**Ejercicio Procesamiento de Lenguaje Natural**

**Carlos Andrés Capachero Martínez​, Javier Enrique Huérfano Diaz,**

**María Fernanda Izquierdo Aparicio**

# Descripción inicial de la córpora (35%):

## ¿Qué campos son los más interesantes desde una perspectiva de procesamiento de lenguaje natural? ¿Por qué?

* **Resúmenes:** Son esenciales para tareas como la clasificación de género, generación de texto, y resumen automático. Analizar los resúmenes puede ayudar a entender cómo se presentan los temas principales de los libros de manera condensada.
* **Autor y géneros:** Facilitan la clasificación del estilo de escritura y permiten realizar análisis temáticos que podrían ser útiles en estudios literarios o de NLP.
* **Fecha de publicación:** Analizar esta variable permite identificar tendencias temporales y cambios lingüísticos a lo largo del tiempo.

## ¿Qué autores son los más comunes en la córpora analizada? ¿Qué porcentaje de los libros tienen un autor asociado?

Los autores más comunes en el CMU Book Summary Corpus incluyen nombres reconocidos de la literatura popular, con un enfoque en autores de géneros como el misterio, la fantasía y la ciencia ficción. A continuación, se presenta una tabla con los autores más representados:

|  |  |  |
| --- | --- | --- |
| Autor | TOTAL LIBROS | % DE LIBROS SOBRE TOTAL |
| Sin Autor | 2382 | 14,385% |
| Agatha Christie | 74 | 0,447% |
| Franklin W. Dixon | 68 | 0,411% |
| K. A. Applegate | 62 | 0,374% |
| Stephen King | 60 | 0,362% |
| Edgar Rice Burroughs | 59 | 0,356% |
| Otros Autores (4709) | 13854 | 83,664% |

Es interesante observar que una gran cantidad de libros (14.385%) no tienen un autor claramente identificado, lo cual puede deberse a diferentes factores: puede tratarse de obras antiguas donde el autor no es conocido, libros que han caído en el dominio público, o errores en la recopilación de metadatos. Esta falta de atribución en los autores podría limitar ciertos análisis que dependen de la consistencia de la información del autor.

Se observa que aproximadamente el 85.62% de los libros en el corpus tienen un autor identificado, lo que sugiere una buena cobertura en términos de metadatos de autor.

## ¿Qué periodo de tiempo cubre los libros analizados? ¿Cuál es el libro más reciente? ¿más antiguo?

El CMU Book Summary Corpus abarca un período extenso, desde el año 398 hasta 2013, lo que proporciona una base de datos histórica valiosa para analizar la evolución de la literatura a lo largo del tiempo. Este rango temporal permite investigar cambios en los estilos narrativos, temas populares en diferentes eras y la evolución de géneros literarios.

* El libro más reciente es Dr. Sleep de Stephen King (2013)
* El libro más antiguo es The Confessions of St. Augustine de Agustín de Hipona (389)

Es importante tener en cuenta que 5610 libros (33.879% del total) no contienen información de fecha de publicación, lo que representa una limitación significativa en el corpus. Esta falta de datos puede afectar los análisis longitudinales y temporales.

## ¿Cuál es el título más largo? ¿más corto?

El libro con el título más largo es:

|  |  |  |
| --- | --- | --- |
| título | autor | fecha de publicación |
| Passionate Minds: The Great Love Affair of the Enlightenment, Featuring the Scientist Emilie du Chatelet, the Poet Voltaire, Sword Fights, Book Burnings, Assorted Kings, Seditious Verse, and the Birth of the Modern World | David Bodanis | 2006-10 |

Los libros con el título más corto son (1 carácter):

|  |  |  |
| --- | --- | --- |
| título | autor | fecha de publicación |
| Q |  | 1999 |
| L |  |  |
| e | Matt Beaumont |  |
| V |  | 2008-02-05 |

