Trabajo practico integrador: PLN

Análisis Comparativo del Lenguaje en los Cuentos de Horacio Quiroga

Materia: Procesamiento de Lenguaje Natural - IFTS24

Profesor: Matías Barreto

Fecha: Septiembre 2025

Hipotesis

"¿Cómo difiere el lenguaje, los temas y el tono en los cuentos de Horacio Quiroga cuando escribe para un público adulto (explorando la muerte, la locura y la crudeza de la naturaleza) frente a cuando escribe fábulas para un público infantil?"

Hipótesis: Esperamos encontrar un vocabulario significativamente más oscuro, violento y complejo en los cuentos para adultos, mientras que las fábulas utilizarán un lenguaje más simple, centrado en animales y con una connotación más positiva o moralizante.

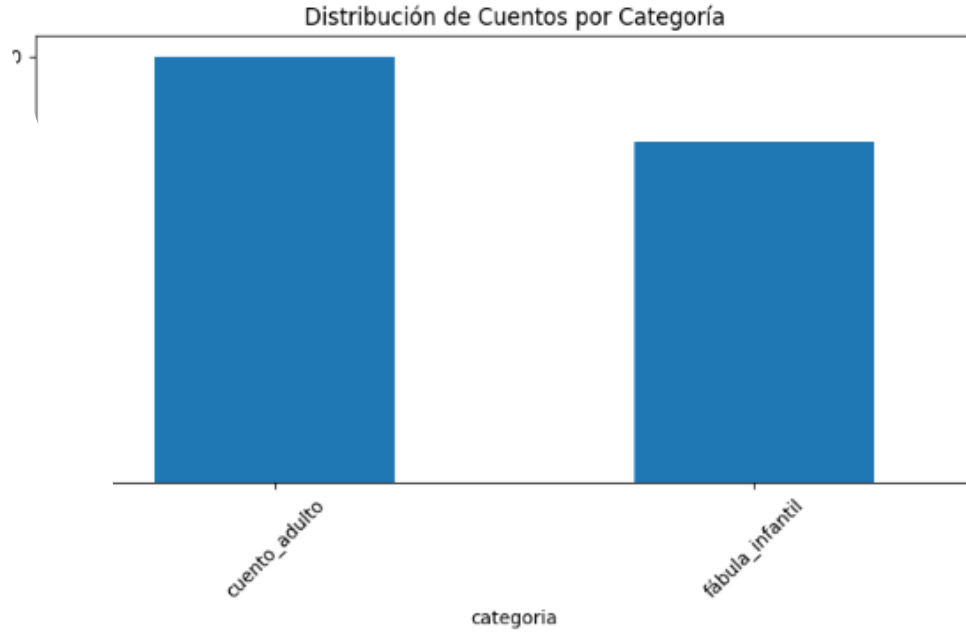
Datos y armado del corpus

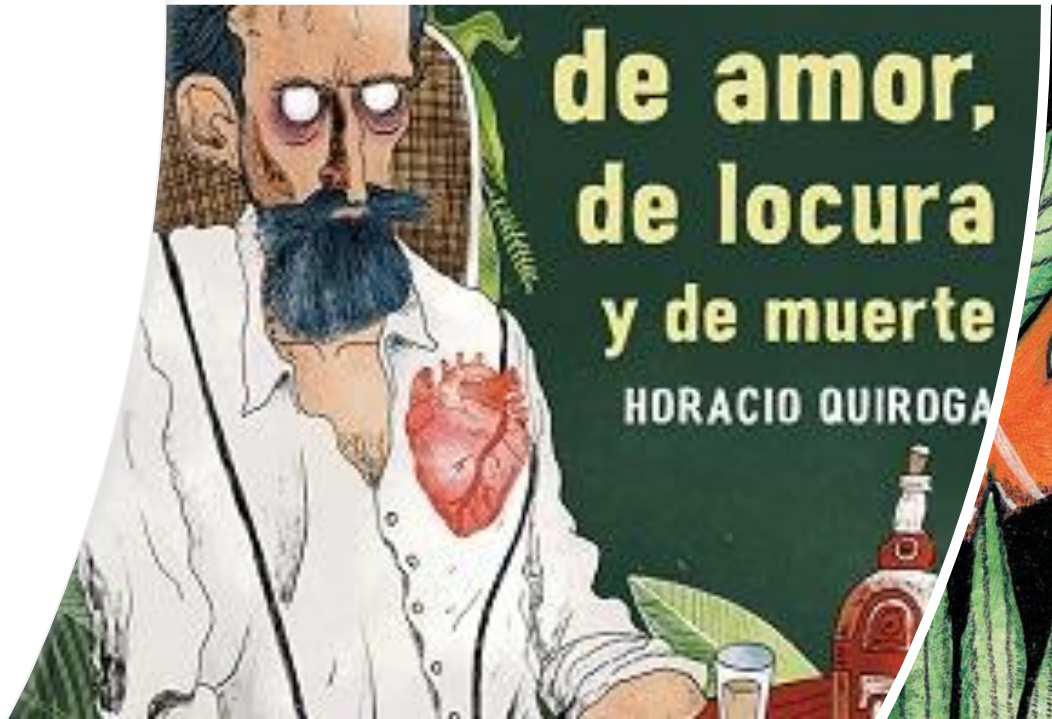
- Textos de Ciudad Seva + metadata propia. Leí archivos, unifiqué en un DataFrame con título, categoría y texto.
- 18 cuentos divididos por categorías, 10 adultos 8 infantiles

```
----------------------------------------------------
ESTADÍSTICAS POR CADA CUENTO INDIVIDUAL:
 'El almohadón de plumas'            : 7,202 caracteres, ~1,215 palabras
'La gallina degollada'               : 14,112 caracteres, ~2,389 palabras
 A la deriva'                        : 6,113 caracteres, ~1,057 palabras
