ISSN 1688-2806

Universidad de la República
Facultad de Ingeniería

Análisis Semántico Latente para minería de datos en ingeniería biomédica

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Gabriel Slomovitz

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Eléctrica.

Director de Tesis
Dr. Eduardo Mizraji ..... Universidad de la República

Tribunal
XXX XXX ..... Universidad 1
XXX XXX ..... Universidad 2
XXX XXX (Revisor Externo) ..... Universidad 3

Director Académico
Prof. Ing. Franco Simini ..... Universidad de la República

Montevideo
lunes 16 junio, 2025

Análisis Semántico Latente para minería de datos en ingeniería biomédica, Gabriel Slomovitz.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 82 páginas.
Compilada el lunes 16 junio, 2025.
`http://iie.fing.edu.uy/`

Quien no añade nada a sus conocimientos, los disminuye.

El Talmud

Esta página ha sido intencionalmente dejada en blanco.

Agradecimientos

Tuve el privilegio de contar como tutores para esta tesis a destacados profesores, grado cinco de la UdelaR, como el Dr. Eduardo Mizraji y el Prof. Franco Simini, a quienes les estoy muy agradecido por su acompañamiento y orientación.

Agradezco también a la Comisión Honoraria de Lucha contra el Cáncer, en particular al Dr. Enrique Barrios y al Ing. Rafael Alonso, por facilitar los datos y colaborar con este trabajo, en lo referente al capítulo de Análisis Estadístico de Distribución de Cáncer.

Por cuestiones personales y laborales, terminó siendo un trabajo distribuido en un largo período, lo que por un lado dificultó el avance pero por otro me permitió asentar las ideas e ir incorporando nuevas, en particular en lo referente a la perspectiva e impacto introducido por la Inteligencia Artificial Generativa. Agradezco a quien me acompañó siempre en este largo trayecto, oficiando como guía y tutor involuntario, también grado cinco de la UdelaR y Doctor en Ingeniería: mi padre.

Al resto de mi familia, y a todos los que me acompañaron directa o indirectamente, ¡muchas gracias!

Esta página ha sido intencionalmente dejada en blanco.

A Emi, Tadeo y Cybel.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

En esta tesis se explora la teoría del Análisis Semántico Latente (LSA) y sus aplicaciones en el campo de la ingeniería biomédica. LSA es una técnica de procesamiento de lenguaje natural basada en la descomposición en valores singulares (SVD) que, no solo funciona como una herramienta de minería de texto, sino que también propone un modelo computacional capaz de simular aspectos del aprendizaje y la cognición humana, particularmente en la adquisición del lenguaje.

En el desarrollo de este trabajo se investiga el estado del arte de las aplicaciones de LSA en medicina. Esto incluye su uso en el análisis de registros médicos para identificar relaciones entre síntomas y enfermedades, mejorar la recuperación de información médica, detectar patrones en datos de pacientes, validar textos simplificados por IA e identificar determinantes sociales de la salud. También se examinan aplicaciones en el análisis del lenguaje para la evaluación de condiciones como la esquizofrenia y el deterioro cognitivo asociado a Alzheimer, mediante la cuantificación de la coherencia del discurso.

Como contribuciones originales, la tesis presenta dos estudios: 1) La aplicación de LSA al análisis de la distribución estadística de tipos de cáncer en Uruguay, demostrando su capacidad para revelar correlaciones geográficas no evidentes con métodos tradicionales; y 2) El uso de LSA para la correlación de mensajes de eventos (syslog), aplicable a dispositivos médicos, mostrando que puede agrupar eventos semánticamente relacionados sin necesidad de reglas predefinidas.

Finalmente, se discute LSA como una teoría de minería de datos, se vincula con las memorias asociativas como modelo de procesos cognitivos y se contrasta con los Large Language Models (LLM) contemporáneos. Se concluye que, si bien los LLM superan a LSA en muchos aspectos, LSA mantiene su relevancia por su interpretabilidad, menor costo computacional y su potencial para explicar procesos cognitivos.

Se proponen futuras líneas de investigación en sistemas híbridos, que combinen ambas tecnologías.

Esta página ha sido intencionalmente dejada en blanco.

Tabla de contenidos

Agradecimientos III

Resumen VII

1. Introducción **1**
1.1. Alcance . . . 1
1.2. Bases de LSA . . . 3
1.3. Perspectiva histórica . . . 6
1.4. Fundamentos matemáticos . . . 9
1.5. LSA y memorias asociativas . . . 13

2. Aplicaciones en medicina **17**
2.1. LSA en registros médicos . . . 17
2.2. LSA en análisis de lenguaje . . . 25
2.2.1. LSA en esquizofrenia . . . 25
2.2.2. LSA en Alzehimer . . . 27
2.2.3. LSA en recuperación de recuerdos . . . 31

3. LSA para análisis estadístico de distribución de cáncer **33**
3.1. Introducción . . . 33
3.2. Metodología . . . 35
3.3. Resultados . . . 40

4. LSA para correlación de eventos **41**
4.1. Introducción . . . 41
4.2. Metodología . . . 43
4.3. Conclusiones y trabajos futuros . . . 47

5. LSA como teoría de Data Mining **49**
5.1. Introducción . . . 49
5.2. Modelos fundacionales y lenguaje . . . 53
5.3. Modelos fundacionales y aprendizaje . . . 55
5.4. Modelos fundacionales y el entendimiento . . . 57
5.5. Análisis Semántico Latente (LSA) vs Modelos de lenguaje extensos (LLMs) . . . 59

Tabla de contenidos

6. Conclusiones **61**

Referencias **63**

Índice de figuras **67**

Capítulo 1

Introducción

1.1. Alcance

La presente tesis de maestría estudia el vínculo entre la teoría conocida como Análisis Semántico Latente (LSA por sus siglas inglés) y aplicaciones en el ámbito médico. Como se verá más adelante, LSA nació como un mecanismo para minería de texto, pero tiene la particularidad de llegar mucho más lejos: permite reproducir, al menos hasta cierto grado, mecanismos cognitivos que ocurren en el cerebro. En este sentido, puede ser vista como una teoría para entender, desde un punto de vista matemático, algunos procesos que tienen que ver con el aprendizaje, y muy en particular con el aprendizaje del lenguaje. Es este un aspecto fascinante que ha intrigado a la humanidad desde Platón hasta nuestros días. LSA permite arrojar algo de luz sobre estas disquisiciones que históricamente se han movido en el ámbito de la filosofía. Por tanto, a diferencia de otros mecanismos de *machine learning* para minería de datos, LSA permite, por un lado, explorar las aplicaciones prácticas que se derivan de su implementación, pero a su vez, acercarnos a un ámbito mucho más ambicioso: la comprensión del funcionamiento de aspectos que ocurren en el cerebro.

El objeto de esta tesis consiste en estudiar la teoría detrás de este mecanismo y explorar el 'estado del arte' en lo que refiere a aplicaciones de LSA en medicina. En la práctica, muchas de estas aplicaciones requerirán de técnicas adicionales para completar aspectos no cubiertos por LSA. Sin embargo, la tesis no pretende abarcar estos mecanismos tecnológicos laterales sino centrarse exclusivamente en el potencial de LSA.

Las siguientes secciones se estructuran del siguiente modo. En la sección 1.2 se describen los conceptos básicos de LSA. En la sección 1.3 se analiza la perspectiva histórica que permitió el surgimiento de la teoría. En la sección 1.4 se detallan los principios básicos del funcionamiento de LSA y su vínculo con las memorias asociativas. En el capítulo 2 se analizan aplicaciones en el ámbito médico, mientras que en los capítulos 3 y 4 se describen contribuciones particulares de esta tesis en

ámbitos como análisis estadístico de distribución del cáncer en nuestro país, y LSA para correlación de eventos. Finalmente, se establece en el capítulo 5 el vínculo entre LSA y otras técnicas de *machine learning* para minería de datos, y en el capítulo 6 se señalan las principales conclusiones del trabajo.

1.2. Bases de LSA

Los autores de LSA [1] parten de la siguiente idea básica: debe existir un 'método computacional' compartido por todos los humanos que nos permite aprender decenas de miles de palabras y entender su significado para cualquier idioma. Y ese método debe ser capaz de aprender por simple inmersión en el lenguaje y sin la necesidad de incorporar reglas y definiciones explícitas, al menos para una importante cantidad de palabras.

Cabe mencionar que existe un interesante campo de discusión en los ámbitos de la filosofía, lingüística y psicología sobre si esta capacidad es innata o aprendida. LSA no incursiona en estas disquisiciones directamente, pero permite arrojar algo de luz sobre el tema. La cuestión fundamental es ¿cómo es posible que el cerebro desarrolle estas habilidades en la escala y rapidez con la que lo hace? ¿Es posible que un sistema expuesto a un cuerpo de texto similar a la que un humano encuentra, aprenda el significado de todas las palabras en cualquier lenguaje?.

El principal descubrimiento de LSA es que con apenas una operación algebraica relativamente sencilla, y sometido a la misma experiencia aproximada que una persona, puede obtener resultados muy similares a los humanos para ciertas tareas cognitivas.

Muchos filósofos y lingüistas han sostenido, y sostienen aún, que el cerebro debe venir 'equipado' con alguna fuente de conocimiento que nos permite la adquisición de lenguaje y significado a gran velocidad. Si esto fuera así, sería imposible para un sistema computacional simular o ser capaz de deducir significado solamente a través de la inmersión en un cuerpo de texto. Es decir, sin recibir reglas ni instrucciones adicionales

Sin embargo, LSA es capaz, con solo un simple algoritmo aplicado a un cuerpo de texto, de reproducir tareas que si hubieran sido realizadas por personas, hubieran implicado el entendimiento del significado de las palabras. Esto fue un descubrimiento sorprendente para los impulsores de LSA. Algunos ejemplos de actividades que validan esta afirmación, en trabajos iniciales sobre el tema, son:

- Test de vocabulario para liceales. [2]
- Comprensión de coherencia entre pasajes de texto. [3]
- Modelado de descubrimientos en sicología cognitiva. [4]
- Detección de mejora en aprendizaje de estudiantes. [5]
- Diagnóstico de esquizofrenia. [6]
- Mejora en sistemas de recuperación de información. [7]
- Simulación de entendimiento humano de significado para distintos idiomas usando exactamente el mismo algoritmo. [8]

Estas simulaciones son un indicio de que LSA es un candidato interesante como mecanismo de explicación del aprendizaje del lenguaje.

Los autores de LSA parten de la premisa de que los humanos primero aprendemos las relaciones entre palabras y luego las mapeamos a experiencias perceptuales, y no al revés. Esto explica que el método propuesto sea efectivo aún sin contar con ninguna información proveniente de algún mecanismo de percepción (vista, oído, tacto). En definitiva, la mayoría de relaciones de significado verbal pueden deducirse a partir de la sola exposición al lenguaje. Y además, pueden ser deducidas a partir de un pequeño conjunto de estas relaciones.

Estos conceptos son los que convierten a LSA en una teoría del 'significado', ya que nos permiten aproximarnos a un modelo computacional que simula cómo los humanos deducimos, o adquirimos, el significado de las palabras y de pasajes de texto. LSA, sin embargo, no es una teoría completa del lenguaje o del significado. En particular, no toma en cuenta el orden de las palabras y puede no representar correctamente metáforas, ánforas, cuantificaciones, proposiciones matemáticas y negaciones. De cualquier modo, aún con sus limitaciones, y siendo un modelo computacional, LSA provee bases para el entendimiento del lenguaje y del significado muy superiores a las teorías filosóficas previas. Dado que el uso adecuado de las palabras, requiere el conocimiento del significado de las mismas, y dado que LSA hace un uso adecuado de las palabras, se infiere [1] que debe poseer dicho conocimiento.

LSA considera que la representación de los pasajes de texto es una función de las palabras que contiene. O dicho de otro modo, que el significado de un pasaje de texto es la suma del significado de las palabras contenidas en dicho pasaje. LSA modela entonces un pasaje de texto como una ecuación lineal y un cuerpo de texto como un conjunto de ecuaciones simultáneas. Para que LSA efectivamente represente el significado de las palabras debe generarse un gran conjunto de texto a ser analizado. Los autores, inicialmente utilizaron el equivalente a lo que un humano promedio pudo haber leído en su vida. Esto puede verse como el cuerpo de entrenamiento del mecanismo. Este cuerpo de texto es luego divido en párrafos que encierran cierta coherencia en cuanto a significado. Se construye, entonces, una matriz cuyas filas corresponden a cada una de las palabras, y las columnas a cada uno de los párrafos. Cada celda indica la cantidad de veces que la palabra aparece en el párrafo, luego de algunas transformaciones que se verán más adelante.

Los autores llegaron a la conclusión de que en base a un cuerpo de texto con 10^5 a 10^9 párrafos y 10^5 a 10^6 palabras distintas se consiguen buenos resultados. Esto produce un conjunto de 10^5 a 10^9 ecuaciones lineales simultáneas.

La solución se compone de un conjunto de vectores, típicamente de 200 a 500 elementos. Cada vector representa una palabra o un pasaje de texto. El significado de cada pasaje de texto, se computa como la suma de los vectores de palabras

que contiene. Y la similitud entre palabras o pasajes de texto, puede medirse como el coseno entre los vectores. El éxito de LSA radica en la reducción de dimensiones que, como se verá más adelante, se realiza a partir de SVD (Singular Value Decomposition). Esto le permite al método aprender el significado de las palabras a partir de la ocurrencia de las mismas en cada pasaje de texto, pero también a partir de la no ocurrencia. Sin esta operación, la comparación entre vectores no sería de mucha utilidad, ya que la matriz original contiene un altísimo porcentaje de celdas vacías. Es la reducción de dimensiones, la que otorga al método su efectividad, permitiendo estimar valores para todas las celdas de la matriz reconstruida. En definitiva es esta reducción de dimensionalidad y la posterior reconstrucción de la matriz, lo que permite medir la similitud entre dos pasajes de texto, aún sin que compartan ninguna palabra común entre ellos. En los siguientes capítulos se profundizan estos conceptos y se exponen las bases matemáticas que los sustentan.

1.3. Perspectiva histórica

Para brindar una perspectiva histórica de LSA, debemos comenzar por describir la evolución de la operación algebraica conocida como Descomposición en Valores Singulares (SVD por sus siglas en inglés). SVD es la base matemática de LSA. El origen del uso de valores singulares parece remontarse al siglo XIX a partir de los intentos de matemáticos de la época por conseguir la reducción de una forma cuadrática a forma diagonal mediante cambios de base ortogonales [9].

Eugene Beltrami, quien fue geómetra italiano, publica el primer trabajo sobre el tema con el afán de promover el estudio de las formas bilineales entre sus alumnos [10]. En dicho artículo propone lo que hoy se conoce como SVD. Pese a que no trata el problema en forma general, y el artículo carece de una demostración completa, Beltrami puede ser considerado el descubridor de la descomposición en valores singulares. Un año más tarde, Camille Jordan publica un trabajo sobre formas bilineales pero desde otro punto de vista: búsqueda de máximos y mínimos de formas bilineales bajo ciertas condiciones [11]. Posteriormente, y al parecer sin tener conocimiento de estos artículos, J. J. Sylvester publica un trabajo sobre la diagonalización de formas bilineales mediante sustituciones ortogonales en una nota en "The Messenger of Mathematics" [12]. En este artículo propone un algoritmo para diagonalizar formas bilineales y nombra a los números reales positivos que aparecen en la forma canónica diagonal como "multiplicadores canónicos".

En 1907, décadas después y desde una perspectiva completamente distinta relacionada con ecuaciones integrales, Erhard Schmidt publica una teoría general de ecuaciones integrales reales con núcleos simétricos [13]. En este trabajo introduce el concepto de valor propio y función propia. En base a esta publicación, H. Nateman [14] introduce la notación de "valores singulares" para definir a los números que son en esencia los recíprocos de los valores propios mencionados por Schmidt. Pero es E. Picard [15] quien específica que para los núcleos simétricos los valores propios de Shcmidt son reales y en ese caso los llama "valores singulares". A pesar de todos estos esfuerzos, la denominación de "valores singulares" recién se estabiliza entre la comunidad matemática años más adelante con el advenimiento de las computadoras y el análisis numérico. En este sentido, SVD comienza a ser utilizado en problemas que requieren reducción de dimensionalidad, tratamiento de imágenes y más recientemente para procesamiento de datos con técnicas de *Machine Learning*.

En lo que refiere a LSA en particular, surge en el ámbito de indexado y búsqueda de información en repositorios de documentos de texto a partir de su precursor: LSI (Latent Semantic Indexing). LSI propone una nueva teoría como propuesta de mejora de los métodos de búsqueda existentes. Los inventores, Deerwester et al, publican en 1990 un artículo en el que explican el novedoso método y muestran los primeros resultados [16]. LSI pretende solucionar las dificultades de las técnicas de recuperación de información de la época basándose en asociar palabras con docu-

mentos. El problema fundamental de estos métodos es que el usuario normalmente realiza la búsqueda en base a contenido conceptual pero las palabras individuales pueden proveer poca información sobre el concepto o el significado del documento. Un mismo concepto puede ser expresado de múltiples maneras, lo que se conoce como sinonimia. Por otro lado, una palabra puede tener múltiples significados e interpretaciones, fenómeno denominado polisemia. Por tanto, la asociación literal entre palabras y documentos resulta poco efectiva como técnica de motores de búsqueda. LSI se basa en que existe una estructura semántica subyacente en los datos, que de alguna manera es oscurecida por la elección de las palabras en particular. SVD es utilizado para estimar esa estructura reduciendo el "ruido" que la oculta.

Deerwaster et al, plantean en [16] que las deficiencias de los motores de búsqueda de la época radican en que las palabras de búsqueda muchas veces no reflejan aquellas a partir de las cuales la información fue indexada en el documento. Esto se debe, más que nada, a los problemas de sinonimia y polisemia mencionados previamente. El grado de variabilidad del uso de distintos términos para describir el mismo objeto es enorme. Por otro lado, la misma palabra puede significar diferentes cosas en distintos contextos. Las fallas de los sistemas de indexado automático pueden reducirse a tres factores, según señalan los autores:

- Los términos utilizados para indexar un documento solo contienen una fracción de los términos que pueden utilizarse para la búsqueda.
- La falta de herramientas automáticas para manejar polisemia.
- Las palabras son tratadas como independientes unas de otras.

El objetivo de LSI es utilizar la matriz de ocurrencias de palabras en documentos para derivar los parámetros de un modelo que permita describir las relaciones entre términos y documentos aún en los casos en que la asociación directa de palabras no es observable. Para ello, analizaron distintas propuestas en base a los siguientes criterios:

- Riqueza representacional ajustable: El modelo debe ser capaz de ajustarse para representar la estructura semántica subyacente. Esto lleva a descartar modelos restrictivos, y orienta la solución hacia modelos dimensionales que permiten controlar el número de dimensiones.
- Representación explícita tanto de términos como de documentos: Para determinar la proximidad entre términos y documentos es necesario poder representar los objetos en la estructura semántica de modo que tanto términos como documentos aparezcan como puntos en el mismo espacio.
- Manejo computacional: Se requiere de técnicas computacionales eficientes para el manejo de la gran cantidad de datos involucrados.

