UNIVERSIDAD AUTONOMA DE QUERETARO

FACULTAD DE INFORMÁTICA

INFORME TÉCNICO

P**ROGRAMACIÓN AVANZADA Y BD**

Minería de datos aplicado al análisis bibliográfico de textos científicos.

Presentan:

Emiliano Cabrera
Alexis Raciel
Jazmin Guadalupe Guiza Aviles

Docente:
Francisco Paulin

28 de Noviembre del 2025

ÍNDICE

Resumen

La minería de datos tiene un conjunto de aplicaciones amplio, para este proyecto se ha aplicado al análisis bibliográfico de textos científicos.

La clave para redactar un texto científico de forma exitosa proviene de la información verídica investigada y referenciar correctamente un texto científico es pieza importante, de lo contrario se podría incurrir en plagio.

Este documento redacta el desarrollo que el equipo realizó para cumplir con los requerimientos solicitados por el docente, aplicando conocimientos adquiridos durante el primer semestre y el dominio de cada uno.

Introducción

¿Qué tanto afecta para el análisis trabajar con datos que no han sido limpiados previamente? Probablemente pensaríamos que los datos pueden analizarse tal cual son obtenidos, sin embargo, hacer esto podría no darnos un panorama real de lo que buscamos conocer debido a que los parámetros que establecemos no son lo suficientemente robustos para abarcar las posibles variantes con las que nos podríamos encontrar.

La minería de datos hace uso del análisis estadístico para descubrir patrones e información importante en nuestros casos de estudio en grandes conjuntos de datos. Las técnicas de minería de datos se pueden desplegar para dos propósitos principales:

- Describir el conjunto de datos objetivo
- Predecir resultados mediante algoritmos de machine learning.

Estos métodos y algoritmos permiten la extracción automática de información que permite caracterizar la relación en los datos; se pretende que la información obtenida posea capacidad predictiva, facilitando el análisis de los datos de forma eficiente [Martinez, B. B. (2001)].

Este trabajo busca evaluar el conocimiento adquirido durante la materia de “Programación avanzada y bases de datos” de la Maestría en Ciencia de Datos, en donde a lo largo del programa se abordaron técnicas básicas de limpieza de datos con el objetivo de facilitar el análisis.

El trabajo se centrará en analizar diferentes archivos en formato .txt que serán utilizados para encontrar y evaluar las referencias del texto, determinar si esta se encuentra en el formato APA 7 y evaluar la recurrencia de los autores, editoriales, nombres, entre otros.

Antecedentes

Minería de datos

El descubrimiento es un tipo de inducción de conocimiento, no supervisado, que implica dos procesos: Búsqueda de regularidades interesantes entre los datos de partida, Formulación de leyes que las describan [Martinez, B. B. (2001)].

Los principales pasos dentro del proceso son los siguientes:

1. Desarrollo y entendimiento del dominio de la aplicación, el conocimiento relevante y los objetivos del usuario final, saber qué partes son susceptibles de un proceso, cuáles son los objetivos

2. Creación del conjunto de datos objetivo, seleccionando el subconjunto de variables sobre los que se realizará el descubrimiento. Implica consideraciones sobre la homogeneidad de los datos, su variación a lo largo del tiempo, estrategia de muestreo.

3. Preprocesado de los datos: eliminación de ruido, estrategias para manejar valores ausentes, normalización de los datos, etc.

4. Transformación y reducción de los datos. Incluye la búsqueda de características útiles de los datos según sea el objetivo final, la reducción del número de variables y la proyección de los datos sobre espacios de búsqueda en los que sea más fácil encontrar una solución.

5. Elección del tipo de sistema para minería de datos. Esto depende de sí el objetivo del proceso es la clasificación, regresión, agrupamiento de conceptos, detección de desviaciones.

6. Elección del algoritmo de minería de datos.

7. Minería de datos. En este paso se realiza la búsqueda de conocimiento con una determinada representación del mismo.

8. Interpretación del conocimiento extraído. Con posibilidad de iterar de nuevo desde el primer paso. La obtención de resultados aceptables dependerá de factores como: definición de medidas del interés que permitan filtrar de forma automática, existencia de técnicas de visualización para facilitar la valoración de los resultados.

9. Consolidación del conocimiento descubierto. Este paso incluye la revisión y resolución de posibles inconsistencias con otro conocimiento extraído previamente

Para este proyecto se ha utilizado minería de datos junto estadística descriptiva.

Análisis de frecuencias

El análisis de frecuencia es un método estadístico utilizado para analizar la frecuencia de puntos de datos dentro de un conjunto de datos. Es particularmente útil para identificar patrones, tendencias y anomalías en datos, lo que puede proporcionar información valiosa para los procesos de toma de decisiones.
Existen varios tipos de análisis de frecuencia:

- Análisis de frecuencia univariante se centra en una única variable y proporciona información sobre su distribución.
- Análisis de frecuencia bivariado examina la relación entre dos variables.
- Análisis de frecuencia multivariado explora las interacciones entre múltiples variables, ofreciendo una visión más completa de los datos.

Interpretar la distribución de frecuencia implica analizar la forma, la tendencia central y la variabilidad de los datos. La forma de la distribución puede indicar si los datos están distribuidos normalmente, están sesgados o tienen valores atípicos.
Las medidas de tendencia central, como la media, la mediana y la moda, brindan información sobre los valores típicos dentro del conjunto de datos, mientras que las medidas de variabilidad, como el rango y la desviación estándar, indican qué tan dispersos están los puntos de datos.

Para el análisis de frecuencias nos podemos basar en herramienta como la estadística descriptiva está constituida por un conjunto de técnicas cuyo objetivo es clasificar, presentar, describir, resumir y analizar los datos relativos a una o más características de los individuos de una población, a partir de la información sobre todos y cada uno de ellos. Para cubrir estos objetivos se usan tablas, gráficos y resúmenes estadísticos. La estadística descriptiva univariante se centra en el análisis de una única característica o cualidad del individuo

Formato APA 7

Todos los trabajos académicos escritos con el Formato APA, deben proporcionar la información necesaria para que el lector localice y recupere cualquier fuente que haya sido citada en tu texto.

Cada entrada en una lista de referencia debe incluir los cuatro elementos básicos de una referencia: el autor, fecha de publicación, título del trabajo y fuente para recuperación, como se aprecia en la **Fig. 1**.

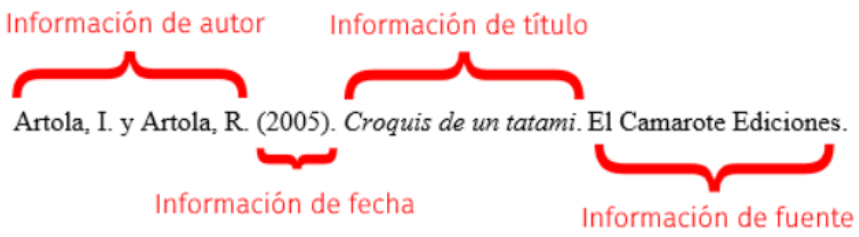


Fig. 1: Formato APA 7..

- La lista de referencias debe ser ordenada alfabéticamente por el primer apellido del autor seguido de las iniciales del nombre del autor.
- En la sexta edición, se hablaba de hasta 7 autores y más de 7 autores. Ahora, en las reglas actualizadas, hablamos de hasta 20 autores y de más de 20 autores.
- Se debe incluir números de página en una referencia, de acuerdo al tipo de fuente que estés citando.

Las información ha sido obtenida a través de Normas APA actualizadas (7ma edición) (https://normas-apa.org/referencias/citar-pagina-web/)

Metodología

Herramientas a utilizar

Para el desarrollo de este sistema se decidió utilizar el lenguaje de programación Python por la sintaxis sencilla, fácil entendimiento del código y versatilidad, la sencillez y rápida adaptabilidad al lenguaje ayudó a gestar la aplicación y a crear la GUI (*Graphical User Interface*, Interfaz Gráfica de Usuario). En el diagrama a bloques de la **Fig. 2** se muestra la metodología que se siguió.

Pasos a seguir

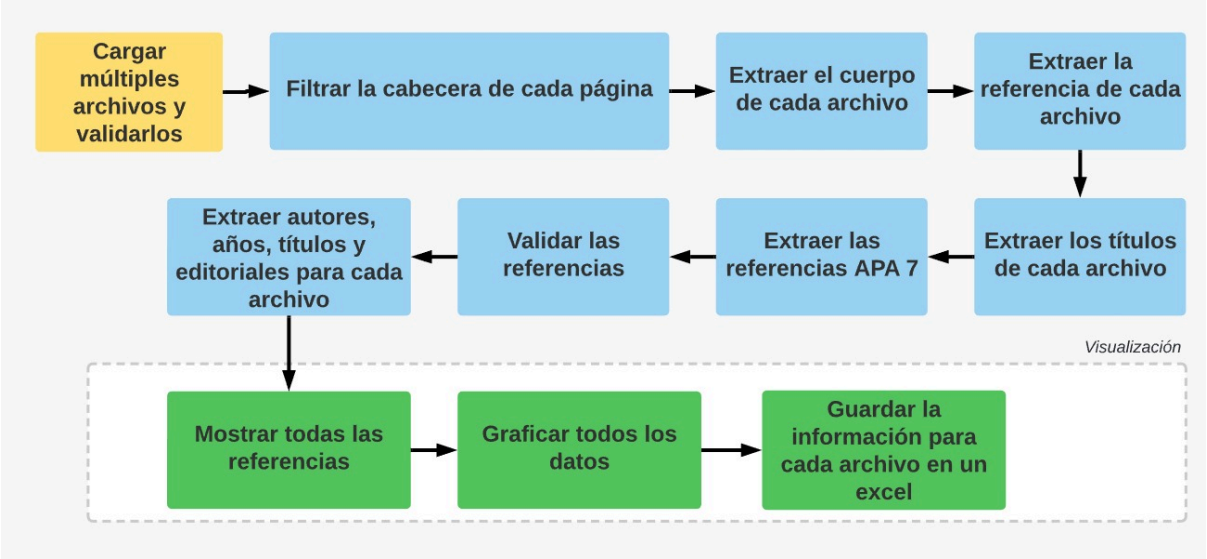


Fig. 2: Arquitectura.

En la **Fig. 1.** podemos observar los pasos que sigue el algoritmo desarrollado los cuales se describen a continuación, el código podrá ser consultado en los archivos adjuntos, adicionalmente se realizarán las referencias a los fragmentos de código que se encuentran en el apéndice.

1. **Cargar múltiples archivos:** Esta sección abre una ventana de archivos y al navegar entre los archivos del sistema podemos seleccionar la carpeta donde se encuentran los documentos a analizar, nuestra aplicación únicamente leerá aquellos con formato txt, devolviendo un arreglo con los nombres de los archivos con el formato determinado y la ruta de la carpeta, para posteriormente ejecutar los pasos siguientes para cada archivo. Véase **Imagen 1 Apéndice A**.

