ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Proximidad documental en repositorios académicos

Exploración intelectual de colecciones mediante análisis léxico

Chris Moreno

Universidad Nacional de San Luis. Facultad de Psicología. Instituto de Ciencias Sociales y Computacionales, Argentina | morenochristian@outlook.com / https://ror.org/oomczdx43 / https://orcid.org/0009-0003-5267-2175

Resumen

Se presenta un análisis lexicométrico de 121 resúmenes de tesis doctorales en psicología obtenidos de repositorios de universidades nacionales de Argentina. El objetivo fue observar si la distribución del vocabulario permite relacionar documentos sin usar descriptores temáticos ni citaciones. Los documentos se organizaron en una matriz término-documento. La asociación entre documentos y vocabulario resultó significativa y de intensidad moderada-alta (χ^2 = 530.239,11; gl = 353.760; V de Cramér = 0,435). El Análisis Factorial de Correspondencias permitió construir un espacio geométrico para localizar documentos próximos y distantes mediante distancia chi-cuadrado. Así se identificaron trabajos relacionados a partir de perfiles de distribución léxica. Los documentos cercanos tienden a compartir repertorios terminológicos asociados a problemas, enfoques teóricos o procedimientos metodológicos. Este hallazgo aborda una dificultad operativa de los repositorios académicos: cómo clasificar, agrupar y recorrer colecciones más allá del acceso individual a los registros. La colección funciona, además de archivo, como una red interna de relaciones que puede orientar la navegación semántica y la identificación de antecedentes.

Palabras clave

Repositorios institucionales
Organización del conocimiento
Lexicometría
Análisis Factorial de Correspondencias
Navegación semántica

Document proximity in academic repositories. Intellectual exploration of collections through lexical analysis

Abstract

A lexicometric analysis was conducted on 121 doctoral thesis abstracts in Psychology retrieved from repositories of Argentine public universities. The aim was to examine whether vocabulary distribution can relate documents without relying on thematic descriptors or citation links. The texts were organized into a term-document matrix. The association between documents and vocabulary was significant and of moderate-to-high magnitude (χ^2 = 530,239.11; df = 353,760; Cramér's V = 0.435). Correspondence Analysis enabled the construction of a geometric space in which documents were located according to χ^2 distance. This made it possible to identify both close and distant works based on lexical distribution profiles. Nearby documents tended to

Keywords

Institutional repositories
Knowledge organization
Lexicometrics
Correspondence Analysis
Semantic navigation

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

share terminological repertoires associated with research problems, theoretical approaches or methodological procedures. These findings address an operational difficulty of academic repositories: how to classify, group and explore collections beyond individual record retrieval. The collection thus functions not only as an archive, but also as an internal network of relations that may support semantic navigation and the identification of prior work.

Artículo recibido: 19-02-2026. Aceptado: 26-05-2026.

Introducción

Un gran avance ha sido el desarrollo de los repositorios institucionales, debido a que han logrado resolver problemas clásicos de acceso y preservación de la información científica. En Argentina, la Ley 26.899 de Repositorios Digitales Institucionales (Argentina, 2013) estableció la obligatoriedad de disponer repositorios de acceso abierto en las instituciones del Sistema Nacional de Ciencia y Tecnología. Como consecuencia, las tesis doctorales, entre otros tipos de documentos académicos, comenzaron a preservarse y publicarse sistemáticamente en repositorios universitarios, lo que generó colecciones documentales accesibles y relativamente comparables para su análisis. No obstante, en la actualidad se ofrecen escasas opciones para orientar la exploración intelectual dentro de la literatura que reúnen estos sistemas. En la organización del conocimiento, la recuperación no debería limitarse al acceso a registros individuales, sino contemplar también la posibilidad de situarlos dentro de un dominio conceptual (Hjørland, 2016).

Los sistemas de recuperación de la información representan los archivos mediante términos indexados o vectores de palabras, y estiman su similitud a partir de coincidencias ponderadas entre sus vocabularios (Baeza-Yates y Ribeiro-Neto, 1999; Manning, Raghavan y Schütze, 2008). Estos procedimientos permiten responder a consultas explícitas. No obstante, resultan menos adecuados para la exploración intelectual, donde el investigador no busca únicamente trabajos que contengan los mismos términos, sino aquellos conceptualmente próximos. La búsqueda académica suele desarrollarse como un proceso iterativo en el que el usuario identifica nuevos documentos a partir de otros previamente encontrados y reformula progresivamente su consulta (Bates, 1989; Marchionini, 2006). En este contexto, la dificultad principal radica menos en la localización que en la posibilidad de situar los documentos dentro de la literatura disponible. Cuando la búsqueda, más que responder a una consulta exacta, avanza como un proceso de indagación progresiva, el investigador necesita reconocer trabajos conceptual o metodológicamente cercanos aun cuando no compartan referencias ni descriptores explícitos.