 l hombre muerto'                    : 8,119 caracteres, ~1,440 palabras
  hijo'                              : 9,207 caracteres, ~1,703 palabras
 insolación'                         : 13,609 caracteres, ~2,348 palabras
  mensú'                             : 20,095 caracteres, ~3,391 palabras
 olitario'                           : 8,830 caracteres, ~1,504 palabras
 el silvestre'                       : 10,084 caracteres, ~1,716 palabras
 mpiro'                              : 5,149 caracteres, ~899 palabras
 tuga gigante'                       : 9,080 caracteres, ~1,703 palabras
 dias de los flamencos'              : 8,199 caracteres, ~1,383 palabras
  pelado'                            : 10,208 caracteres, ~1,834 palabras
 a de los yacarés'                   : 16,779 caracteres, ~2,941 palabras
 ciega'                              : 10,300 caracteres, ~1,856 palabras
 de dos cachorros de coatí           : 12,680 caracteres, ~2,270 palabras
 haragana'                           : 11,398 caracteres, ~1,984 palabras
 el yabebirí'                        : 18,166 caracteres, ~3,224 palabras
----------------------------------------

 DEL CORPUS:
l de cuentos: 18
l de caracteres: 199,330
l de palabras (aprox.): 34,857
```

```
- cuento_adulto: 10 documentos
- fábula_infantil: 8 documentos
```