Posterior a cargar los archivos los siguientes pasos se realizan para la limpieza del texto contenido en cada uno:

1. **Filtrar cabecera de cada página:** Una vez abierto el documento se lee fila por fila buscando coincidencias a una variable determinada que tiene similitud al Header de cada página, así como los saltos de línea con la finalidad de borrar las filas que contengan estas concordancias. Véase **Imagen 2 y 3 Apéndice A**.

2. **Extraer cuerpo del archivo:** Cuando el texto ya no cuenta con espacios en blanco o cabeceras se almacena en una nueva variable para continuar el análisis (**Véase Imagen 2 Apéndice A**).
3. **Extraer la referencia de cada archivo:** Ahora con el documento almacenado en una variable buscaremos la sección de las referencias, almacenando esta vez desde donde empiezan estas hasta el final del documento, utilizando una variable con el string “REFERENCIAS” para determinar cuándo hemos encontrado esta sección. Véase **Imagen 4 Apéndice A**.
4. **Validar las referencias:** Se evalúa el array con la sección de las referencias encontradas y utilizando ahora la el string “http”, almacenado en una variable recorreremos las filas buscando las referencias. Véase **Imagen 5 Apéndice A**. Posteriormente evaluarlas con el método “Validar APA 7”, el cual por medio de un patrón establecido devolverá true cuando se detecta que cumple con las características de APA 7 y false cuando no cumple con estas, almacenando las referencias devueltas cómo verdadero en nuestro último arreglo con las referencias encontradas. Véase **Imagen 6 Apéndice A**.
5. **Extraer contadores:** Para terminar con la evaluación de los archivos utilizamos nuevamente las características del formato APA para buscar, autores, editoriales y títulos entre las referencias que detectamos cómo buenas almacenando cada una en vectores y su periodicidad, los cuales se agrega si el input no existe y se suma en uno la cantidad cuando este existe. Véase **Imagen 7 Apéndice A**.

Los cinco pasos anteriormente descritos se repiten para cada file encontrado en la carpeta que se seleccionó al principio, los vectores contadores de autores, títulos, editoriales, etc. concatena la información de los nuevos archivos incrementando sus contadores o tamaños por cada archivo que se evalúa.

Para finalizar se muestran los datos almacenados posterior a la evaluación con tablas y gŕaficas correspondientes a cada contador realizado, de igual manera cómo salida se genera un archivo tipo excel con la información de frecuencias obtenidas, lo anterior utilizando librerías cómo matplotlib, openpyxl y pandas. Véanse **Imágenes 8, 9 y 10 Apéndice A**. Asimismo, en la **Gráfica 1 y Tabla 1** se muestran las frecuencias de los autores.

d. Análisis de frecuencias absolutas.

C = característica cualitativa. **[Autores, Títulos, Años y Editorial]**

n = número de individuos que presentan la característica.

f = frecuencia relativa. **[n/N]**

Ni = frecuencia absoluta acumulada**.[sumatoria de n]**

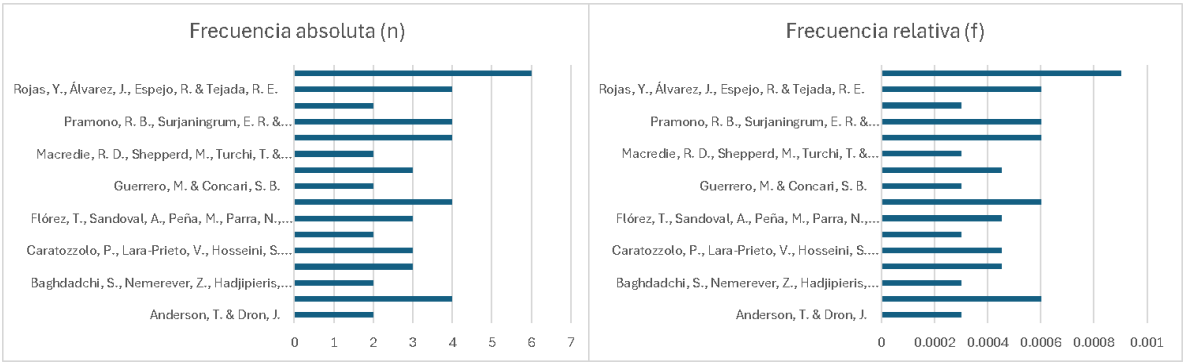
F = frecuencia relativa acumulada. **[sumatoria de Ni/N]**

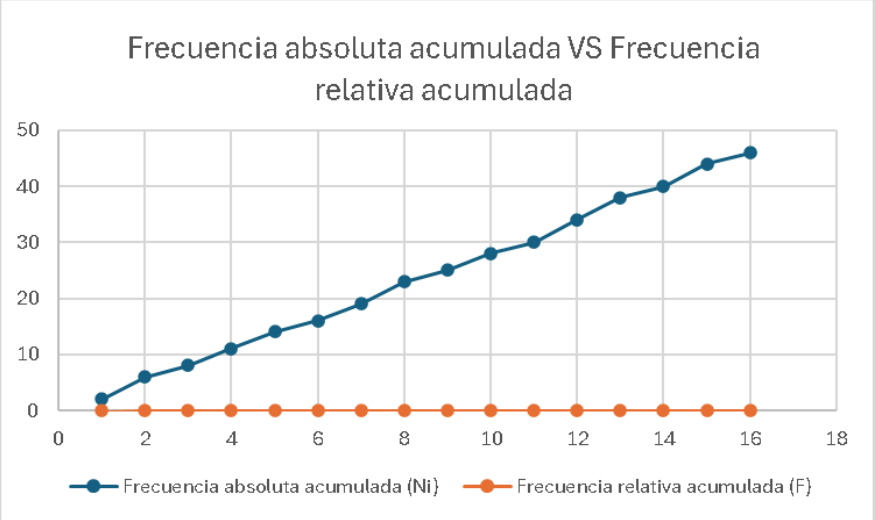
N = tamaño de la muestra (Cantidad de palabras) **[6639]**

AUTORES EN REFERENCIAS E1

Autores	Frecuencia absoluta (n)	Frecuencia relativa (f)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa acumulada (F)
Anderson, T. & Dron, J.	**2**	**0.0003012**	**2**	**0.0003012**
Arancibia, S., Maréchal	**4**	**0.0006025**	**6**	**0.0009037**
Baghdadchi, S., Nemerever, Z., Hadjipieris, P., Serslev, S. & Sandoval, C.	**2**	**0.0003012**	**8**	**0.0012049**
Cabero-Almenara, J. & De Los Ríos, J. L. P. D.	**3**	**0.0004518**	**11**	**0.0016567**
Caratozzolo, P., Lara-Prieto, V., Hosseini, S. & Membrillo-Hernández, J.	**3**	**0.0004518**	**14**	**0.0021085**
Cárdenas-Oliveros, Rodriguez-Borges, J. A. ;, ; Pérez-Rodríguez, C. G., ValenciaZambrano, J. A. ; & Xavier H.	**2**	**0.0003012**	**16**	**0.0024097**
Flórez, T., Sandoval, A., Peña, M., Parra, N., Ramírez, C., Garzón, P. & Cortés, J.	**3**	**0.0004518**	**19**	**0.0028615**
Greener, S.	**4**	**0.0006025**	**23**	**0.003464**
Guerrero, M. & Concari, S. B.	**2**	**0.0003012**	**25**	**0.0037652**
Macedo, A.	**3**	**0.0004518**	**28**	**0.004217**
Macredie, R. D., Shepperd, M., Turchi, T. & Young, T.	**2**	**0.0003012**	**30**	**0.0045182**
Mayor, S., Tarma, W. N. & Mayor, J.	**4**	**0.0006025**	**34**	**0.0051207**
Pramono, R. B., Surjaningrum, E. R. & Yoenanto, N. H.	**4**	**0.0006025**	**38**	**0.0057232**
Putra, P. D. A., Sulaeman, N. F., Supeno & Wahyuni, S.	**2**	**0.0003012**	**40**	**0.0060244**
Rojas, Y., Álvarez, J., Espejo, R. & Tejada, R. E.	**4**	**0.0006025**	**44**	**0.0066269**
Watson, Goodwin. & Glaser, Edwin.	**6**	**0.0009037**	**46**	**0.0075306**

Tabla 1: Análisis de frecuencias en autores para referencias E1.





Gráfica 1. Frecuencias para autores.

Como se puede observar en las **Tabla 1,** la frecuencia absoluta es el número de veces que aparece cada autor en el documento donde el autor más citado antes de validar APA 7 es Rojas con un total de 6 apariciones en el documento.

La frecuencia relativa es la proporción o el porcentaje de veces que ocurre la búsqueda de cada autor en relación con el total de palabras en el documento, el autor más citado antes de validar APA 7 es Rojas con un aproximado de 0.0008.

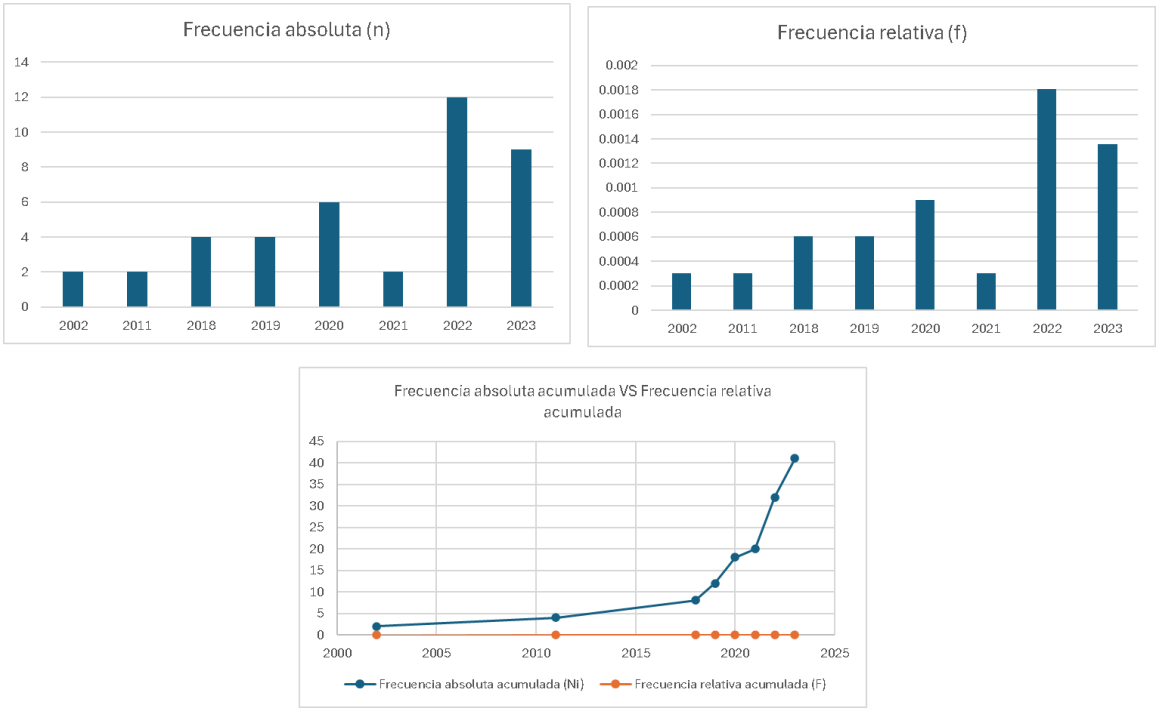
En la tabla comparativa de frecuencia absoluta acumulada contra frecuencia relativa cumulada se observa como existe una gran diferencia en entre la repetición de palabras contra la cantidad de palabra en el texto en general, no existe un límite específico sobre cuántas veces se debe citar un autor o fuente en un trabajo según las normas APA, sin embargo de acuerdo al índice de la frecuencia relativa podemos indicar que el número de referencias es aceptable.

