cruz argentina	21									
216	jujuy argentina	21									
217	lama guanicoe	21									
226	southern patagonia	21									
237	cruz province	20									
244	northwest argentina	20									
246	province cordoba	20									
		2460									

Figura 5 : Agrupamiento manual de los bigramas. Ocurrencia mínima n=20

Una interpretación de los clusters originados con *k-means* a la luz de los agrupamientos anteriores, nos permite comprobar la pertinencia de los resultados que nos arroja la técnica. En la tabla siguiente se muestra una selección de términos provenientes de cada cluster, que de manera cualitativa se identifican con alguno de los agrupamientos de la figura 5. Dentro del color correspondiente, el tono intenso indica, además, que el término forma parte de un bigrama de la figura 5. En el caso en el que el término perteneciera a bigramas de dos agrupamientos distintos, se optó por indicar el segundo color a la derecha de la columna. Esto permite observar que la conformación de los 3 clusters responde a los agrupamientos denominados Sociología/Política (cluster 1), Salud y Educación (cluster 2) y Antropología/Arqueología (cluster 3). Así como también se puede observar el nivel de "contaminación" con términos de otros agrupamientos. El porcentaje al final de cada tabla muestra la proporción que significa esta selección respecto al total de términos del cluster.

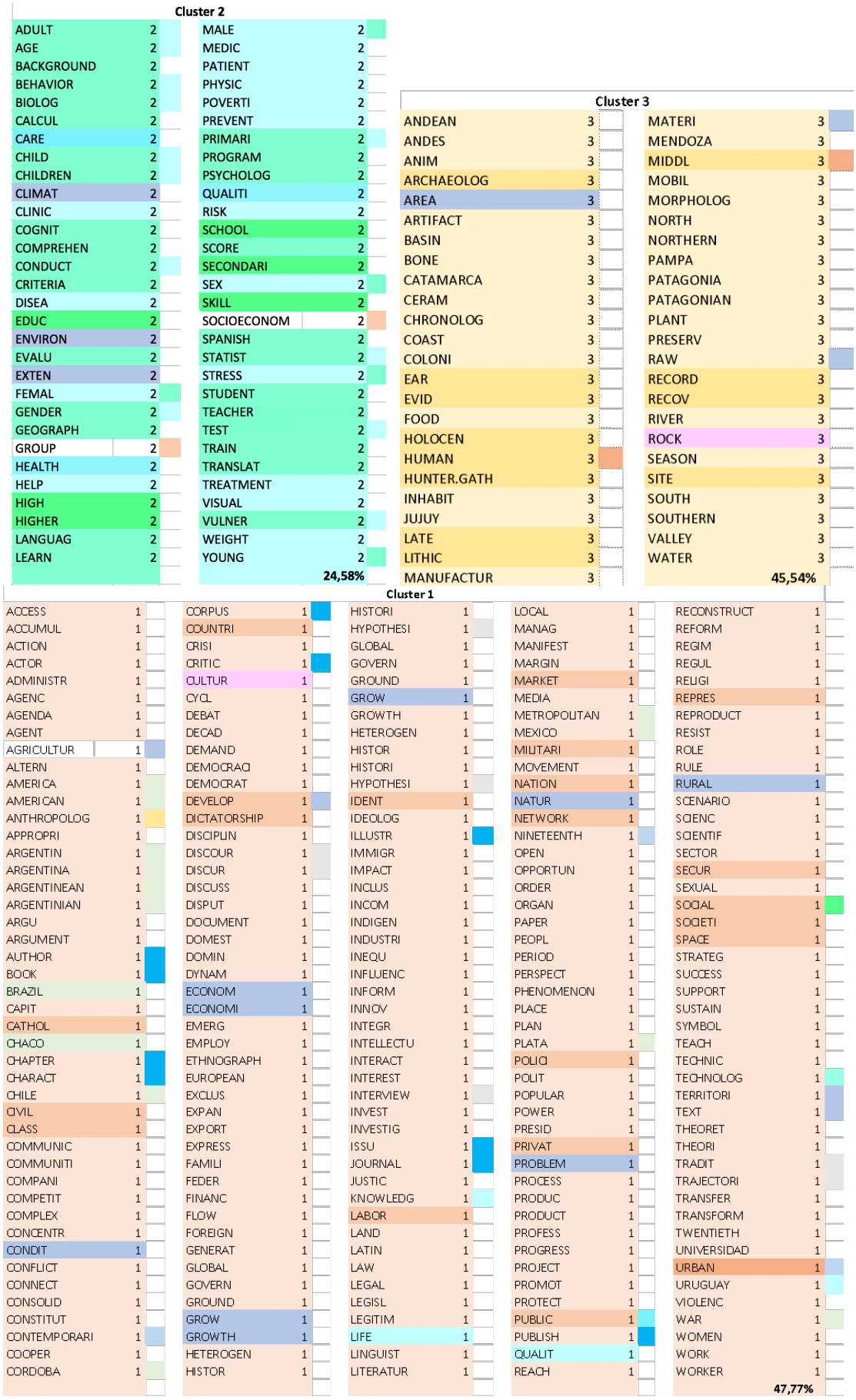


Fig 4 : Agrupamiento manual de los bigramas. Ocurrencia mínima n=20

En quinto término, para la aplicación de la técnica de modelado de tópicos LDA, se utilizó el paquete *topicmodels* con apoyatura del código propuesto por Contador Pachón (2016). Se aplicó utilizando stemming, definiendo 6 tópicos y utilizando dos métodos diferentes de ajuste, el *Variational Expectation Maximization* (LDA-VEM) y *Correlated Topic Model* (CTM-VEM). El primero asume que los tópicos no correlacionan y en el segundo la correlación está permitida (Blei & Lafferty, 2007). Se muestran los 10 primeros términos en cada caso y se realiza una interpretación similar a la realizada con los clusters.