El modelo que cumple con estos criterios es el denominado *Two-mode factor analysis*, basado en SVD. El método más simple *One-mode factor analysis* genera

una matriz que permite asociar pares de elementos del mismo tipo, por ejemplo: documentos. Esta matriz se descompone luego en dos matrices de componentes linealmente independientes que contienen los valores propios y los vectores propios. Ignorando los componentes más pequeños, puede reconstruirse la matriz original pero con mucha menor cantidad de elementos. Es posible entonces comparar los documentos calculando el coseno o producto escalar de los vectores que los representan. Por otro lado, *Two-mode factor analysis* utiliza matrices rectangulares, no cuadradas, permitiendo asociar elementos de distinta entidad, por ejemplo: términos y documentos. La matriz es descompuesta utilizando SVD y las matrices resultantes se componen de vectores singulares y valores singulares. Del modo ya descripto, puede aproximarse a la matriz original pero con menos dimensiones. En este espacio reducido es posible medir la similitud de término-término, documento-documento o término-documento. Es así que SVD puede utilizarse para recuperación de información cumpliendo los tres requisitos mencionados previamente.

Desde otra perspectiva, LSA nos acerca a un problema que ha intrigado a filósofos desde la antigüedad: cómo adquirimos el conocimiento del lenguaje los humanos. Dicho de otro modo, cómo somos capaces sin instrucción alguna, problema al que se enfrenta cualquier niño al aprender su lengua materna, de entender el significado de las palabras. Para muchos filósofos, Platón en particular, la respuesta radica en que el conocimiento se encuentra de algún modo incorporado de forma innata en las personas y el aprendizaje del lenguaje se reduce a un proceso de contemplación y asociación.

Curiosamente, LSA permite simular este proceso de aprendizaje, al menos hasta cierto punto, sin utilizar para ello ningún conocimiento previo gramatical ni semántico, tomando como única entrada texto crudo. Y aún así, es capaz de evaluar relaciones de similitud entre palabras o pasajes de texto, que permiten inferir su significado. En definitiva LSA plantea una explicación distinta al problema de Platón, tal como se conoce al misterio por el cual los humanos parecemos tener mucho más conocimiento de aquel al que hemos sido expuestos. Esta nueva explicación radica en que, explotando correctamente las sutiles relaciones entre palabras, somos capaces de inferir el significado de muchas otras, amplificando sustancialmente el proceso de aprendizaje.

En este sentido existen trabajos que vinculan la representación matricial en la que se basa LSA, con "memorias asociativas"(ver sección 1.3) que modelan memorias biológicas . En consecuencia LSA puede ser interpretado no solo como un método para análisis de texto en al ámbito de (*Information Retrieval)* sino también como una posible teoría sobre adquisición del conocimiento que ayude a entender los procesos cognitivos en humanos.

1.4. Fundamentos matemáticos

LSA representa el texto a analizar en una matriz que se conforma del siguiente modo. Cada fila de la matriz corresponde a una palabra en particular, y cada columna corresponde al pasaje o contexto en el que aparece. Las celdas de la matriz contienen la cantidad de veces que la palabra aparece en el pasaje. Luego se aplica el método de descomposición en valores singulares (SVD) que permite descomponer una matriz rectangular M (de t columnas por p filas) en el producto de tres matrices:

$$M = USV^T \tag{1.1}$$

Donde M es la matriz original, U es una matriz t x r de columnas ortonormales, V es una matriz p x r de columnas ortonormales y S es una matriz de r x r diagonal ordenada de manera decreciente. Los componentes de la matriz S son los valores singulares, y las matrices U y V corresponden a los vectores de palabras y de pasajes, respectivamente. Puede demostrarse que es posible descomponer cualquier matriz rectangular de esta manera [17].

La dimensión de la matriz original puede ser reducida durante su reconstrucción utilizando solamente algunos de los coeficientes de la matriz diagonal S. Tal como se muestra en (1.2), la matriz es reconstruida usando solo los primeros k valores singulares.

$$M_k = U_k S_k V_k^T \tag{1.2}$$

Esto es equivalente al método de aproximación k-dimensional en el sentido de mínimos cuadrados; M_k es la mejor aproximación a M utilizando mínimos cuadrados. El proceso de reducción de dimensiones es punto clave del método. En base a ello, LSA es capaz de extraer los principales componentes del texto a analizar.

Se describe a continuación un ejemplo simple de aplicación del método [17], utilizando pasajes de texto que corresponden a nueve artículos técnicos sobre "*Human Computer Interaction*" y cuatro sobre "*Graph theory*". La figura 1.1 muestra el texto original, los títulos de ambos temas aparecen agrupados en c y m respectivamente. La matriz M se conforma según se muestra en la figura 1.2.

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for 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, ordered *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*

Figura 1.1: Texto usado para la aplicación del método propuesto por Landauer y otros en [17]

	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

Figura 1.2: Matriz M, Tomado de la propuesta de Landauer y otros en [17]

En las figuras 1.3, 1.4, 1.5 se presenta el resultado de la descomposición en valores singulares y en la figuran 1.6 se muestra la reconstrucción de la matriz M utilizando solamente dos dimensiones, es decir utilizando solo las dos primeras columnas de las matrices.

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

Figura 1.3: Matriz U. Tomado de la propuesta de Landauer y otros en [17]

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

Figura 1.4: Matriz S. Tomado de la propuesta de Landauer y otros en [17] (las celdas no mostradas tienen valor nulo)

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

Figura 1.5: Matriz V traspuesta. Tomado de la propuesta de Landauer y otros en [17]

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Figura 1.6: Matriz M reconstruida. Tomado de la propuesta de Landauer y otros en [17]

En términos generales, la matriz reconstruida permite describir un segmento de texto a partir de la cuantificación de las palabras que lo componen. A su vez, cada palabra puede verse como la suma de sus aportes en cada segmento. A partir de estas cuantificaciones, pueden estimarse relaciones de similitud entre palabras o entre pasajes de texto. Esta idea se describe con más detalle a continuación comparando las filas correspondientes a las palabras *human*, *user* y *minors*. En la matriz original se observa que la palabra *human* nunca aparece en el mismo pasaje que *user* y *minors*. El coeficiente de correlación Spearman [17] entre *human* y *user* es -0.38 y entre *human* y *minors* 0.29. Sin embargo, en la matriz reconstruida los coeficientes de correlación se transforman en 0.94 para *human-user* y -0.83 para *human-minors*. Pese a que *human* y *user* nunca aparecen en el mismo pasaje, el procedimiento de reducción de dimensiones permite descubrir relaciones subyacentes. Como los contextos en los que aparecen las palabras tienen un significado similar, LSA las representa como similares y lo opuesto ocurre para *human* y *minors*.

En lo que refiere a comparaciones de similitud de palabras o pasajes, la manera más simple es calcular el coseno entre los vectores que resultan de la descomposición en valores singulares y su posterior reducción de dimensiones. Adicionalmente, pueden implementarse otras medidas de similitud como distancia Euclideana o "city block distance".

La primera pregunta que surge es ¿por qué la reducción de dimensiones permite inferir relaciones de similitud con mejores resultados que utilizando los datos

originales?. Los impulsores de LSA sostienen que esto se debe a que utilizando para la reconstrucción un espacio de las mismas dimensiones que las utilizadas para la generación del texto, existe una mayor probabilidad que palabras cercanas sean similares. Dicho de otro modo, la reducción de dimensiones logra reducir el ruido asociado a la variabilidad y a la información innecesaria agregada, revelando la estructura semántica subyacente. Palabras similares en su significado se encuentran cercanas en el nuevo espacio vectorial de dimensiones reducidas, más allá de su co-ocurrencia en el texto original. Y lo mismo aplica para documentos. Si su significado conceptual es similar, aparecerán cercanos en el nuevo espacio vectorial aunque difieran en las palabras que los conforman.

La elección del número de dimensiones a utilizar para la reconstrucción debe determinarse empíricamente, y permanece como un área de investigación. Por tanto, a la hora de utilizar LSA, se requiere de una validación externa para determinar el valor óptimo de dimensiones.

Existen estudios que han intentando encontrar un sustento matemático a los sorprendentes resultados de LSA, vinculando las bases matemáticas del procedimiento con estructuras que permiten modelar algunos procesos cognitivos que ocurren en el cerebro. En la siguiente sección se explica en detalle esta teoría.

1.5. LSA y memorias asociativas

Las memorias asociativas son un tipo de memoria (en biología y en computación) que se basan en la relación entre estímulos, ideas o experiencias. Es decir, se recuerda algo porque se lo asocia a otro recuerdo. Pese a que utilizando LSA puede reproducirse con alto grado de similitud el comportamiento humano para actividades como el aprendizaje del lenguaje, existe poca investigación sobre el vínculo matemático entre la teoría algebraica que sustenta el método y los procesos cognitivos que ocurren en el cerebro. Este capítulo resume estudios que proponen justamente una nueva perspectiva para relacionar memorias asociativas con métodos de búsqueda de información del tipo de LSA.

De manera aún más amplia, esta línea de investigación permite acercarnos a responder cuestiones que han intrigado a la humanidad desde siempre, en particular: el origen del pensamiento. Esto es, qué procesos ocurren en nuestro cerebro que nos permiten desarrollar pensamientos, que nos permiten conformar la 'mente'. Las primeras ideas desarrolladas en el siglo XX en este sentido giran en torno a las redes neuronales. Es decir, modelos matemáticos que simulan las características básicas de una neurona. Las memorias asociativas son parte fundamental de esta teoría.

Hacia 1943 Warren McCulloch y Walter Pitts establecen el primer modelo matemático neuronal basado en un funcionamiento binario con umbrales. El uso conjunto de muchas de estas unidades permitía conseguir redes neuronales para funciones complejas. Curiosamente, en la misma época se desarrollaron las bases de la computación siguiendo las mismas ideas, probablemente inspirados en ellas. De cualquier modo, no fue hasta la década de 1970 que estas ideas florecieron y sentaron las bases de las redes neuronales modernas. El elemento fundamental fue la posibilidad de simular matemáticamente el funcionamiento simultáneo de una gran cantidad de redes neuronales, y obtener con ello una memoria distribuida. Esto se consiguió a partir del uso del álgebra de matrices. La representación matricial permite asociar datos distribuyendo los componentes que los vinculan, entre todos los coeficientes de la matriz, exactamente lo que busca una memoria distribuida. Adicionalmente, estos coeficientes representan actividades o propiedades biológicas relacionadas con las sinapsis. Son estas ideas las que impulsaron los trabajos cuya base matemática se detalla a continuación, con el objetivo de vincular LSA con las memorias asociativas y en definitiva sentar las bases que permitan explicar las razones de los sorprendentes resultados obtenidos cuando esta teoría se utiliza para simular proceso cognitivos.

En el trabajo de E. Mizraji [18] se comienza estudiando memorias neuronales que pueden modelarse usando espacios vectoriales y álgebra matricial, tal como se mencionó previamente. Una memoria neuronal puede verse como un operador que vincula dos espacios vectoriales. Los elementos de los vectores representan las frecuencias de los potenciales de acción transportados por los axones. Esta

representación vectorial es apropiada para la actividad cerebral, ya que miles de axones actúan a la vez generando un procesamiento que ocurre esencialmente en paralelo. Si f y g son respectivamente la entrada y salida de la memoria asociativa *Mem*, esta relación puede expresarse como se observa en la ecuación 1.3.

$$\hat{g}_i = Mem(\hat{f}_i), \qquad i = (1, 2, ..., K) \tag{1.3}$$

Si bien existen fenómenos no lineales relacionados con los potenciales de acción, puede simplificarse el estudio considerando solamente la región lineal. En este caso, la memoria *Mem* puede representarse tal como se expresa en la ecuación 1.4.

$$M = \sum_{i=1}^{K} \hat{g}_i \hat{f_i^T} \tag{1.4}$$

La memoria M se comporta del modo descrito en la ecuación 1.5.

$$M\hat{f} = \sum_{i=1}^{K} \langle \hat{f}_i, \hat{f} \rangle \hat{g}_i, \tag{1.5}$$

Donde $\langle \hat{a}, \hat{b} \rangle = a^T b$ es el producto escalar entre los vectores **a** y **b**. Si el conjunto de vectores $\hat{\mathbf{f}}_i$ es ortogonal y la entrada $\hat{\mathbf{f}} = \hat{\mathbf{f}}_k$ entonces:

$$\mathbf{M}\hat{\mathbf{f_k}} = v_k \hat{\mathbf{g_k}} \tag{1.6}$$

Siendo $v_k = \left\| \hat{f}_k \right\|^2$, esto es la norma Euclidiana de vector $\hat{f}_k$ al cuadrado. En general si el vector $\hat{f}$ pertenece al espacio S_M (subespacio generado por el conjunto de vectores de entrada $\{\hat{f}_i\}$) la salida es una combinación lineal de los vectores $\hat{g}_i$. Si f es ortogonal a S_M entonces la salida es nula (no hay reconocimiento). Las memorias de matrices muestran aceptables niveles de reconocimiento cuando los vectores guardados son de grandes dimensiones y *sparse*. Esto es porque vectores de este tipo, bajo ciertas condiciones, pueden ser quasi-ortognonales. Por ello el producto escalar de la ecuación 1.5 es muy pequeño excepto para i = k.

Usualmente, el significado de un patrón no depende del largo de los vectores. Patrones completamente distintos se mapean en vectores ortogonales, patrones idénticos en vectores paralelos y patrones similares forman ángulos pequeños. Se puede normalizar los vectores sin perder la estructura de la matriz de memoria.

Si $f_i = \hat{f}_i \|\hat{f}_i\|^{-1}$ y $g_i = \hat{g}_i \|\hat{g}_i\|^{-1}$ se puede definir $\mu_i = \|\hat{g}_i\| \|\hat{f}_i\|$ y se puede reescribir la memoria matricial de la ecuación 1.4 como:

$$M = \sum_{i=1}^{K} \mu_i g_i f_i^T \tag{1.7}$$

Si los conjuntos de vectores de entrada $\hat{f}$ y $\hat{g}$ son ortonormales, la estructura de la ecuación 1.7 se corresponde con la descomposición en valores singulares de M. Los escalares μ_i son los valores singulares y los vectores $\hat{g}_i$ y $\hat{f}_i$ son los vectores

singulares de la matriz rectangular M. En general, la matriz la ecuación 1.7 no se corresponde con la descomposición en valores singulares. Pero si g_i y f_i son de dimensiones grandes, sparse y quasi ortogonales hay una similitud formal y numérica con SVD, que es la base de LSA.

Modelos más efectivos se han basado en operaciones multiplicativas que ofrecen una representación más flexible de procesos cognitivos complejos. En particular, mediante productos tensoriales, se pueden representar memorias asociativas dependientes del contexto. Estos pueden adaptarse a cambios funcionales y estructurales en las redes neuronales, lo cual hace que sean especialmente adecuados para simular funciones cognitivas como la memoria y la atención [19].

Los modelos neuronales tensoriales, al permitir asociaciones adaptativas y sensibles al contexto, amplían significativamente la capacidad representacional de anteriores modelos de memoria matricial. Sus vínculos con LSA son mostrados en [20]

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 2

Aplicaciones en medicina

2.1. LSA en registros médicos

La acumulación de datos en registros médicos ha sido históricamente desaprovechada, ya sea por las dificultades que implicaba salir del archivo manual de cada paciente o porque los sistemas digitales no lograban capturar toda la riqueza de la interacción médico-paciente. Los mecanismos de procesamiento de lenguaje natural previos al advenimiento de la Inteligencia Artificial (IA) generativa buscaban analizar y combinar los datos de muchos pacientes, para extraer patrones que revelasen relaciones entre condiciones médicas imposibles de observar manualmente. La utilidad de la identificación de patrones por estos medios son bien diversas: desde análisis de relaciones de patologías, hasta identificación de efectos secundarios en testeo de drogas. Por otro lado, suponen un magnífico soporte para las consultas médicas, ayudando a identificar posibles relaciones subyacentes y tratamientos apropiados. Este capítulo presenta una revisión de los principales trabajos de investigación realizados en las últimas dos décadas en torno al uso del análisis semántico latente (LSA) con fines clínicos. Los primeros estudios analizados corresponden a la etapa previa a la irrupción de la IA generativa; mientras que los más recientes exploran el potencial de LSA cuando se lo combina con tecnologías basadas en modelos generativos, abriendo así un novedoso campo de investigación.

En [21] se señalan las limitaciones de los sistemas de asistencia computacional para la toma de decisiones clínicas (CDS por sus siglas en inglés). En dicha época dependían en general de bases de datos construidas con estructuras hechas a mano y aprovechaban muy poco la riqueza de información que se presenta en los registros médicos. Dichas limitaciones se mantienen hasta hoy en día en muchos de los sistemas utilizados. Los autores sostienen que un sistema CDS que utilice lenguaje natural, y permita realizar inferencias sin intervención humana es de gran interés en esta área. Los CDS son en general extremadamente rígidos y restrictivos, mientras que la interacción doctor-paciente no lo es. Los síntomas rara vez son formulados de forma precisa y es habitual que no encajen en las estructuras de bases de datos médicas [22]. En consecuencia, se propone en [21] el uso de

LSA para medir la relación semántica entre síntomas y enfermedades. Dado que LSA no distingue algunas estructuras del lenguaje, tales como negaciones, muy importantes en el ámbito médico, los autores plantean una mejora al mecanismo tradicional de LSA. El texto es pre-procesado a los efectos de agregar una etiqueta a las palabras negadas. Por ejemplo, si las palabras "sin fiebre"figuran en el texto, son sustituidas por "fiebre-negada". De modo que cualquier palabra y su negación pasan a ser dos palabras distintas para LSA. Esto se consigue implementando un algoritmo que detecta en primera instancia condiciones clínicas y luego determina si estas aparecen negadas o no.

Para el estudio realizado en [21] se genera un espacio semántico en base a más de 3000 documentos médicos sobre enfermedades (unos 250 MB) extraído de `www.healthline.com`. A los efectos de evaluar el método, se eligen combinaciones de enfermedades y síntomas tomadas aleatoriamente de *Professional Guide to Signs and Symptoms* [23]. Se crean 1000 muestra aleatorias como base para la comparación (ver ejemplo en la figura 2.1). Luego de aplicar LSA, se mide la similitud del par enfermedad-síntoma real en base al coseno y se lo compara con el coseno de los pares enfermedad-síntoma aleatorios. Se utiliza un clasificador del tipo Naive-Bayesian para determinar el mejor umbral y poder así clasificar automáticamente pares reales y pares aleatorios.

Para evaluar los resultados se realiza un análisis de varianza (Anova) entre las medidas de similitud de los pares reales y aleatorios. Tal como se muestra en la Figura 2.1 , los resultados reflejan una diferencia significativa, con altos valores de coseno para los pares reales y más bajos para los aleatorios. Esto prueba que LSA es capaz de relacionar con buen grado de precisión los síntomas con las enfermedades correspondientes. Este trabajo es una muestra de las capacidades del método y del gran potencial de técnicas de este tipo para inferir conclusiones automáticamente a partir de texto médico sin la necesidad de bases de datos estructuradas o intervenciones manuales.