Como se aprecia en la **Gráfica 2 y Tabla 2**, el año más citado antes de validar APA 7 es 2022 con un total de 12 apariciones en el documento y una frecuencia relativa de 0.0018.

AÑOS EN REFERENCIAS E1

Años	Frecuencia absoluta (n)	Frecuencia relativa (f)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa acumulada (F)
2002	2	0.0003012	2	0.0003012
2011	2	0.0003012	4	0.0006024
2018	4	0.0006025	8	0.0012049
2019	4	0.0006025	12	0.0018074
2020	6	0.0009037	18	0.0027111
2021	2	0.0003012	20	0.0030123
2022	12	0.0018075	32	0.0048198
2023	9	0.0013556	41	0.0061754

Tabla 2: Análisis de frecuencias en años para referencias E1.



Gráfica 2. Frecuencias para años.

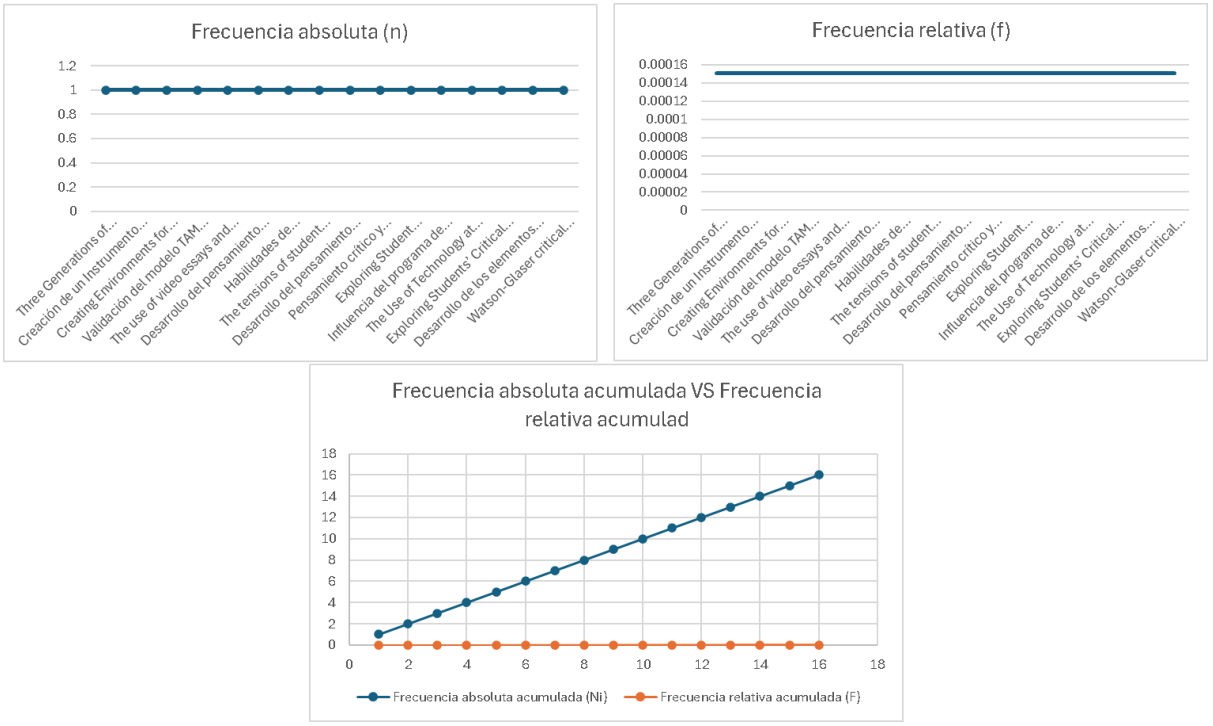
Cada título como cada editorial aparece 1 sola vez en el documento por lo cual podemos asegurar que el procedimiento para referenciar es correcto, se observan los resultados de esto en la **Gráfica 3 y Tabla 3.**

TÍTULOS EN REFERENCIAS E1

Títulos	Repeticiones (n)	Frecuencia relativa (f)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa acumulada (F)
Three Generations of Distance Education Pedagogy.	**1**	**0.0001506**	**1**	**0.0001506**
Creación de un Instrumento de medición pensamiento crítico a través de la matemática: Una aplicación a estudiantes de ingeniería de primer año universitario	**1**	**0.0001506**	**2**	**0.0003012**
Creating Environments for Critical Thinking: Building Upon Multiple Choice Problems in Electrical Engineering Education	**1**	**0.0001506**	**3**	**0.0004518**
Validación del modelo TAM de adopción de la Realidad Aumentada mediante ecuaciones estructurales.	**1**	**0.0001506**	**4**	**0.0006024**
The use of video essays and podcasts to enhance creativity and critical thinking in engineering	**1**	**0.0001506**	**5**	**0.000753**
Desarrollo del pensamiento crítico: Metodología para fomentar el aprendizaje en ingeniería	**1**	**0.0001506**	**6**	**0.0009036**
Habilidades de pensamiento crítico en estudiantes de Ingeniería de Sistemas en modalidad virtual	**1**	**0.0001506**	**7**	**0.0010542**
The tensions of student engagement with technology. In Interactive Learning Environments	**1**	**0.0001506**	**8**	**0.0012048**
Desarrollo del pensamiento crítico en estudiantes de Ingeniería mediante una estrategia didáctica que integra laboratorios remotos sobre circuitos eléctricos: primera intervención	**1**	**0.0001506**	**9**	**0.0013554**
Pensamiento crítico y rendimiento académico en estudiantes de último ciclo en FIEECS-UNI	**1**	**0.0001506**	**10**	**0.001506**

Exploring Student Engagement and Outcomes: Experiences from Three Cycles of an Undergraduate Module	**1**	**0.0001506**	**11**	**0.0016566**
Influencia del programa de pensamiento crítico en el rendimiento académico en el área de matemática de los estudiantes de la Facultad de Ingeniería de Minas de la UNCP	**1**	**0.0001506**	**12**	**0.0018072**
The Use of Technology at the Higher Education Level and Student Engagement: A MetaAnalysis	**1**	**0.0001506**	**13**	**0.0019578**
Exploring Students' Critical Thinking Skills Using the Engineering Design Process in a Physics Classroom	**1**	**0.0001506**	**14**	**0.0021084**
Desarrollo de los elementos del pensamiento crítico y su incidencia en la formación universitaria	**1**	**0.0001506**	**15**	**0.002259**
Watson-Glaser critical thinking appraisal, UK edition : practice test. Psychological Corporation	**1**	**0.0001506**	**16**	**0.0024096**

Tabla 3: Análisis de frecuencias en títulos para referencias E1.



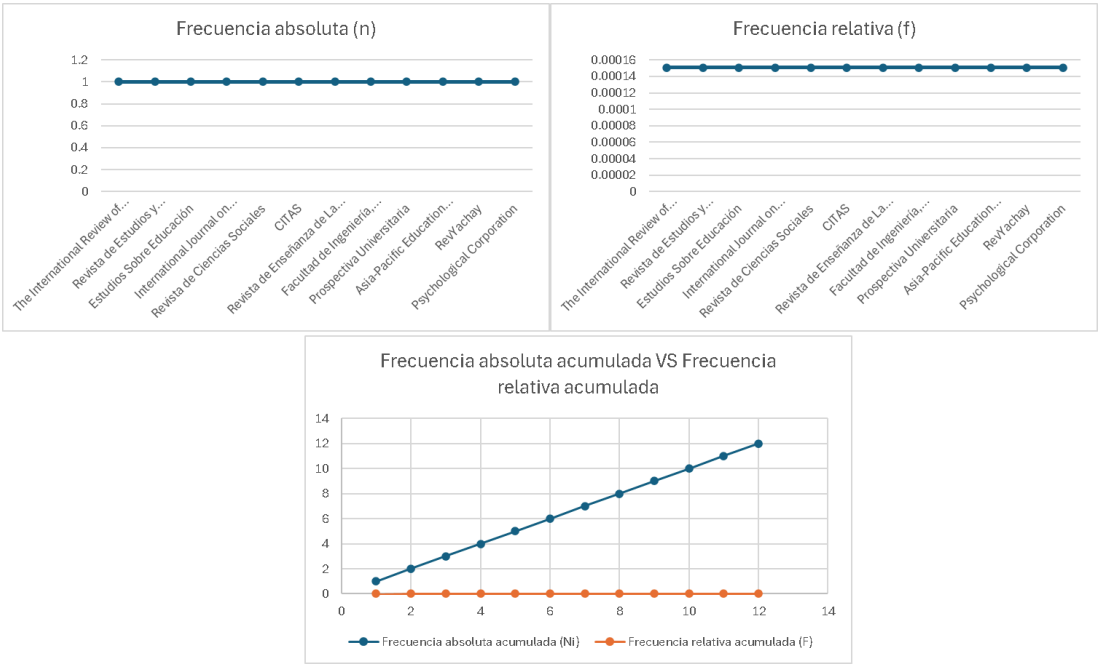
Gráfica 3. Frecuencias para Títulos.

Las frecuencias de los editoriales para cada documento se muestran en la **Tabla 4 y Gráfica 4,** así como su frecuencia relativa, absoluta acumulada y relativa acumulada.