La bibliometría y la indización temática han permitido establecer relaciones entre documentos a través de citaciones, referencias o descriptores. Entre esas formas de vinculación queda un aspecto menos explorado: la organización interna de una colección derivada del propio lenguaje de los textos. Los documentos académicos presentan regularidades terminológicas y retóricas que no siempre coinciden con las categorías asignadas por los sistemas de metadatos. Así, la distribución diferencial del vocabulario puede revelar afinidades discursivas entre trabajos. Diversos enfoques han señalado que la estructura conceptual de un campo puede inferirse a partir de regularidades terminológicas compartidas. En particular, el análisis de co-palabras propuso que la co-ocurrencia sistemática de términos permite reconstruir redes de problemas de investigación aun cuando no existan vínculos bibliográficos explícitos entre los trabajos (Callon, Courtial y Laville, 1991). Asimismo, la cartografía de la ciencia ha mostrado que la cercanía entre documentos puede aproximarse

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

mediante mapas basados tanto en citaciones como en similitud de contenido (Callon, Courtial y Laville, 1991; Noyons, 2012). En el contexto de la ciencia abierta, esta cuestión adquiere otra relevancia: el objetivo excede el acceso e incluye facilitar el descubrimiento y la reutilización de la información. En este sentido el Análisis Factorial de Correspondencias (AFC) es una herramienta útil ya que permite representar simultáneamente documentos y términos a partir de sus perfiles de frecuencia relativa (Benzécri, 1973; Lebart, Salem y Berry, 1998; Greenacre, 2017). La proximidad en este espacio no implica necesariamente tratar el mismo tema, sino compartir formas comparables de construcción del discurso académico.

En este trabajo se examina la organización de resúmenes doctorales recuperados de repositorios institucionales de una misma disciplina y se evalúa si sus relaciones pueden observarse a partir de la estructura léxica del corpus. Se plantea que el repositorio puede considerarse, además de un sistema de acceso, un espacio estructurado de conocimiento cuya organización emerge del propio lenguaje académico. Para lograr estos objetivos, primero se evalúa la asociación entre términos y documentos mediante la prueba chi-cuadrado de Pearson aplicada a la matriz término-documento, complementada con su tamaño de efecto (V de Cramér). Luego se aplica el AFC como técnica multivariante textual que elabora una representación geométrica de los perfiles léxicos. Esta representación posibilita observar visualmente la organización del vocabulario y de los documentos dentro del corpus. A partir del mismo espacio factorial se calculan distancias que permiten estimar la proximidad entre documentos y seleccionar casos de relación documental. Así, la contribución del estudio consiste en mostrar que las relaciones entre resúmenes pueden observarse directamente en la organización del vocabulario, permitiendo recorrer los registros como un espacio estructurado para la exploración académica.

Estado del arte

Relaciones documentales y análisis léxico

Las relaciones entre documentos tienen una larga trayectoria de investigación a través de vínculos bibliográficos, coincidencias terminológicas o patrones de uso. Una primera tradición corresponde a la bibliometría, centrada en los vínculos de citación. A partir de los trabajos pioneros sobre redes de publicaciones científicas, se propuso que la estructura de la literatura puede reconstruirse mediante las referencias explícitas entre artículos (Price, 1965). Posteriormente, la co-citación permitió identificar especialidades científicas al observar que ciertos trabajos son citados conjuntamente (Small, 1973), mientras que el acoplamiento bibliográfico vinculó trabajos que comparten fuentes de referencia.

La bibliometría relacional permitió describir la estructura social de la comunicación científica mediante redes de citación; sin embargo, investigaciones posteriores señalaron que estas relaciones no agotan la organización conceptual de la literatura. Los mapas basados en co-citación y acoplamiento bibliográfico reconstruyen comunidades de investigación, pero dependen de prácticas editoriales, visibilidad y tradiciones de referencia. Por ello, se desarrollaron aproximaciones orientadas a captar la proximidad intelectual entre trabajos a partir de su contenido, integradas dentro del denominado *science mapping*. Estos enfoques conciben la literatura científica como un espacio estructurado donde la posición de los ítems puede estimarse mediante similitudes semánticas o terminológicas (Van Raan, 2005; Börner, Chen y Boyack, 2003). Por ende, la relación entre documentos excede la existencia de referencias compartidas pues también puede derivarse de regularidades terminológicas observables en los textos.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Un segundo grupo de aproximaciones se desarrolló en el ámbito de la recuperación de información. Los modelos clásicos usan representaciones mediante términos indexados y han calculado su similitud a partir de coincidencias ponderadas (Salton y McGill, 1983). Este enfoque permitió automatizar la búsqueda documental, limitada y orientada solo a responder consultas explícitas. En este marco, la relación depende de categorías temáticas asignadas o de coincidencias literales de términos, lo que podría no capturar afinidades en la formulación de los problemas investigativos.

Por otra parte, los estudios sobre comportamiento informativo introdujeron otra perspectiva al observar que la actividad académica no siempre responde a una consulta previamente definida. Bates (1989) describió la búsqueda como un proceso de *berrypicking*, en el cual, el investigador avanza mediante desplazamientos sucesivos, mientras que Marchionini (1995) caracterizó estos procesos como búsqueda exploratoria. Sin embargo, aunque estos estudios describen cómo los usuarios recorren la literatura, los sistemas de información continúan apoyándose principalmente en descriptores temáticos, citas y referencias.

Mientras los mapas de citación describen el aspecto social de la comunicación científica, las aproximaciones basadas en contenido buscan captar su organización conceptual.

La estadística textual ofrece una vía particular: en lugar de representar los trabajos por enlaces bibliográficos o coincidencias literales, modeliza sus perfiles de frecuencia relativa. El AFC sitúa simultáneamente documentos y términos en un mismo espacio geométrico, donde la distancia entre perfiles expresa semejanzas de uso del lenguaje (Benzécri, 1973; Lebart y Salem, 1994). Esta representación puede interpretarse como una forma de cartografía documental basada en regularidades discursivas y no en vínculos de referencia explícitos.

