Análisis semántico latente usando tres documentos de Pierre Bourdieu

Latent semantic analysis using three documents by Pierre Bourdieu

Cristina Restrepo Arango
Institución Universitaria Pascual Bravo. Biblioteca (Colombia).
Correo electrónico: crestrepoarango@gmail.com

Resumen

Objetivo: Aplica el análisis semántico latente a tres documentos publicados por el sociólogo francés Pierre Bourdieu y traducidos al español.

Método: Utiliza el análisis semántico latente para aplicar esta técnica se usó el lenguaje de programación en R en el entorno de desarrollo integrado (EDI) de RStudio, en el cual se utilizó el paquete pdftools, tm, lsa y LSAfun.

Resultado: La matriz de términos-documentos está compuesta por 2.646, la matriz Tk está compuesta por 3.138 palabras en total para los tres documentos. La palabra "campos" tiene una relación semántica entre el documento dos y el documento tres. La palabra "capital" tiene una relación semántica entre el documento uno y el documento dos, mientras que no se evidencia relación semántica entre los tres documentos y la palabra "científico", la palabra "cultural" y la palabra gusto. La tercera matriz Sk muestran que hay relación semántica entre el documento dos (Los tres estados del capital cultural) y el documento tres (Sobre el poder simbólico). Este análisis no muestra una relación con el documento uno (Criterios y bases sociales del gusto). La similitud de las palabras "poder", "social" y "cultural" contenidas en los tres documentos que se usaron en este análisis se les aplicó la similitud de coseno.

Abstract:

Objective: It applies latent semantic analysis to three published by the French sociologist Pierre Bourdieu and translated into Spanish.

Methods: It uses latent semantic analysis to apply this technique, the R programming language was used in the RStudio integrated development environment (EDI), in which the pdftools, tm, lsa and LSAfun package was used.

Results: The term-document matrix is composed of 2,646, the Tk matrix is composed of 3,138 words in total for the three documents. The word "fields" has a semantic relationship between document two and document three. The word "capital" has a semantic relationship between document one and document two, while no semantic relationship is evident between the three documents and the word "scientific", the word "cultural" and the word taste. The third matrix Sk shows that there is a semantic relationship between document two (The three states of cultural capital) and document three (On symbolic power). This analysis does not show a relationship with document one (Criteria and social bases of taste). The similarity of the words "power", "social" and "cultural" contained in the three documents that were used in this analysis, cosine similarity was applied to them.

Conclusiones: Las palabras "campo", "poder", "producción", "social", etc. están relacionadas, mientras que las palabras "capital" y "cultural" no están relacionadas con la totalidad de las palabras; sin embargo, todas estas palabras forman parte de los conceptos ampliamente usados por Pierre Bourdieu.

Palabras clave: Análisis semántico latente; Bourdieu, Pierre 1930-2002

Conclusions: The words "field", "power", "production", "social", etc. are related, while the words "capital" and "cultural" are not related to the entirety of the words; However, all these words are part of the concepts widely used by Pierre Bourdieu.

Keywords: Latent semantic analysis; Bourdieu, Pierre 1930-2002

Fecha de recepción: 20/11/2024
Fecha de aceptación:17/12/2024

Cita sugerida: Restrepo Arango, C. (2024). Análisis semántico latente usando tres documentos de Pierre Bourdieu. *Revista Prefacio*, 8(13),92-102. DOI: https://doi.org/10.58312/ 2591.3905.v8.n13.47619

Introducción

El análisis semántico latente(ASL) denominado en inglés latent semantic analysis(LSA), es una técnica utilizada en el procesamiento del lenguaje natural (PLN) para analizar y representar las relaciones semánticas de las palabras y documentos. Es un método estadístico creado a finales de los años 80 en el Laboratorio Bell por Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum y Lynn Streeter y lo definieron como "una teoría y un método para extraer y representar el significado de uso contextual de las palabras mediante cálculos estadísticos aplicados a un gran corpus de texto" (Lam; Toai; Vaclav, 2023, p. 1190).