EDITORIAL EN REFERENCIAS E1

Editorial	Repeticiones (n)	Frecuencia relativa (f)	Frecuencia absoluta acumulada (Ni)	Frecuencia relativa acumulada (F)
The International Review of Research in Open and Distributed Learning	**1**	**0.0001506**	**1**	**0.0001506**
Revista de Estudios y Experiencias En Educación	**1**	**0.0001506**	**2**	**0.0003012**
Estudios Sobre Educación	**1**	**0.0001506**	**3**	**0.0004518**
International Journal on Interactive Design and Manufacturing,	**1**	**0.0001506**	**4**	**0.0006024**
Revista de Ciencias Sociales	**1**	**0.0001506**	**5**	**0.000753**
CITAS	**1**	**0.0001506**	**6**	**0.0009036**
Revista de Enseñanza de La Física	**1**	**0.0001506**	**7**	**0.0010542**
Facultad de Ingeniería, Estadística y Ciencias Sociales- Universidad Nacional de Ingeniería	**1**	**0.0001506**	**8**	**0.0012048**
Prospectiva Universitaria	**1**	**0.0001506**	**9**	**0.0013554**
Asia-Pacific Education Researcher	**1**	**0.0001506**	**10**	**0.001506**
RevYachay	**1**	**0.0001506**	**11**	**0.0016566**
Psychological Corporation	**1**	**0.0001506**	**12**	**0.0018072**

Tabla 4: Análisis de frecuencias en editoriales para referencias E1.



Gráfica 4. Frecuencias para Editoriales.

Resultados

En la siguiente sección se detallarán los resultados obtenidos del estudio de los textos presentados, así cómo el entregable anexo cómo lo son el código fuente y la aplicación de escritorio.

El código fuente fue desarrollado en python y se utilizaron librerías de creación de graficas asi cómo el entorno gráficos tkinter para la parte visual del sistema, el código y la aplicación se encontrarán adjuntos a la entrega del presente reporte, es importante detallar que la aplicación solo considera cómo formato valido aquellos que cumplan con las características del patrón de la **Fig. 3.**

```
1  # Patrón básico: Apellido, Inicial. (Año). Título...
2          patron = r"^[A-Z][a-z]+, [A-Z]\. \((\d{4})\)\. .+$"
```

Fig. 3: Patrón detección APA 7.

el cual está diseñado para validar un formato de referencias bibliográficas estilo APA 7 en donde se lista el autor y el año de publicación al principio. Véase **Tabla 1 Apéndice A.**

Cómo resultado se obtienen las gráficas mostradas en la aplicación (**Fig. 4**) y un documento en formato excel con las tablas de los datos que se encontraron

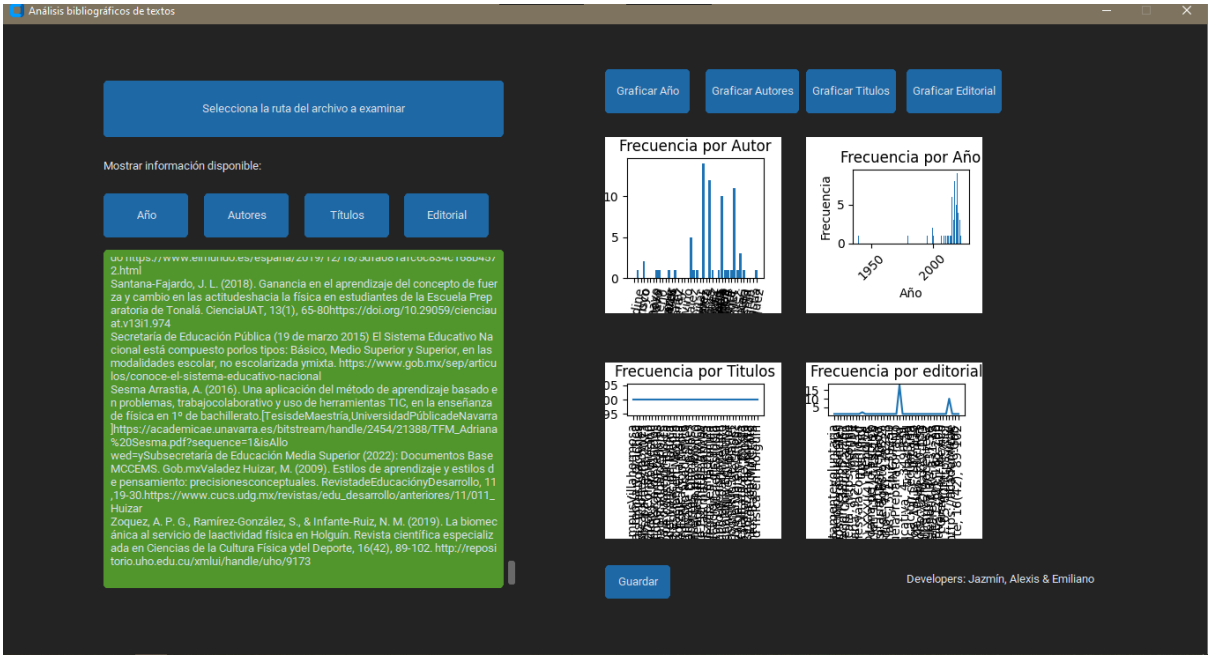
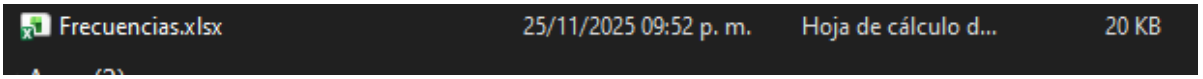
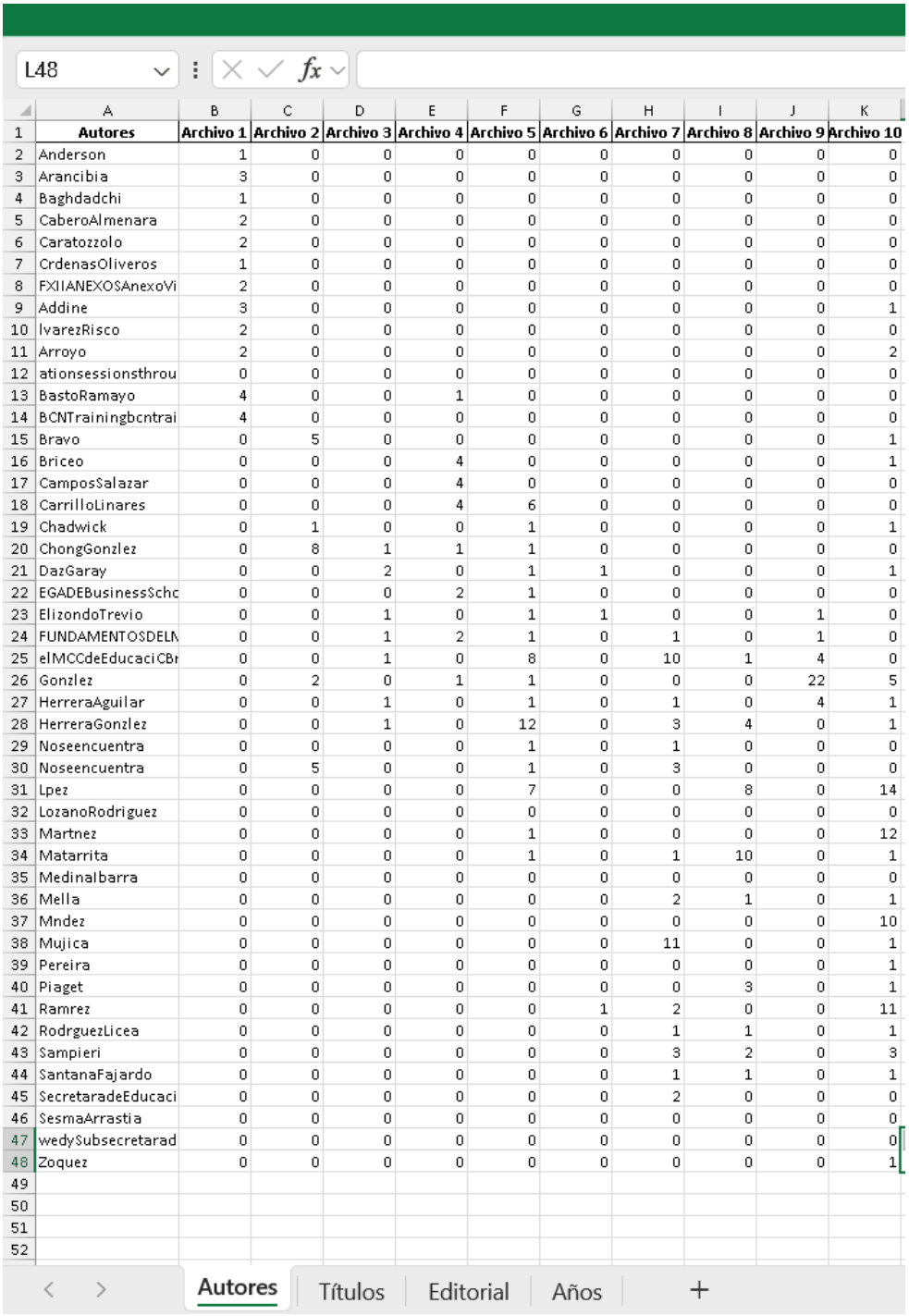


Fig. 4: Interfaz gráfica.