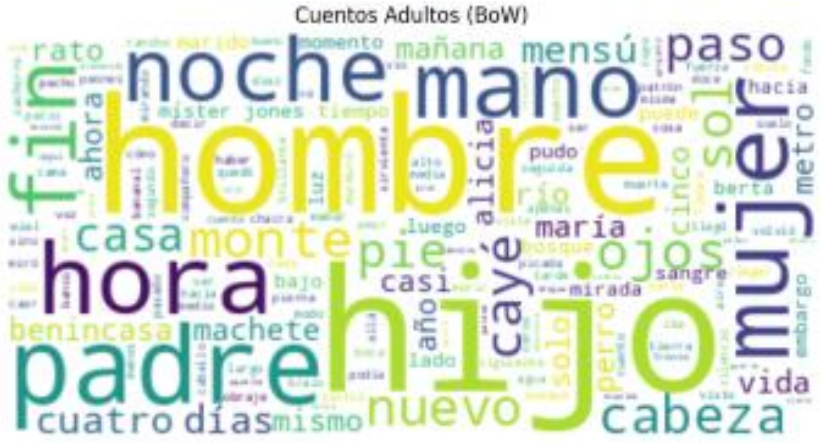


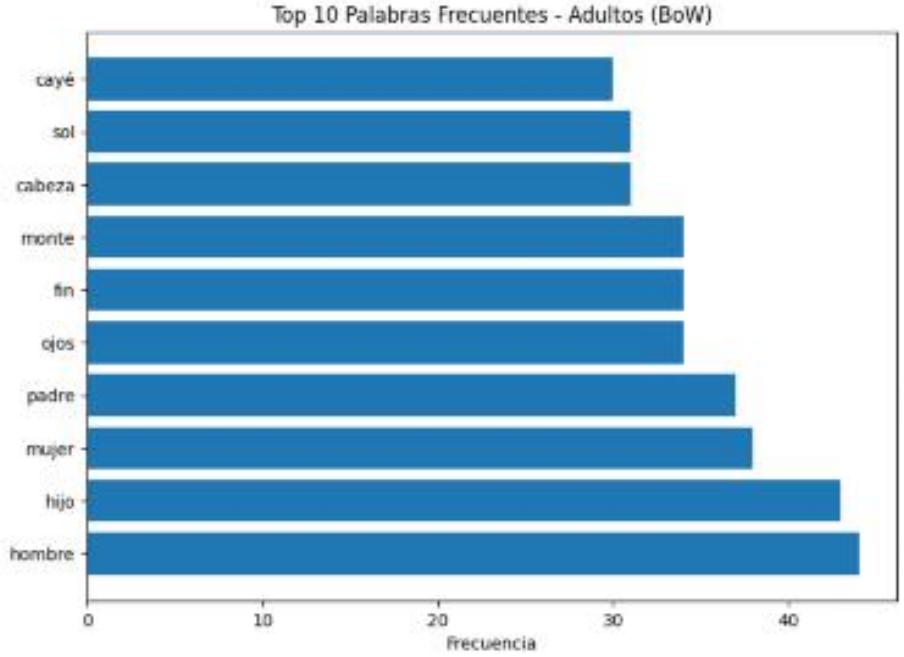
Pipeline de NLP y Preprocesamiento

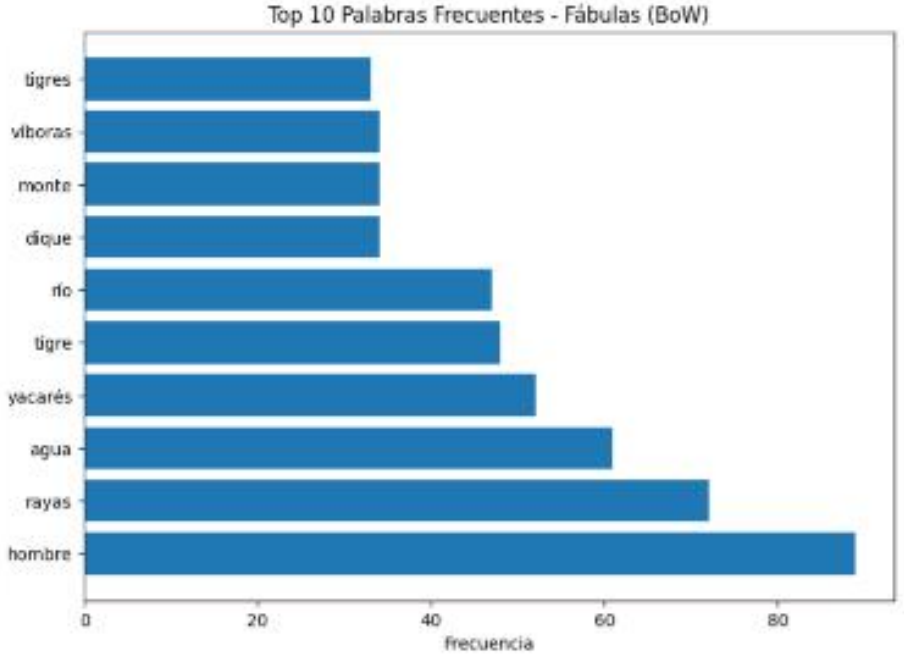
- **Diagrama de Flujo :**
Texto Original -> Limpieza (minúsculas, sin puntuación) -> Tokenización -> Filtrado (Stop Words) -> Texto Listo para Analizar

- **Ejemplo "Antes y Después“:**
  - **Antes:** "El hombre pisó algo blancuzco, y en seguida sintió la mordura en el pie."
  - **Después:** "hombre piso algo blancuzco seguida sintió mordura pie"