El ASL se basa en que las palabras que aparecen en el mismo contexto tienden a tener significados similares.Su objetivo es mostrar el significado semántico de palabras o documentos más allá de su significado literal. Representa las palabras o los documentos como vectores en un espacio de alta dimensión, de acuerdo con sus propiedades de distribución. También muestra con qué frecuencia coocurren con otras palabras o documentos en un corpus de texto y proporciona una medida continua de similitud semántica, es decir, cero representa que no hay ninguna similitud y uno representa la similitud perfecta (Fischer; Voracek; Tran, 2023)

El ASL es un análisis computacional que permite determinar y cuantificar la similitud semántica entre piezas textuales (por ejemplo, palabras, documentos o palabras y documentos) de un corpus de textos de un mismo dominio de conocimiento. Utiliza un modelo matemático que funciona con la técnica de descomposición en valores singulares (DVS), en inglés singular valuedecomposition (SVD). El DVS se utiliza para crear un modelo matemático que describe la relación entre los documentos y las palabras, a partir del cual se genera una representación vectorial del corpus o espacio semántico que permite identificar que tan similares son los documentos en función de los términos que contienen, o bien, las relaciones entre palabras basadas en el contexto (GUTIÉRREZ, 2005).

Según Landauder, Foltz y Laham (1998, p. 263), "ASL es una técnica matemática y estadística [automatizada] para extraer e inferir relaciones de uso contextual esperado de las palabras en pasajes de discurso".Extrae y revela conocimiento implícito, por medio del análisis de texto que ayuda a identificar la coocurrencia de palabras, las cuales pueden revelar relaciones de sinonimia, o bien, términos que podrían estar relacionados en un contexto. Para Landauder, Foltz y Laham (1998, p. 263) el ASL "toma como entrada solo texto sin procesar dividido en palabras definidas como cadenas de caracteres únicos y separadas en pasajes significativos o muestras, como oraciones o párrafos".Analiza y extrae las palabras de los textos y muestra las relaciones entre palabras y documentos.

Revisión de literatura

El análisis semántico latente (ASL) fue patentado en 1988 por Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum y Lynn Streeter. El ASL también se denomina indexación semántica latente (ISL) que aplica la descomposición de valores singulares (DVS) a un corpus de documentos para extraer una matriz de término-documento basado en la semántica de los documentos.Su principal propósito es mejorar la recuperación de la información en cuanto a los problemas de polisemia (palabras con múltiples significados) y sinonimia (varias palabras pueden significar lo mismo) (Mamani Roque, 2018).

Dos años más tarde Deerwester, et al. (1990) publicaron un artículo que aplicó el ASL en la indexación de la recuperación de la información. A partir de esta publicación se ha usadoen la bibliotecología y ciencia de la información para la extracción de palabras clave (Deena; Raja, 2022); en la evaluación de resúmenes en español

(Vanegas, 2011); en la detención de las similitudes entre documentos de patentes y publicaciones científicas (Magerman; VanLooy; Song, 2010); en el mapeo bibliométrico (Van Eck, 2010); en el análisis del discurso (Landauder; Foltz; Laham, 1998); y en la coherencia textual (Foltz; Kintsch; Landauer, 1998), entre otros aplicaciones.

También el ASL lo han aplicado en otras áreas del conocimiento como es la regulación de las emociones y la salud mental (Fischer; Voracek; Tran, 2023); en el análisis de los comunicados del Banco de la República de Colombia (Arango; Pantoja; Velásquez, 2023); en el análisis de las falsas noticias en la campaña presidencial de Estados Unidos (Mayopu; Wang; Chen, 2023); en la búsqueda de enfermedades (Lam; Toai; Vaclav, 2023); en el análisis de las emociones en la revista The Reader (Zhang; Xia, 2024); en la gestión verde de recursos humanos (Sharma; Sakhuja; Nijjer, 2022); en la cadena de suministro turístico (Sanguri; Bhuyan; Patra, 2020); y en la ciencia cognitiva (Contreras Kallens; Dale, 2018), entre otros estudios que están presentes en la literatura. Cabe destacar que los artículos que se encontraron en esta revisión sobre aplicaciones de la ASL en diferentes áreas que fueron publicados en el siglo XXI, a pesar de que el primer artículo fue publicado en 1990 (Deerwester, et al.). Esto significa que las potencialidades del ASL han sido redescubiertas recientemente. Lo importante de esta investigación es que el ASL ha sido poco aplicado en el idioma español, puesto que en la revisión de literatura sólo se encontró un estudio.