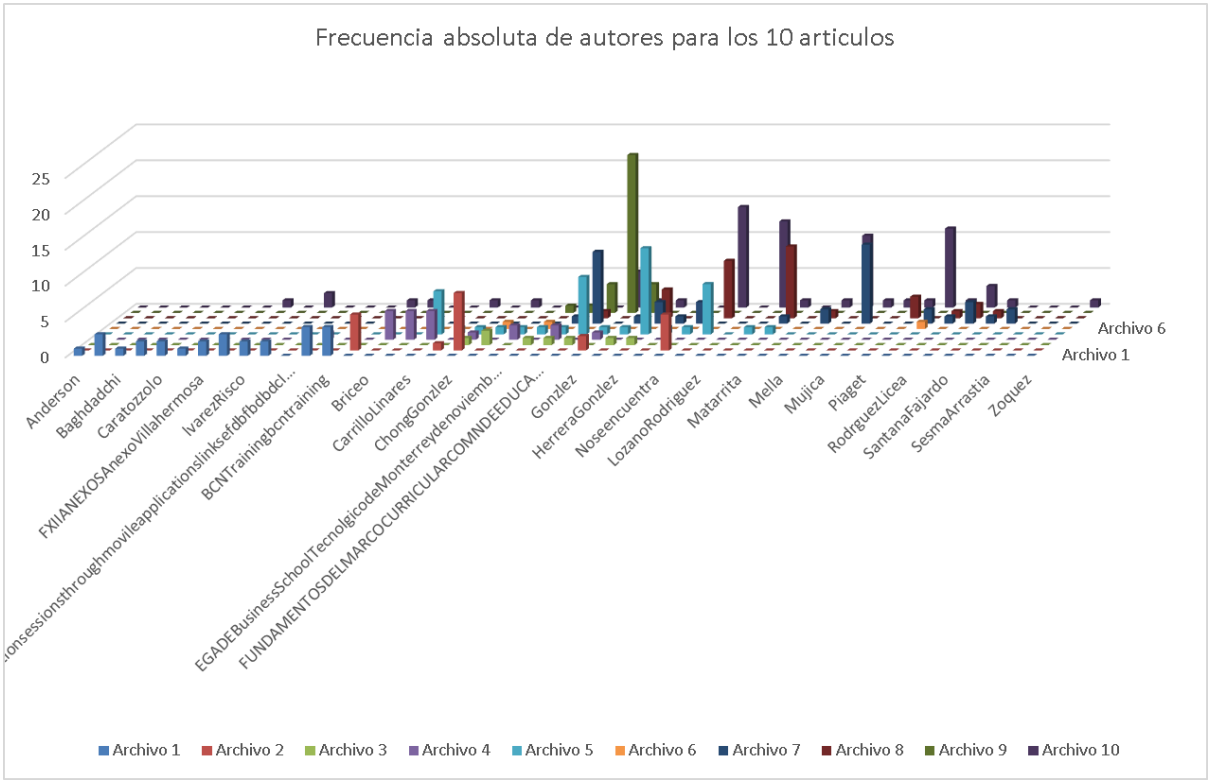
En el archivo de frecuencias se extraen las tablas para cada categoría, graficando en excel se observa la siguiente gráfica con la comparación de frecuencia absoluta entre los diez artículos, en el apartado de metodología se obtuvo como resultado para Autores a Rojas como el más citado, ahora en este análisis comparativo entre los 10 artículos ni siquiera aparece, y esto es debido a que ya se ha validado en APA 7. El contenido del excel al dar click en el botón «Guardar» se ve en la **Fig. 5.**



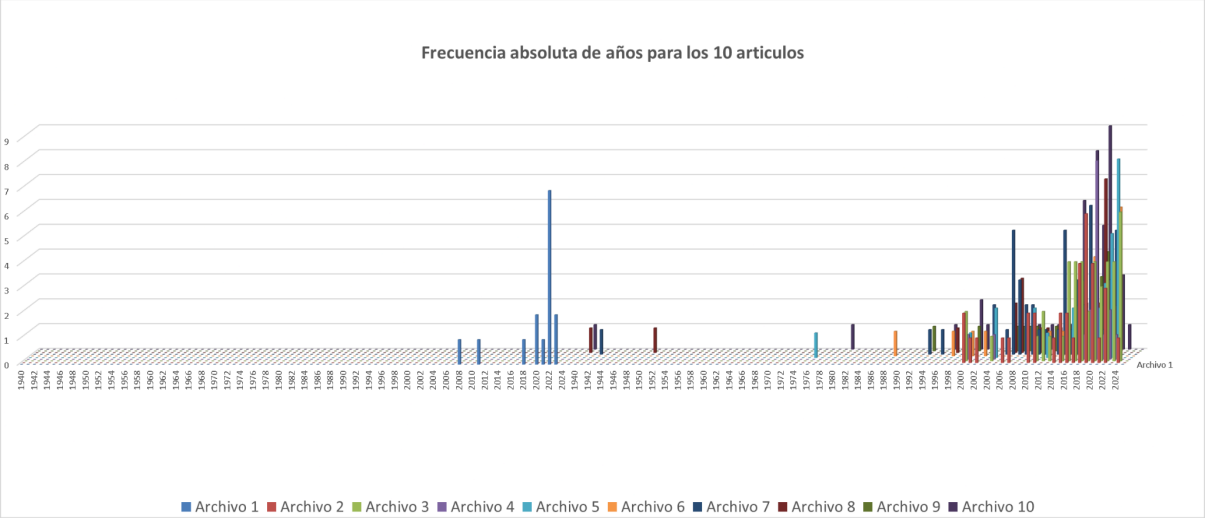
	A	B	C	D	E	F	G	H	I	J	K
1	**Autores**	**Archivo 1**	**Archivo 2**	**Archivo 3**	**Archivo 4**	**Archivo 5**	**Archivo 6**	**Archivo 7**	**Archivo 8**	**Archivo 9**	**Archivo 10**
2	Anderson	1	0	0	0	0	0	0	0	0	0
3	Arancibia	3	0	0	0	0	0	0	0	0	0
4	Baghdadchi	1	0	0	0	0	0	0	0	0	0
5	CaberoAlmenara	2	0	0	0	0	0	0	0	0	0
6	Caratozzolo	2	0	0	0	0	0	0	0	0	0
7	CrdenasOliveros	1	0	0	0	0	0	0	0	0	0
8	FXIIANEXOSAnexoVi	2	0	0	0	0	0	0	0	0	0
9	Addine	3	0	0	0	0	0	0	0	0	1
10	IvarezRisco	2	0	0	0	0	0	0	0	0	0
11	Arroyo	2	0	0	0	0	0	0	0	0	2
12	ationsessionsthrou	0	0	0	0	0	0	0	0	0	0
13	BastoRamayo	4	0	0	1	0	0	0	0	0	0
14	BCNTrainingbcntrai	4	0	0	0	0	0	0	0	0	0
15	Bravo	0	5	0	0	0	0	0	0	0	1
16	Briceo	0	0	0	4	0	0	0	0	0	1
17	CamposSalazar	0	0	0	4	0	0	0	0	0	0
18	CarrilloLinares	0	0	0	4	6	0	0	0	0	0
19	Chadwick	0	1	0	0	1	0	0	0	0	1
20	ChongGonzlez	0	8	1	1	1	0	0	0	0	0
21	DazGaray	0	0	2	0	1	1	0	0	0	1
22	EGADEBusinessSchc	0	0	0	2	1	0	0	0	0	0
23	ElizondoTrevio	0	0	1	0	1	1	0	0	1	0
24	FUNDAMENTOSDELM	0	0	1	2	1	0	1	0	1	0
25	elMCCdeEducaciCBr	0	0	1	0	8	0	10	1	4	0
26	Gonzlez	0	2	0	1	1	0	0	0	22	5
27	HerreraAguilar	0	0	1	0	1	0	1	0	4	1
28	HerreraGonzlez	0	0	1	0	12	0	3	4	0	1
29	Noseencuentra	0	0	0	0	1	0	1	0	0	0
30	Noseencuentra	0	5	0	0	1	0	3	0	0	0
31	Lpez	0	0	0	0	7	0	0	8	0	14
32	LozanoRodriguez	0	0	0	0	0	0	0	0	0	0
33	Martnez	0	0	0	0	1	0	0	0	0	12
34	Matarrita	0	0	0	0	1	0	1	10	0	1
35	Medinalbarra	0	0	0	0	0	0	0	0	0	0
36	Mella	0	0	0	0	0	0	2	1	0	1
37	Mndez	0	0	0	0	0	0	0	0	0	10
38	Mujica	0	0	0	0	0	0	11	0	0	1
39	Pereira	0	0	0	0	0	0	0	0	0	1
40	Piaget	0	0	0	0	0	0	0	3	0	1
41	Ramrez	0	0	0	0	0	1	2	0	0	11
42	RodrguezLicea	0	0	0	0	0	0	1	1	0	1
43	Sampieri	0	0	0	0	0	0	3	2	0	3
44	SantanaFajardo	0	0	0	0	0	0	1	1	0	1
45	SecretaradeEducaci	0	0	0	0	0	0	2	0	0	0
46	SesmaArrastia	0	0	0	0	0	0	0	0	0	0
47	wedySubsecretarad	0	0	0	0	0	0	0	0	0	0
48	Zoquez	0	0	0	0	0	0	0	0	0	1
49											
50											
51											
52											

Fig. 5: Archivo xlsx de salida.

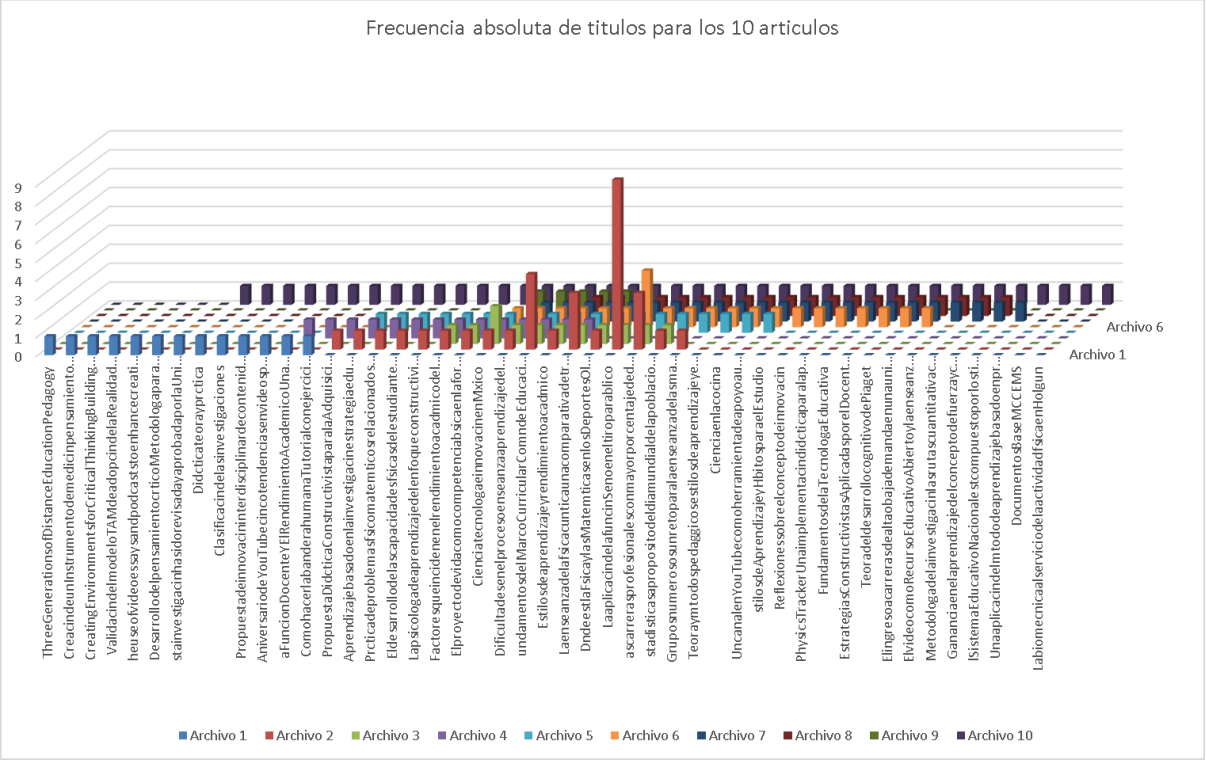
Las frecuencias absolutas por autor se muestran en la **Gráfica 5**; por año, en la **Gráfica 6**; por título, en la **Gráfica 7**; por editorial, en la **Gráfica 8**.