Estudios previos sobre similitud documento-documento en s*cience mapping* mostraron que la proximidad puede estimarse desde el contenido textual, además de las citaciones (Ahlgren y Colliander, 2009). Asimismo, el análisis de correspondencias se ha utilizado para visualizar corpus como espacios de relación entre documentos y rasgos lingüísticos; en esos usos, la visualización cumple una función exploratoria para observar relaciones entre documentos y vocabulario, antes que una clasificación cerrada (Petrović et al., 2009).

Las tradiciones mencionadas describen la literatura científica desde perspectivas complementarias: las citaciones reconstruyen la estructura social de la comunicación científica, la recuperación de información organiza contenidos a partir de términos o descriptores, y los estudios de comportamiento informativo describen los recorridos de búsqueda y lectura. El análisis léxico introduce un nivel intermedio: permite observar cómo los textos distribuyen su vocabulario dentro de una colección y cómo esa distribución aproxima o separa documentos. En ese sentido, dos resúmenes pueden aproximarse por la distribución relativa de su vocabulario, por la recurrencia de ciertos procedimientos o por modos semejantes de formular un problema, aun cuando no compartan exactamente los mismos términos principales.

En repositorios institucionales, esta capa de lectura cumple una función complementaria. No reemplaza la descripción bibliográfica ni la clasificación temática, pero permite explorar relaciones que emergen del propio corpus, especialmente cuando los metadatos disponibles no bastan para orientar una navegación intelectual más fina. Desde esta lógica, el análisis léxico se entiende aquí como una estrategia de organización y exploración documental, no como una taxonomía disciplinar cerrada.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Metodología

Para analizar la distribución del vocabulario se construyó una base de datos conformada por 121 resúmenes en español de tesis doctorales en psicología, recuperados de repositorios institucionales de tres universidades nacionales argentinas: Universidad Nacional de La Plata (n = 74), Universidad Nacional de Mar del Plata (n = 24) y Universidad Nacional de Córdoba (n = 23). La selección de estas instituciones respondió a una delimitación que no buscó representar exhaustivamente la producción doctoral argentina en psicología, sino construir un corpus comparable de registros de metadatos completos disponibles, identificables y procesables en repositorios académicos. Los resúmenes de tesis se tomaron como unidad de análisis porque constituyen un género académico relativamente estandarizado, con estructuras recurrentes orientadas a la formulación del problema, la descripción metodológica y la exposición de resultados (Swales, 1990; Hyland, 2000).

El procedimiento analítico se desarrolló en seis momentos: recolección de datos; pretratamiento textual y elaboración de la matriz término-documento (DTM); evaluación de la asociación entre términos y documentos mediante la prueba chi-cuadrado de Pearson y V de Cramér; aplicación del AFC y definición del subespacio factorial de 30 dimensiones; cálculo de distancias entre documentos y; visualización de proximidades derivadas del AFC.

El preprocesamiento textual se construyó en tres etapas: a) normalización formal de los textos, mediante conversión a minúsculas, eliminación de acentos, signos de puntuación, caracteres no alfabéticos y palabras vacías (*stop words*), que incluyeron artículos, preposiciones, conjunciones, pronombres y otras palabras funcionales; b) normalización léxica selectiva de variantes ortográficas, flexivas o de género y número cuando remitían a una misma unidad interpretativa, por ejemplo, psicoanalítica, psicoanalítico y sus plurales y; c) uso de filtros de frecuencia, reteniendo términos presentes en al menos dos resúmenes y en no más del 92 % de los documentos. No se aplicó una lematización automática general, con el fin de conservar variaciones terminológicas potencialmente relevantes en psicología.

En la DTM, cada fila representa un resumen doctoral y cada columna un término del vocabulario procesado. En la expresión siguiente, n_{ij} representa la frecuencia del término j en el documento i; I corresponde al número de documentos y J al número de términos retenidos:

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

La asociación entre documentos y términos se evaluó mediante la prueba chi-cuadrado de Pearson aplicada a la matriz término-documento. Esta prueba permitió examinar si la distribución del vocabulario varía de manera sistemática entre los resúmenes. El tamaño de efecto V de Cramér se calculó con el fin de valorar la intensidad de la asociación entre términos y documentos. En esta fórmula, n es el total de ocurrencias de la matriz, I el número de documentos y J el número de términos:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(I-1, J-1)}}$$

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Una vez aplicado el AFC, los documentos quedaron representados como perfiles relativos de vocabulario. La proximidad entre dos resúmenes se estimó mediante la distancia chi-cuadrado entre perfiles, entendida como una medida geométrica de semejanza o diferencia en la distribución relativa de términos. A diferencia de la prueba chi-cuadrado anterior, aquí no funciona como contraste de hipótesis, sino como medida geométrica de proximidad en el espacio factorial, el cual quedó compuesto por 120 dimensiones posibles. Los primeros ejes del AFC facilitan la interpretación visual de las principales oposiciones de una tabla; sin embargo, en aplicaciones textuales también se han utilizado subespacios más amplios para calcular proximidades entre documentos. Morin (2006), por ejemplo, conservó 30 dimensiones en un análisis de corpus. En este estudio, las distancias entre resúmenes se calcularon en un subespacio también de 30 dimensiones, dado que en tablas léxicas la inercia tiende a distribuirse entre numerosos ejes por la alta dimensionalidad y esparsidad del vocabulario (Lebart, Salem y Berry, 1998; Murtagh, 2005). La estabilidad del subespacio se evaluó mediante correlaciones de Spearman entre la matriz principal de 30 dimensiones y matrices alternativas de 20, 40, 60 y 120 dimensiones. Las soluciones de 20 y 40 dimensiones mantuvieron alta correspondencia ($\rho = 0{,}90$ y $\rho = 0{,}96$), mientras que la asociación disminuyó al ampliar el cálculo a 60 y 120 dimensiones ($\rho = 0{,}84$ y $\rho = 0{,}51$).