## ¿Qué géneros son los más comunes?

El análisis de los géneros en el CMU Book Summary Corpus revela una amplia variedad que refleja las tendencias literarias a lo largo del tiempo. De los 227 géneros identificados, se presenta un resumen de los géneros más comunes:

|  |  |  |
| --- | --- | --- |
| Género | Conteo de libros | % de libros con este género |
| Fiction | 4747 | 28,67% |
| Speculative Fiction | 4314 | 26,05% |
| Science Fiction | 2870 | 17,33% |
| Novel | 2463 | 14,87% |
| Fantasy | 2413 | 14,57% |

## ¿Hay problemas de calidad en la córpora? ¿Qué tipo de problemas?

El CMU Book Summary Corpus tiene varios problemas de calidad que pueden afectar los análisis realizados sobre él:

* **Inconsistencias en los resúmenes:** Los resúmenes varían en longitud y estilo, lo que dificulta comparaciones entre ellos y afecta la uniformidad del corpus.
* **Exactitud de la información:** Dado que los resúmenes fueron extraídos de Wikipedia, pueden contener errores o estar desactualizados, lo que compromete la confiabilidad de los datos.
* **Falta de estandarización:** La falta de un formato uniforme para los resúmenes dificulta su uso en estudios formales y puede no cumplir con criterios académicos.
* **Problemas de metadatos:** Aunque se incluyen detalles como el autor y el género, estos pueden no ser siempre son consistentes o precisos, afectando la calidad de los análisis.
* **Sesgo en la selección de libros:** El corpus está compuesto en su mayoría por libros con resúmenes disponibles en Wikipedia, lo que puede excluir obras menos conocidas pero relevantes.

## ¿Algún otro análisis qué nos permita entender mejor la córpora? (Bono por propuestas interesantes)

Para comprender mejor el CMU Book Summary Corpus, se pueden realizar varios análisis que abordan diferentes aspectos de la calidad y la utilidad del corpus. A continuación, se presentan algunas metodologías y enfoques que pueden ser útiles:

* Análisis de Contenido
  + Evaluación de la Semántica: Un análisis semántico puede ayudar a determinar la precisión y relevancia del contenido de los resúmenes en relación con los libros originales. Esto implica comparar los resúmenes con las tramas y temas principales de los libros para identificar discrepancias o omisiones significativas.
  + Análisis Estructural: Examinar la estructura de los resúmenes (longitud, organización, uso de lenguaje) puede revelar patrones que afectan la comprensión del texto. Esto incluye verificar si los resúmenes siguen una estructura narrativa coherente.
  + Pragmática y Uso: Un estudio sobre cómo se utilizan estos resúmenes en diferentes contextos (educativos, informativos) puede proporcionar información sobre su efectividad y aplicabilidad en situaciones del mundo real.
* Análisis Comparativo
  + Comparación con Otros Corpora: Comparar el CMU Book Summary Corpus con otros corpora similares puede ayudar a identificar fortalezas y debilidades. Por ejemplo, se podría analizar cómo se diferencian los resúmenes de este corpus en términos de calidad y estilo en comparación con resúmenes generados automáticamente o por humanos.
  + Estudio de Sesgos: Realizar un análisis sobre el sesgo en la selección de libros podría revelar si ciertos géneros o autores están sobrerrepresentados, lo que podría afectar la diversidad del corpus.
* Análisis Cuantitativo
  + Estadísticas Descriptivas: Generar estadísticas sobre la longitud promedio de los resúmenes, la frecuencia de palabras clave y la distribución de géneros literarios puede ofrecer una visión general del corpus.
  + Modelos de Aprendizaje Automático: Implementar modelos de aprendizaje automático para clasificar o predecir géneros basados en los resúmenes puede proporcionar insights sobre las características lingüísticas que definen cada género, aunque esto también depende de la calidad del corpus
* Análisis Cualitativo
  + Revisión Crítica: Realizar una revisión crítica por parte de expertos en literatura o educación puede ofrecer perspectivas sobre la calidad literaria y pedagógica del corpus.
  + Encuestas a Usuarios: Recoger opiniones de investigadores y estudiantes que utilizan el corpus puede ayudar a identificar problemas prácticos que no son evidentes a través del análisis técnico.