DISEASES, SYMPTOMS (ACTUAL RELATIONSHIPS), AND INCORRECT SYMPTOMS(RANDOM RELATIONSHIPS)

Disease	Actual Symptoms	Random Symptoms
Cirrhosis	abdominal pain, anorexia, constipation, diarrhea, edema, fever, hepatomegaly, jaundice, nausea, vomiting, weight change	tachycardia, blood, urinary frequency, malaise, pain, oliguria, decreased consciousness, abdominal weight
Erysipelas	cervical adenopathy, fever, headache, malaise, sore throat, vomiting	breast nodule, chills, bowel sounds hyperactive, cough, breath sounds decreased, abdominal tenderness
Gastritis	anorexia,fever,nausea,vomiting,weight change	abdominal pain, pain buttock and sacral, dysuria, dyspnea, urinary stream changes
Uremia	abdominal tenderness, anorexia, chest pain, diarrhea, nausea, oliguria or anuria, vomiting	fatigue, abdominal pain, rhonchi, deep tendon reflexes hypoactive, abdominal tenderness, crackles, vomiting
Ankylosing spondylitis	arm pain, back pain, decreased range of motion, fatigue, fever, malaise, nuchal rigidity	crackles, weight loss, skin mottling, anorexia, sore throat, eye pain, pain perineal

Figura 2.1: Comparación entre síntomas reales y aleatorios, según cada enfermedad. Tomado de la propuesta de Dada y otros [21]

Una línea de investigación distinta es planteada en [24], donde se propone un sistema de búsqueda de información médica denominado "modelo de enriquecimiento semántico". La idea es identificar el contexto de la búsqueda para luego comparar el o los términos del *query* con una matriz generada a partir de LSA. El proceso, representado en la figura 2.2, consta de los siguientes pasos. En primer lugar, se realiza una consulta relacionada con el contexto elegido para el estudio: cáncer de cerebro, en la ontología del *National Cancer Institute*. Como resultado se tiene un conjunto de términos que representan las clases relacionadas con la búsqueda. Luego estos términos son comparados con la matriz semántica de similitud (MSS). Dicha matriz fue obtenida a partir de la aplicación de LSA a artículos de un repositorio médico sobre el tema en particular (cáncer de cerebro). Cada fila y cada columna de MSS corresponde a un término. De modo que cada celda MSSij indica el grado de similitud (entre 0 y 1) entre el término i y el término j. Para determinar las palabras con mayor grado de similitud con un término cualquiera t, se buscan los valores más altos dentro de la fila o la columna correspondiente a t. En el caso que t no figure en la matriz, se realiza la búsqueda en base a sinónimos obtenidos de *WordNet*.

A los efectos de evaluar el sistema, los autores utilizaron como repositorio 1125 artículos de la revista *Plos One* publicada por *Public Library of Science*. Se comparan los resultados obtenidos de distintos *queries* entre el modelo propuesto y métodos existentes. Para medir la efectividad de la búsqueda se utilizan dos conceptos del ámbito de *Information Retrieval*: precisión y sensibilidad (o *recall*). La precisión se define como la cantidad de instancias recuperadas relevantes sobre la cantidad total de instancias recuperadas. Mientras que la sensibilidad corresponde a la cantidad de instancias relevantes recuperadas sobre la cantidad total de instancias relevantes. Si bien el trabajo no especifica claramente los distintos métodos contra los que se comparan los resultados, ver figuras 2.3 y 2.4, se demuestra, al menos cualitativamente, la relevancia de tomar en cuenta no sólo la sintaxis sino también la semántica de los términos de búsqueda en sistemas de recuperación de información.

Existen otros modelos de procesamiento de lenguaje natural basados en LSA que incorporan elementos estadísticos: pLSA (*probabilistic LSA*) y LDA (*Latent Dirichlet Allocation*). Estas teorías se basan en que los documentos pueden verse como un conjunto de temas distribuidos según alguna distribución de probabilidades (distribución de Dirichlet para LDA). En [25] y [26] se utilizan estas teorías para analizar registros médicos. Se muestra que estos sistemas pueden ser aplicados exitosamente en el análisis de historias clínicas, diagnóstico, prescripciones, así como en informes de imágenes médicas.

En línea con lo mencionado al inicio del presente capítulo, artículos más recientes abordan el desafío que implica aprovechar gran cantidad de datos médicos acumulados, más allá del registro puntual de un examen médico. La idea es que si se combinan datos de muchos pacientes con la misma afección, se podrían obser-

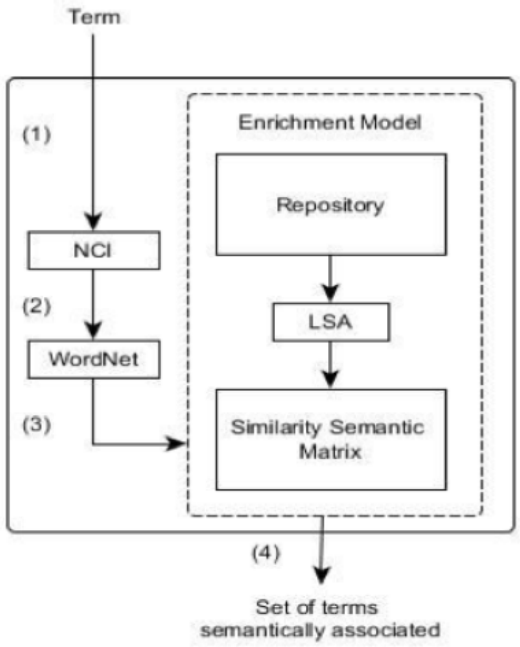

Figura 2.2: Proceso de búsqueda de información médica. Tomado de la propuesta de García y otros en [24]

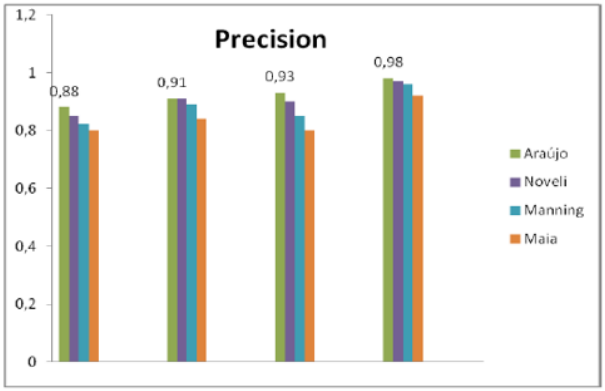

Figura 2.3: Precisión en la efectividad de la búsqueda. Tomado de la propuesta de García y otros en [24]

var patrones generales. Estos patrones pueden revelar relaciones entre condiciones médicas en un amplio conjunto de pacientes y cómo estas se relacionan con los códigos ICD-10-CM (*International Classification of Diseases, Tenth Revision, Clinical Modification*) publicados por la Organización Mundial de la Salud. A través de este análisis, se podrían identificar nuevas áreas de investigación médica y de salud pública, así como posibles efectos secundarios en las pruebas de medicamentos de fase IV.

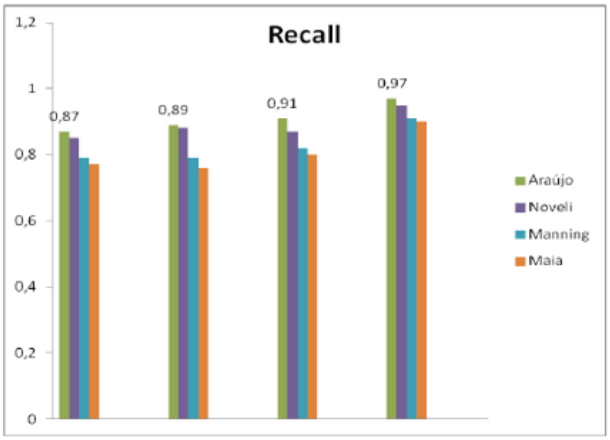

Figura 2.4: Sensibilidad en la búsqueda. Tomado de la propuesta de García y otros en [24]

En [27] se usa LSA en registros médicos relacionados con la insuficiencia cardíaca congestiva para identificar patrones entre términos y códigos ICD que aparecen para el mismo paciente. El estudio aplicó LSA para identificar asociaciones en registros médicos, de acuerdo al proceso indicado en la figura 2.5. Algunas asociaciones identificadas, como la relación entre hipertensión y obesidad, eran esperadas, validando la credibilidad del método. Sin embargo, se encontraron asociaciones menos esperadas, como entre 'hipertensión' y 'ciática'. Estas asociaciones inusuales pueden revelar casos que requieren atención especial o relaciones desconocidas anteriormente.
Este trabajo complementa la literatura existente sobre LSA en contextos médicos, al demostrar su potencial para asociar términos médicos y códigos ICD en informes médicos. Además, introduce una escala que muestra cuán cercanos están los términos entre sí en comparación con otros términos. Por ejemplo, los resultados sugieren que la hipertensión está más relacionada con el término benigna que crónica, por la cercanía de los términos hipertensión esencial benigna, y menos relacionada con el hipotiroidismo. Los hallazgos también indican que este método podría ayudar en la gestión del tratamiento médico al identificar casos inusuales que requieren atención especial.

Los datos utilizados en el estudio fueron proporcionados por IBX (Independence Blue Cross) en un proyecto conjunto con la Universidad de Drexel. Estos comprenden 32 124 archivos de texto obtenidos mediante OCR ('Reconocimiento Óptico de Caracteres' por sus siglas en inglés) de 1009 transcripciones médicas escaneadas correspondientes a 416 pacientes distintos con sospecha o diagnóstico de insuficiencia cardíaca congestiva durante 2013 y 2014. El análisis de los datos se realizó tal como estaban, sin corregir errores de ortografía o de OCR. Esta decisión se tomó intencionadamente para demostrar la capacidad del LSA incluso con datos sin procesar, ya que la corrección manual de los informes médicos puede ser costosa y propensa a introducir más errores.

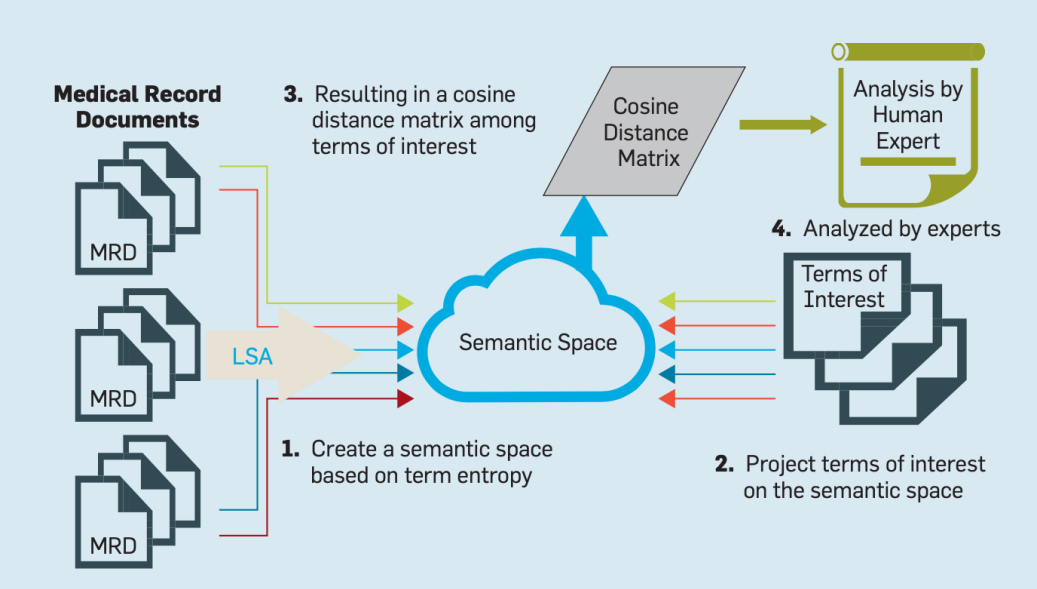

Figura 2.5: Proceso de aplicación de LSA relacionados con la insuficiencia cardíaca congestiva. Tomado de la propuesta de Gefen y otros en [27])

Este método crea una Matriz de Frecuencia de Términos y Documentos (TDM) donde las palabras se convierten a minúsculas, se eliminan las puntuaciones y palabras de parada comunes, y se conservan números que podrían representar códigos ICD. Posteriormente, se somete la TDM a una transformación TF-IDF y luego a una Descomposición de Valor Singular (SVD), reteniendo 100 dimensiones. A través de este método, se identificaron y analizaron términos relacionados con 'cardíaco' e 'hipertensión'. Los resultados mostraron asociaciones esperadas, como 'hipertensión' con 'hiperlipidemia', y otras menos obvias. Las asociaciones indirectas, que son difíciles de identificar manualmente, también se revelaron utilizando LSA. Por ejemplo, se encontraron conexiones entre hipertensión' y 'reflujo gastroesofágico'.

El análisis también señaló que el término 'cardíaco' estaba fuertemente asociado con partes específicas del cuerpo y procedimientos, mientras que 'hipertensión' tenía conexiones con una variedad más amplia de términos y condiciones coexistentes. El mapa de calor demostró cómo ciertos términos se relacionan entre sí, y también resaltó la diferencia entre problemas cardíacos agudos y la naturaleza crónica de la hipertensión. En resumen, el uso de LSA en este contexto demostró ser una herramienta valiosa para identificar asociaciones tanto directas como indirectas en registros médicos, lo que podría tener aplicaciones significativas en la investigación médica y la práctica clínica.

Los resultados de este tipo de investigaciones demuestran el potencial de aplicar LSA en tales contextos. Estos análisis se centran en identificar relaciones conocidas entre términos médicos y vincular diagnósticos y tratamientos. A partir de datos

más amplios, se podría desarrollar un modelo que permita la detección temprana de enfermedades y la identificación de casos excepcionales que requieran atención médica inmediata. Estos modelos podrían ser útiles para el monitoreo de medicamentos o para identificar efectos indirectos de medicamentos. Además, el análisis de registros médicos podría permitir comparaciones de diagnóstico y pronóstico entre poblaciones y sugerir investigaciones sobre enfermedades raras. La ventaja clave del LSA es que permite la clasificación ordenada de términos relacionados, ofreciendo así una perspectiva sobre la relación entre síntomas y enfermedades.

Trabajos más recientes utilizan las capacidades de LSA en conjunto con técnicas modernas de IA generativa. Por ejemplo, en [28] se analizan textos médicos simplificados mediante inteligencia artificial, y se evalúa la fidelidad semántica de los resultados ulilizando LSA. El trabajo tiene por objeto determinar si modelos generativos como ChatGPT -4.0 pueden reducir el nivel de complejidad sintáctica de textos clínicos sin comprometer su contenido. Para ello, se utilizaron como corpus 100 resúmenes de artículos científicos y 340 materiales educativos para pacientes (PEMs, por sus siglas en inglés). Estos textos fueron procesados por ChatGPT con la instrucción de mantener la información clave, pero adaptándola a un nivel de lectura de quinto grado. Se recurrió a LSA como método para cuantificar el nivel de preservación semántica entre los textos originales y los reformulados. Los resultados obtenidos mediante LSA fueron consistentes con las valoraciones subjetivas de expertos clínicos, confirmando que las transformaciones realizadas por el modelo conservaban el contenido temático esencial. En consecuencia, se demostró la viabilidad de LSA como mecanismo de validación automática en tareas realizadas por IA generativa, al menos en el ámbito en cuestión. Se concluye que LSA puede posicionarse como una solución de validación complementaria de bajo costo computacional, y alta interpretabilidad para integrarse en flujos de trabajo que combinan generación automática de texto con evaluación humana.

En [29] se presenta una evaluación comparativa de métodos de aprendizaje automático para a detección de determinantes sociales y conductuales de la salud (SBDH, por sus siglas en inglés) a partir de notas clínicas no estructuradas del conjunto de datos MIMIC-III. Los SBDH son factores del entorno social y del comportamiento individual que influyen de manera significativa en el estado de salud de una persona, en sus patrones de uso de servicios médicos y en los resultados clínicos que puede alcanzar. Estos factores incluyen aspectos como: inseguridad habitacional, inseguridad financiera, consumo de sustancias, violencia, abuso, etc. Muchas veces no están registrados en los campos estructurados de la historia clínica del paciente, pero sí pueden estar documentados de forma narrativa no estructurada. La identificación automatizada y sistemática de los SBDH permite comprender mejor las necesidades de los pacientes y diseñar intervenciones más efectivas. Los autores aplicaron Latent Semantic Indexing (LSI), la técnica predecesora de LSA, a más de dos millones de notas clínicas de 46 146 pacientes. Se construyeron representaciones vectoriales de pacientes y términos, y se evaluó la similitud entre ellos mediante coseno de ángulo, clasificando a los pacientes según 15 categorías

SBDH. Paralelamente, se emplearon modelos de IA generativa como GPT-3.5 y GPT-4 para realizar inferencias sobre la presencia de dichas categorías, utilizando prompts diseñados para retornar respuestas estructuradas. El rendimiento de ambos enfoques se evaluó utilizando conjuntos de referencia validados manualmente y métricas como precisión, recall y F1. LSI demostró un rendimiento robusto y comparable con modelos más recientes de IA generativa, destacándose por su capacidad de procesar la totalidad de los documentos sin limitaciones de contexto ni costos adicionales. Además, LSI fue particularmente ventajoso al no requerir fine-tuning ni entrenamiento externo.

Uno de los descubrimientos más interesantes es que, si bien los modelos de IA generativa ofrecen mayor precisión contextual y reconocimiento de lenguaje natural, enfrentan restricciones técnicas como la capacidad limitada del tamaño de entrada y la variabilidad en las respuestas [29]. En contraste, LSI, al operar de manera determinística y reproducible, superó a GPT-4 en seis de las nueve categorías evaluadas. El estudio concluye que la integración de métodos como LSI con modelos generativos como GPT puede constituir una estrategia costo-efectiva para la identificación masiva de SBDH en entornos clínicos reales. Proponen un enfoque híbrido en el cual LSI actúe como filtro inicial para reducir el corpus de texto a subconjuntos de interés, que luego puedan ser refinados por GPT, maximizando así la eficiencia computacional y la sensibilidad analítica. Este tipo de sinergia metodológica permitiría mejorar los modelos de predicción de riesgo y de necesidades sociales a nivel poblacional, al aprovechar tanto los campos estructurados como las notas de la historia clínica electrónica.

2.2. LSA en análisis de lenguaje

2.2.1. LSA en esquizofrenia

LSA ha sido utilizado inicialmente como método de búsqueda de información y categorización de texto. Sin embargo, su habilidad para descubrir estructuras escondidas y asociar objetos similares lo convierten en una técnica adecuada para modelar ciertos procesos que ocurren en el cerebro [30]. Esto abre un enorme campo de investigación en el ámbito de ciencias cognitivas. En lo que refiere a aplicaciones médicas basadas en análisis de lenguaje en particular, LSA ha sido propuesto como método para evaluar pacientes con esquizofrenia. Este análisis permite cuantificar el grado de coherencia del discurso del paciente, constituyendo una herramienta de evaluación objetiva del nivel de enfermedad. La siguiente reseña, resume algunos trabajos de investigación en esta línea.

En [31] se establece que pacientes con desórdenes de pensamiento (ThD por sus siglas en inglés) y en particular con esquizofrenia, muestran anormalidades en el uso del lenguaje. Los métodos tradicionales de evaluación de estos pacientes, comprenden análisis clínicos basados en entrevistas, incluyendo cuestionarios y preguntas de final abierto. Existen estándares para calificar el desempeño del paciente por parte del médico que realiza el diagnóstico, según se establece en [32]. De estas evaluaciones se puede derivar una calificación global que indica el grado de desorden en las habilidades de comprensión, lenguaje y comunicación (TLC por sus siglas en inglés). El valor 0 significa ausencia de desorden, y el valor 4 indica desorden extremo.