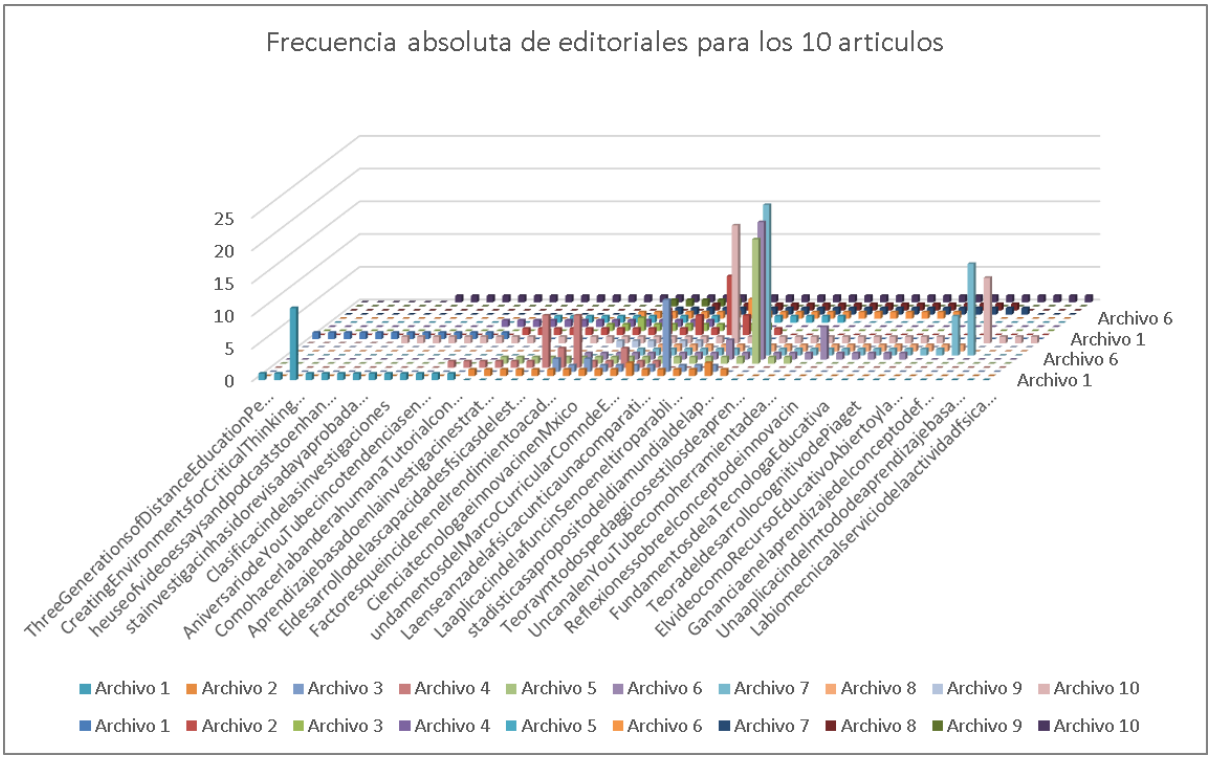
Gráfica 5. Frecuencias absolutas por Autor para los 10 artículos.



Gráfica 6. Frecuencias absolutas por año para los 10 artículos.



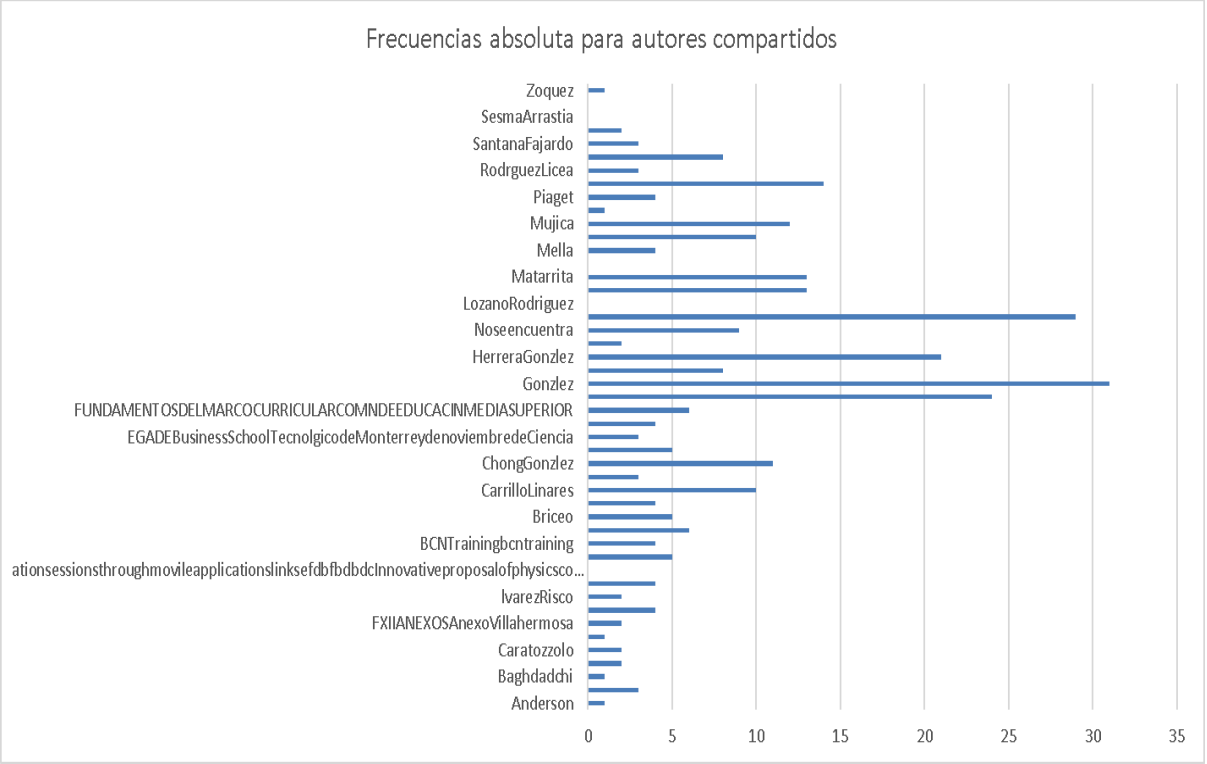
Gráfica 7. Frecuencias absolutas por títulos para los 10 artículos.



Gráfica 8. Frecuencias absolutas por editorial para los 10 artículos.

Tres autores más compartidos, véase **Gráfica 9**:

1. Gonzalez
2. López.
3. Herrera Gonzalez.

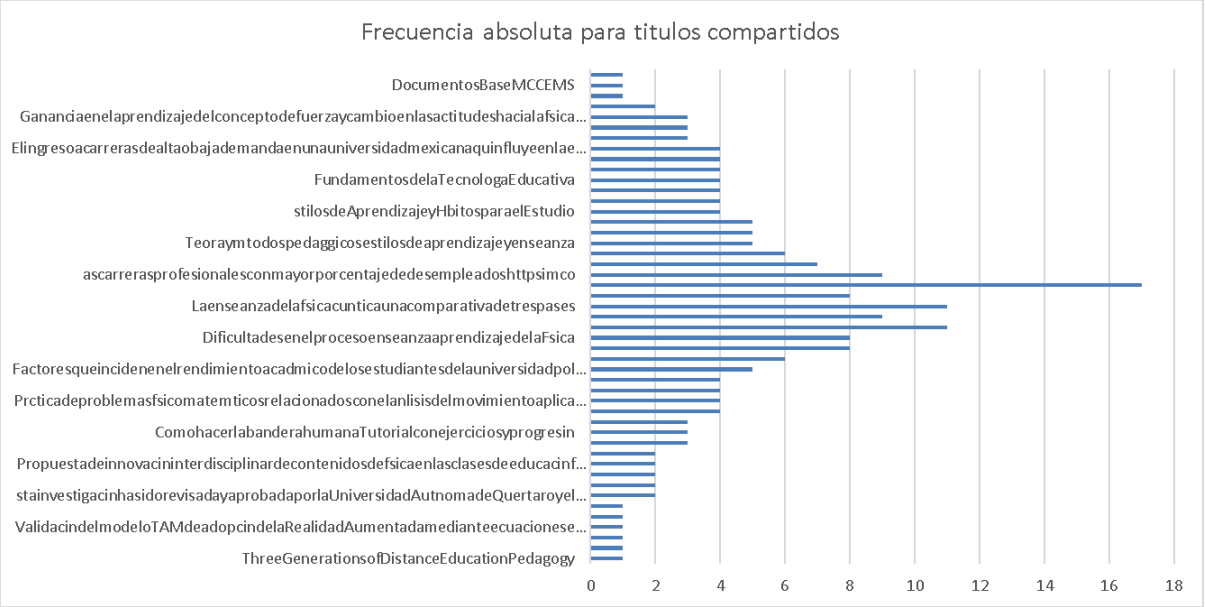


Gráfica 9. Frecuencia absoluta autores para los 10 artículos.

Dentro de los resultados del análisis encontramos 48 autores diferentes de los cuales el que más se repitió fue Gonzalez, aunque este puede no ser información importante ya que es un apellido bastante común. Fue el archivo número 10 el que más referencias encontró.