Análisis con BoW y TF-IDF







Nubes de Palabras con Stop Words Refinadas


```
# --- 2. Crear la lista de stopwords refinada ---
stopwords_espanol = stopwords.words('spanish')
stopwords_personalizadas = ['dos', 'vez', 'siguiente', 'dijo','así', 'si', 'tan',
                            'entonces', 'aún', 'lado', 'hacia', 'tres', 'pues', 'tal', 'día', 'allí', 'pronto', 'instante','gran', # cuantificadores
    "todas","todos","todo","toda",# funcionales
    "tras","mientras","después","siempre","bien",]
stopwords_contextuales = ['kassim', 'jordán', 'mazzini', 'quiroga', 'podeley', 'maría', 'mensú', 'alicia', 'míster', 'jones']
stopwords_completas = list(set(stopwords_espanol + stopwords_personalizadas + stopwords_contextuales))
print(f"Lista de stopwords completa tiene {len(stopwords_completas)} palabras.")
```

Comparación BoW vs TF-IDF

--- SIMILITUD ENTRE DOCUMENTOS (EMBEDDINGS) ---

Top 3 pares más similares:

- 'La gama ciega' & 'Historia de dos cachorros de coatí' (Similitud: 0.970)
- 'Los mensú' & 'La miel silvestre' (Similitud: 0.969)
- 'El hombre muerto' & 'La miel silvestre' (Similitud: 0.962)

Top 3 pares más diferentes:

- 'El hijo' & 'Las medias de los flamencos' (Similitud: 0.748)
- 'El solitario' & 'Las medias de los flamencos' (Similitud: 0.786)
- 'El hijo' & 'La guerra de los yacarés' (Similitud: 0.792)

--- SIMILITUD ENTRE DOCUMENTOS (TF-IDF) ---

Top 3 pares más similares:

1. 'El solitario' & 'El vampiro' (Similitud: 0.399)
2. 'La gallina degollada' & 'El hijo' (Similitud: 0.274)
3. 'A la deriva' & 'El hombre muerto' (Similitud: 0.197)

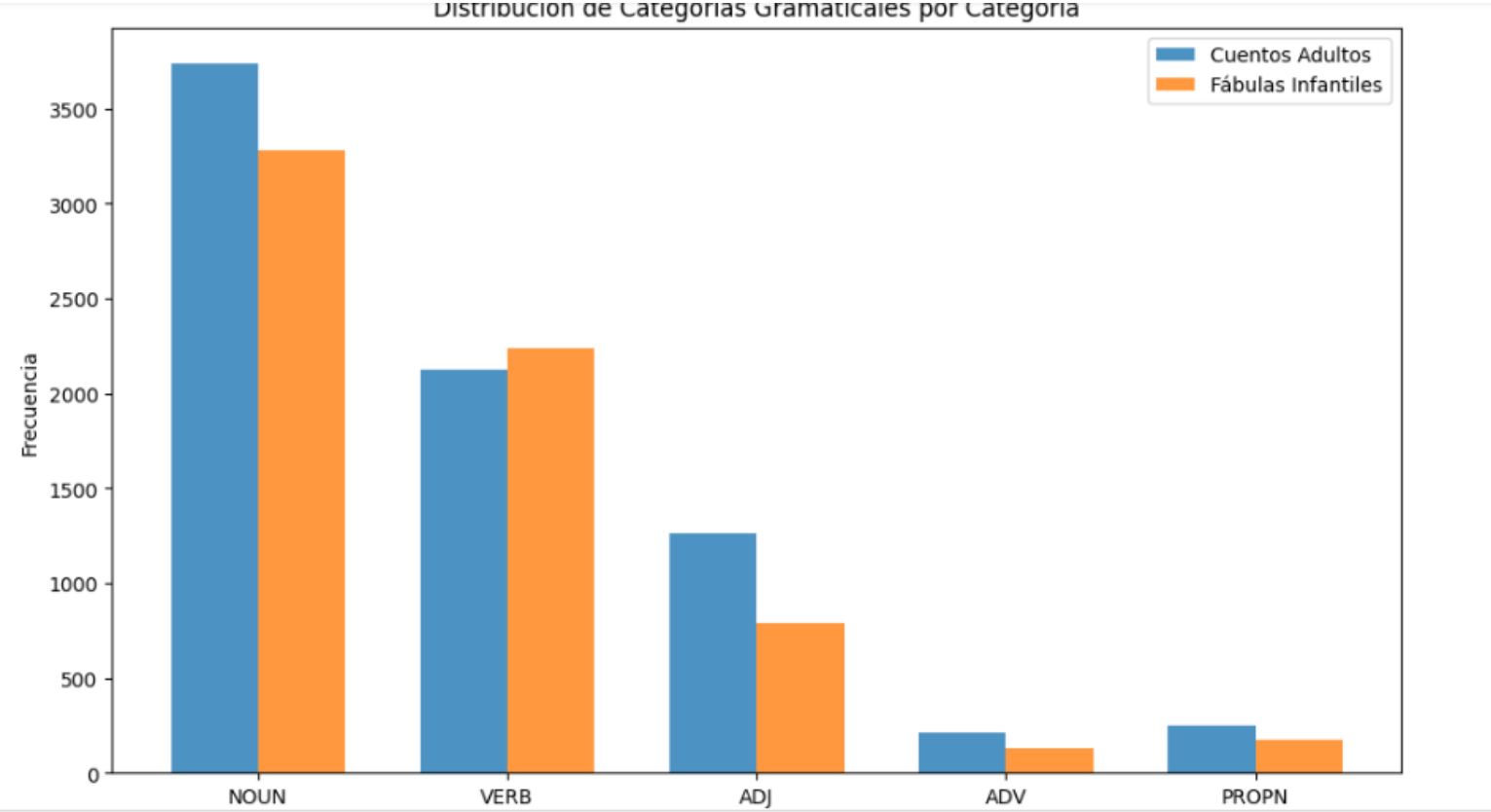
Top 3 pares más diferentes:

1. 'El solitario' & 'Las medias de los flamencos' (Similitud: 0.015)
2. 'El hijo' & 'Las medias de los flamencos' (Similitud: 0.016)
3. 'Las medias de los flamencos' & 'El loro pelado' (Similitud: 0.019)

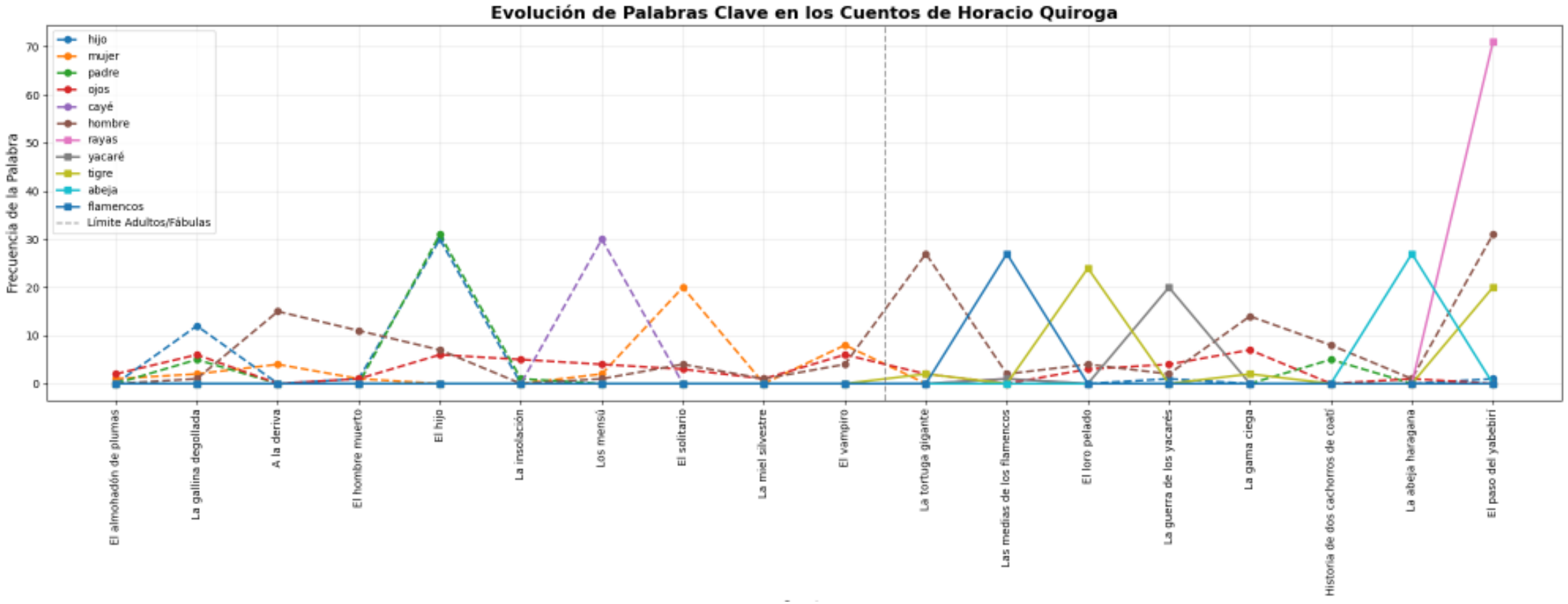
En este caso, BoW/TF-IDF parece más efectivo para distinguir entre las categorías temáticas elegidas.