Metodología

Se usó el análisis semántico latente con tres documentos[1] que fueron previamente seleccionados teniendo en cuenta las dos características que deben tener los documentos para poder efectuar la aplicación de esta técnica: primero, pertenecer al mismo ámbito del conocimiento; segundo, estar en formato pdf, no digitalizados como imagen, sino en el formato pdf que permitan la edición. Los documentos seleccionados fueron: 1) Criterios y bases sociales del gusto; 2) Los tres estados del capital cultural; y 3) Sobre el poder simbólico. Los tres documentos fueron publicados por el sociólogo francés Pierre Bourdieu y son traducciones en español del texto original publicado en francés.

El ASL aplicó la descomposición en valores singulares (DVS) a la matriz en la que cada fila representa una palabra única y cada columna representa la frecuencia de aparición de cada palabra. Este análisis se denomina análisis de factores que descompuso esta matriz en el producto de otras tres matrices. Una matriz componente que describe las entidades originales de las filas como vectores de valores de factores ortogonales derivados; una segunda matriz describe las entidades originales de las columnas de la misma manera; y la tercerama-triz una diagonal que contiene valores de escala. La ecuación uno representa a ASL:

$$X_{¡} = TkSkDk' \quad (1)$$

Donde:
Tk = muestra la relación entre los términos y documentos.
Sk = es una matriz diagonal con los valores singulares de A en diagonal.
Dk = es la relación entre cada documento.

La primera dimensión es una matriz que muestra la relación entre las palabras y los documentos en el espacio vectorial latente (dimensión Tk). La segunda matriz es la vectorización y cada valor es la suma de los valores que representan sus palabras componentes (dimensión Sk). La tercera dimensión son los productos de puntos cosenos o métricas que se utilizan para representar similitudes entre palabras y pasajes (dimensión Dk).
También se empleó la similitud decoseno para identificar los términos relacionados y su grado de similitud entre las palabras que conforman un vector. Se considera que un vector tiene similitud cuando el ángulo es menor. Se usaron las palabras: cultural, poder y social.

El ASL recogió las palabras que emplearon los traductores de los tres documentos y estableció la frecuencia de palabras por contexto. Para poder aplicar esta técnica se usó el lenguaje de programación en R en el entorno de desarrollo integrado (EDI) de RStudio. Por ello se utilizó el paquete pdftools que permite extraer textos y datos (Ooms, 2023), con el fin de manipular documentos en ese formato; el paquete tm permite la limpieza del texto y el procesamiento (Feinerer and Hornik 2023) con el propósito de realizar minería de textos; lsa(Wild, 2022) para poder hacer el análisis semántico latente; y LSAfun proporciona funciones para emplear modelos de espacio vectorial (Guenther, 2023) para poder aplicar el análisis factorial.

Resultados

La matriz de términos-documentosrepresenta en cada filauna palabra y en cada columna un documento. Incluye 2.646 palabras diferentes que contienen el número de veces que cada palabra aparece en el documento correspondiente. Las palabras de cada documento no son concurrentes en los tres documentos a la vez, algunas si aparecen mínimo en dos documentos, por ejemplo, "beneficio" está en el artículo dos y en el artículo tres, así sucesivamente como se muestra en el extracto de la matriz (Ver Tabla 1).