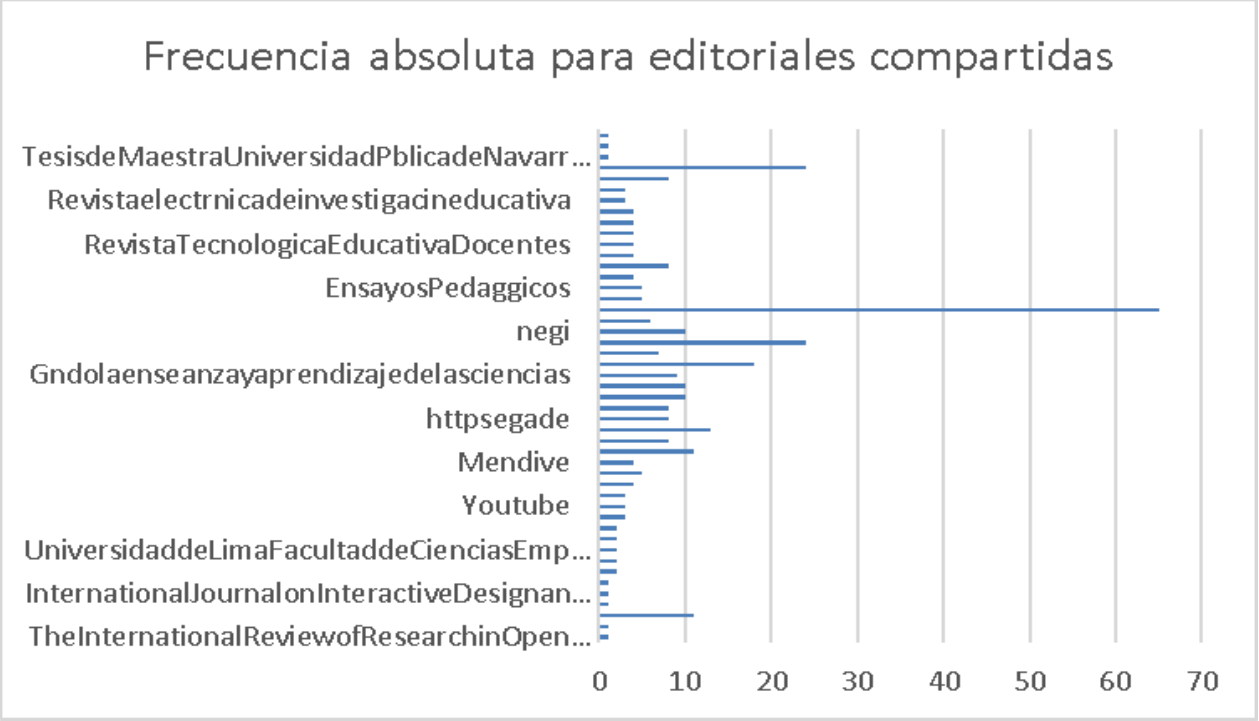
Tres títulos más compartidos, véase **Gráfica 10**:

1. La aplicación de la función Seno en el tiro parabólico.
2. La enseñanza de la física cuántica es una comparativa de tres pasos.
3. Fundamentos del Marco Curricular Común de Educación Media



Gráfica 10. Frecuencia absoluta títulos para los 10 artículos.

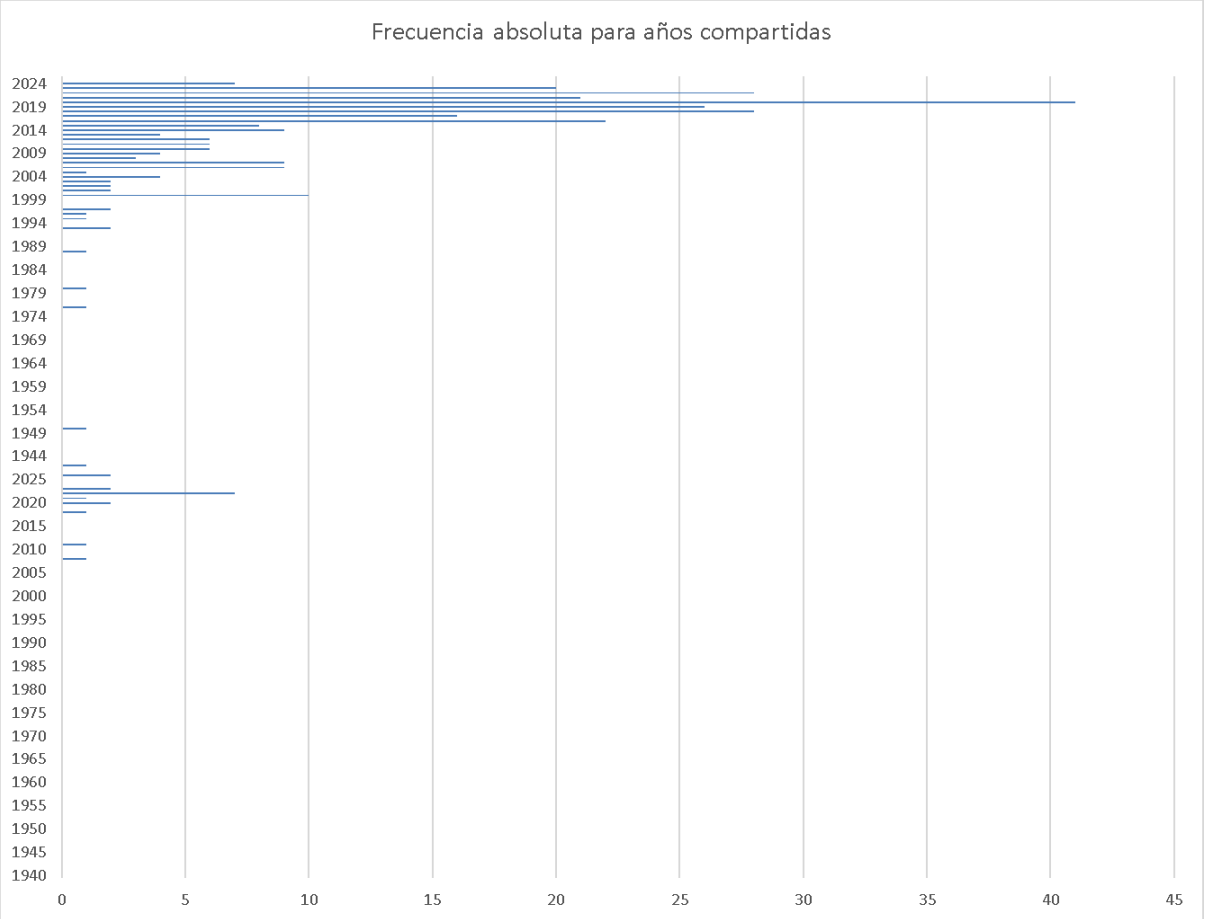
En la **Gráfica 11** se aprecia que, en los títulos de los archivos, el string "LaaplicacindelafuncinSenoeneltiroparablico" que se puede interpretar cómo "La aplicación de la función seno en el tiempo parabólico" fue la más repetida apareciendo 9 veces en el texto 2 y 17 veces en total, aunque nuevamente fue en el archivo número 10 en el que más referencias detectó.



Gráfica 11. Frecuencia absoluta editoriales para los 10 artículos

La editorial más común encontrada se guarda y presenta con el string "Trillased", **Gráfica 12**, apareciendo en 65 ocasiones aunque predomina en los archivos 5, 6 y 7, nuevamente

el archivo 10 detecto más referencias y para esta parte del análisis de los resultados se podría empezar a pensar que influye la longitud del archivo o la correcta utilización del formato APA.



Gráfica 12. Frecuencia absoluta por años para los 10 artículos

La referencia más antigua que se referenció fue de 1940 y la frecuencia de documentos consultados aumenta conforme se acerca al año actual, aunque no hay referencias para este año 2025, siendo la más reciente en el 2024 y el año de publicación más común 2020.

Para este estudio se utilizaron las siguientes 10 tesis:

1. "Desarrollo de una plataforma de aprendizaje interactivo para el desarrollo del pensamiento crítico en estudiantes de primer año de ingeniería".
2. Modelo de resolución de conflictos en infancias del programa UAQ-PERAJ a través del desarrollo de habilidades socioemocionales con recursos digitales, estrategias de comunicación y aprendizaje colaborativo.
3. Diseño de una metodología fundamentada en competencias digitales para la inclusión de adultos mayores de la Universidad Autónoma de Baja California.
4. Estrategia didáctica basada en Chat GPT para escritura de textos académicos en estudiantes de nivel superior

5. Estrategia didáctica con recursos digitales en la enseñanza de español segunda lengua, para enriquecer el aprendizaje significativo de estudiantes de Cultura mexicana en la Facultad de Lenguas y Letras.
6. La impresión 3D como herramienta educativa en el aprendizaje y el proceso de diseño en los estudiantes de educación superior
7. Tecnologías del Aprendizaje y el Conocimiento en la enseñanza de lenguas y Technology-Enhanced Language Learning (TELL) .
8. Modelo basado en análisis estructural para mejorar la calidad educativa de los programas educativos virtuales del TecMN Campus Villahermosa.
9. Propuesta Didáctica para la Enseñanza de la Física General a través del Deporte
10. Implementación de una estrategia para la escritura de ensayos argumentativos a través del uso de las TIC y el STEAM en Educación Media Superior

La mayoría de los artículos analizados están relacionados con estrategias didácticas, de acuerdo a nuestro análisis el artículo: "***Fundamentos del Marco Curricular Común de Educación Media"*** es una excelente opción para utilizar como referencia en investigaciones de la misma área.

Conclusiones

En resumen este estudio tuvo cómo objetivo utilizar herramientas de minería de datos en textos reales para encontrar información de nuestro interés, en este caso las referencias APA, se puede ver con los resultados obtenidos en las gráficas cómo en las tablas presentadas que se logró detectar exitosamente una gran cantidad de datos, sin embargo, el procesamiento de los datos para la detección de las referencias puede carecer de robustez para arrojar información más especifica, tal es el caso del apellido “González” que es el más frecuente en el análisis pero también es uno de los más comunes en México. Haciendo una comparación entre las referencia antes y después de validar APA 7 se perdió un dato importante como fue el autor Rojas, quien era el autor más referenciado en el artículo E1, validando entonces la importante de la correcta referenciación en artículos académicos.

En este sentido los datos sugieren que podríamos realizar un trabajo de minería de datos más robusto, pero vuelve en este punto la pregunta de qué tanto, en qué momento habríamos que parar de robustecer el sistema pues los cambios o diferencias entre los resultados dejan de ser significativos.

Este análisis es bastante útil aplicado para en el desarrollo de tesis para artículos con temas similares, ya que nos permite conocer las referencias más utilizadas por año que podríamos utilizar y obtener información contundente

Apéndices

Apéndice A Código Fuente.