En [31] se realizan tres tipos de experimentos con pacientes catalogados con diferentes niveles de esquizofrenia, así como con un grupo de control. El primer experimento tiene que ver con asociación de palabras. El paciente tiene que expresar la primera palabra que le viene en mente, luego de leer 10 palabras consecutivas (“Dios”, “comida”, “niño”, “oscuro”, “difícil”, “alto”, “rey”, “mesa”, “lento” y “hombre”). Se espera que pacientes con ThD muestren una asociación de palabras menos usual que el grupo de control. El grado de similitud entre las palabras propuestas y la respuesta del paciente es medido utilizando LSA, en particular usando los algoritmos desarrollados en [33]. Se compara el resultado de los cosenos entre los vectores, haciendo análisis de varianza (Anova). Los resultados muestran que LSA es capaz de evaluar diferencias sutiles en grado de ThD. Adicionalmente, estos resultados se correlacionan significativamente con los obtenidos por evaluación clínica de ThD.

El segundo experimento busca medir la fluidez verbal. Se le pide al paciente que exprese la mayor cantidad de “animales” que se le ocurra, en un minuto. Esta es una prueba muy utilizada en pacientes con esquizofrenia, ya que habitualmente muestran una capacidad reducida en las respuestas esperadas. El experimento mide, utilizando LSA, la similitud entre palabras consecutivas. Los resultados pueden observarse gráficamente en la figura 2.6. El grupo de pacientes con ThD bajo

muestra un grado de coherencia más alto que el grupo de ThD alto. El método logra tener una mejor sensibilidad para este tipo de pruebas, que las técnicas tradicionales.

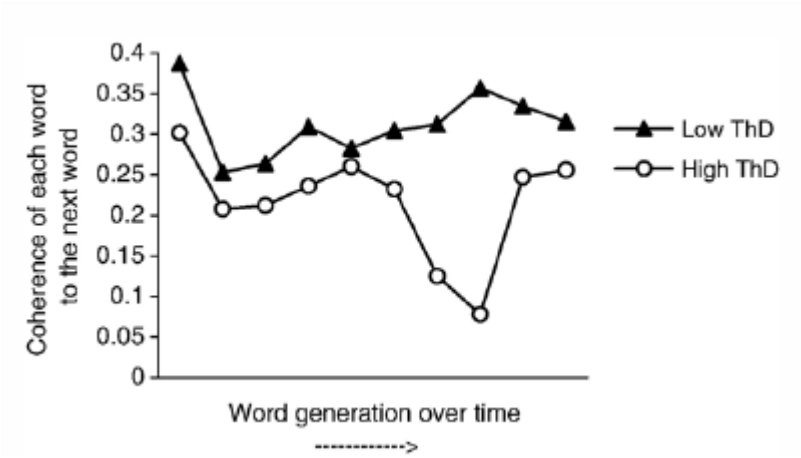

Figura 2.6: Coherencia entre palabras consecutivas de experimento que busca medir la fluidez verbal. Tomado de la propuesta de Elvevag y otros en [31]

El tercer experimento consiste en hacer que el paciente hable por algunos minutos, respondiendo ciertas preguntas o contando una pequeña historia. Se utiliza una ventana deslizante de unas pocas palabras y se mide, utilizando LSA, el grado de coherencia del relato. A medida que la ventana aumenta de 2 a 8 palabras, la coherencia disminuye en los pacientes altamente afectados por ThD.comparada con grupos de control, tal como se muestra en la figura 2.7.

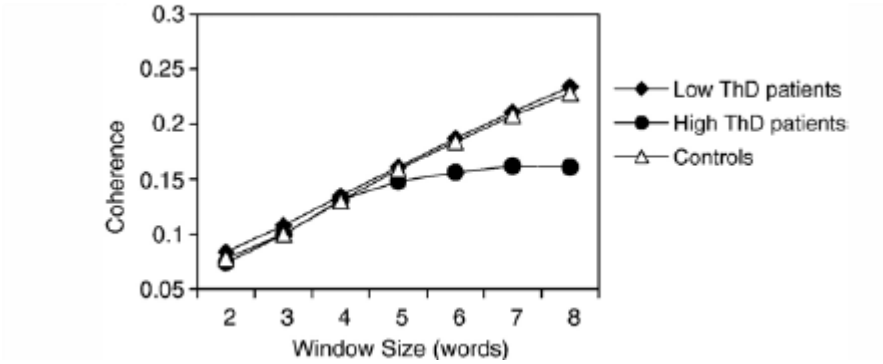

Figura 2.7: Coherencia de texto usando ventana deslizante. Tomado de la propuesta de Elvevag y otros en [31]

Como conclusión de estos estudios puede afirmarse que LSA es capaz de detectar pacientes con esquizofrenia con una exactitud razonable, así como evaluar la severidad de incoherencia en el uso del lenguaje de estos pacientes.

En [34] se analiza otra orientación del tema que destaca la importancia de detectar desviaciones o desórdenes sutiles en el discurso de los pacientes pero en etapas tempranas, previas a la aparición de los primeros síntomas. Para ello, se estudian grupos de pacientes con esquizofrenia pertenecientes a una misma familia, grupos de pacientes no familiares y grupos de control. Se utiliza LSA en conjunto con otras técnicas, con el objetivo de demostrar que los experimentos pueden diferenciar de manera exitosa discursos pertenecientes a pacientes y familiares de esos pacientes, de discursos pertenecientes a otros grupos. De esta manera se busca establecer que es posible detectar diferencias psicológicas sutiles que se expresan en el lenguaje, correspondientes al carácter hereditario de la enfermedad.

En [35] se plantea la posibilidad de utilizar LSA para realizar evaluaciones de narrativa psiquiátrica que simulen un nivel de experto. Esto es, realizar un sistema que automáticamente pueda extraer los conceptos clínicos del texto, sin la necesidad de un médico experto. Para ello, se usa LSA para generar un espacio semántico, del que se pueden extraer asociaciones entre términos psiquiátricos. Los resultados obtenidos, comparados contra evaluaciones de médicos expertos, muestran que se aproximan a los que se obtendrían por evaluación médica.

En resumen, puede afirmarse que las aplicaciones para detección de síntomas de esquizofrenia, son una interesante muestra de las capacidades de LSA y sus relaciones con los procesos cognitivos que se desarrollan en el cerebro. Existe un vasto campo de trabajo futuro en esta área, por ejemplo en detección de anomalías en el uso del lenguaje y la memoria causados por Alzheimer, desorden bipolar, o demencia semántica.

2.2.2. LSA en Alzehimer

En lo que refiere a Alzheimer (AD por sus siglas en inglés) es de particular relevancia reseñar algunos trabajos cuyo objetivo es detección temprana de síntomas. Esta enfermedad es la más común de las demencias caracterizada por un deterioro progresivo de las capacidades funcionales del individuo. Teniendo en cuenta los números de prevalencia de trastornos cognitivos leves (MCI: *Mild Cognitive Impairment*), dentro de los que se encuentran los primeros síntomas de AD, es de vital importancia contar con mecanismos automáticos de diagnóstico que permitan una detección en etapas tempranas. MCI tiene una incidencia de 13.2/1000 por año para personas mayores de 60 años [36] y el 50 % de estos individuos desarrollan algún tipo de demencia en un período de 5 años [37]. Dentro de la literatura de análisis de discurso para cuantificación de desordenes de lenguaje, los trabajos que emplean LSA son relativamente pocos. Entre de ellos, se destaca el *paper* presentado en [38]. En este artículo se utiliza la herramienta denominada *Coh-Metrix* desarrollada por *Memphis University* para adaptarla al portugués y adicionalmente agregarle funcionalidades basadas en LSA. Se busca distinguir individuos con AD o MCI y grupos de control (HC: *Healthy Control*) utilizando la historia de

"Cenicienta"por medio de parámetros cuantitativos. El estudio se realizó sobre 60 individuos divididos en 3 grupos: mAD (AD leve por sus siglas en inglés), aMCI (MCI amnésica, por sus siglas en inglés) y un grupo de control sano. Para la evaluación se utilizó un libro con 22 escenas secuenciales de la historia, sin subtítulos. Se instruyó a los individuos a narrar la historia para alguien que no la conociera sin límite de tiempo, con acceso a las imágenes del libro en todo momento.

Todos los discursos fueron transcritos y se utilizó la herramienta *Coh-Metrix* para extraer 73 diferentes métricas correspondientes a distintos aspectos lingüísticos. Se definieron previamente 28 proposiciones que describen las principales ideas de la historia (Figura 2.8).

List with the 28 propositions of the narrative

1. Cinderella's mother dies
2. Cinderella's father marries again
3. Cinderella and her father/her father's death
4. Rich girl
5. Envy (Stepmother and Daughters)
6. Cleaning the attic/Being a servant
7. Debauchery and wickedness
8. Invitation to the ball (dance)
9. They do not let Cinderella go to the dance
10. Animals help make the dress
11. Cinderella is happy with the dress
12. Stepmother's daughters tear Cinderella dress
13. Refuge in the forest/crying
14. Fairy godmother appears
15. Fairy godmother measuring Cinderella for new dress
16. Moment of transformation/pumpkin-carriage
17. Fairy godmother makes/gives a dress to Cinderella
18. Fairy Godmother warns Cinderella to return before midnight
19. Went to the dance
20. Prince meets Cinderella
21. Prince dances with Cinderella
22. Midnight/Cinderella loses shoe on ladder
23. Prince picks up shoe and looks for Cinderella
24. Stepmother holds Cinderella in the attic
25. The stepmother's daughter tries the shoe and does not fit
26. Animals free Cinderella
27. Cinderella tries the shoe and it fits
28. Marriage

Figura 2.8: Investigación de desórdenes del lenguaje, usando LSA, con textos de "La Cenicienta". Tomado de la propuesta de Borges y otros en [38]

Utilizando la herramienta y también análisis manual, se extrajeron las características macroestructurales del discurso presentadas en la Figura 2.9.

En lo que refiere a número de proposiciones reportados, los individuos con mAD presentaron cantidades menores en comparación los grupos de aMCI y HC, indicando discursos menos informativos. En los promedios de similitud entre todos

Features of Coh-Metrix-Dementia	
Latent semantic analysis (LSA)	
Average between adjacent sentences	Mean of similarity between pairs of adjacent sentences present in the text
Standard deviation between adjacent sentences	Standard deviation of the similarity between the pairs of adjacent sentences present in the text
Average similarity between all sentence pairs in the text	Mean of similarity between all sentence pairs in the text, not just the adjacent pairs
Standard deviation between sentences, all sentence pairs	Standard deviation of similarity between all sentence pairs in text
Average between adjacent paragraphs	Average similarity between adjacent paragraphs in the text
Standard deviation between adjacent paragraphs	Standard deviation of similarity between adjacent paragraphs in the text
Mean givenness of sentences	Average similarity between each sentence and all the text that precedes it. Average givenness of each sentence of the text from the second sentence onward. If the text has only one sentence, the metric is set to 0.0. Givenness of a sentence is defined as the LSA similarity between the sentence and all the text that precedes it.
Standard deviation of sentences givenness	Standard deviation of the similarity between each sentence and all the text that precedes it. Standard deviation of the givenness of each sentence of the text from the second sentence onward. If the text has only one sentence, the metric is set to 0.0. The givenness of a sentence is defined as the LSA similarity between the sentence and all the text that precedes it.
Mean span of sentences	Mean span of each sentence of the text from the second onward. If the text has only one sentence, the metric is set to 0.0. The span of a sentence, as well as givenness, is a way of measuring the closeness between a sentence and the context that precedes it. The difference, in simple terms, is that span seeks to capture similarity not only with the explicit content presented earlier in the text but also with everything that can be inferred from that content.
Standard deviation of sentence span	The standard deviation of the span of each sentence of the text, from the second onward. If the text has only one sentence, the metric is set to 0.0.
Semantic density	
Total idea density	Number of propositions present in the text, per every 10 words. For the calculation of the propositions, empty or disfluent propositions are not taken into account, and the calculation is done on the revised text for better performance of the extraction tool.

Figura 2.9: Características macroestructurales del discurso. Tomado de la propuesta de Borges y otros en [38]

los pares de oraciones se encontraron diferencias entre aMCI y mAD. El grupo de mAD presentó los valores más altos de desviación estándar entre oraciones. Asimismo los individuos con mAD presentaron una gran proporción de sentencias vacías. En la figura 2.10 se presentan los resultados relacionados con el análisis de LSA.

En resumen, este estudio presenta un método innovador para analizar la coherencia global utilizando métricas extraídas automáticamente y marcado de emisiones vacías. Se encontró que los individuos con Alzheimer tienen una mayor alteración en la coherencia global.

En el estudio, los individuos con mAD mostraron dificultades en la organización de ideas, errores en la estructura del texto y un aumento de oraciones vacías. Se encontró que había una mayor cantidad de 'modalizaciones' en el discurso mAD, lo que indica interrupciones en la estructura discursiva. Estas modalizaciones pueden reflejar esfuerzos del sujeto para interactuar o introducir contenido irrelevante debido a problemas en el componente semántico-pragmático del lenguaje.

El uso de herramientas computacionales, como Coh-Metrix-Dementia, reduce el tiempo y el esfuerzo requerido en comparación con los análisis manuales tradicionales, y puede ser útil en el diagnóstico y evaluación de individuos con declive cognitivo. En este estudio, se utilizó una herramienta informatizada para analizar discursos, encontrando que los individuos con mAD presentaban discursos con mayor deterioro macroestructural, menor coherencia y más modalizaciones en comparación con otros grupos.

Empty emissions analysis, total idea density analysis, latent semantic analysis, and number of modalizations

Item	Group			Kruskal-Wallis test (*P*)	Tukey multiple comparison test (*P*)	Results
	aMCI	mAD	HC			
Empty emissions						
Mean	11.40	27.10	12.55		(aMCI × mAD) (*P*) = .002*	
Median	8.50	22.50	9.00	.001*	(aMCI × HC) (*P*) = .964	mAD > aMCI = HC
Standard deviation	8.57	19.55	11.99		(mAD × HC) (*P*) = .005*	
n	20	20	20			
Total idea density						
Mean	0.38	0.32	0.37		(HC × aMCI) (*P*) = .799	
Median	0.39	0.33	0.38	.003*	(HC × mAD) (*P*) = .006*	HC = aMCI > mAD
Standard deviation	0.05	0.06	0.04		(aMCI × mAD) (*P*) = .001*	
n	20	20	20			
Average between adjacent sentences						
Mean	0.26	0.33	0.29		(HC × aMCI) (*P*) = .631	HC = aMCI
Median	0.27	0.33	0.30	.009*	(HC × mAD) (*P*) = .186	HC = mAD
Standard deviation	0.06	0.11	0.05		(aMCI × GDA) (*P*) = .025*	aMCI < mAD
n	20	20	20			
Average similarity between all sentence pairs in the text						
Mean	0.22	0.28	0.24		(HC × aMCI) (*P*) = .427	HC = aMCI
Median	0.22	0.28	0.25	.022*	(HC × mAD) (*P*) = .171	HC = mAD
Standard deviation	0.05	0.09	0.04		(aMCI × mAD) (*P*) = .009*	aMCI < mAD
n	20	20	20			
Standard deviation between all pairs of sentences						
Mean	0.21	0.24	0.21		(HC × aMCI) (*P*) = .812	
Median	0.21	0.24	0.21	.022*	(HC × mAD) (*P*) = .017*	HC = aMCI < mAD
Standard deviation	0.03	0.05	0.02		(aMCI × mAD) (*P*) = .075	
n	20	20	20			
Modalizations						
Mean	0.90	5.90	0.40		(aMCI × mAD) (*P*) = .002*	
Median	0.00	2.50	0.00	<.001*	(aMCI × HC) (*P*) = .934	mAD > aMCI = HC
Standard deviation	1.25	7.62	0.99		(mAD × HC) (*P*) = .001*	
n	20	20	20			

Abbreviations: aMCI, amnestic mild cognitive impairment; mAD, mild Alzheimer's disease; *n*, number of individuals; HC, healthy control.
*Statistical difference.

Figura 2.10: Resultados de la comparación entre diferentes grupos con distintas patologías, basado en el análisis de LSA. Tomado de la propuesta de Borges y otros en [38]

2.2.3. LSA en recuperación de recuerdos

En el trabajo [39] se investiga cómo el sueño posterior al aprendizaje afecta la recuperación de recuerdos, específicamente en términos de su coherencia semántica, utilizando LSA. Se cuantifica la relación semántica entre enunciados consecutivos producidos durante tareas de recuerdo libre. Los participantes visualizaron videos protagonizados por animales y fueron distribuidos en grupos con intervalos de retención de 12 o 24 horas, algunos incluyendo sueño y otros exclusivamente vigilia. De esta forma se evaluó cómo el sueño posterior al aprendizaje influye en la coherencia semántica del recuerdo.

LSA se aplicó para el análisis de textos de recuerdo libre, tras los períodos de vigilia o de sueño. Se introdujeron dos métricas: la coherencia semántica secuencial (SSC) que evalúa la similitud conceptual entre oraciones consecutivas, y la coherencia semántica temática (TSC) que refiere a la coherencia conceptual general entre todas las oraciones (ver figura 2.11). La primera se calculó como el coseno del ángulo entre los vectores de dos oraciones adyacentes. La segunda, calculando el promedio de coherencia semántica entre todas las combinaciones posibles de pares de oraciones dentro del texto.

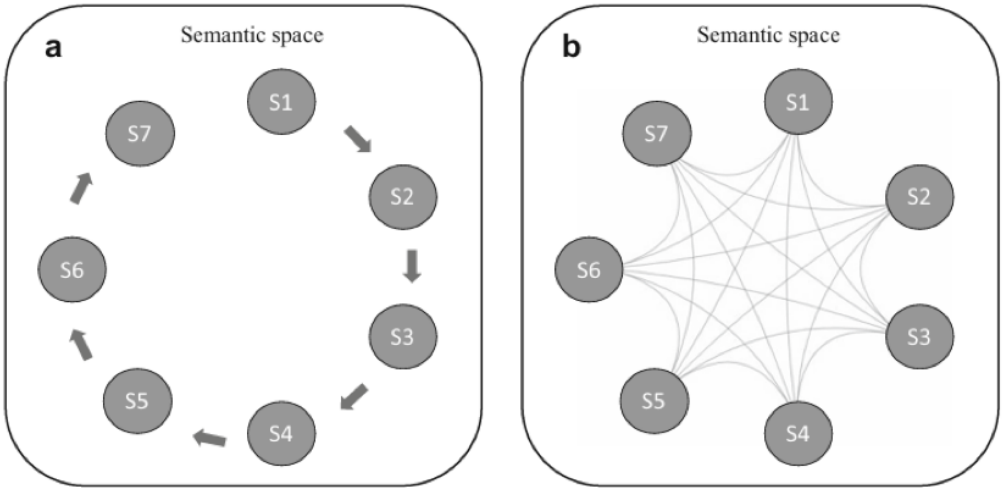

Figura 2.11: Efecto del sueño en los recuerdos, utilizando LSA. a) coherencia semántica entre oraciones consecutivas, b) coherencia semántica de todo el texto. Tomado de la propuesta de Ren y otros en [39]

LSA dio como resultados los niveles de coherencia secuencial y temática, señalando qué grado de coherencia y continuidad temática existía en los textos. Los resultados mostraron que el sueño redujo significativamente la coherencia semántica entre las ideas evocadas. Esta pérdida de coherencia semántica sugiere que el sueño no solo consolida memorias en términos cuantitativos, sino que también las transforma cualitativamente, generando una reorganización cognitiva de los recuerdos. Estos hallazgos refuerzan el papel del sueño como un proceso activo en la reestruc-

turación de la memoria, y demuestran la utilidad de LSA para analizar coherencia de textos.