Tabla 1. Matriz términos-documentos

Palabras	Documento 1	Documento 2	Documento 3
1. cosa	1	0	0
2. nélements	1	0	0
3. abierto'	1	0	0
4. abolir	1	0	0
5. acceso	1	0	0
6. aceptada	1	0	0
7. aceptadas	1	0	0
8. actitud	1	0	0
9. activande	1	0	0
10. acto	3	0	0

Fuente: *Elaboración propia*

La dimensión Tk está compuesta por 3.138 palabras en total para los tres documentos,[2] en esta matriz aparece el valor del vector de cada palabra, este valor puede ser positivo, negativo, menor o mayor a uno o al cero. En esta matriz aparecen las diferentes formas de una palabra (verbo, plural, adjetivo, adverbio, etc.). En el caso de la palabra "cultura", "cultural" y "culturales"que están localizadas en un punto del plano, cada plano se le asigna un valor representado en un vector que servirá para graficarlo en un plano cartesiano y dependiendo de la ubicación de ese valor en el plano se reflejará la relación entre documentos y palabras.

Por ejemplo, al revisar los valores de las palabras se puede evidenciar que la palabra "campos" tiene una relación semántica entre el documento dos y el documento tres. La palabra "capital" tiene una relación semántica entre el documento uno y el documento dos, mientras que no se evidencia relación semántica entre los tres documentos y la palabra "científico", la palabra "cultural" y la palabra "gusto" (Ver Tabla 2).

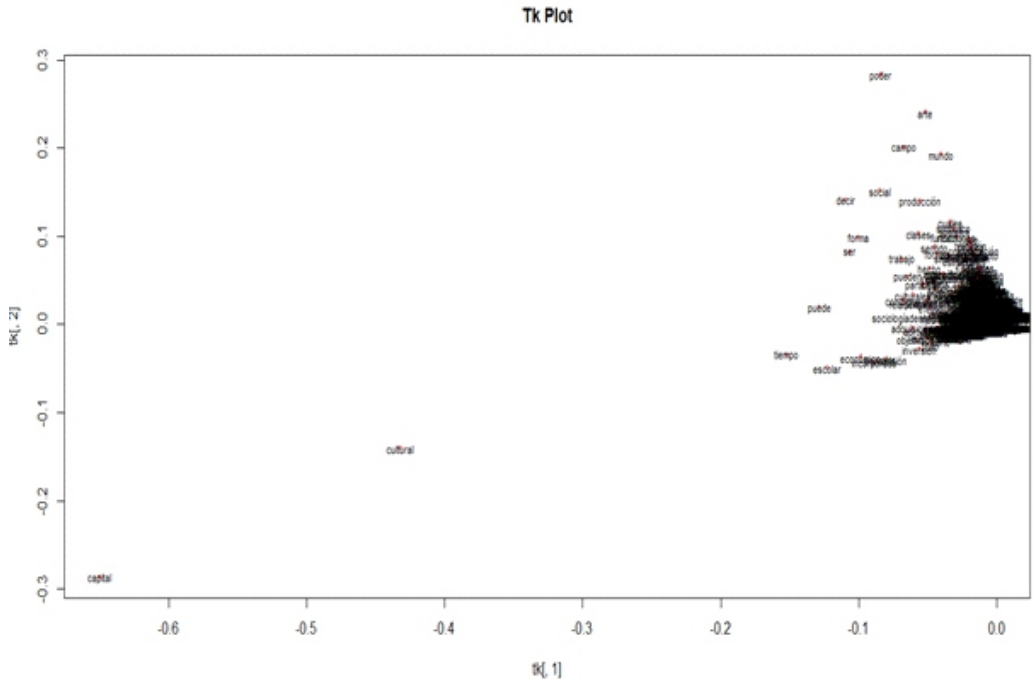
Tabla 2. *Extracto de la matriz Tk*

Palabra	Documento 1	Documento 2	Documento 3
campos	-0.060438349	33173,97	-0.063441976
capital	-0.009976665	0.005546468	0.0105504876
científic o	-0.649621749	-0.287169318	0.0424773111
cultural	-0.432347472	-140429,6	-0.069423827
gusto	-0.016095890	-0.007891136	0.0002811909

Fuente: *Elaboración propia*

La relación entre las palabras que conforman los tres documentos muestra que las palabras "campo", "poder", "producción", "social", etc. están relacionadas con el agrupamiento de la mayoría de las palabras, mientras que las palabras "capital" y "cultural" no están conectadas con la mayoría de las palabras que aparecen en laFigura 1.Es decir, son dos palabras aisladas, aunque forman el concepto "capital cultural" ampliamente usado por Bourdieu en su obra.