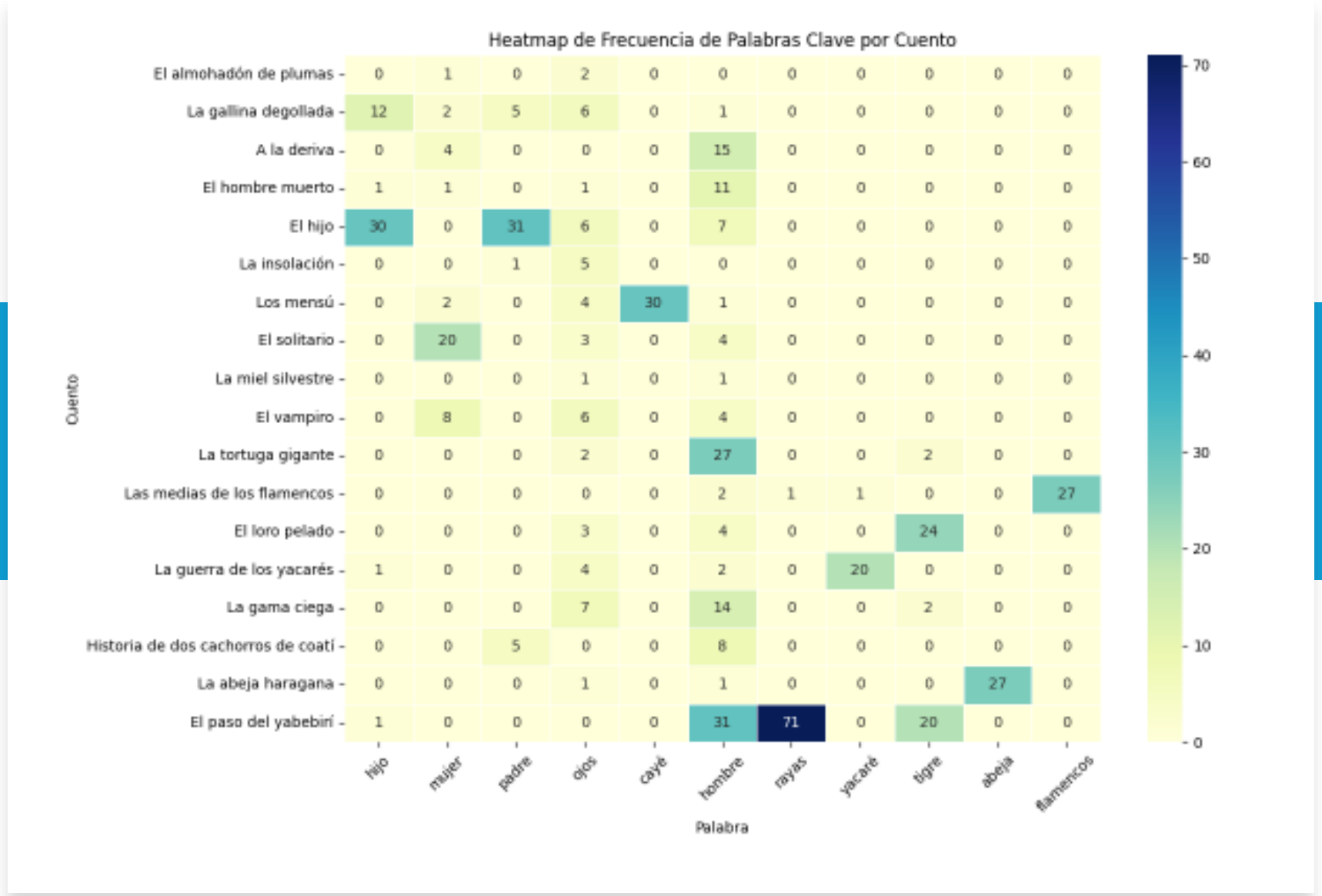
Análisis Complementario

Los cuentos para adultos podrían tener una mayor proporción de sustantivos y adjetivos relacionados con emociones y entornos oscuros. Las fábulas infantiles podrían tener una mayor proporción de verbos de acción y sustantivos relacionados con animales y objetos cotidianos.