Las visualizaciones en red incluidas en los resultados se construyeron a partir de las coordenadas factoriales y de las distancias calculadas en el subespacio de 30 dimensiones como apoyo gráfico para leer proximidades entre términos y documentos.

Se empleó el entorno de programación Python en su versión 3. Las librerías utilizadas fueron spaCy, NLTK, scikit-learn y prince. Para la manipulación de datos y figuras se utilizaron pandas, NumPy y Matplotlib.

La Tabla 1 sintetiza los conceptos operativos utilizados y la forma en que se los interpreta en este estudio.

Tabla 1. Conceptos operativos utilizados en el estudio

Concepto	Definición operativa	Explicación en el estudio
Matriz término-documento	Documentos por términos con frecuencias	Representa el corpus como colección analizable
Prueba chi-cuadrado de Pearson	Evalúa la asociación entre términos y documentos	Permite observar si el vocabulario varía de manera sistemática entre resúmenes
V de Cramér	Tamaño de efecto asociado al chi-cuadrado	Resume la intensidad de esa asociación
AFC	Representa filas y columnas en un espacio factorial	Ubica documentos y vocabulario según sus perfiles léxicos
Distancia chi-cuadrado	Diferencia entre perfiles de frecuencia relativa	Permite estimar proximidad textual entre documentos
Inercia	Variabilidad respecto del modelo de independencia	Expresa la dispersión léxica de la colección
Eje factorial	Dimensión derivada de la descomposición de la tabla	Resume una tendencia de organización léxica o discursiva

Fuente: elaboración propia basada en Benzécri (1973), Greenacre (1984), Lebart y Salem (1994), Husson, Lê y Pagès (2017) y Petrović et al. (2009).

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Resultados

Descripción del conjunto de datos

El corpus de 121 resúmenes quedó representado en una DTM compuesta por 2.949 términos únicos y 23.326 ocurrencias, entendidas como apariciones totales de esos términos en el conjunto del corpus. Estas cifras corresponden al vocabulario retenido después del preprocesamiento textual. En promedio, cada resumen reunió 192,78 ocurrencias (DE = 159,90; mediana = 165; RIC = 102). Además, los documentos presentaron una media de 128,39 términos distintos por resumen (DE = 74,96; mediana = 118). La prueba chi-cuadrado de Pearson mostró una asociación significativa entre documentos y términos, $\chi^2(353.760) = 530.239{,}11$, $p < .001$, con un tamaño de efecto moderado-alto, V de Cramér = 0,435. A partir de esa variación, el AFC organizó el léxico en un espacio de 120 dimensiones factoriales. Estos ejes resumen la relación entre documentos y vocabulario, y permiten ubicar los resúmenes según sus perfiles léxicos.

La Figura 1 muestra la distribución de la inercia por dimensión factorial, mientras que la Tabla 2 resume los valores de las primeras dimensiones y del subespacio de cálculo. Las figuras siguientes trasladan ese espacio factorial a proximidades entre términos y documentos.

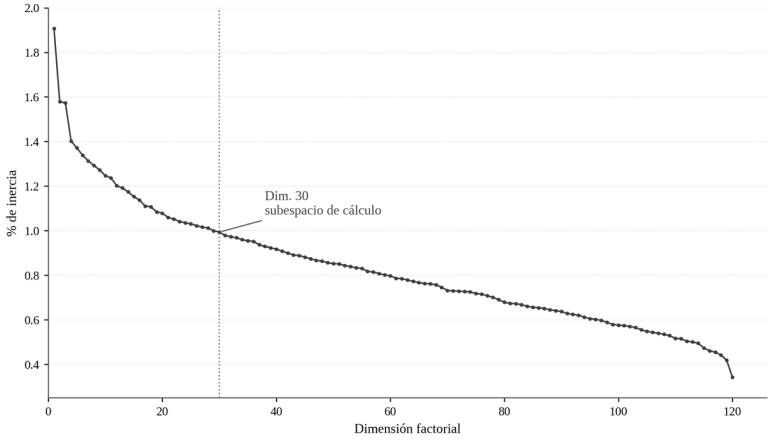


Figura 1. Distribución de la inercia en las 120 dimensiones del AFC

Fuente: elaboración propia.

Tabla 2. Inercia explicada por dimensión y subespacio de cálculo

Dimensión o conjunto	Inercia (%)	Inercia acumulada (%)
1	1,91	1,91
2	1,58	3,49
3	1,57	5,06
1–30	—	36,04

Fuente: elaboración propia.