Figura 1. Relación entre las palabras



Fuente: *Elaboración Propia*

La tercera matriz Skestá compuesta por tres valores para los tres documentos (Ver Tabla 3). Estos valores muestran que hay relación semántica entre el documento dos (Los tres estados del capital cultural) y el documento tres (Sobre el poder simbólico). Este análisis no muestra una relación con el documento uno (Criterios y bases sociales del gusto). A pesar de que los documentos fueron producidos por Pierre Bourdieu, quien es considerado uno de los sociólogos franceses del siglo XX más importante e influyente por el uso del análisis de correspondencias y la inclusión de conceptos como "capital cultural", todos los documentos usados en este análisis no están relacionados.

Tabla 3. MatrizSk

Dimensión	Artículo 1	Artículo 2	Artículo 3
1	118.35203	77.16116	61.61941

Fuente: *Elaboración propia*

La matrizDk está compuesta por tres dimensiones y por los tres documentos (Ver Tabla 4) que mide las distancias entre los documentos.En la dimensión uno y en la dimensión dos las distancias entre los documentos son lejanas, mientras que en la dimensión tres hay cercanía entre los documentos dos y tres.

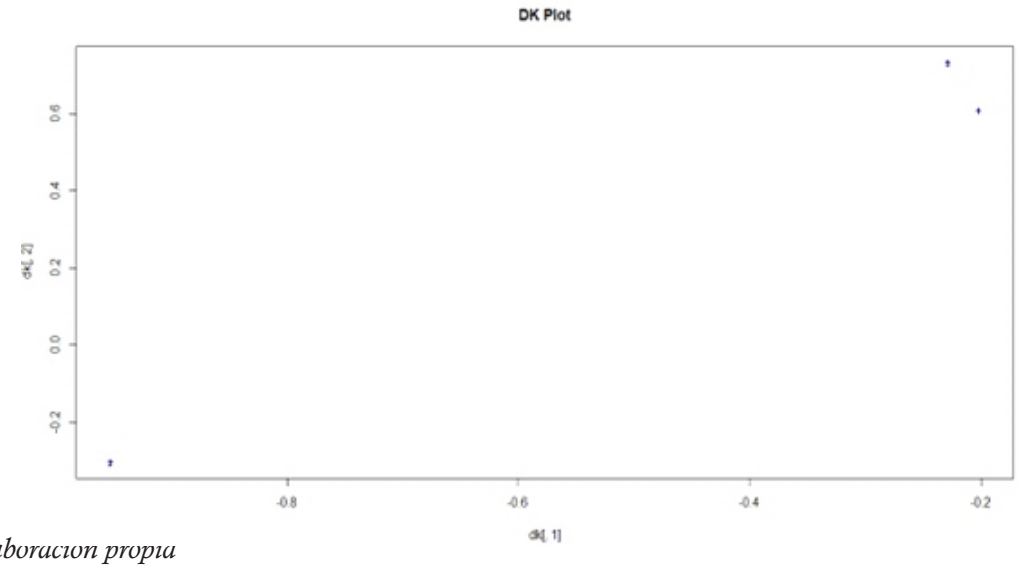
Tabla 4. MatrizDk

Dimensión	**Documento1**	**Documento2**	**Documento 3**
1	-0.2016316	0.6089989	-0.767114781
2	-0.9524906	-0.3044446	0.008663407
3	-0.2282679	0.7324165	0.641451369

Fuente: *Elaboración propia*

Los valores obtenidos por el documento dos (Los tres estados del capital cultural) y el documento tres (Sobre el poder simbólico) muestra que están relacionados, mientras que el documento uno (Criterios y bases sociales del gusto) no está relacionado con los otros dos documentos(Ver Figura 2).