# Preprocesamiento de Texto:

## ¿Cuántas palabras tiene en total la sección de ‘Plot Summary’? ¿Cuál es el promedio por libro? Realice el mismo análisis con palabras únicas. Tip: Recuerde trabajar todo en minúsculas o mayúsculas y eliminar signos de puntuación.

* El total de palabras de todos los resúmenes es: **7.096.809**.
* El promedio de palabras por resumen es: **429,58**.

Para el análisis de palabras únicas, tras normalizar el texto (conversión a minúsculas y eliminación de signos de puntuación):

* El número total de palabras únicas es: **156.201**

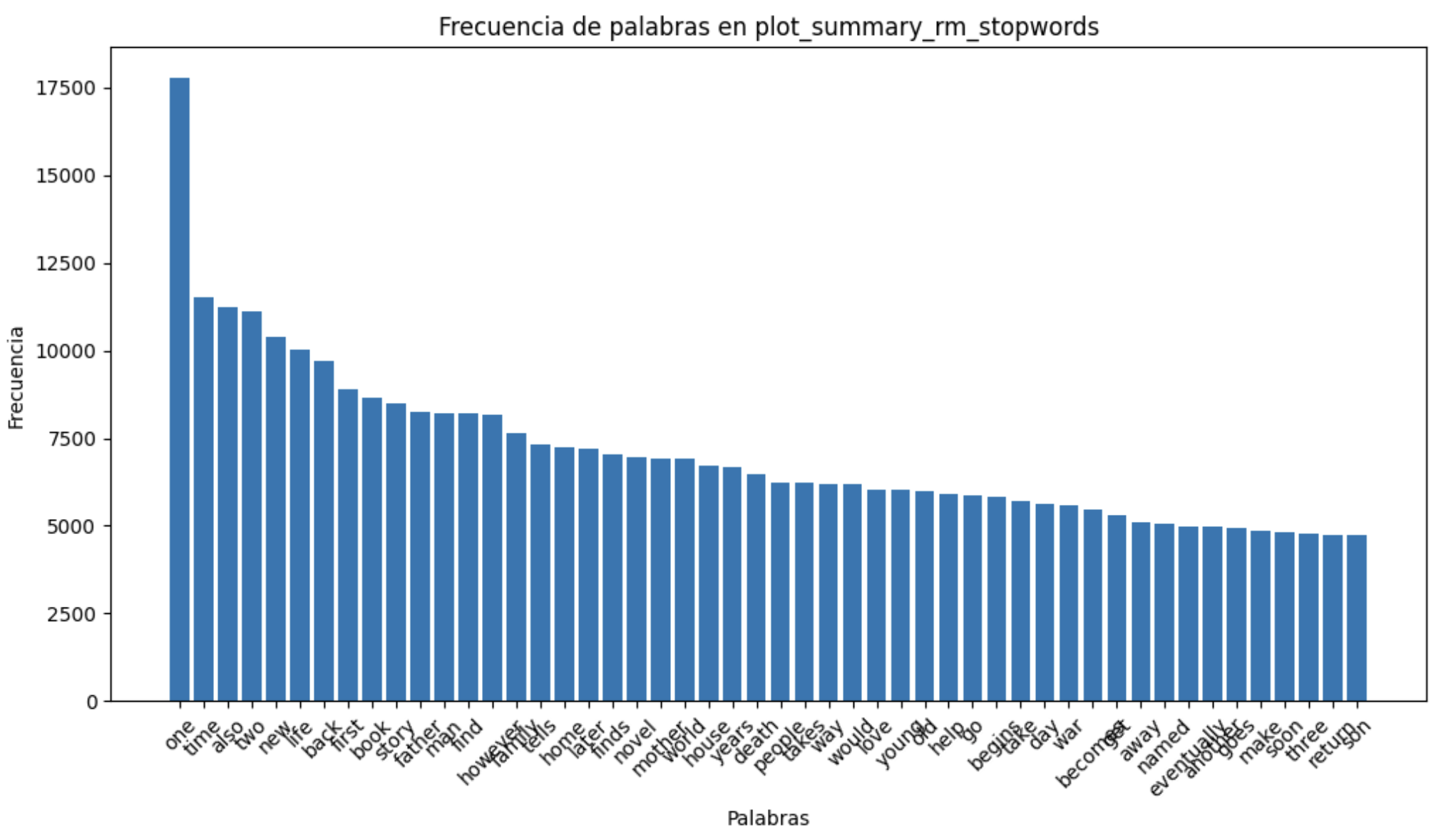
El alto número de palabras totales y únicas sugiere que el corpus contiene una variedad considerable de términos, lo cual es esencial para un análisis más profundo sobre la diversidad de lenguaje utilizado en los resúmenes de los libros.