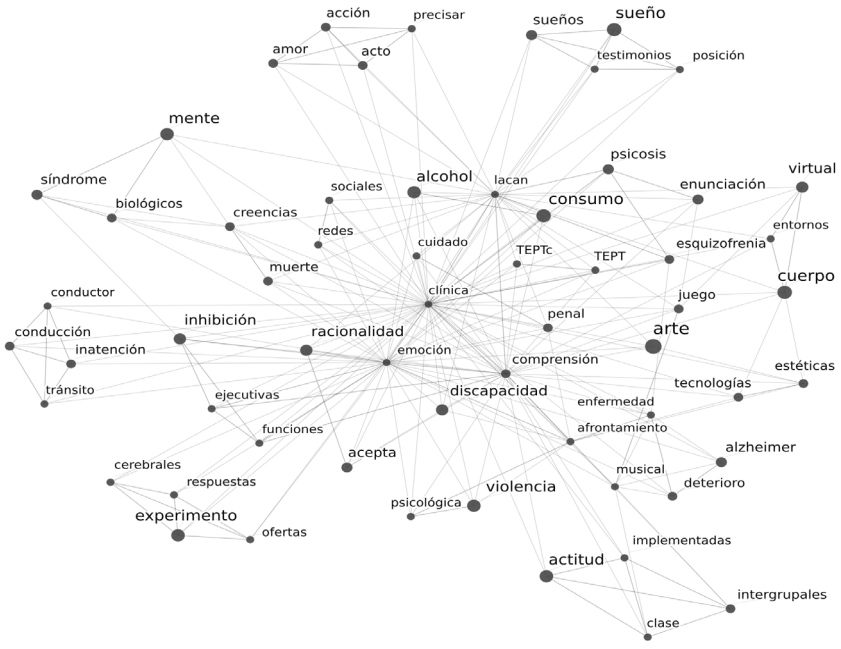
Como se observa en la Figura 1 y en la Tabla 2, la primera dimensión concentró el 1,91 % de la inercia, mientras que la segunda y la tercera reunieron valores similares, de 1,58 % y 1,57 %. En conjunto, las tres primeras dimensiones acumularon el 5,06 % y el subespacio de 30 dimensiones reunió el 36,04 %. La distribución presenta

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

una disminución progresiva, sin un corte abrupto, consistente con la dispersión de la variabilidad léxica en múltiples ejes.

Las primeras dimensiones tienen un valor descriptivo, ya que permiten reconocer las oposiciones léxicas iniciales del AFC. Sin embargo, no se utilizaron como base exclusiva para estimar la proximidad documental; las distancias entre resúmenes se calcularon sobre el subespacio factorial de 30 dimensiones.

Figura 2. Red léxica de términos contribuyentes en 30 dimensiones



Fuente: elaboración propia.

La Figura 2 muestra la visualización derivada del AFC. Se muestran 60 términos seleccionados por su contribución. Los enlaces se trazaron entre términos próximos, usando como criterio gráfico los cinco más cercanos de cada unidad léxica. La red quedó compuesta por 60 nodos y 230 enlaces. Esta imagen permite reconocer repertorios diferenciados, entre ellos una zona vinculada con clínica psicoanalítica (acto, Lacan, psicosis, sueño, amor), otra asociada a funciones ejecutivas y procesos cognitivos (inhibición, inatención, control, ejecutivas), un conjunto referido a consumo y conductas asociadas, un sector neuropsicológico (Alzheimer, deterioro, mente, síndrome), una zona relacionada con trauma y estrés (trastorno por estrés postraumático (TEPT), trastorno de estrés postraumático complejo (TEPTc), violencia, afrontamiento) y un repertorio ligado a corporalidad y tecnologías (virtual, cuerpo, arte, tecnologías).

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

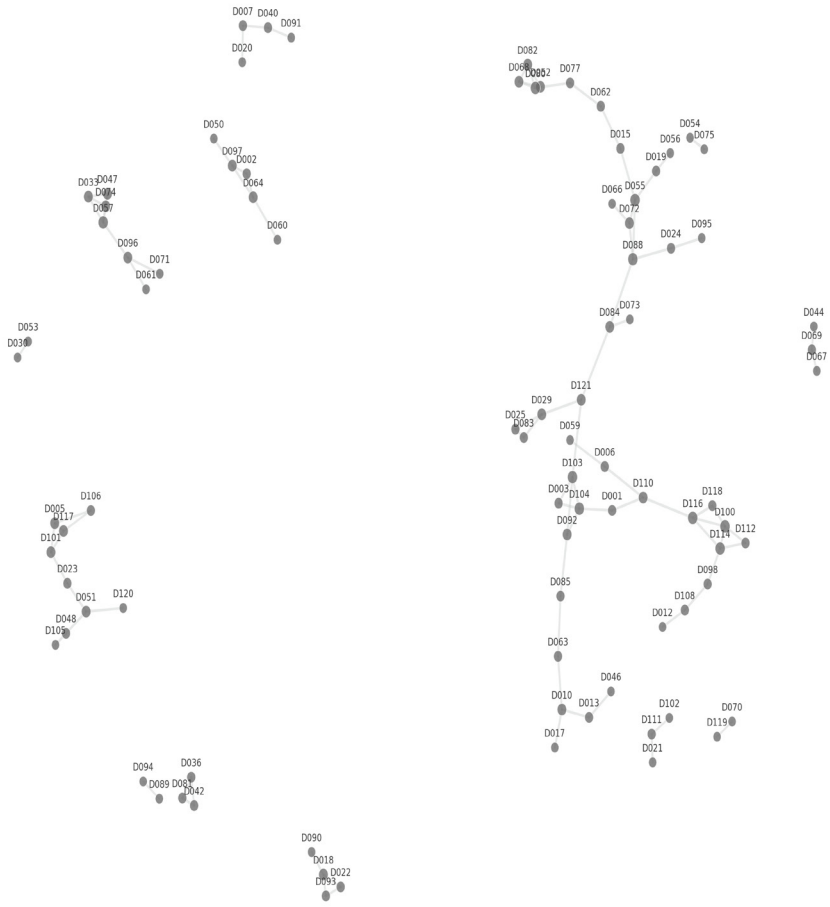


Figura 3. Red de proximidad documental en 30 dimensiones

Fuente: elaboración propia.