Figura 2. Relación entre los documentos



Fuente: *Elaboracion propia*

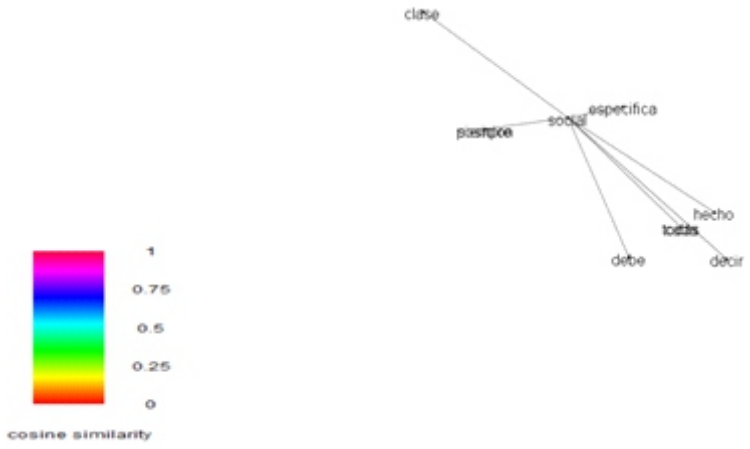
La similitud de las palabras "poder", "social" y "cultural" contenidas en los tres documentos que se usaron en este análisis se les aplicó la similitud de coseno. De acuerdo con el modelo de análisis semántico latente, la palabra poder tiene una estrecha similitud con simbólico, aunque también hay similitud con otras palabras como "fuerza", "estructura", etc., pero esa similitud es lejana por el tamaño de las líneas que representan esa relación (Ver Figura 3)

Figura 3. Similitud de coseno de la palabra "poder"

Fuente: *Elaboración propia*

La palabra "social" tiene similitud con "clase", mientras que las palabras "debe", "decir", "hecho", etc. no forman parte de los conceptos de lateoría sociológica de Pierre Bourdieu (Ver Figura 4).

Figura 4. Similitud de coseno de la palabra "social"



Fuente: *Elaboración propia*

La palabra "cultural" tiene similitud con "capital", también hay semejanza con las palabras "mercado", "acumulación", etc. Se destaca que "capital cultural" es uno de los conceptos principales de la teoría Pierre Bourdieu ampliamente usados en la literatura (Ver Figura 5).

Figura 5. Similitud de coseno de la palabra "cultural"

Fuente: *Elaboración propia*

Conclusiones

El análisis semántico latente muestra la relación entre documentos y palabras usando la técnica de la descomposición de valores singulares. El ASL permite extraer matrices que muestran la coocurrencia entre las palabras, su relación y la de los documentos. Para usar el ASL es necesario que los documentos que se exploren con esta técnica desarrollen en su contenido el mismo campo del conocimiento. Además, ayuda a identificar y comprender un documento o conjunto de documentos, a partir de la frecuencia de los términos en los documentos. Según Sharma, Sakhuja & Nijjer (2022), es considerado uno de los mejores métodos para extraer relaciones significativas de las palabras contenidas en un documento.

El ASL es un método de indexación no supervisado en la programación natural del lenguaje, con el fin de extraer las palabras que están relacionadas semánticamente (Deena; Raja, 2022). Con este método se analizaron tres documentos del sociólogo francés Pierre Bourdieu, este estudio mostró que hay una relación semántica entre el documento dos (Los tres estados del capital cultural) y el documento tres (Sobre el poder simbólico). Este análisis no muestra una relación con el documento uno (Criterios y bases sociales del gusto).

Este método muestra las relaciones entre los términos que forman las oraciones del documento, estos términos se descomponen en la DVS. La DVS clasifica los términos y los documentos semánticamente, lo que permite inferir en este estudio que las palabras "campo", "poder", "producción", "social", etc. están relacionadas, mientras que las palabras "capital" y "cultural" no están relacionadas con la totalidad de las palabras. Cabe resaltar que estas palabras forman parte de conceptos ampliamente usados por Pierre Bourdieu.

Es importante señalar el aporte de este tipo de técnicas; sin embargo, los resultados en español requieren refinarse para obtener mejores análisis que permitan evidenciar las relaciones entre documentos y palabras.