Capítulo 3

LSA para análisis estadístico de distribución de cáncer

3.1. Introducción

El cerebro humano normalmente posee la capacidad de extraer y clasificar datos a partir de informaciones empíricas. Estos datos pueden ser organizados en tópicos con propiedades comunes (v.g. animales cuadrúpedos, teléfonos celulares, automóviles híbridos). Esta capacidad de clasificar la complejidad del mundo exterior en temas ha sido una guía en el desarrollo de modelos neuronales. En particular, las memorias matriciales poseen capacidades de clasificación. Es interesante que un procedimiento de minería de datos, el Análisis Semántico Latente (LSA), utilice una estructura matemática análoga a la de las memorias matriciales para procesar extensas colecciones de datos codificados. La posterior aplicación del LSA como modelo del aprendizaje en niños, refuerza aún más esta analogía. El poder computacional actual permite al LSA extraer tópicos a partir de información masiva cuya complejidad excede las capacidades usuales del cerebro humano.

En esta sección se muestra el trabajo que he realizado para aplicar LSA para el análisis de la distribución de distintas formas de cáncer en diferentes regiones de Uruguay (información cordialmente cedida por el Dr. E. Barrios y el Ing. Rafael Alonso de la Comisión Honoraria de Lucha contra el Cáncer de Uruguay). Una primera constatación es que los métodos usuales (v.g. correlaciones estadísticas) quedan en parte sesgados y dominados por la alta prevalencia de ciertas formas de cáncer que oscurecen la distribución de formas menos frecuentes. Esto también limita la habilidad del LSA para determinar si existen agrupaciones regionales de las formas de cáncer de menor frecuencia. Para ello, se estudió un sistema reducido con una matriz de datos que sólo retiene las categorías diagnósticas más raras. En base al uso del LSA en esta matriz reducida, pueden inferirse ciertas variaciones en la correlación entre distribuciones de cáncer no visibles con otros métodos de estudio. Este estudio constituye un trabajo inicial que demuestra la viabilidad de la propuesta planteada, la cual deberá ser complementada por futuras investiga-

ciones para afinar los resultados y realizar evaluaciones más exhaustivas.

3.2. Metodología

Los datos utilizados para el análisis se basaron en la tasa, ajustada por edad, obtenidos del Atlas de mortalidad por la enfermedad cáncer en el Uruguay para el período 2009 - 2013 [40]. Las tasas de mortalidad en países con distinta esperanza de vida brindan una idea falsa de mayor riesgo de muerte en los países con mayor número de adultos mayores. La estandarización por edad anula el efecto de esas distribuciones demográficas y permite comparaciones más equitativas. Para calcular la tasa ajustada por edad se calculan las tasas específicas de cada clase etaria; esto es el número de casos de muerte observados (d_i) en la iésima clase etaria sobre la población expuesta a riesgo en ese intervalo de edades en el período considerado (n_i):

$$t_i = d_i/n_i \qquad i = 1, 2, ..,18 \tag{3.1}$$

La tasa ajustada resulta de sumar las tasas específicas de cada intervalo etario ponderándolas por la proporción de individuos que integran ese intervalo en la población de referencia (distribución determinada por la OMS que puede observarse en la figura 3.1):

$$t_s = \sum_{i=1}^{18} w_i t_i \tag{3.2}$$

siendo t_s la tasa ajustada y w_i las proporciones de individuos en cada intervalo de edades.

En la figura 3.2 puede observarse la interfaz gráfica del desarrollo, programado en la plataforma Labview de National Instruments, para implementar los cálculos requeridos. Los distintos tipos de cáncer se encuentran codificados según la Clasificación Internacional de Enfermedades para Oncología [41]. Las tablas de datos indican la tasa ajustada por edad para más de 50 tipos de cáncer por Departamentos del Uruguay (ver figura 3.3).

La propuesta para aplicar LSA a estos datos implica tomar a cada tipo de cáncer como si fuera una palabra de la matriz término-documento y cada departamento como si fuera un pasaje de texto. Luego, es sencillo comparar incidencia de tipos de cáncer entre departamentos, es decir comparar pasajes de texto, en la idea original de LSA, o comparar tipos de cáncer entre sí, es decir comparar palabras. El estudio presentado en este capítulos se restringe a los tipos de cáncer con menor incidencia; se toman en cuenta solamente aquellos tipos que presentan tasas ajustadas pequeñas. El primer paso para implementar LSA es determinar empíricamente la cantidad de dimensiones a utilizar para la reconstrucción de la matriz. Para ello se ejecuta el algoritmo que implementa LSA con todas las dimensiones posibles y se calcula la correlación entre columnas de la matriz reconstruida para todos los casos. El resultado se muestra en la figura 3.4. En el eje vertical se indica el número de coeficientes de correlación mayor que 0.8 y en el eje horizontal las dimensiones utilizadas para la reconstrucción. Como es de esperar, utilizando

Intervalo de edad	Población mundial estándar
0 - 4	120
5 -9	100
10 - 14	90
15 - 19	90
20 - 24	80
25 - 29	80
30 - 34	60
35 - 39	60
40 - 44	60
45 - 49	60
50 - 54	50
55 - 59	40
60 - 64	40
65 - 69	30
70 - 74	20
75 - 79	10
80 - 84	5
85 y más	5
Total	**1000**

Figura 3.1: Mortalidad por cáncer en población mundial estándar. Extraído de Barrios y otros en [40]

pocas dimensiones todas las columnas están altamente correlacionadas y utilizando todas las dimensiones la correlación de las columnas es relativamente baja. Para buscar posibles datos correlacionados debe elegirse un punto intermedio. Para este estudio, en forma empírica, se eligieron 6 dimensiones.

Del mismo modo que en el uso típico de LSA, es posible realizar búsquedas o *queries*, que no son otra cosa que comparaciones entre vectores que representan términos o documentos. En este estudio esos vectores representan tipos de cáncer y departamentos. Por tanto puede realizarse un *query* directo de un tipo de cáncer para intentar determinar si existen estructuras relacionadas no visibles en los datos originales. En las figuras 3.5 y 3.6 se presenta un ejemplo de *query* directo para el tipo de cáncer C0 (cáncer de labio). En el resultado puede verse cómo se distribuye ese tipo de cáncer en los distintos departamentos según la reconstrucción de LSA. Valores nulos en los datos originales, se convierten en no nulos y esto puede ser un indicio de relaciones no visibles sin la aplicación de este método.

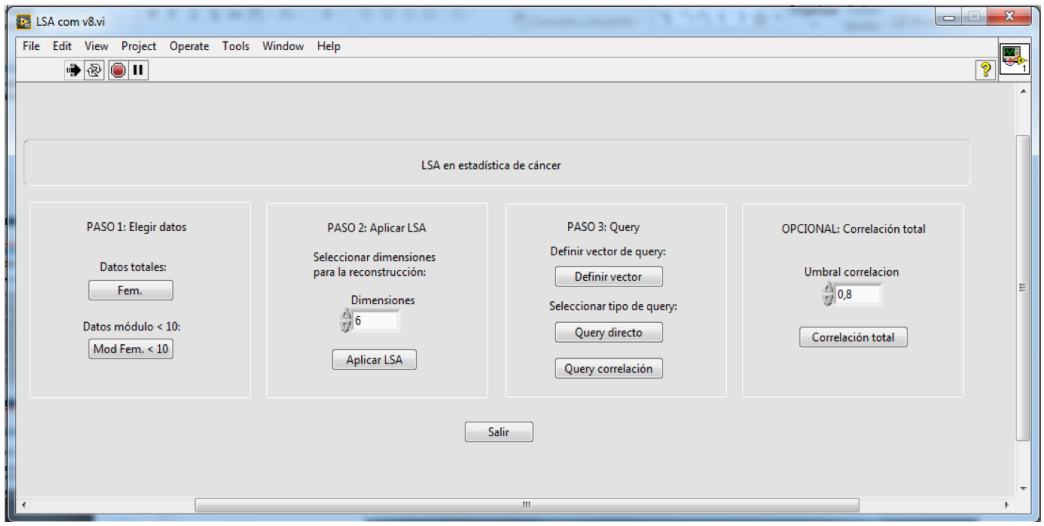

Figura 3.2: Interfaz gráfica del desarrollo programado en Labview.

	Departamentos													
Tipos de cancer	ARTIGAS	CANELO	CERRO	COLONIA	DURAZN	FLORES	FLORIDA	LAVALLEJ	MALDON	MONTEVI	PAYSAN	RIO	RIVERA	ROCHA
C0	0	0,0216441	0,430632	0	0	0,759013	0	0,113714	0	0,246053	0	0	0	0
C1-C2	0	0,551594	0	0,346396	0,15873	0	0,284333	1,02127	0,758326	0,679532	0,0988924	0,237643	0,306748	0,614419
C3.C6	0,363769	0,51108	0,24118	0,878479	0,694364	0,44843	0,522106	0,531379	0,194024	0,787376	0,379939	0	0,344274	0,565113
C7-C8	1,15953	0,492687	2,4679	0,496625	0	0	0,119104	1,03683	0,658906	0,413187	0,11409	0	0,445757	0,385356
C9-C14	0,475059	0,880037	0	0,539063	0,905379	1,10416	0	0,522944	0,92778	0,864888	1,34033	0,237643	0,677759	0,668002
C17	0,962013	0,543031	0	0,712603	1,57435	1,20744	0,111259	0,751554	0,279979	0,705295	0,599012	1,06853	0,227583	0,456204
C21	1,16526	0,657129	0,736272	1,32477	0,15873	1,19296	1,16333	0	2,4753	1,09208	0,655559	2,22619	1,02907	0,966527
C22	0,798967	1,18613	1,20371	0,563806	1,01182	1,79125	1,56617	1,34221	1,24219	0,74486	0,11409	0,922722	1,1585	0,737283
C26	0	0,950521	1,00858	0,286511	0,789446	0	0,399169	0,62825	0,769025	0,732215	2,00847	0	1,11935	1,60104
C30-C31	0	0,135415	0,115741	0	0,15873	0	0	0,63784	0,194024	0,188877	0,962771	0	0	0
C32	0,892744	0,585229	0,125439	0,46706	0,884564	1,49365	0,284333	1,00107	0,362086	0,751075	0,673974	1,05554	0,797066	0
C38-C39	0,156789	0,0389129	0,670279	0	0	0	0	0	0,492682	0,17333	0,560931	0	0,227583	0,236168
C40.C41	0	0,933453	0,873241	0,634898	1,32953	0	0,749766	0,113714	0,09997	0,838168	0	0	0,741198	0,58117
MEL EXCUT	0	0,404353	0,33389	0,470663	0	0	0	0	0	0,263636	0	1,32209	0,438452	0
PIELNM-	1,97398	0,552387	2,82637	1,28184	1,42016	0	1,9873	0,797938	0,320984	0,763919	0,677821	1,9956	0,25111	0,127259
MESOTELIOM	0	0	0	0	0	0	0	0	0,439301	0,159342	0,286512	0	0,726424	0,724966
SARCOMA	0	0,0702856	0	0	0	0	0	0,123305	0	0,163106	0	0	0	0

Figura 3.3: Tasas ajustadas por edad para distintos tipos de cáncer de pequeñas tasas, por departamento.

El segundo tipo de análisis consiste en comparar los vectores columnas entre sí para buscar posibles relaciones entre incidencia de cáncer en distintas zonas geográficas. Se busca analizar zonas geográficas de correlación baja en los datos originales. En la figura 3.7 se muestra, a modo de ejemplo, un caso de este tipo. La correlación entre los vectores correspondientes a Lavalleja y Treinta y Tres es 0.023 originalmente, pero aumenta a 0.540 en los datos reconstruidos. Este método muestra un potencial interesante, aplicado en zonas geográficas más grandes y con distribución de datos más granulares.

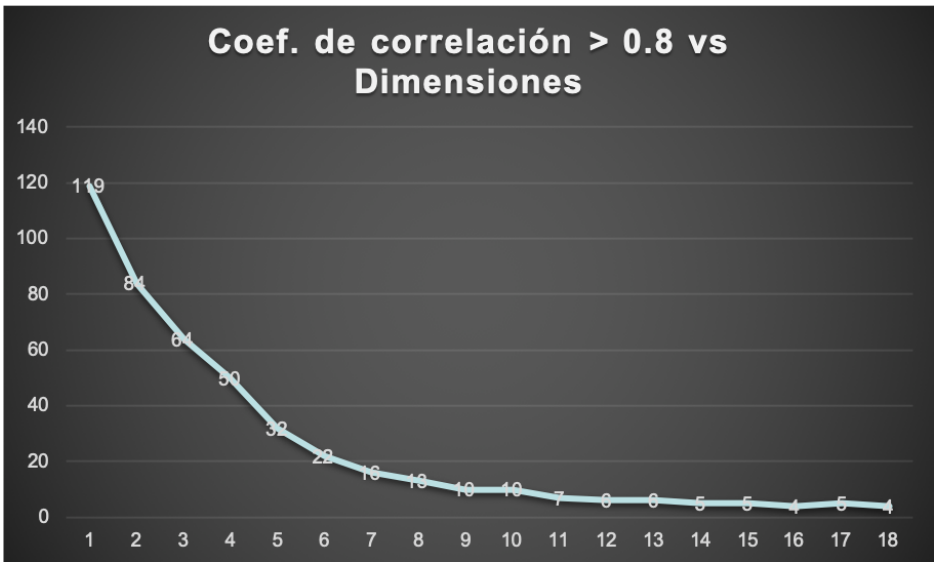

Figura 3.4: Relación entre el número de coeficientes de correlación mayor que 0.8 y el número de dimensiones para la reconstrucción.

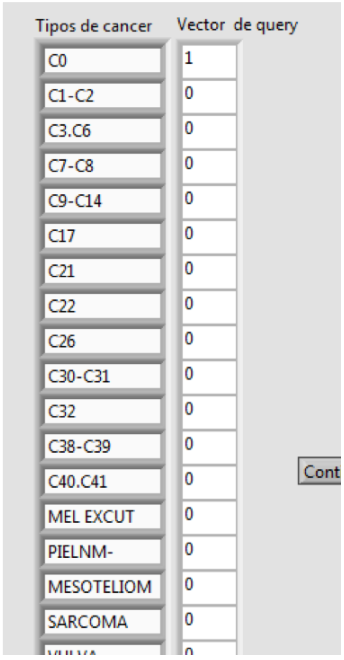

Figura 3.5: Comparaciones entre vectores que representan términos o documentos.

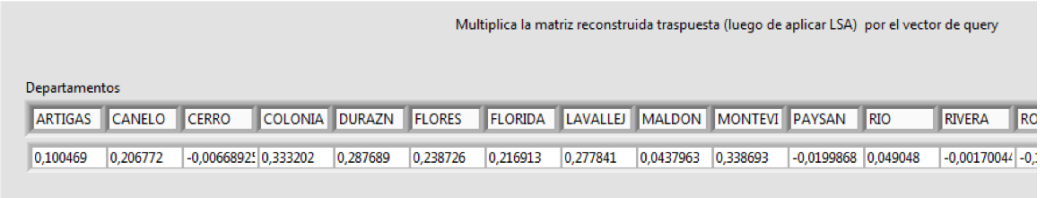

Figura 3.6: Resultado del query

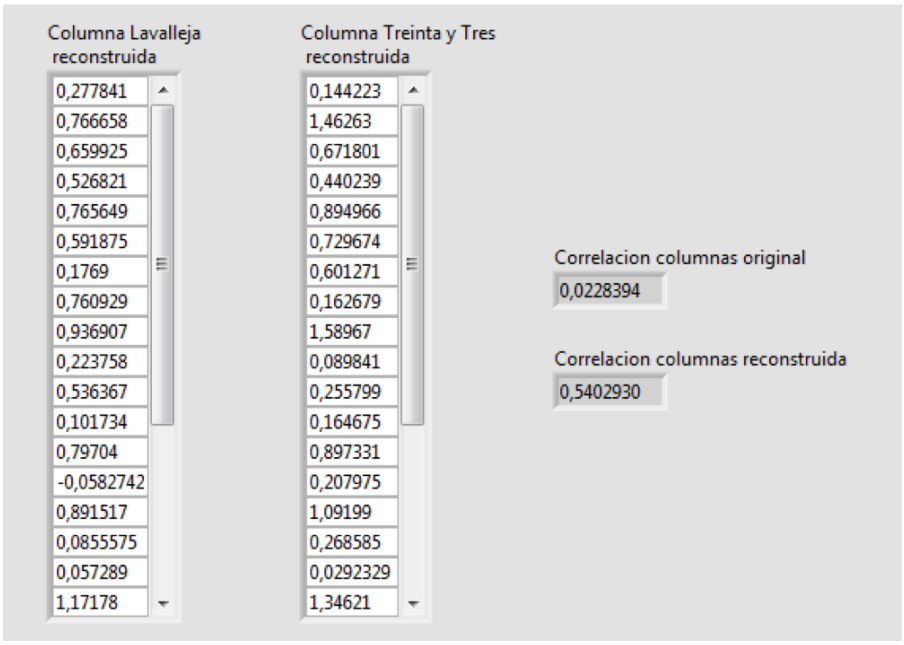

Figura 3.7: Departamentos no correlacionados originalmente

3.3. Resultados

El estudio demostró que LSA puede revelar relaciones no visibles en los datos originales, sugiriendo que ciertos tipos de cáncer poco frecuentes pueden tener distribuciones regionales específicas que pasan desapercibidas sin este análisis. Un ejemplo destacado se observa en los departamentos de Lavalleja y Treinta y Tres, donde la correlación entre ambos departamentos para el tipo de cáncer analizado, aumenta significativamente luego de aplicar LSA. Este resultado sugiere que ambos departamentos comparten un patrón latente en la distribución relativa del tipo de cáncer analizado, que no era detectable mediante métodos convencionales. Esto, indudablemente, debe complementarse con análisis desde el lado médico, para confirmar y determinar posibles causas de estas altas correlaciones. El enfoque planteado no solo facilita la identificación de patrones ocultos, sino que también abre la puerta a estudios más profundos y exhaustivos que consideren datos de geografías más amplias en otros países. Se plantea, también, la necesidad de recibir retroalimentación de expertos en estadísticas de cáncer para afinar los puntos de interés y mejorar la precisión de los resultados.

En resumen, la aplicación de LSA en este contexto tiene un gran potencial para avanzar en la comprensión de la distribución geográfica del cáncer y optimizar las estrategias de prevención y tratamiento. El trabajo propuesto se plantea como una primera aproximación metodológica. Para validar e interpretar estos patrones con mayor profundidad, se requiere de un análisis interdisciplinario junto a epidemiólogos y expertos en salud pública.

En particular, se proponen los siguientes puntos para trabajos futuros:

- Análisis más exhaustivo de los posibles clusters en la matriz reconstruida (enfocándose en los valores nulos en la matriz original).
- Comparación de resultados con datos de otros países (por ej.: Estados Unidos de Norteamérica). Se espera datos originales menos correlacionados.
- Análisis por filas, es decir definiendo a cada tipo de cáncer según su distribución geográfica.
- Feedback de especialistas en estadísticas de cáncer para definir puntos de interés.