La Figura 3 corresponde a la visualización derivada del AFC. Cada nodo representa un resumen doctoral. Se muestran relaciones recíprocas entre documentos próximos. La red quedó compuesta por 89 nodos visualizados y 94 enlaces. Revela cómo se conectan los resúmenes entre sí a partir de sus perfiles léxicos y de las distancias calculadas en el AFC, en la red, los documentos conectados tienden a compartir un uso más parecido del vocabulario. Se observan cadenas con varios enlaces y nodos con pocas relaciones. Un caso visible corresponde al conjunto D110-D116-D118-D100-D114-D112-D098-D108-D012, que reúne tesis asociadas a funciones ejecutivas, inhibición, memoria de trabajo, flexibilidad cognitiva y procesos de control. También se reconocen otros segmentos de proximidad: D033-D047-D057-D074-D096-D061, vinculado con comprensión de textos, lectura, escritura y evaluación educativa; D007-D020-D035-D037-D040-D091, asociado a formulaciones psicoanalíticas y discusiones teórico-clínicas; y D002-D050-D064-D097, relacionado con violencia, familia y dinámicas subjetivas.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Exploración de documentos

Tabla 3. Documentos próximos en casos seleccionados

Documento de referencia	Documento próximo	Distancia chi-cuadrado
D104	D103	0,955
D104	D003	1,499
D055	D072	0,931
D055	D088	1,269
D089	D094	3,461
D089	D024	5,328

Fuente: elaboración propia.

En la Tabla 3 se presentan ejemplos de documentos próximos a partir de documentos de referencia seleccionados. D104 se vincula con trabajos próximos sobre intervención, evaluación y programas psicosociales. D055 se asocia con registros sobre prácticas profesionales y promoción de la salud. En el caso de D089, relacionado con maltrato infantil, depresión y TEPT, el documento más cercano corresponde a otro trabajo centrado en TEPT, aunque las distancias posteriores aumentan.

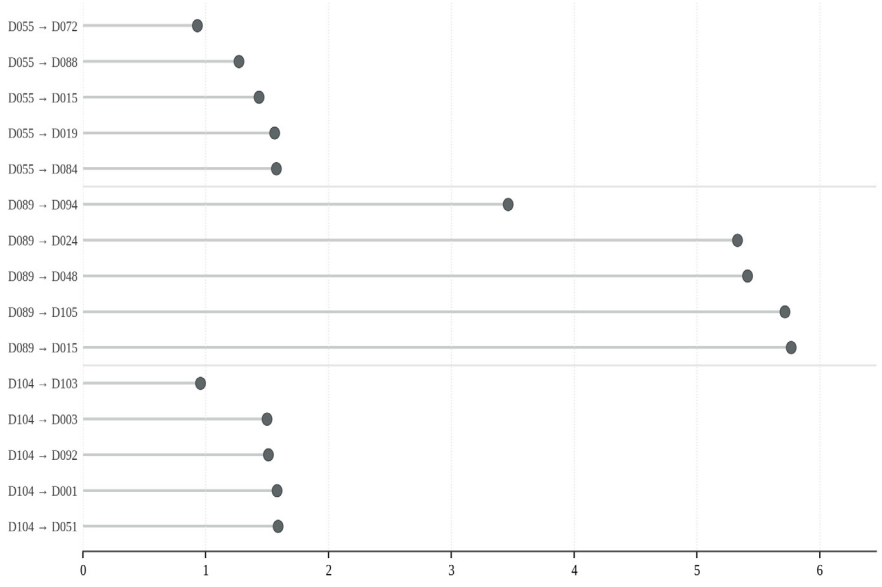


Figura 4. Distancias de documentos próximos en tres documentos de referencia

Fuente: elaboración propia.

En la Figura 4 se seleccionaron tres documentos de referencia y se identificaron los cinco documentos más próximos a cada uno. Cada punto representa la distancia chi-cuadrado entre el documento de referencia y otro documento del corpus. Se aprecia que D104 y D055 presentan distancias bajas con sus documentos

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

más próximos, lo que indica que ocupan posiciones cercanas a otros registros con perfiles léxicos semejantes. D089 muestra una separación mayor desde el primer documento y un aumento progresivo de las distancias. Esta variación muestra que la colección no ofrece el mismo grado de cercanía para todos los documentos: algunos resúmenes se vinculan con registros cercanos, mientras que otros conducen rápidamente hacia zonas más distantes.