Método para Obtener el Path de Carpeta, devuelve el path y un arreglo que contiene los nombres de los archivos txt

```python
#Método que regresa la ruta y los files txt contenidos
def obtener_path_carpeta(self):
    archivos_txt = []
    root = Tk()
    root.withdraw()  # Oculta la ventana principal de Tkinter
    carpeta_seleccionada = filedialog.askdirectory()

    #archivos_txt.append(carpeta_seleccionada)
    elementos = os.listdir(carpeta_seleccionada)
    for elemento in elementos:
    # La función 'endswith' verifica si el string termina con el sufijo dado
        if elemento.endswith(".txt"):
            archivos_txt.append(elemento)

    return carpeta_seleccionada, archivos_txt
```

Imagen 6, Apéndice A: Método para Obtener el Path de Carpeta

Se busca el Header principal en cada página que debido al formato de los documentos tiene cómo palabras clave “PROTOCOLO DE INVESTIGACIÓN”, una vez que se encuentra una similitud se envía el número de la línea al método para borrar el header, también se borran las líneas que se encuentran vacías

```python

if self.filename != "":
    self.file_name_t = os.path.basename(self.filename)   # Obtener solo el nombre del archivo sin todo el path
    self.nombre_header = "PROTOCOLO DE INVESTIGACIÓN"    # Nombre que aparece en la cabecera de cada página
    self.lineas_archivo_actual = []                      # "Barrer" el archivo línea por línea
    with open(self.filename, "r", encoding="utf-8-sig", errors="replace") as archivo:
        for linea in archivo:
            # linea.isspace(): la línea contiene puros espacios en blanco
            if linea != "" and linea.isspace() == False and self.eliminarHeader(self.nombre_header, linea):
                self.lineas_archivo_actual.append(linea.rstrip()) # rstrip() elimina el salto de línea al final
                self.cont += 1                                    # Imprimir en consola la fila "cont"
    self.archivos_procesados.append(self.lineas_archivo_actual)

```

Imagen 7, Apéndice A: Búsqueda del Header en el texto.

```python

# Método para eliminar el header que tiene una imagen
def eliminarHeader(self, text_b, lin):
    if text_b in lin:
        # Eliminar el header completo porque aparecen 3 líneas:
        # el número de páf, espacio y nómbre del header
        if(self.cont_headers != 0):
            # Eliminar las filas que pertenecen al header
            self.filas.pop() # Espacios en blanco
            self.filas.pop() # Número de página
            self.cont -= 2
            self.cont_headers += 1

        return False
    else:
        return True
```

Imagen 8, Apéndice A: Método para eliminar el header

Método que obtiene el título del archivo procesado y encuentra las referencias del documento.

```python
    def getTituloArchivo(self):
        #ENCONTRANDO SECCIÓN DE REFERENCIA
        self.content_archivo = []   # nuevo diccionario para guardar resultados
        self.referencias_encontradas = []
        self.referencias_para_conteo = []
        self.nombres_t_archivos = []

        for filas_ in self.archivos_procesados:
            ini_ref = 0
            for i, fila in enumerate(filas_):
                if "REFERENCIAS" in fila.upper():
                    ini_ref = i
                    break
# Guardar contenido antes de las referencias en diccionario content_archivo
            contenido = filas_[:ini_ref]
            self.content_archivo.extend(contenido)
            self.referencias_encontradas.append(filas_[ini_ref + 1:])
```

Imagen 9 Apéndice A: Método que extrae referencias.

Método para extraer las referencias APA 7

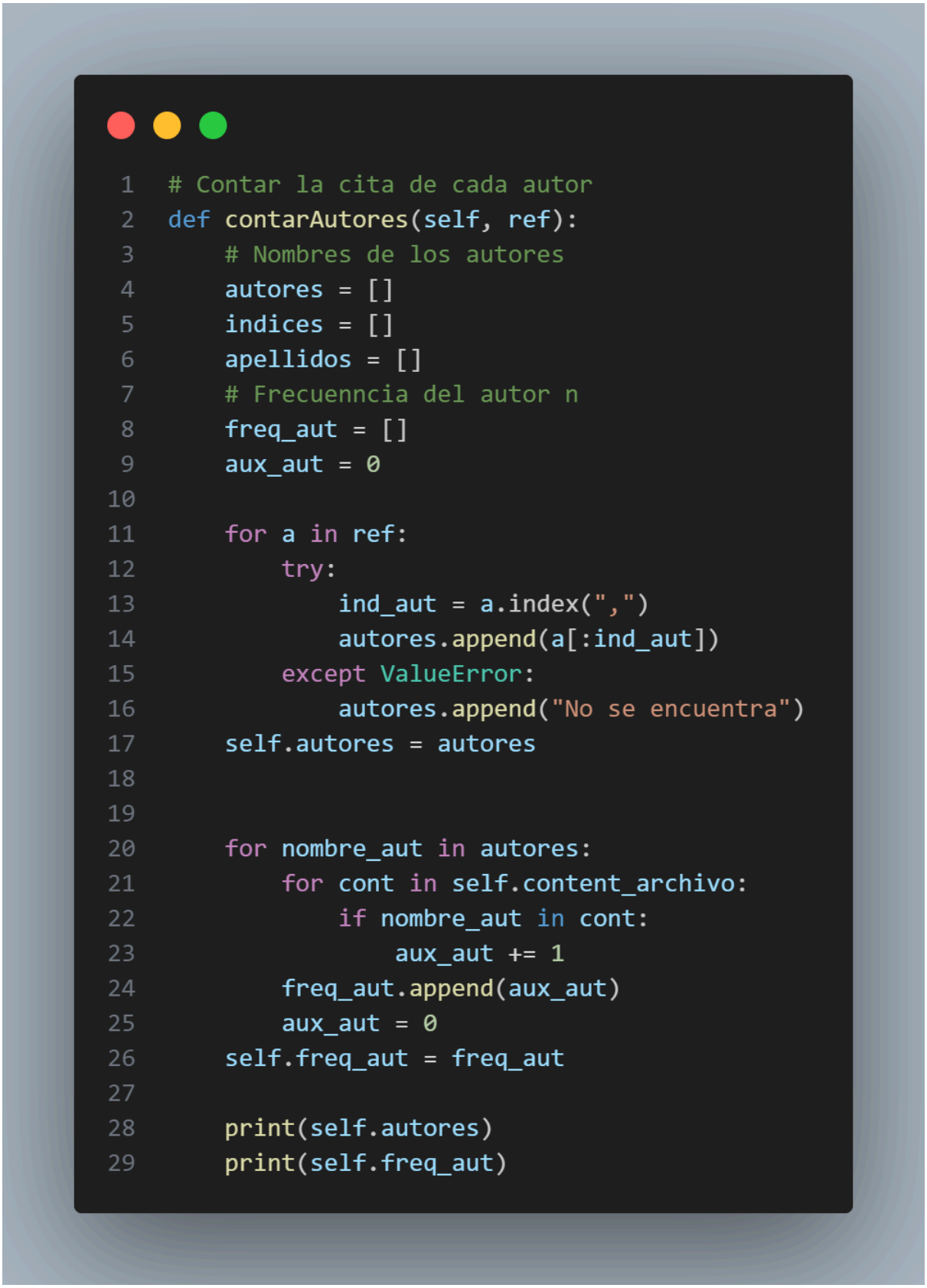
```python
# Método para extraer las referencias en formato APA7
def extraerReferenciasAPA7(self):
    self.referencias_separadas = []   # lista final de referencias ya separadas
    aux_referencias = ""
    self.archivos_ref = []
    todas_ref = []

    for i in range (0, 10):
        for fila in self.referencias_encontradas[i]:
            aux_referencias = aux_referencias + fila
            # Separar por DOI
            if "HTTP" in fila.upper():
                #if self.validarReferencia(aux_referencias):
                print(fila)
                self.referencias_separadas.append(aux_referencias)
                todas_ref.append(aux_referencias)
                aux_referencias = ""
                print(len(self.referencias_separadas))
        self.archivos_ref.append(self.referencias_separadas)
        print(self.archivos_ref)
        self.referencias_separadas = [
```

Imagen 10 Apéndice A: Método que extrae las referencias.

```python
#
def validarReferencia(self, referencia):
    # Patrón básico: Apellido, Inicial. (Año). Título...
    patron = r"^[A-Z][a-z]+, [A-Z]\. \((\d{4})\)\. .+$"
    if re.match(patron, referencia):
        return True
    else:
        return False
```

Imagen 11 Apéndice A: Método Validar Referencia



```python
# Contar la cita de cada autor
def contarAutores(self, ref):
    # Nombres de los autores
    autores = []
    indices = []
    apellidos = []
    # Frecuenncia del autor n
    freq_aut = []
    aux_aut = 0

    for a in ref:
        try:
            ind_aut = a.index(",")
            autores.append(a[:ind_aut])
        except ValueError:
            autores.append("No se encuentra")
    self.autores = autores


    for nombre_aut in autores:
        for cont in self.content_archivo:
            if nombre_aut in cont:
                aux_aut += 1
        freq_aut.append(aux_aut)
        aux_aut = 0
    self.freq_aut = freq_aut

    print(self.autores)
    print(self.freq_aut)
```

Imagen 12 Apéndice A: Método Contar autores

```python
def TablaAutores(self):
    self.graf_Autores_btn.place(x = self.ancho/12*7, y = self.alto/15)

# Método que genera la tabla de tiempos
def TablaTiempo(self):
    self.graf_Tiempo_btn.place(x = self.ancho/12*6, y = self.alto/15)

# Método que genera la tabla de titulos
def TablaTitulos(self):
    self.graf_Titulos_btn.place(x = self.ancho/12*8, y = self.alto/15)

# Método que genera la tabla de Editoriales
def TablaEditorial(self):
    self.graf_Editorial_btn.place(x = self.ancho/12*9, y = self.alto/15)

```

Imagen 13 Apéndice A: Métodos para generar las tablas

```python
# Método que genera grafica de autores
def graficarAutores(self):
    fig, ax = plt.subplots(figsize=(2,2), dpi = 100)      #DPI son puntos por pulgada

    ax.bar(self.autores, self.freq_aut)
    ax.set_title("Frecuencia por Autor")
    ax.set_xlabel("Autores")
    ax.set_ylabel("Frecuencia")
    plt.xticks(rotation=80)
    plt.subplots_adjust(bottom=0.20)
    plt.tight_layout()

    canvas = FigureCanvasTkAgg(fig, master=self.master)
    canvas.draw()

    widget = canvas.get_tk_widget()
    widget.place(x = self.ancho/6 * 3, y = self.alto/6)
```

Imagen 14 Apéndice A: Métodos para generar gráficas

```python
def guardarDatos(self):
    with pd.ExcelWriter('Frecuencias.xlsx') as writer:
        self.df1.to_excel(writer, sheet_name='Autores', index=False)
        self.df2.to_excel(writer, sheet_name='Títulos', index=False)
        self.df3.to_excel(writer, sheet_name='Editorial', index=False)
        self.df4.to_excel(writer, sheet_name='Años', index=False)

```

Imagen 15 Apéndice A:Método que guarda datos en archivo de xlsx

Carácter	Descripción
^	Inicio de la cadena al inicio del texto
[A-Z]	Coincide con una letra mayúscula
[a-z]+	Coincide con una o más minúsculas
,	Coincide con una coma
[A-Z]	Coincide con una letra mayúscula
\.	Coincide con un punto
\	Coincide con un espacio
\((\d{4})\)	Coincide con un año entre paréntesis
\.	Coincide con un punto
$	Fin de la cadena

Referencias

Martinez, B. B. (2001). Minería de datos. *Cómo hallar una aguja en un pajar. Ingenierías*, *14*(53), 53-66.

Normas APA actualizadas, Formatos.https://normas-apa.org/referencias/