## Use la lista de stopwords de NLTK. De las palabras que no fueron filtradas por esta lista, ¿qué palabras se deberían eliminar? Justifique su decisión.

Después de remover las stopwords, quedan **147.438** palabras únicas. Entre estas, algunas palabras altamente frecuentes y genéricas se proponen para eliminación:

|  |  |
| --- | --- |
| Palabra | Frecuencia |
| one | 17.765 |
| time | 11.500 |
| also | 11.223 |
| two | 11.125 |
| new | 10.408 |

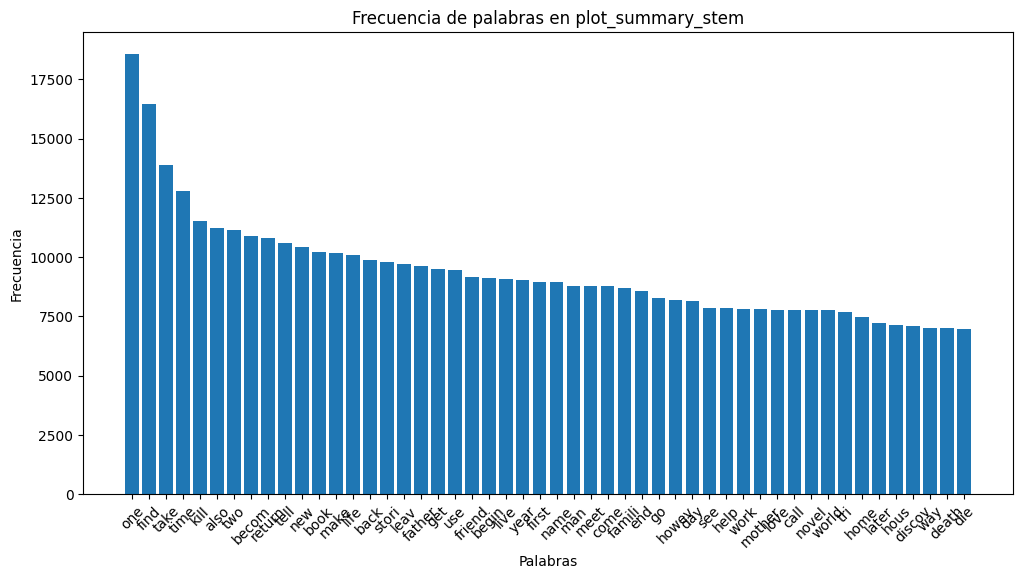
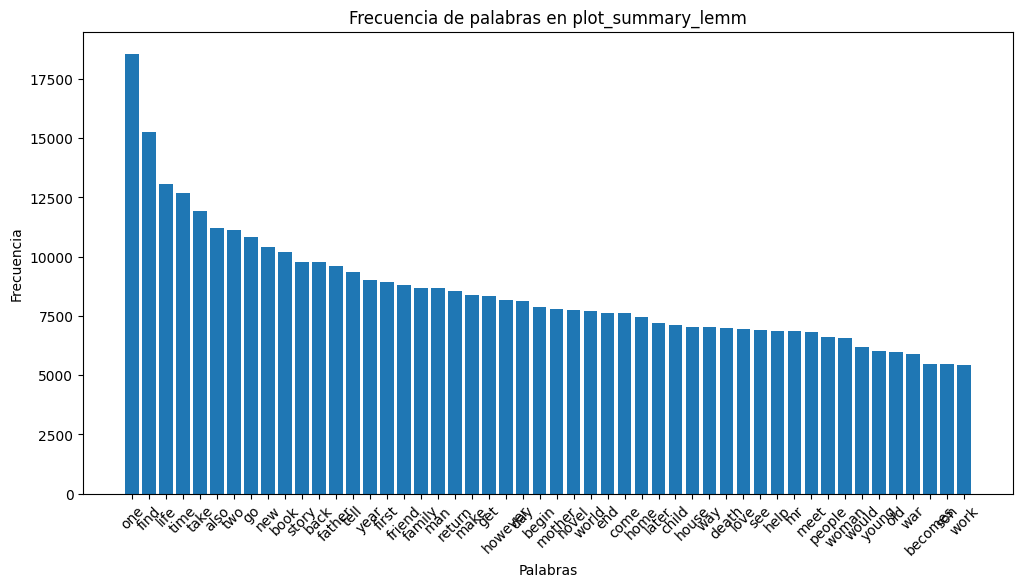
Estas palabras, aunque no son parte de las stopwords estándar, son genéricas y no aportan un valor significativo al contenido de los resúmenes. "One", "two", "time", entre otras, aparecen con alta frecuencia, pero no contribuyen a una comprensión más profunda del argumento del libro. Eliminarlas permitirá enfocar el análisis en palabras que ofrezcan información más específica y relevante para tareas como clasificación o resumen automático.