a) *LDA-VEM*

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
bodi	patient	women	argentina	holocen	educ
indigen	treatment	gender	polit	record	school
death	use	age	militari	region	student
human	result	group	dictatorship	date	univers
argentina	medic	sexual	articl	chang	teacher
popul	clinic	femal	state	lake	learn
studi	studi	sex	religi	southern	teacher
individu	cost	men	institut	human	argentina
age	argentina	use	church	glacier	studi
collect	hospit	differ	cathol	level	research

Grupos

b) *CTM-VEM*

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
region	polit	worker	urban	social	use
knowledg	cultur	argentina	citi	practic	differ
process	articl	product	area	famili	site
read	research	worker	air	argentina	area
word	nation	articl	bueno	work	studi
differ	argentina	place	social	rural	shell
comprehens	studi	social	space	legal	sampl
develop	public	wine	rural	play	concentr
state	univers	experi	hous	right	argentina
paper	polici	church	chang	public	analysi

Conclusión

- 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.
- Las dos técnicas utilizadas en este estudio, clustering, basado en k-means y el modelado de tópicos basado en Latent Dirichlet Allocation (LDA) necesitan que se les dé un valor de entrada (la cantidad de clusters o la cantidad de tópicos). El caso del k-means, tal como se lo aplicó aquí, no parece presentar problemas para los tamaños de conjuntos de registros bibliográficos referenciales extraídos de bases de datos. La herramienta por nosotros utilizada permite un máximo de 8 y nuestros resultados muestran que no forma más de 3 clusters aún con un mínimo

de frecuencias de 2. En la técnica de LDA hay que determinar el número de tópicos, y esto es más problemático.

- La técnica de clustering resultó adecuada para tener un panorama general de la cantidad de grupos que conforman temáticamente el corpus. Los bigramas aportan más contenido informativo, no encontrando, salvo por la densidad, mayor diferencia en las redes con o sin stemming. El resultado de la técnica LDA-VEM resulta más razonable que CTM-VEM.
- El uso de stopwords estandarizadas no resulta suficiente para eliminar el vocabulario que produce ruido.
- La semántica presente en los títulos, palabras clave y resúmenes de los registros bibliográficos podría ser limitada para el objetivo propuesto. Debiéramos realizar pruebas con textos completos para explorar si los resultados son o no similares.
- El software utilizado y los paquetes estadísticos seleccionados resultaron sólidos y fiables.
- Para el caso analizado que involucra el corpus documental de la producción de Ciencias Sociales de Argentina y sobre Argentina la calidad de la información que brinda Scopus no es 100% aprovechable. Deberíamos realizar pruebas con otros corpus de dominios temáticos y países diferentes para determinar si la fuente presenta similares limitaciones.

Referencias

Blei, D.M., Ng, A.Y. y Jordan, M.I. (2003). Latent dirichlet allocation. *J. Mach. Learn,* 993-1022.

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. Recuperado de http://http://r-es.org/7jornadasR/ponencias/sergio_contador_pachon.pdf

Deerwester, S. (1988). *Improving Information Retrieval with Latent Semantic Indexing.* Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, pp. 36–40.

Hofmann, T., (1999). *Probabilistic latent semantic indexing*, in Proceedings of the 22nd annual International ACM SIGIR conference on Research and development in information retrieval. ACM: Berkeley, California, United States. pp. 50-57.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297. Berkeley, CA: University of California Press.

R Core Team (2017). *R: A language and environment for statistical computing*. R. Foundation for Statistical Computing, Vienna, Austria. Recuperado de https://www.R-project.org/.

Salton, G. y McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Co.

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

Autores

Claudia M. González

Licenciada en Bibliotecología y Documentación por la Universidad Nacional de La Plata, 2008 y Máster en Documentación Digital por la UniversitatPompeuFabra, Barcelona, 2013. Se encuentra actualmente realizando el doctorado TIC en la Universidad de Granada. Profesora adjunta de la cátedra Tratamiento Automático de la Información I de la carrera de Bibliotecología y del Taller de Trabajo Final de la Especialización en Gestión de Información Científica y Tecnológica, ambos de la Fac. de Humanidades y Cs de la Educación de la UNLP, donde además coordina el Campus Virtual FaHCE. Es profesional principal del CONICET y docente investigadora FaHCE-IdIHCS.

Sebastián Varela

Licenciado en Sociología por la UNLP, Máster en Metodología de la Investigación Social por la Università di Bologna/UNTREF (2006) y Doctor en Ciencias Sociales por la UBA (2009). En la actualidad es Profesor-investigador del

Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCS) y del Departamento de Sociología (FaHCE-UNLP). Su actividad de investigación y producción académica se orientan hacia temas de educación superior, estadística, y métodos de investigación social.

Sandra Miguel

Licenciada en Bibliotecología y Documentación por la Universidad Nacional de La Plata (UNLP), Argentina (1995). Doctora en Documentación por la Universidad de Granada, España (2008).Directora del Departamento de Bibliotecología de la Facultad de Humanidades y Ciencias de la Educación de la UNLP. Docente de la Licenciatura y Profesorado en Bibliotecología y Ciencia de la Información de la FAHCE-UNLP, y en carreras de posgrado de la propia institución y de otras instituciones argentinas y extranjeras. Directora de la Especialización en Gestión de Información Científica y Tecnológica de la FAHCE-UNLP (en acreditación). Investigadora del Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCs), UNLP-CONICET.

Directora de proyectos de investigación acreditados por instituciones del sistema científico y tecnológico argentino. Se especializa en estudios de la comunicación científica, bibliometría y acceso abierto.