Capítulo 4

LSA para correlación de eventos

4.1. Introducción

LSA ha sido utilizado inicialmente para la categorización y la búsqueda de información. Sin embargo, debido a los resultados significativos obtenidos, similar al procesamiento humano, LSA se ha convertido en mucho más que un simple método para analizar texto. En este trabajo, proponemos utilizar LSA para inferir el grado de similitud de los mensajes de eventos (syslog) descubriendo relaciones ocultas entre ellos. Usando ejemplos reales de mensajes de syslog, mostramos que LSA puede resaltar los mensajes más correlacionados por tema. Este método se puede utilizar para evitar sistemas de correlación de eventos complejos que generalmente necesitan firmas o definiciones de conjuntos de reglas y una gran experiencia para su configuración.

El protocolo Syslog (RFC 5424 [42]) se usa ampliamente para manejar mensajes de notificación de eventos en el ámbito de TI (Tecnologías de la Información). Los enrutadores, conmutadores, servidores y otros dispositivos de TI generalmente recopilan eventos, errores e información de depuración en el formato estandarizado por el protocolo. Los mensajes se almacenan localmente en los dispositivos o se envían a través de la red a un servidor centralizado. En las redes modernas se puede generar una gran cantidad de mensajes, lo que hace que el análisis de esta información sea una tarea difícil. Aunque existen correlacionadores de eventos sofisticados, en la mayoría de los casos los operadores de red analizan manualmente los mensajes de syslog después de que ocurre una falla para encontrar pistas sobre la causa del problema. Un solo evento o falla en la red puede afectar a muchos dispositivos, y en cada uno de ellos puede afectar diferentes características dependiendo de la función del dispositivo en la red. La correlación adecuada de esos mensajes no es una tarea trivial, incluso para los expertos en redes.

En este estudio (ver [43]) investigamos la posibilidad de utilizar LSA para correlacionar los mensajes de Syslog enfocándonos en una implementación simple de los algoritmos. El trabajo se centra en mensajes de errores y eventos en el ámbito

de redes de telecomunicaciones, pero es aplicable también al ámbito médico, en cualquier sistema de dispositivos que envíe mensajes de eventos para reportar estados de funcionamiento.

Los sistemas de correlación de eventos se utilizan ampliamente en operadores, redes empresariales y sistemas de seguridad. Hay muchas soluciones comerciales e incluso soluciones de código abierto como SEC [44]. La mayoría de las soluciones disponibles se basan en algún tipo de definición de conjunto de reglas. Se necesita información de sintaxis previa sobre los mensajes para encontrar patrones del evento. Además, se necesita una amplia experiencia para la configuración de esos sistemas. En [45], los autores proponen un nuevo sistema para transformar y comprimir automáticamente mensajes syslog de bajo nivel y mínimamente estructurados en eventos de red significativos y prioritarios, de alto nivel, que no requieren conocimiento previo del dominio o comportamiento esperado de la red. Sin embargo, usan firmas y diccionarios para agrupar eventos. Se presenta una propuesta diferente en [46], donde el grado de apariencia global del tipo de mensaje en un conjunto de datos dado se utiliza para resaltar los mensajes importantes. La misma idea es utilizada por muchas investigaciones en el campo de recuperación de información. Las funciones de ponderación global se utilizan para preprocesar los datos y asignar valores más bajos a información menos importante. Luego, se implementan técnicas de minería de texto para la recuperación de información. Siguiendo esta línea, proponemos en este trabajo la aplicación de LSA.

Nuestro estudio se centra en el uso de LSA para descubrir relaciones ocultas entre los mensajes de syslog, que pueden usarse para automatizar el proceso de correlación y ayudar en la detección de eventos. Las redes actuales enfrentan nuevos desafíos con la implementación de las tecnologías SDN (*Software Defined Networking*). SDN propone un nuevo paradigma que desacopla el plano de control del plano de datos de las redes. La centralización de la inteligencia de red propuesta por estas nuevas ideas requiere altas capacidades de análisis en los controladores (los dispositivos a cargo del control de la red). En consecuencia, las técnicas de minería de datos de texto, como LSA, se pueden utilizar para expandir el conocimiento del controlador sobre la red, proporcionando una perspectiva significativa para los operadores y administradores de red.

4.2. Metodología

Los mensajes de Syslog son generados por diferentes tipos de dispositivos informáticos y de comunicaciones para indicar el estado de algún proceso interno o cualquier evento que deba ser reportado. Estos eventos pueden estar relacionados con el funcionamiento normal, fallas, alarmas, información de depuración y otros. Hay algunos RFC (*Request for comments*) que proporcionan un marco para la estandarización del protocolo (RFC 3195, RFC 5424, RFC 5425, RFC 5427 y otros). En particular, RFC 5424 describe el formato estándar para los mensajes de syslog y las formas de transporte que deben ser compatibles con la transmisión. En la figura 4.1 se muestra un mensaje típico de syslog enviado por un enrutador. Está compuesto por un campo de marca de tiempo, nombre de *host* del dispositivo, módulo que generó el mensaje dentro del dispositivo, gravedad (que indica el nivel de importancia del evento), descripción y mensaje. La estructura de cada campo está bien descrita en el RFC. El campo del mensaje es un texto de forma libre que proporciona información sobre el evento. Cada proveedor define la información transportada en este campo. Por lo general, es una oración no estructurada que proporciona información importante relacionada con el evento.

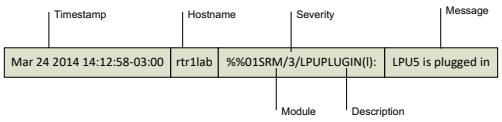

Figura 4.1: Formato de mensaje Syslog.

En este trabajo solo nos interesan los campos de "descripción"z "mensaje"del mensaje de syslog. Aunque los otros campos contienen información importante que debería usarse en sistemas de detección y correlación de eventos reales, queremos mostrar que es posible agrupar los mensajes basándonos exclusivamente en los campos de texto libre.

Mensajes de syslog reales de un fabricante en particular fueron seleccionados aleatoriamente para tres tecnologías de red distintas: ISIS, BGP y SSH. Las muestras de mensajes BGP pueden observarse en la figura 4.2. Antes de aplicar LSA, se eliminan palabras comunes como: *the, that, a, to, from, on, is* que no llevan información útil.

La matriz M se construye como se describe en la sección 1.4. Las columnas de la matriz corresponden a cada mensaje de syslog, y las filas a cada palabra. En las celda se indica la frecuencia con que aparece la palabra en el mensaje. Se aplica posteriormente SVD a la matriz M. Luego se reconstruye M tomando solo dos dimensiones (k = 2). Las primeras ocho filas de la matriz original y la matriz reconstruida M_k se muestran en las figuras 4.3 y 4.4. Las columnas de la matriz están formadas por tres mensajes de registro del sistema ISIS (I1 - I4), tres mensajes BGP (B1 - B3) y cuatro mensajes SSH (S1 - S4). El objetivo es comparar

Module	Severity	Description	Message
BGP	6	RECV_NOTIFY	The router received NOTIFICATION message from peer [neighbor-address]. (ErrorCode=[ULONG], SubErrorCode=[ULONG], BgpAddressFamily=[STRING], ErrorData=[STRING])
BGP	6	SNMP_PEER_SHUTDOWN	An SNMP command was received to suspend the peer session for [peer-address]. (InstanceName [STRING])
BGP	3	STATE_CHG_UPDOWN	The status of the peer [peer-address] changed from [previous-state] to [current-state]. (InstanceName=[STRING], StateChangeReason=[STRING])

Figura 4.2: Syslog de tecnología de red BGP.

la correlación entre los mensajes del mismo grupo con los mensajes de los otros grupos.

	I1	I2	I3	I4	B1	B2	B3	S1	S2	S3	S4
ISIS	1	3	1	1	0	0	0	0	0	0	0
manual	1	0	0	0	0	0	0	0	0	0	0
area	2	0	0	0	0	0	0	0	0	0	0
address	2	0	0	0	0	0	0	0	0	0	0
instance	1	0	0	0	0	0	0	0	0	0	0
invalid	2	0	0	0	0	0	0	0	0	0	0
failed	0	2	0	0	0	0	0	2	0	0	0
initialize	0	1	0	0	0	0	0	0	0	0	0

Figura 4.3: Reconstrucción de la matriz M

	I1	I2	I3	I4	B1	B2	B3	S1	S2	S3	S4
ISIS	1,36	2,61	0,97	0,86	0,12	0,06	0,10	0,80	-0,04	-0,03	-0,06
manual	0,18	0,34	0,10	0,07	-0,04	-0,08	-0,01	0,11	-0,02	-0,01	-0,01
area	0,36	0,67	0,21	0,14	-0,07	-0,16	-0,03	0,22	-0,03	-0,01	-0,03
address	0,36	0,67	0,21	0,14	-0,07	-0,16	-0,03	0,22	-0,03	-0,01	-0,03
instance	0,18	0,34	0,10	0,07	-0,04	-0,08	-0,01	0,11	-0,02	-0,01	-0,01
invalid	0,36	0,67	0,21	0,14	-0,07	-0,16	-0,03	0,22	-0,03	-0,01	-0,03
failed	0,89	1,66	0,54	0,40	-0,11	-0,27	-0,04	0,53	-0,06	-0,03	-0,06
initialize	0,34	0,63	0,21	0,16	-0,03	-0,08	-0,01	0,20	-0,02	-0,01	-0,02

Figura 4.4: Matriz M_k

Para comparar la correlación entre mensajes, calculamos el coeficiente de correlación Pearson (r), comúnmente utilizado para la comparación de vectores, entre las columnas de la matriz original y la reconstruida. Estas columnas representan cada uno de los mensajes de syslog. La figura 4.5 muestra la correlación original,

y la figura 4.6 la reconstrucción bidimensional. En aras de la claridad, solo mostramos las correlaciones de los dos primeros grupos (ISIS y BGP). La correlación de mensajes del mismo grupo en la matriz reconstruida es significativamente mayor que la original. El promedio de (r) para los mensajes originales (comparando mensajes del mismo grupo), es 0.07 pero aumenta a 0.78 después de aplicar LSA. Particularmente, el promedio r para mensajes ISIS aumentó de 0.03 a 0.67, y para BGP aumentó de 0.16 a 1.00. El promedio (r) entre mensajes de diferentes grupos es -0.05 en los datos originales, y permanece en un valor bajo (0.06) después de la reconstrucción.

	I1	I2	I3	I4	B1	B2
I2	0,09					
I3	-0,05	0,21				
I4	-0,08	0,09	-0,08			
B1	-0,15	-0,14	-0,05	0,14		
B2	-0,17	-0,16	0,01	-0,04	0,20	
B3	-0,13	-0,12	0,02	0,14	0,02	0,24

Figura 4.5: Correlación de la matriz original.

	I1	I2	I3	I4	B1	B2
I2	1,00					
I3	0,75	0,81				
I4	0,28	0,37	0,84			
B1	-0,43	-0,34	0,27	0,75		
B2	-0,48	-0,40	0,21	0,71	1,00	
B3	-0,39	-0,30	0,31	0,78	1,00	0,99

Figura 4.6: Correlación de la matriz reconstruida.

Este aumento significativo en el coeficiente de correlación tras aplicar LSA, de un promedio de 0.07 a 0.78 para mensajes del mismo grupo, indica que la técnica es capaz de identificar de forma efectiva patrones semánticos latentes en los mensajes syslog, incluso cuando estos no comparten una estructura léxica obvia. Dicho de otro modo, LSA permite revelar relaciones entre mensajes que, a simple vista, no parecerían similares, pero que en realidad se refieren a eventos conceptualmente relacionados.

El aumento de un orden de magnitud refleja que, en el espacio semántico reducido, los mensajes de un mismo tipo (por ejemplo, BGP o ISIS) se agrupan de manera mucho más coherente. Esto valida la capacidad de LSA de agrupación en forma automática sin necesidad de reglas explícitas.

Por otro lado, el hecho de que la correlación entre mensajes de distintos grupos permanezca baja antes y después del análisis es también importante. Esto sugiere

que LSA preserva las diferencias entre mensajes de contextos diferentes, evitando agrupaciones artificiales o falsas asociaciones. Esto es especialmente relevante para aplicaciones prácticas donde se desea minimizar los falsos positivos en la detección de eventos correlacionados.

4.3. Conclusiones y trabajos futuros

LSA es una teoría prometedora y se puede utilizar con éxito para la correlación de syslog y la detección de eventos de red. Se obtuvieron buenos resultados para este trabajo como se muestra en la sección anterior. El método puede distinguir entre mensajes de diferentes temas generales. Incluso para un pequeño conjunto de muestras, está claro que LSA puede contribuir significativamente en el análisis de la información de syslog. Como continuación de esta línea de investigación se propone como trabajo futuro determinar la capacidad de LSA para la detección de relaciones más sutiles. Por ejemplo, para la detección de diferentes mensajes de syslog de diferentes temas pero relacionados con el mismo problema.

En particular, se propone extender las pruebas preliminares realizadas para esta investigación en los siguientes puntos:

- Extensión del espacio semántico: LSA permite analizar una gran cantidad de texto y generar grandes espacios semánticos. En el caso del análisis de syslog, esta base se puede construir incluyendo la documentación técnica de los fabricantes de dispositivos.
- Ponderación automática: funciones de ponderación de frecuencia de término como *tf-idf* pueden usarse para el preprocesamiento, lo que permitiría obtener mejores resultados después de aplicar LSA.
- Comparación de vectores: se deben probar diferentes técnicas para la comparación de vectores a fin de determinar a qué grupo pertenece el mensaje syslog. Los algoritmos de agrupación de aprendizaje automático, como *Support Vector Machines (SVM)*, pueden usarse para ese propósito.

Adicionalmente a estas líneas de trabajo, LSA puede integrarse en arquitecturas de SIEM (*Security Information and Event Management*) en ámbito hospitalario. En términos generales, estas soluciones permiten recolectar datos de eventos y registros desde múltiples dispositivos de red, servidores, aplicaciones, etc., y correlacionarlos para detectar amenazas, eventos anómalos y actividades sospechosas. En los últimos tiempos, este tipo de soluciones se han vuelto críticas para las organizaciones y se han transformado en los motores de analítica de las infraestructuras de TI. En el contexto médico conviven múltiples tipos de dispositivo que generan eventos no solo en formato Syslog, sino también en sus propios formatos, por ejemplo HL7 o DICOM. La capacidad de agrupar y correlacionar eventos semánticamente relacionados puede resultar de mucha utilidad tanto para detección de fallas o de incidentes de ciberseguridad. LSA podría utilizarse como herramienta de pre-análisis para aliviar la carga sobre los motores de analítica de los SIEM y adicionalmente facilitar la detección de patrones o relaciones complejas entre los datos.

Esta estrategia, aunque no en particular para LSA, ha sido validada en estudios sobre ciberseguridad médica, tal como lo plantea el proyecto SAFECARE [47],

donde se propone un modelo para la recolección y análisis de logs médicos para detección temprana de amenazas. Del mismo modo, trabajos como el de [48] muestran que el uso combinado de técnicas de minería de datos o *machine learning* y motores de SIEM basados en reglas o aprendizaje automático permiten mejorar la precisión de los sistemas.

En suma, existe una prometedora línea de investigación que combina herramientas y soluciones de SIEM existentes con LSA, para mejorar la eficiencia de los sistemas tanto en lo que implica detección de anomalías que puedan formar parte de amenazas de ciberseguridad, así como enriquecimiento y mejoras en el análisis de eventos en tiempo real, en particular en el ámbito médico hospitalario.

Capítulo 5

LSA como teoría de Data Mining

5.1. Introducción

El lenguaje es la base de la mayoría de las comunicaciones e interacciones humanas. No es solo un medio para lograr objetivos compartidos: tiene un papel central en el pensamiento humano, en la formación de relaciones sociales y emocionales, en la identidad individual y social, y en el registro y desarrollo del conocimiento. Las lenguas, ya sean habladas o escritas, emergen en todas las sociedades humanas y, a pesar de su diversidad, presentan similitudes sorprendentes. Son sistemas complejos pero eficientes, que los niños adquieren rápidamente y que evolucionan según las necesidades de las comunidades. El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es una subárea de la inteligencia artificial enfocada en el lenguaje. Su objetivo es permitir que las computadoras comprendan y generen lenguaje humano.

En la Figura 5.1 se detallan los principales algoritmos de *Machine Learning* divididos en cinco grandes grupos: *Supervised Learning* (Aprendizaje Supervisado), *Unsupervised Learning* (aprendizaje no supervisado), *Neural Networks* (Redes Neuronales), *Reinforcement Learning* (Aprendizaje por refuerzo), y *Ensamble Learning* (Aprendizaje de conjunto).

Los algoritmos de aprendizaje supervisado parten de un conjunto de datos de entrenamiento como entrada y un conjunto de etiquetas o 'respuestas correctas', para cada conjunto de entrenamiento, como salida. Luego se entrena el modelo para mapear correctamente la entrada a la salida, es decir hacer una predicción correcta. El propósito final es encontrar los mejores parámetros del modelo de tal modo que el mapeo entrada→salida sea el correcto incluso para nuevos ejemplos de entrada.

Los algoritmos de aprendizaje no supervisado se entrenan con datos no etiquetados. El objetivo es descubrir patrones, relaciones o estructuras subyacentes dentro de los datos sin ninguna orientación explícita sobre qué buscar. Algunas de

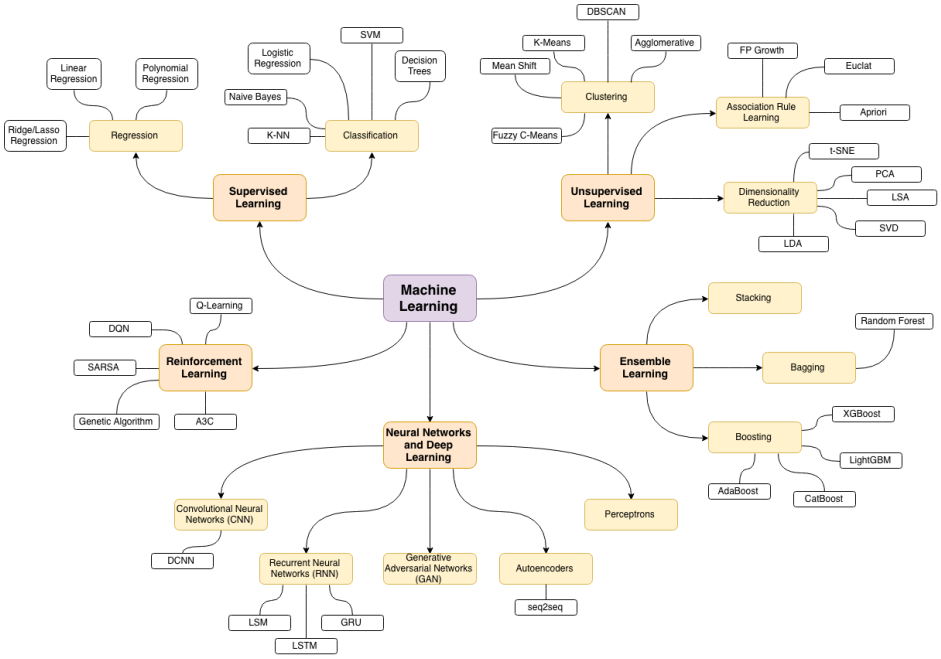

Figura 5.1: Algoritmos de Machine Learning, tomado de [49]

las aplicaciones comunes del aprendizaje no supervisado incluyen:

- Clustering o agrupamiento: Donde los datos se dividen en grupos o clústeres basados en características similares.
- Reducción de dimensionalidad: Como el Análisis de Componentes Principales (PCA), donde se busca reducir la cantidad de características o dimensiones de los datos conservando la mayoría de la información.
- Reglas de asociación: Identifica patrones interesantes entre variables en grandes conjuntos de datos.