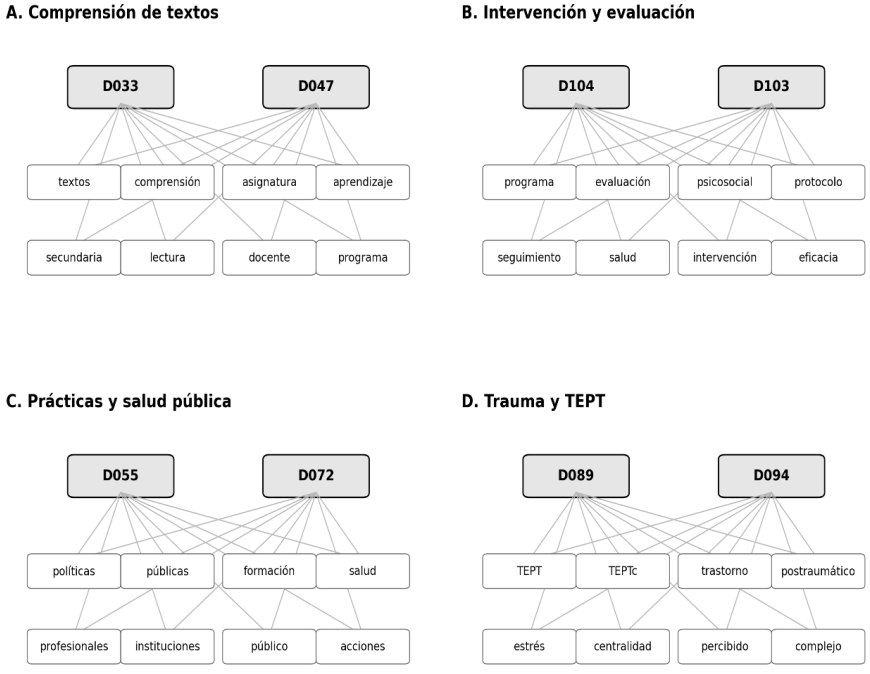


Figura 5. Pares de documentos próximos y vocabulario en común

Fuente: elaboración propia.

La Figura 5 muestra cuatro pares de documentos próximos junto con vocabulario presente en ambos resúmenes. Cada panel permite observar qué tipo de repertorio léxico vuelve interpretable la proximidad calculada en el espacio factorial. En el Panel A, la cercanía se apoya en términos vinculados con comprensión de textos, lectura y escuela secundaria. En el Panel B, los documentos comparten vocabulario asociado a programas de intervención, evaluación, protocolo, eficacia y seguimiento. El Panel C muestra una proximidad sostenida por términos referidos a prácticas profesionales, políticas públicas, formación, instituciones y salud. En el Panel D, la relación se organiza alrededor de trauma, TEPT, TEPTc, estrés y afrontamiento.

Discusión

El AFC representó cada documento como un perfil léxico relativo y permitió calcular distancias entre ellos. Las proximidades reportadas se basan en composiciones de vocabulario, no en coincidencias literales (Greenacre, 2017).

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

En la visualización los términos contribuyentes no aparecieron como unidades independientes: quedaron agrupados en zonas de proximidad dentro del espacio factorial, esta representación ayudó a interpretar las asociaciones documentales posteriores, debido a que los documentos no se aproximan solo por palabras aisladas, sino por la posición que ocupan sus vocabularios en una estructura en común.

De esta forma, los resultados sugieren que las colecciones académicas presentan una organización interna observable a partir de relaciones de proximidad léxica. Cada documento quedó definido por un perfil relativo de vocabulario, y las distancias permitieron reconstruir relaciones posibles entre registros. En algunos casos, los documentos quedaron vinculados con elementos muy cercanos; en otros, mostraron relaciones más dispersas. Esta diferencia coincide con los estudios de *science mapping*: las citaciones describen interacciones dentro de la comunidad científica, mientras que el contenido textual permite aproximarse a la organización intelectual del campo (Van Raan, 2005; Leydesdorff, 2001).

Este tipo de relación no sustituye a la bibliometría tradicional ni a los modelos de similitud basados en coincidencias de términos, su aporte reside en incorporar una capa complementaria de relación documental derivada de la distribución empírica del vocabulario. En este estudio, la cercanía entre documentos no fue definida por citaciones, descriptores o categorías previamente asignadas, sino por regularidades léxicas. La visualización de proximidad y las tablas de documentos próximos permiten observar este comportamiento de manera directa. Como ya se mencionó, desde la perspectiva de la exploración documental, esta estructura se aproxima a lo que Bates (1989) describió como *berrypicking*: el investigador avanza por desplazamientos sucesivos dentro de un espacio donde cada ítem abre nuevas relaciones posibles. Algunos resúmenes se asociaron con registros muy cercanos, mientras que otros se ubicaron en zonas más distantes del espacio factorial. En un sistema de exploración, esa información podría orientar al usuario sobre el grado de cercanía o dispersión de los documentos vinculados, así como llegar a descubrir nuevos trabajos por coincidencia léxica. Asimismo, los casos observados como pares próximos sugieren que la cercanía entre documentos no responde a un único criterio. En algunos casos, los términos compartidos remiten al objeto de estudio; en otros, la relación se sostiene en vocabulario metodológico, institucional o aplicado. Esta diferencia sugiere que el AFC puede complementar la recuperación basada en descriptores temáticos convencionales, al ordenar documentos por perfiles léxicos relativos y no solo por etiquetas explícitas.

Limitaciones

Este estudio no pretende describir la producción científica de la psicología ni caracterizar de manera exhaustiva los repositorios de los que fueron recuperados los resúmenes. El interés fue observar la existencia de relaciones documentales derivadas de la distribución del léxico. Por ende, no implicó desarrollar un sistema de recuperación, realizar pruebas con usuarios ni alcanzar exhaustividad nacional. Se limitó a mostrar relaciones de proximidad entre documentos dentro de una colección delimitada y exploratoria. Como también la red léxica presentada en los resultados cumple una función exploratoria: visualiza relaciones entre unidades, pero no pretende agotar la organización ni exploración temática del corpus.

Por otra parte, como es característico del AFC sobre tablas léxicas de alta dimensionalidad (Lebart, Salem y Berry, 1998; Murtagh, 2005), la inercia se distribuyó entre numerosos ejes. En este sentido, futuras investigaciones podrían comparar y

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

utilizar otras técnicas para la selección de dimensiones a retener. Como también podrían incorporar técnicas de agrupamiento jerárquico o detección de comunidades para examinar si las proximidades léxicas observadas forman segmentos estables.