## Realice stemming y lematización. ¿En cuánto se reduce el número de palabras únicas en cada caso? ¿Por qué la diferencia?

* Después de aplicar lematización, el número de palabras únicas se reduce de 147.438 a 135.246.
* Después de aplicar stemming, el número de palabras únicas se reduce de 147.438 a 105.190.

La diferencia en la reducción se debe a que el stemming corta las palabras a su raíz más básica, lo que puede llevar a reducir varias formas de una palabra a una sola representación (e.g., "running", "ran" y "runner" se convierten en "run"). Sin embargo, esta técnica puede ser menos precisa, ya que no siempre se conserva el significado correcto de las palabras. En cambio, la lematización considera el contexto y transforma las palabras a su forma base o "lema", manteniendo la integridad semántica, lo que resulta en una menor reducción de palabras pero mayor precisión.



## Con base en el tipo de análisis a realizar en la próxima sección y otras restricciones, ¿preferiría usar stemming o lematización para trabajar con la columna de ‘Plot Summary’? Justifique su respuesta.

Entre stemming y lematización, preferiríamos usar lematización para trabajar con los resúmenes de los libros. La razón principal es que la lematización conserva mejor el significado de las palabras, lo cual es crucial en un corpus de resúmenes donde el contexto y la precisión del lenguaje son importantes para el análisis. Dado que los resúmenes están diseñados para capturar los detalles clave de cada libro, es esencial que el análisis semántico mantenga la coherencia entre palabras similares sin perder su contexto o significado original. Aunque el stemming reduce más el número de palabras únicas, podría generar ambigüedades al combinar palabras que no necesariamente tienen el mismo sentido.

# Text Classification:

El proceso seguido para clasificar los libros como “Science Fiction” se presenta a continuación:

## 3.1. Preparación de los datos

En la preparación de los datos, se utilizó la columna plot\_summary\_lem como características (X), la cual representa el resumen lematizado del argumento de cada libro. La etiqueta de clasificación (Y) se obtuvo de la columna ScienceFiction, donde 1 indica que un libro es de ciencia ficción y 0 que no lo es. A continuación, el conjunto de datos se dividió en un 80% para el entrenamiento y un 20% para la prueba.

## 3.2. Transformación TF-IDF

Se utilizó *TfidfVectorizer* para convertir el texto de X\_train y X\_test en representaciones numéricas de TF-IDF. Esto captura la importancia relativa de cada palabra dentro de cada documento (resumen), y elimina las palabras comunes en todo el conjunto de datos, produciendo X\_train\_tfidf y X\_test\_tfidf.

## 3. Entrenamiento del modelo:

Se entrenó un clasificador de Máquina de Vectores de Soporte lineal (LinearSVC) usando los datos transformados (X\_train\_tfidf) y sus etiquetas (y\_train). El modelo ahora está entrenado para predecir si un libro pertenece al género de ciencia ficción.