Referencias bibliográficas

- **Arango, L. E.; Pantoja, J.; Velásquez, C.**(2023).A content analysis of the Central Bank's press releases in Colombia. *Latin American Journal of Central Banking*, 4(3),100097,
- **Contreras Kallens, P.; Dale, R.** (2018) Exploratory mapping of theoretical landscapes through word use in abstracts. *Scientometrics* 116(3), 1641-1674.
- **Deena, G.; Raja, K.** (2022). Keyword extraction using latent semantic analysis for question generation. *Journal of Applied Science and Engineering*, 26(4),501-510.
- **Deerwester, S.;Dumais, S. T.; Furnas, G. W.;Landauer, T. K.; Harshman, R.** (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
- **Feinerer, I.; Hornik, K.**(2023).*tm: Text Mining Package*, 2023.https://CRAN.R-project.org/package=tm
- **Fischer, A., Voracek, M., Tran, U. S.** (2023). Semantic and sentiment similarities contribute to construct overlaps between mindfulness, Big Five, emotion regulation, and mental health. *Personality and Individual Differences*,(210), 112241.
- **Foltz, P. W.; Kintsch, W., Landauer, T. K.** (1998). The measurement of textual coherence with latent semantic analysis.*Discourse Processes*, 25(2-3), 285-307. DOI: 10.1080/01638539809545029
- **Guenther, F.** (2024). *LSAfun: Applied Latent Semantic Analysis (LSA) Functions*, 2023.https://cran.r-project.org/web/packages/LSAfun/index.html.
- **Gutiérrez, R. M.** (2005). Análisis Semántico Latente: ¿teoría psicológica del significado?*RevistaSignos*, 38(59), 303-323.
- **Lam, L. C. Q., Toai, T. K.; Vaclav, S. (2023)**A latent semantic analysis method for ranking the results of human disease search engine. *Bulletin of Electrical Engineering and Informatics*, 12(2), 1189-1195**.**
- **Landauer, T. K.; Foltz, P. W., Laham, D.** (1998). An introduction to latent semantic analysis. *Discourseprocesses*, 25(2-3), 259-284.
- **Magerman, T., Van Looy, B., Song, X.** (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*,82(), 289-306.
- **Mayopu, R.G., Wang, Y.Y., Chen, L.S.** (2024). Analyzing Online Fake News Using Latent Semantic Analysis: Case of USA Election Campaign. *Big Data CognitiveComputation*, 7(81). https://doi.org/10.3390/bdcc7020081
- OOMS, J. (2023). *Package pdftools*, 2023.https://cran.r-project.org/web/packages/pdftools/pdftools.pdf
- **Mamani Roque, M.** (2018). Descomposición en valores singulares y análisis semántico latente. 44h. Tesis (Máster) -- Universidad Politécnica de Valencia. Departamento de Matemática Aplicada.
- **Sanguri, K.; Bhuyan, A.; Patra, S.** (2020). A semantic similarity adjusted document co-citation analysis: a case of tourism supply chain. *Scientometrics*, 125(1),233-269.
- **Sharma, C.; Sakhuja, S.; Nijjer, S.** (2022). Recent trends of green human resource management: Text mining and network analysis. *Environmental Science and Pollution Research*, 29(56), 84916-84935.
- **Van Eck, N.; Waltman, L.; Noyons, E.; Buter, R.** (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 8(3),581-596.
- **Venegas, R.**(2011). Evaluación de resúmenes en español con Análisis Semántico Latente: una implementación posible. *Revista signos*, 44(75), 85-102
- **Wild, F.** (2022). *LatentSemanticAnalysis*. https://cran.r-project.org/web/packages/lsa/lsa.pdf
- **Zhang, L.; Xia, Y.**(2024). Text Study of Reader Magazine in the Context of Big Data. *Applied Mathematics and Nonlinear Sciences*, 9(1), 1-14. https://sciendo.com/article/10.2478/amns.2023.2.00284

Notas al pie de página

1.- Los tres documentos se obtuvieron en pdf de la página web: https://www.bloghemia.com/2018/12/pierre-bourdieu-coleccion-de-libros-en.html.

2.- La matriz términos-documentos no se incluye, ya que tiene una extensión de 60 páginas.

3.- La matriz Tk tiene una extensión de 75 páginas, por eso no se incluyó completa en este artículo.