Asimismo, la eventual aplicación de estos resultados a interfaces de exploración requiere estudios adicionales que escapan al presente estudio, así como se vuelve necesaria una extensión en diversas disciplinas para generalizar los hallazgos.

Conclusiones

La posibilidad de recorrer la literatura mediante afinidades textuales basadas en análisis léxico ofrece un enfoque complementario para la exploración en repositorios académicos. Las colecciones académicas pueden describirse como conjuntos de relaciones internas observables. El AFC permite representar estas conexiones y facilita una tarea exploratoria: vincular documentos sin depender exclusivamente de metadatos, descriptores o citaciones.

Este cambio de perspectiva tiene implicaciones para la organización del conocimiento. Además de recuperar y clasificar registros, los sistemas podrían orientar una navegación intelectual en la literatura científica a partir de relaciones internas entre documentos. Así, este enfoque se aproxima a los objetivos tradicionales de la organización del conocimiento: facilitar la interpretación de conjuntos documentales más allá de la mera recuperación de registros.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

Referencias bibliográficas

» Ahlgren, Per y Cristian Colliander. 2009. Document-document similarity approaches and science mapping: Experimental comparison of five approaches. En *Journal of Informetrics*. Vol. 3, no. 1, 49–63. <https://doi.org/10.1016/j.joi.2008.11.003>

» Argentina. 2013. Ley 26.899. Repositorios digitales institucionales de acceso abierto, propios o compartidos. En *Boletín Oficial de la República Argentina*, 13 de noviembre de 2013. <https://www.argentina.gob.ar/normativa/nacional/ley-26899-222648> [Consulta: 10 mayo 2025].

» Baeza-Yates, Ricardo y Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Reading: Addison-Wesley.

» Bates, Marcia J. 1989. The design of browsing and berrypicking techniques for the online search interface. En *Online Review*. Vol. 13, no. 5, 407–424. <https://doi.org/10.1108/eb024320>

» Benzécri, Jean-Paul. 1973. *L'analyse des données. Tome 2: L'analyse des correspondances*. Paris: Dunod.

» Börner, Katy, Chaomei Chen y Kevin W. Boyack. 2003. Visualizing knowledge domains. En *Annual Review of Information Science and Technology*. Vol. 37, no. 1, 179–255. <https://doi.org/10.1002/aris.1440370106>

» Callon, Michel, Jean-Pierre Courtial y Françoise Laville. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. En *Scientometrics*. Vol. 22, no. 1, 155–205. <https://doi.org/10.1007/BF02019280>

» Greenacre, Michael. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

» Greenacre, Michael. 2017. *Correspondence analysis in practice*. 3rd ed. Boca Raton: Chapman & Hall/CRC. <https://doi.org/10.1201/9781315369984>

» Hjørland, Birger. 2016. Knowledge organization (KO). En *Knowledge Organization*. Vol. 43, no. 6, 475–484. <https://doi.org/10.5771/0943-7444-2016-6-475>

» Husson, François, Sébastien Lê y Jérôme Pagès. 2017. *Exploratory multivariate analysis by example using R*. 2nd ed. Boca Raton: Chapman & Hall/CRC. <https://doi.org/10.1201/b21874>

» Hyland, Ken. 2000. *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

» Lebart, Ludovic y André Salem. 1994. *Statistique textuelle*. Paris: Dunod.

» Lebart, Ludovic, André Salem y Lisette Berry. 1998. *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

» Leydesdorff, Loet. 2001. *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Boca Raton: Universal Publishers.

» Manning, Christopher D., Prabhakar Raghavan y Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 54 (junio 2026)
doi: 10.34096/ics.i54.17955

» Marchionini, Gary. 1995. *Information seeking in electronic environments*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511626388>

» Marchionini, Gary. 2006. Exploratory search: From finding to understanding. En *Communications of the ACM*. Vol. 49, no. 4, 41–46. <https://doi.org/10.1145/1121949.1121979>

» Morin, Annie. 2006. Intensive use of Factorial Correspondence Analysis for text mining: application with statistical education publications. En *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7)*. Estados Unidos: International Association for Statistical Education.

» Murtagh, Fionn. 2005. *Correspondence analysis and data coding with Java and R*. Boca Raton: Chapman & Hall/CRC. <https://doi.org/10.1201/9781420034943>

» Noyons, Ed C. M. 2012. Using bibliometric maps of science in a science policy context. En *Em Questão*. Vol. 18, edición especial, 15–27.

» Petrović, Saša, Bojana Dalbelo Bašić, Annie Morin, Blaž Zupan y Jean-Hugues Chauchat. 2009. Textual features for corpus visualization using correspondence analysis. En *Intelligent Data Analysis*. Vol. 13, no. 5, 795–813. <https://doi.org/10.3233/IDA-2009-0393>

» Price, Derek J. de Solla. 1965. Networks of scientific papers. En *Science*. Vol. 149, no. 3683, 510–515. <https://doi.org/10.1126/science.149.3683.510>

» Salton, Gerard y Michael J. McGill. 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill.

» Small, Henry. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. En *Journal of the American Society for Information Science*. Vol. 24, no. 4, 265–269. <https://doi.org/10.1002/asi.4630240406>

» Swales, John M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

» Van Raan, Anthony F. J. 2005. For your citations only? Hot topics in bibliometric analysis. En *Measurement: Interdisciplinary Research and Perspectives*. Vol. 3, no. 1, 50–62. <https://doi.org/10.1207/s15366359mea0301_7>