## 4. Evaluación del modelo:

Se evaluó el modelo usando classification\_report y la matriz de confusión, que proporcionan las métricas de rendimiento del modelo en el conjunto de prueba (X\_test\_tfidf y y\_test). El resultado es:

- Precisión (precision): Qué tan bien el modelo evita los falsos positivos.

- Sensibilidad(recall): Qué tan bien el modelo captura los positivos reales.

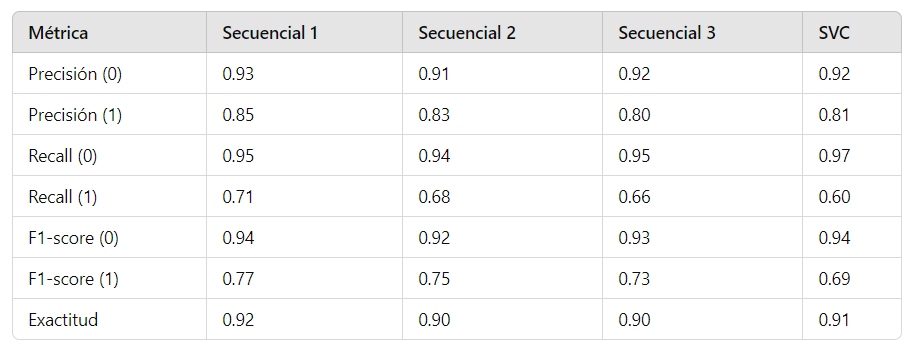
- F1-score: Un equilibrio entre precisión y sensibilidad.

- La precisión general fue del 91%, con el modelo funcionando mejor en identificar los libros que no son de ciencia ficción (0) en comparación con los que sí lo son (1).

|  |  |
| --- | --- |
|  |  |

Posteriormente se implementaron otros tres modelos secuenciales para abordar el problema de clasificación binaria, cada uno con diferentes arquitecturas y enfoques con el propósito de verificar otras opciones.

El Modelo 1 consistió en capas densas con activación ReLU y una capa de salida con activación sigmoide, optimizado para maximizar la precisión. El Modelo 2 amplió esta estructura, incorporando más capas y técnicas de regularización como Dropout, con el objetivo de mejorar la recuperación, lo cual es crucial para identificar correctamente las instancias positivas. Por último, el Modelo 3 adoptó una arquitectura más simple, equilibrando precisión y recuperación, lo que lo hizo menos propenso al sobreajuste. Cada modelo fue evaluado en función de métricas como precisión, recuperación, F1-score y exactitud, permitiendo una comparación exhaustiva para determinar cuál se adaptaba mejor a los requerimientos del problema. A continuación se presentan los resultados:



Los modelos secuenciales (1, 2 y 3) presentan un rendimiento sólido en la clasificación de libros de ciencia ficción, con métricas de precisión y recall que indican su capacidad para identificar correctamente tanto los libros de ciencia ficción como los no de ciencia ficción. En particular, el Secuencial 1 muestra la mejor precisión para la clase positiva (ciencia ficción) con 0.85 y un recall de 0.71, lo que sugiere que es eficaz en la identificación de estos libros.

El SVC tuvo una precisión de 0.81 en la clase positiva y un recall de 0.60, lo que significa que, aunque pudo clasificar la mayoría de los libros no de ciencia ficción con un recall muy alto de 0.97, su capacidad para identificar libros de ciencia ficción fue inferior a la de los modelos secuenciales.

En términos de F1-score, el SVC logró un puntaje de 0.69 para la clase positiva, que es notablemente más bajo que los modelos secuenciales, que oscilaron entre 0.73 y 0.77. A pesar de esto, el SVC logró un F1-score de 0.94 para la clase negativa, equiparando el rendimiento de los secuenciales.

En conclusión, aunque el SVC presenta buenos resultados, los modelos secuenciales ofrecen un rendimiento superior en la identificación de libros de ciencia ficción, lo que sugiere que podrían ser más adecuados para esta tarea específica.