LSA, como algoritmo de *Machine Learning* utilizado para NLP, se ubica dentro de las técnicas de aprendizaje no supervisado, en el grupo de algoritmos de reducción de dimensionalidad. En esta sección, analizaremos las últimas evoluciones y tendencias en lo que refiere a algoritmos de NLP y su posibles puntos de contacto con LSA.

La mayoría de los sistemas de Inteligencia Artificial (IA) actuales operan mediante técnicas de *Machine Learning*, donde modelos predictivos se entrenan con datos históricos para realizar predicciones futuras. Esta tendencia comenzó en la década de 1990 [50], marcando un cambio significativo respecto al modo de construcción de los sistemas de IA anteriores. En lugar de especificar 'cómo' resolver una tarea, el algoritmo lo deduce basándose en los datos. Esto también representó un paso hacia la homogeneización, permitiendo que una amplia variedad de aplicaciones se alimentaran de un algoritmo de aprendizaje genérico, como la regresión

logística por citar un ejemplo dentro de los mecanismos incluidos en la figura 5.1. Sin embargo, a pesar de la omnipresencia del aprendizaje automático en la IA, tareas semánticamente complejas en procesamiento de lenguaje natural como la respuesta a preguntas, aún requerían la intervención de expertos en el dominio para realizar *'feature engineering'*. Esto implica escribir lógicas específicas del dominio para convertir datos brutos en características de nivel superior que sean más aptas para los métodos de aprendizaje automático populares.

Desde la década de 2010, las redes neuronales profundas (*deep neural networks*), revitalizadas como 'aprendizaje profundo' [51], comenzaron a destacar en el aprendizaje automático, potenciadas por datasets más grandes y una mayor capacidad de computación, gracias a la disponibilidad de GPUs. Estas redes permitieron que las características de alto nivel surgieran naturalmente durante el proceso de entrenamiento. Esto condujo a notables mejoras en el rendimiento. Además, el aprendizaje profundo facilitó la homogeneización en la industria, al hacer posible utilizar una única arquitectura de red neuronal profunda para diversas aplicaciones, eliminando la necesidad de pipelines específicos de *feature engineering* para cada una de ellas.

En 2017 el rubro de NLP en particular y de IA en general sufriría un cambio radical a partir de la publicación del paper titulado 'Attention is All You Need' [52] por parte de investigadores de Google. Este trabajo introdujo el modelo Transformer, una nueva arquitectura que revolucionó el campo de NLP y se convirtió en la base de los LLM (*Large Language Models*) que ahora conocemos, como GPT, PaLM y otros. El documento propone una arquitectura de red neuronal que reemplaza las redes neuronales recurrentes tradicionales (RNNs) y las redes neuronales convolucionales (CNNs) con un mecanismo completamente basado en la 'atención'.

El modelo Transformer utiliza la 'autoatención' para calcular representaciones de secuencias de entrada, lo que le permite capturar dependencias a largo plazo y paralelizar eficazmente el cálculo. En la figura 5.2 puede observarse la arquitectura general del modelo. Sin entrar en detalles, cabe mencionar que la gran diferencia con los modelos anteriores basados en RNN y CNN está compuesta por las capas de *Multi-Head Attention*. Estas capas permiten al modelo enfocarse en los aspectos más relevantes de una secuencia de palabras de entrada, y además en lugar de procesar estas palabras de modo secuencial, hacerlo en paralelo. El resultado es una representación contextualizada de las palabras, que capturan las relaciones importantes entre ellas. La segunda capa del modelo que cabe destacar es la de *Embedding*. Dicho bloque se encarga de representar a las palabras en un espacio vectorial de dimensiones reducidas, tal como lo hace LSA. Si bien el modelo Transformer utilza redes neuronales para esta tarea, LSA es mencionado como el precursor de estas técnicas [53].

Los autores demostraron que su modelo logra un rendimiento de vanguardia en varias tareas de traducción automática y supera a los modelos anteriores que

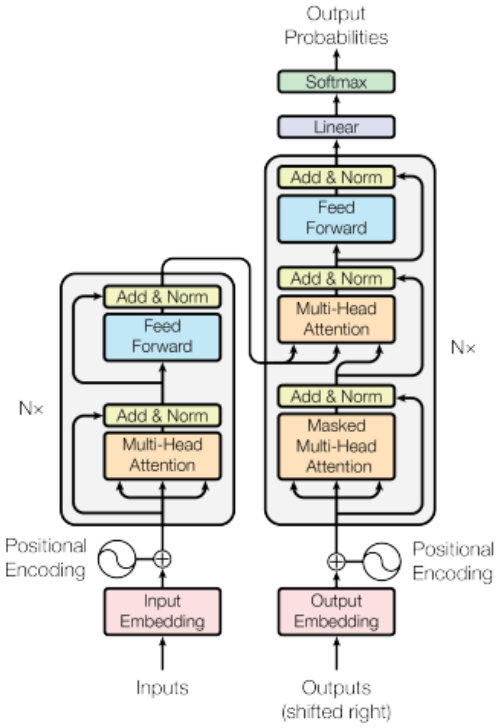

Figura 5.2: Arquitectura de Transformer. Tomado de la propuesta de Vaswani y otros en [52]

dependen de RNNs o CNNs. Investigadores de Stanford llamaron a los Transformers 'Modelos fundacionales' en un artículo de agosto de 2021 [50], definiéndolos como la base para una nueva era en el terreno de IA. En las siguientes secciones analizaremos las características generales de los Modelos Fundacionales en lo que refiere a su relación con el lenguaje, el aprendizaje y el entendimiento.

5.2. Modelos fundacionales y lenguaje

Los Modelos Fundacionales, que han encontrado una fuerte presencia en el procesamiento del lenguaje natural, están siendo vistos como un nuevo paradigma en el campo de la IA. Estos modelos han surgido gracias a dos factores cruciales: el aprendizaje por transferencia y la escala. El aprendizaje por transferencia permite aplicar el conocimiento adquirido de una tarea a otra distinta, siendo una técnica predominante en el aprendizaje profundo. La escala, que ha potenciado estos modelos, ha sido posible debido a los avances en hardware, como la mejora en las GPU; al desarrollo de la arquitectura del modelo Transformer, que aprovecha el paralelismo del hardware para entrenar modelos más expresivos; y a la disponibilidad de una gran cantidad de datos de entrenamiento.

Los modelos fundacionales han tenido un gran impacto en el campo del NLP, convirtiéndose en esenciales para la mayoría de los sistemas y la investigación en NLP. En un primer nivel, muchos de estos modelos son hábiles generadores de lenguaje, al punto que es difícil para los no expertos distinguir entre textos cortos escritos por ejemplo por GPT-4 y humanos. Sin embargo, la característica más destacada de estos modelos de NLP no es su capacidad de generación, sino su sorprendente generalidad y adaptabilidad, permitiendo que un solo modelo se adapte a diversas tareas lingüísticas. Históricamente, las técnicas de NLP se han centrado en definir sistemas para tareas lingüísticas complejas, esperando que los modelos competentes en estas tareas sean útiles en aplicaciones posteriores. Estas tareas abarcan desde análisis de sentimientos en reseñas hasta la identificación de relaciones entre palabras.

En el pasado, las tareas de NLP tenían comunidades de investigación distintas que desarrollaban arquitecturas específicas para cada tarea, a menudo basadas en cadenas de diferentes modelos. Cada uno de estos modelos realizaba una sub-tarea lingüística, como la segmentación de tokens o el análisis sintáctico.

En contraste, el enfoque moderno dominante para realizar cada tarea consiste en usar un único Modelo Fundacional y adaptarlo ligeramente utilizando pequeñas cantidades de datos anotados específicos para cada tarea (clasificación de sentimientos, etiquetado de entidades, traducción, resumen) para crear un modelo adaptado. Este enfoque ha demostrado ser extremadamente exitoso para la gran mayoría de las tareas mencionadas, un Modelo Fundacional ligeramente adaptado supera ampliamente a los modelos anteriores o a las cadenas de modelos diseñados específicamente para esa tarea. Por ejemplo, el mejor sistema para responder preguntas científicas abiertas en 2018, antes de los modelos fundacionales, alcanzaba el 73 % en el examen de ciencias de 8º grado de NY Regents. Un año después, en 2019, un Modelo Fundacional adaptado logró el 91 % [50].

Los Modelos Fundacionales son versátiles en términos de conocimiento lingüístico, pero su capacidad para manejar variaciones del lenguaje aún no está clara. El

lenguaje cambia no solo entre idiomas, sino también dentro de un solo idioma, y esto se ve influenciado por factores sociales y políticos. Los Modelos Fundacionales tienen el potencial de representar una gran diversidad lingüística. Además, tras el éxito de estos modelos en inglés, se han desarrollado modelos multilingües que se entrenan en múltiples idiomas a la vez, ayudando a incluir idiomas con menos recursos al beneficiarse de las estructuras compartidas entre diferentes lenguas. La robustez multilingüe de estos modelos aún es una cuestión sin resolver. No está claro cuánto pueden representar los modelos entrenados con estos datos sobre lenguajes muy distintos al inglés o con pocos recursos lingüísticos disponibles. A pesar de que los modelos multilingües muestran un mejor desempeño en lenguas similares a las de alto recurso en su entrenamiento, aún es un problema para idiomas muy distintos. Esto se debe a los datos utilizados para entrenar estos modelos: en muchos corpus multilingües, los datos en inglés son más abundantes y de mayor calidad que los de lenguas con menos recursos. Sin embargo, la solución no es simplemente crear corpus más equilibrados, ya que hay muchas variaciones lingüísticas y sería inviable crear un corpus balanceado en todos los aspectos. El futuro y la equidad de los Modelos Fundacionales dependen de manejar adecuadamente la variación lingüística a pesar de los datos desequilibrados. Los modelos multilingües actuales pueden no capturar completamente las sutilezas de los idiomas y sus variedades. Aun así, son útiles para ciertas aplicaciones. Algunos resultados muestran que estos modelos superan a los basados en un solo idioma, especialmente para idiomas menos representados.

5.3. Modelos fundacionales y aprendizaje

Aunque los modelos fundacionales han propiciado grandes avances en la creación de sistemas NLP que se comportan como humanos, aún hay diferencias significativas en cómo adquieren el lenguaje y en su proceso de aprendizaje, comparado con el lenguaje humano. Mientras que la adquisición del lenguaje humano es eficiente (los modelos como GPT-4 se entrenan con mucha más información lingüística de la que una persona recibirá en su vida), una diferencia clave es que el lenguaje humano está vinculado al mundo real. Por ejemplo, los bebés y sus cuidadores se señalan objetos durante el desarrollo lingüístico. En cambio, muchos modelos NLP aprenden a partir de textos no vinculados al mundo real. Aunque los bebés también presentan aprendizaje estadístico no vinculado, mejorar el aprendizaje del lenguaje vinculado en los Modelos Fundacionales es un campo importante para acercarse a la eficiencia humana en adquisición del lenguaje. Otra dirección relevante es investigar los sesgos inductivos en los modelos base. Aunque existen diferencias claras en los sesgos inductivos entre el cerebro humano y estos modelos, las formas en que aprenden el lenguaje también son muy distintas. Especialmente, los humanos interactúan con un mundo físico y social, teniendo diversas necesidades y deseos, mientras que los modelos base principalmente observan y modelan datos producidos por otros.

La eficiencia en la adquisición del lenguaje radica en que los humanos adquieren un sistema lingüístico sistemático y generalizable. A pesar de las diversas teorías sobre las abstracciones teóricas del lenguaje humano, se acuerda generalmente que aprendemos el lenguaje de manera que nos permite integrar fácilmente nuevo conocimiento y crear nuevas oraciones gramaticales. Por ejemplo, un niño de diez años ya ha adquirido muchas abstracciones sobre cómo funciona su idioma, aunque las palabras y construcciones que usa cambiarán drásticamente con el tiempo. En contraste, los Modelos Fundacionales a menudo no adquieren las abstracciones sistemáticas que esperamos de los humanos. Por ejemplo, cuando un Modelo Fundacional produce una construcción lingüística correctamente una vez, no garantiza que usos futuros serán consistentes, especialmente tras cambios significativos en el tema tratado. El desafío para la NLP es desarrollar una especie de sistematicidad en la adquisición para Modelos Fundacionales, sin depender demasiado de reglas lingüísticas rígidas.

El aprendizaje del lenguaje es un proceso continuo a lo largo de la vida de una persona: la gramática de los idiomas humanos evoluciona y los humanos se adaptan con flexibilidad a situaciones lingüísticas nuevas. Por ejemplo, los adultos incorporan fácilmente nuevos términos en sus frases gramaticales y adaptan sus patrones gramaticales según diferentes grupos sociales. En contraste, el sistema lingüístico de los Modelos Fundacionales es mayormente determinado por los datos de entrenamiento, siendo bastante estático. Aunque se pueden adaptar para diferentes tareas, no está claro cómo cambiar su base lingüística sin un entrenamiento extenso.

5.4. Modelos fundacionales y el entendimiento

Existe una interesante discusión sobre la capacidad de entendimiento de los Modelos Fundacionales. Estos modelos pueden producir lenguaje con una fluidez sorprendente, pero a menudo caen en incoherencias. En consecuencia, surge la duda, si esto es evidencia de limitaciones inherentes al propio modelo o si modelos futuros realmente podrían llegar a entender los símbolos que procesan.

Todos los Modelos Fundacionales comparten una característica definitoria: son auto-supervisados. En la auto-supervisión, el único objetivo del modelo es aprender patrones abstractos de co-ocurrencia en las secuencias de símbolos con las que fue entrenado. Esto permite que muchos de estos modelos generen secuencias plausibles de símbolos. No hay una razón obvia por la cual esta auto-supervisión le indique al modelo el significado real de los símbolos. La única información que se le proporciona directamente es sobre qué palabras tienden a co-ocurrir con otras. A simple vista, saber que 'El sándwich contiene queso' probablemente continúe con jamón, no dice nada sobre qué es un sándwich, qué es el queso, cómo se combinan estos objetos, etc. Esto podría sugerir una limitación inherente en lo que un Modelo Fundacional podría lograr. No obstante, el modelo no tiene que limitarse solo a entrada textual. Podría ser entrenado en una amplia gama de símbolos: no solo lenguaje, sino también código informático, archivos de bases de datos, imágenes, audio y lecturas de sensores. Durante este aprendizaje, el modelo podría llegar a representar asociaciones entre un texto dado y una lectura por ejemplo de algún sensor específico, o entre una secuencia de valores de píxeles y una entrada en la base de datos. Estas asociaciones podrían reflejar aspectos importantes del mundo que habitamos y el lenguaje que usamos para describirlo.

Más allá de qué es la 'comprensión', existen algunas características que parecerían alcanzarse solo si el modelo es capaz de comprender [50]:

- Confianza: Podríamos argumentar que no podemos confiar en un sistema lingüístico a menos que comprenda el lenguaje que utiliza. Aunque confiamos en sistemas para realizar tareas sin que comprendan el proceso, el lenguaje, por ser inherentemente humano, podría ser una excepción. Además, el lenguaje puede ser usado para engañar. Por lo tanto, la comprensión es esencial para confiar en el uso del lenguaje.

- Interpretabilidad: Si la comprensión real del lenguaje implica mantener y actualizar un modelo interno del mundo, y si podemos analizar cómo el lenguaje interactúa con este modelo, podríamos mejorar la interpretabilidad, predictabilidad y control de estos sistemas.

- Rendición de cuentas: En el futuro, podríamos querer responsabilizar a los agentes artificiales por el lenguaje que producen, y la comprensión podría ser un requisito previo.

La pregunta central es si un Modelo Fundacional puede llegar a comprender un lenguaje natural. Para abordar esto, necesitamos definir qué entendemos por comprensión. Se han propuesto diversas perspectivas sobre lo que significa entender el lenguaje natural [50]:

- Internalismo: La comprensión lingüística implica recuperar las estructuras representacionales internas adecuadas en respuesta a una entrada lingüística. Por lo tanto, no es posible entender el lenguaje sin un repertorio conceptual interno amplio.
- Referencialismo: Un agente comprende el lenguaje cuando sabe lo que se necesita para que las diferentes oraciones en ese lenguaje sean coherentes (relativo a un contexto). Es decir, las palabras tienen referentes y comprender implica evaluarlas en relación con una situación.
- Pragmatismo: Comprender no requiere representaciones internas o cálculos, y la verdad y referencia no son fundamentales. Lo importante es que el agente utilice el lenguaje de la manera correcta, lo que puede incluir inferencias, razonamientos y conversaciones adecuados.

Dependiendo de cuál de estas perspectivas adoptemos, la respuesta sobre si un Modelo Fundacional puede comprender el lenguaje será diferente.

Tanto el internalismo como el referencialismo se pueden describir como un problema de mapeo: asociar un signo lingüístico con un "significado". Para el internalismo, podría ser una representación interna; para el referencialismo, una palabra podría mapearse a un referente externo. La cuestión es si la auto-supervisión podría lograr el mapeo deseado en un Modelo Fundacional. Si un modelo recibe solo entradas lingüísticas, su capacidad para aprender este mapeo podría estar limitada. Sin embargo, si recibe diversa información externa (imágenes, audio, sensores, etc.), los patrones de co-ocurrencia podrían contener suficientes datos para inducir mecanismos que permitan realizar el mapeo requerido.

El acercamiento planteado en los párrafos anteriores, propuesto por los autores del artículo que define los Modelos Fundaciones [50], es absolutamente cualitativo y no se basa en ningún análisis matemático del tema. En contraste, los autores de LSA plantean un modelo del 'Significado´que es capaz de obtener resultados similares a los producidos por humanos, pero además siendo entrenado con la misma cantidad y tipo de información que un humano. Existen adicionalmente, trabajos de investigación que vinculan LSA con teorías que permiten explicar los modelos cognitivos que ocurren en el cerebro y además vislumbrar su similitud con los procesos algebraicos base de LSA, tal como se presentó en la sección 1.5. Por tanto, si bien los resultados obtenidos a partir del uso de los Modelos Fundaciones son asombrosos, el modelo en sí mismo, a diferencia de LSA, no parece de gran utilidad para explicar los procesos cognitivos relacionados con aprendizaje y entendimiento.

5.5. Análisis Semántico Latente (LSA) vs Modelos de lenguaje extensos (LLMs)

Como hemos analizado, Latent Semantic Analysis (LSA), es una técnica pionera en el procesamiento de lenguaje natural (NLP) y la recuperación de información, que captura relaciones semánticas entre palabras y textos al reducir la dimensionalidad de grandes conjuntos de datos. Esta técnica, basada en la descomposición en valores singulares (SVD) de una matriz de términos por documentos, ha jugado un papel crucial en el entendimiento de cómo los modelos computacionales pueden procesar el lenguaje humano.

En comparación, los Large Language Models (LLMs), como GPT-4, han revolucionado el procesamiento del lenguaje y la generación de texto, alcanzando logros sin precedentes, empleando arquitecturas de redes neuronales y aprendizaje profundo.

Sin embargo, a pesar de sus impresionantes capacidades, la predominancia de estos modelos ha introducido un problema no trivial: la homogeneización de la investigación en NLP. Esto puede llevar a un estancamiento en la exploración y desarrollo de otras metodologías, como LSA, que se basa en la comprensión estadística de las relaciones semánticas.

Los sorprendentes resultados obtenidos por los LLMs, pueden derivar en concentrar la atención y la investigación científica solo en este tipo de modelos alejando la investigación y los recursos de métodos como el LSA, que no dependen de enormes conjuntos de datos ni de grandes capacidades computacionales. Esta homogeneización de enfoques en NLP es preocupante por varias razones. Primero, la concentración en una única metodología puede crear puntos ciegos donde ciertas características del lenguaje o del procesamiento pueden ser ignoradas o malinterpretadas. En segundo lugar, la diversidad metodológica es crucial para fomentar la innovación. Además, hay preocupaciones prácticas sobre la sostenibilidad de la investigación centrada en los LLMs. Estos modelos requieren una gran cantidad de energía y recursos computacionales, lo que cuestiona su viabilidad a largo plazo y su impacto ambiental. Por otro lado, el riesgo de dejar técnicas como el LSA en el olvido podría significar perder perspectivas valiosas en la comprensión semántica. LSA puede ofrecer ventajas en escenarios donde los modelos más pequeños y menos consumidores de recursos son preferibles, o donde la facilidad de interpretación y explicación de los procesos internos son críticas.

Por otro lado, a través de la reducción de la dimensionalidad, LSA extrae patrones que pueden simular algunos aspectos de la cognición humana, como la asociación y la recuperación de la información. La fortaleza del LSA radica en su capacidad para modelar cómo los humanos pueden interpretar el significado en contextos amplios, incluso con información limitada o ruido. Tal como fue analizado, LSA ha sido utilizado para dar una interpretación sobre cómo las personas

pueden reconocer sinónimos o cómo teoría para explicar el vínculo entre memorias asociativas y procesos cognitivos. Esta capacidad de LSA de ofrecer un acercamiento a la explicación de la cognición humana es muy valiosa, especialmente en campos como la psicolingüística y la educación, donde entender cómo los humanos procesan y comprenden el lenguaje es crucial.

Los LLMs, utilizan redes neuronales profundas para modelar el lenguaje. A diferencia del LSA, los LLMs no están diseñados específicamente para explicar procesos cognitivos humanos, sino para replicar el desempeño lingüístico humano. Tienen gran capacidad para generar texto coherente y realizar tareas lingüísticas complejas.

Sin embargo, esta gran capacidad también tiene desventajas. Los procesos internos de los LLMs son difíciles de interpretar, lo que significa que ofrecen poca ayuda en el entendimiento sobre la cognición humana en comparación con LSA. La interpretación de cómo un LLM llega a una conclusión particular es opaca y no necesariamente alineada con los procesos cognitivos humanos. La explicación de los procesos cognitivos humanos en la interacción con el lenguaje es donde LSA tiene una ventaja. La estructura matemática subyacente del LSA puede modelar y explicar cómo los conceptos y significados se relacionan en la mente humana de una manera más transparente que los modelos de LLMs.

Otro posible uso de LSA en conjunto con modelos generativos como los LLMs, es su utilización para validación de resultados. LSA puede determinar la proximidad semántica de los términos dentro de un corpus. Podría ser utilizado para evaluar si el texto generado por un LLM mantiene la coherencia semántica en relación con un conjunto de documentos o un corpus de referencia.

En consecuencia, mientras que los LLMs superan a LSA en muchas tareas de NLP en términos de precisión y fluidez, la relevancia de LSA radica en su contribución al entendimiento cognitivo. Tal como se plantea en [16], uno de los desarrollos originales de LSA fue precisamente su aplicación como modelo cognitivo, destacando su capacidad para simular asociaciones semánticas humanas. Esta propiedad sigue siendo valiosa, especialmente en psicología cognitiva, educación y contextos donde se requiere interpretabilidad.

Una línea futura de investigación sería combinar LSA con LLMs de forma complementaria, usándolo como herramienta externa de análisis semántico para auditar, validar o interpretar las salidas de los LLMs. Este enfoque híbrido permitiría aprovechar la precisión de los modelos generativos junto con la capacidad explicativa de LSA. Lo vemos especialmente útil en áreas sensibles como la educación, la salud o la toma de decisiones, donde la trazabilidad del razonamiento semántico es tan importante como la exactitud del contenido generado.

Capítulo 6

Conclusiones

En este trabajo se analizó la técnica Latent Semantic Analysis (LSA), pionera en el procesamiento de lenguaje natural (NLP) y en la recuperación de información; basada en la captura de relaciones semánticas entre palabras y textos. Se fundamenta en la reducción de la dimensionalidad de grandes conjuntos de datos, usando la descomposición en valores singulares (SVD) de una matriz de términos por documentos.

Se realizó un análisis crítico del funcionamiento de LSA y su vínculo con Memorias Asociativas, con foco en las aplicaciones que esta técnica puede tener en el campo de la medicina, en particular en registros médicos. Se analizó bibliografía reciente que trata sobre el aprovechamiento de grandes cantidades de datos médicos acumulados, y no solo el registro específico de un examen médico. La idea fundamental es combinar información de numerosos pacientes con una misma afección para identificar patrones generales. Esto revelaría conexiones ocultas entre condiciones médicas en un extenso grupo de pacientes y la forma en que estas se entrelazan.

Inicialmente, LSA se utilizó como un método de búsqueda de información y categorización de texto. Sin embargo, se mostró que tiene capacidad para descubrir estructuras subyacentes y asociar objetos similares, lo que lo convierte en una técnica idónea para modelar ciertos procesos cerebrales. Esta capacidad abre un amplio campo de investigación en el ámbito de las ciencias cognitivas. Específicamente en el contexto de aplicaciones médicas basadas en el análisis del lenguaje. En particular, se analizó el uso de LSA para la evaluación de pacientes con esquizofrenia.

Por nuestra parte, desarrollamos un estudio sobre la utilización de LSA para inferir el grado de similitud de los mensajes de Syslog mediante el descubrimiento de relaciones ocultas entre ellos. Usando ejemplos reales de mensajes de Syslog, mostramos que LSA es capaz de detectar los mensajes más correlacionados por tema. Este método se puede utilizar para evitar el uso de sistemas complejos de correlación de eventos que generalmente necesitan firmas o definiciones de con-

juntos de reglas y una gran experiencia para su configuración; lo cual puede ser aplicado en equipamiento médico en ámbitos hospitalarios.

Adicionalmente, desarrollamos un estudio utilizando LSA para examinar la distribución de diferentes tipos de cáncer en las distintas regiones del Uruguay. A partir de este estudio, concluimos que los enfoques convencionales basados en correlaciones estadísticas están parcialmente sesgados y están dominados por la alta prevalencia de ciertas formas de cáncer. Esto dificulta la visualización de la distribución de formas menos comunes. Para identificar posibles agrupaciones regionales de las formas menos frecuentes de cáncer implementamos un sistema reducido con una matriz de datos que únicamente conserva las categorías de diagnóstico menos frecuentes. Mediante el uso del LSA en esta matriz reducida, se pudo inferir correlaciones entre las distribuciones de cáncer, aspectos que no eran perceptibles mediante otros métodos de estudio.

Finalmente analizamos LSA como teoría de Data Mining, su relación con los actuales modelos de lenguaje denominados LLMs (Large Language Models) y se establecieron posibles líneas de trabajo futuro vinculando ambas teorías.

Referencias

[1] S. Dennis W. Kintsch T. K. Landauer, D. S. Mcnamara. Handbook of latent semantic analysis. *Routledge*, 2007.

[2] P. W. Foltz T. K. Landauer, D. Laham. Automated essay assessment. *Assessment in Education: Pinciples, Police and Practice*, 10:295–308, 2007.

[3] T. K. Landauer P. W. Foltz, W. Kintsch. Analysis of text coherence using latent semantic analysis. *Discourse Processes*, 25:285–307, 1998.

[4] T. K. Landauer. On the computational basis of congnition: Arguments from lsa. *The psychology of learning and motivation*, pages 43–84, 2002.

[5] M. Wolf D. Laham T. Landauer W. Kintsch B. Redher, M. Schreiner. Using latent semantic analysis to assess knowledge: Some technical considerentions. *Discourse Processes*, 25:337–354, 1998.

[6] D. Weinberg T. Goldberg B. Elvevag, P. Foltz. Quantifying incoherence in speech: An automated methodology and novle application to schizophrenia. *Schizophr Res*, 25:304–316, 2007.

[7] Landauer T. K. Littman M. L. Dumais, S. T. Automatic cross-linguistic information retrieval using latent semantic indexing. *Automatic cross-linguistic information retrieval using Latent Semantic Indexing*, pages 16–23, 1996.

[8] D. Laham. Latent semantic analysis approaches to categorization. *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, page 979, 1997.

[9] Ion Zaballa. Valores singulares. ¿qué son?.¿para qué sirven? *Departamento de Matemática Aplicada y EIO, Universidad del País Vasco.*

[10] Eugene Beltrami. Sulle funcioni bilineari. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, 11:98–106, 1873.

[11] C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées, Deuxieme Série*, 19:35–54, 1874.

[12] J. J. Sylveste. A new proof that a general quadric may be reduced to its canonical form (that is, a linear function of squares) by means of a real orthogonal substitution. *The Messenger of Mathematics*, 19:1–5, 1889.

[13] E. Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. i teil. entwicklung willk¨urlichen funktionen nach system vorgeschriebener. *Mathematische Annalen*, 63:433–476, 1907.

[14] H. Nateman. A formula for the solving function of a certain integral equation of the second kind. *Transactions of the Cambridge Philosophical Society*, 20:179–187, 1908.

[15] E. Picard. Sur un th´eor‘em g´en´eral relatif aux int´egrales de premi‘er esp‘ece et sur quelques probl‘emes de physique math´ematique. *Rendicondi del Circulo Matematico di Palermo*, 25:79–97, 1910.

[16] Susan T; Furnas George W; Landauer Thomas K; Harshman Richard Deerwester, Scott; Dumais. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[17] T. K. Landauer; P. W. Foltz; D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[18] E. Mizraji. Neural memories and search engines. *International Journal of General Systems*, 37, no.6:715–732, 2008.

[19] J. C. Valle-Lisboa, A. Pomi, and E. Mizraji. Multiplicative processing in the modeling of cognitive activities in large neural networks. *Biophysical Reviews*, 15(4):767–785, 2023.

[20] E. Mizraji. Towards the neural modelling of mental spaces. In *2017 Computing Conference*, pages 692–696, 2017.

[21] V. Datla; King-Ip; M. Louwerse. Capturing disease-symptom relations using higher-order co-occurrence algorithms. *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 816–821, 2012.

[22] M. Galnares; S. NesmachnoW; F. Simini. Instance-based learning following physician reasoning for assistance during medical consultation. *Appl. Sci*, 11(13):5886, 2021.

[23] Lippincott Williams Wilkins. Professional guide to signs and symptoms. 2010.

[24] L. Garcia J.C. de Araujo, J.M. Parente. Recovery medical articles using semantic enrichment method. *11th International Conference on Signal-Image Technology Internet-Based Systems*, pages 705–711, 2015.

[25] Xinyu Jin; Wentao Ma; Yunze Li. Medical record text analysis based on latent semantic analysis. *8th International Symposium on Computational Intelligence and Design (ISCID*, pages 108–110, 2015.

[26] Ke Wang Bo Li. Computer aided diagnosis semantic model for the report of medical image via lda and lsa. *IT in Medicine and Education (ITME), 2011 International Symposium on*, 1:699–703, 2011.

[27] D. Gefen; J. Miller; J. K. Armstrong; F. H. Cornelius; N. Robertson; A. Smith-McLallen; J. A. Taylor. Identifying patterns in medical records through latent semantic analysis. *Communications for the ACM*, 61(6), 2018.

[28] Bryce Picton, Saman Andalib, Aidin Spina, Brandon Camp, Sean S. Solomon, Jason Liang, Patrick M. Chen, Jefferson W. Chen, Frank P. Hsu, and Michael Y. Oh. Assessing ai simplification of medical texts: Readability and content fidelity. *International Journal of Medical Informatics*, 195:105743, 2025.

[29] Sujoy Roy, Shane Morrell, Lili Zhao, and Ramin Homayouni. Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of latent semantic indexing and generative pretrained transformer (gpt) models. *BMC Medical Informatics and Decision Making*, 24(296), 2024.

[30] J. C.Valle-Lisboa; E. Mizraji. The uncovering of hidden structures by latent semantic analysis. *Information Sciences*, 177:4122–4147, 2007.

[31] B. Elvevag; P. W. Foltz; D. R. Weinberger; T. E. Goldberg. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93:304–316, 2007.

[32] N. C. Andreasen. Scale for the assessment of thought, language and communication (tlc). *Schizophrenia Bulletin*, 12:474–482, 1986.

[33] http://lsa.colorado.edu/.

[34] B. Elvevag; P. W. Foltz; M. Rosenstein; L. E. DeLisi. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, 23:270–284, 2010.

[35] T. Cohen; B. Blatter; V. Patel. Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, 41:1070–1087, 2008.

[36] Brucki. Epidemiology of mild cognitive impairment in brazil. *Demen Neuropsychol*, 7:363–366, 2013.

[37] Zaudig M Petersen RC Ritchie K Broich K et al Gauthier S, Reisberg B. Mild cognitive impairment. *Lancet*, 15:1262–1270, 2006.

[38] L. Borges S.M. Dozzi E. Sturzeneker M. Okada L. Lesa C. Matsuda, S. M. Aluísio. Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, 10:31–40, 2018.

[39] X. Ren and M. N. Coutanche. Sleep reduces the semantic coherence of memory recall: An application of latent semantic analysis to investigate memory reconstruction. *Psychonomic Bulletin & Review*, 28(4):1336–1343, 2021.

[40] E. Barrios; C. Musetti; R. Alonso; M. Garau. V atlas de mortalidad por cáncer en el uruguay. *Registro nacional de cáncer, Comisión honoraria de lucha contra el cáncer*, 2015.

[41] A. Fritz; C. Percy; A Jack; S. Kanagarathnam; L. Sobin; DN Parkin; S. Whelan. International classification of diseases for oncology. *World Health Organization*, 3rd ed., 2000.

[42] https://tools.ietf.org/html/rfc5424.

[43] G. Slomovitz. Latent semantic analysis (lsa) for syslog correlation. *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 1–4, 2017.

[44] E. Çalışkan R. Vaarandi, B. Blumbergs. Simple event correlator - best practices for creating scalable configurations. *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 96–100, 2015.

[45] What happened in my network: mining network events from router syslogs. *IMC '10 Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 472–484, 2010.

[46] K. Fukuda. On the use of weighted syslog time series for anomaly detection. *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium*, 2011.

[47] SAFECARE Project Consortium. Security analytics and monitoring of medical devices. 2021.

[48] A. Sebbar, O. Cherqi, K. Chougdali, and M. Boulmalf. Real-time anomaly detection in sdn architecture using integrated siem and machine learning for enhancing network security. pages 1795–1800, 2023.

[49] GitHub contributors. Machine learning algorithms. *GitHub Topics*, 2025. Available at: `https://github.com/topics/machine-learning-algorithms`, Accessed: June 15, 2025.

[50] R Bommasani et all. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

[51] G-Hinton Y. LeCun, Y. Bengio. Deep learning. *Nature*, 521:436–444, 2015.

[52] A. Vaswani et all. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

[53] J. Atkinson-Abutridy. Grandes modelos de lenguaje: Conceptos, técnicas y aplicaciones. *Marcombo; N.º 1 edición*, ISBN-10: 8426736793, 2023.

Índice de figuras

1.1. Texto usado para la aplicación del método propuesto por Landauer y otros en [17] 9
1.2. Matriz M, Tomado de la propuesta de Landauer y otros en [17] . . 10
1.3. Matriz U. Tomado de la propuesta de Landauer y otros en [17] . . 10
1.4. Matriz S. Tomado de la propuesta de Landauer y otros en [17] (las celdas no mostradas tienen valor nulo) 10
1.5. Matriz V traspuesta. Tomado de la propuesta de Landauer y otros en [17] 11
1.6. Matriz M reconstruida. Tomado de la propuesta de Landauer y otros en [17] 11

2.1. Comparación entre síntomas reales y aleatorios, según cada enfermedad. Tomado de la propuesta de Dada y otros [21] 18
2.2. Proceso de búsqueda de información médica. Tomado de la propuesta de García y otros en [24] 20
2.3. Precisión en la efectividad de la búsqueda. Tomado de la propuesta de García y otros en [24] 20
2.4. Sensibilidad en la búsqueda. Tomado de la propuesta de García y otros en [24] 21
2.5. Proceso de aplicación de LSA relacionados con la insuficiencia cardíaca congestiva. Tomado de la propuesta de Gefen y otros en [27]) . . . 22
2.6. Coherencia entre palabras consecutivas de experimento que busca medir la fluidez verbal. Tomado de la propuesta de Elvevag y otros en [31] 26
2.7. Coherencia de texto usando ventana deslizante. Tomado de la propuesta de Elvevag y otros en [31] 26
2.8. Investigación de desórdenes del lenguaje, usando LSA, con textos de “La Cenicienta”. Tomado de la propuesta de Borges y otros en [38] 28
2.9. Características macroestructurales del discurso. Tomado de la propuesta de Borges y otros en [38] 29
2.10. Resultados de la comparación entre diferentes grupos con distintas patologías, basado en el análisis de LSA. Tomado de la propuesta de Borges y otros en [38] 30

Índice de figuras

2.11. Efecto del sueño en los recuerdos, utilizando LSA. a) coherencia semántica entre oraciones consecutivas, b) coherencia semántica de todo el texto. Tomado de la propuesta de Ren y otros en [39] . . . 31

3.1. Mortalidad por cáncer en población mundial estándar. Extraído de Barrios y otros en [40] . . . 36
3.2. Interfaz gráfica del desarrollo programado en Labview. . . . 37
3.3. Tasas ajustadas por edad para distintos tipos de cáncer de pequeñas tasas, por departamento. . . . 37
3.4. Relación entre el número de coeficientes de correlación mayor que 0.8 y el número de dimensiones para la reconstrucción. . . . 38
3.5. Comparaciones entre vectores que representan términos o documentos. 38
3.6. Resultado del query . . . 39
3.7. Departamentos no correlacionados originalmente . . . 39

4.1. Formato de mensaje Syslog. . . . 43
4.2. Syslog de tecnología de red BGP. . . . 44
4.3. Reconstrucción de la matriz M . . . 44
4.4. Matriz M_k . . . 44
4.5. Correlación de la matriz original. . . . 45
4.6. Correlación de la matriz reconstruida. . . . 45

5.1. Algoritmos de Machine Learning, tomado de [49] . . . 50
5.2. Arquitectura de Transformer. Tomado de la propuesta de Vaswani y otros en [52] . . . 52

Esta es la última página.
Compilado el lunes 16 junio, 2025.
http://iie.fing.edu.uy/