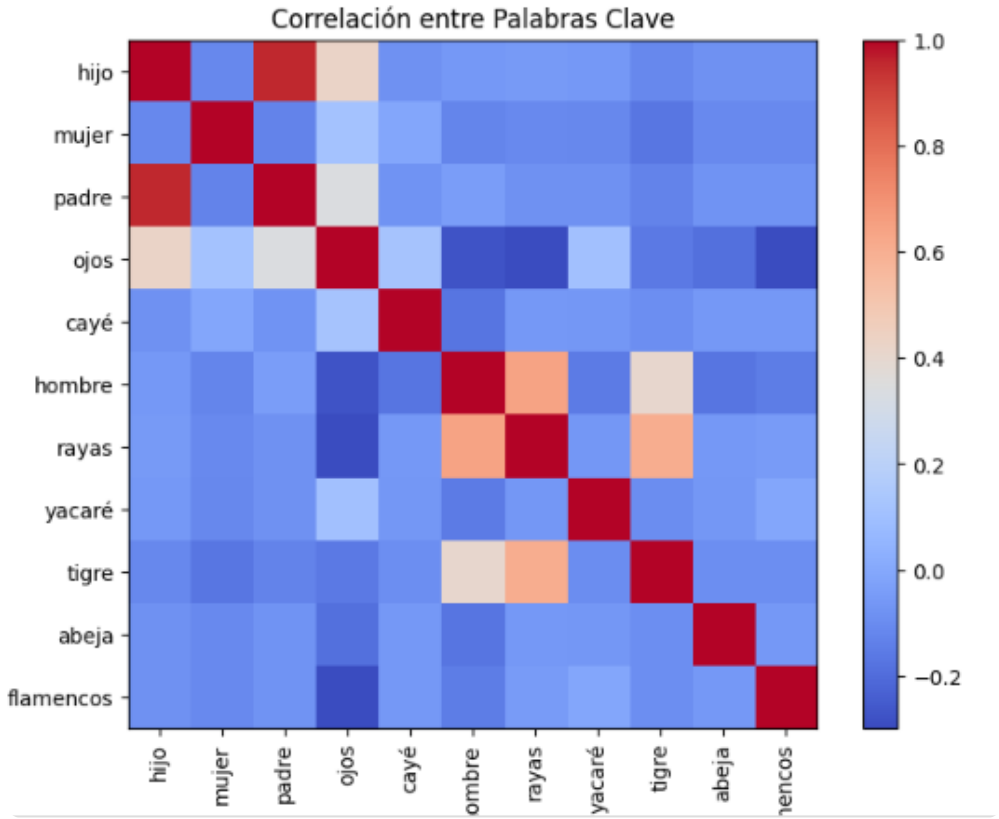


Se puede ver cómo ciertas palabras (como “hijo”, “mujer”, “ojos”) tienen picos en los primeros cuentos (adultos), mientras que otras (“rayas”, “yacaré”, “abeja”, “flamencos”) aparecen más en los últimos (fábulas).





Heatmap de frecuencias



✅ Las palabras típicas de cuentos para adultos están correlacionadas entre sí (tema familiar, emocional, dramático).

✅ Las palabras típicas de fábulas están correlacionadas entre sí (tema animal, naturalista).

✅ No hay correlación significativa entre ambas categorías, lo que valida la separación temática.

Hallazgos pricipales

- **Diferenciación Léxica**
- Cuentos para adultos: Se caracterizan por un vocabulario centrado en relaciones humanas (familiares, emocionales), psicología interna y entornos dramáticos. Palabras como “hijo”, “mujer”, “padre”, “ojos”, “cabeza”, son frecuentes.
- Fábulas infantiles: Se distinguen por un lenguaje más simple y centrado en animales, naturaleza y acciones concretas. Palabras como “rayas”, “yacaré”, “tigre”, “abeja”, “río”, “dique”, “flamencos”, “gamita” dominan este grupo.

Eficacia de las Técnicas de NLP

- BoW y TF-IDF: Fueron muy efectivos para identificar términos característicos y distinguir entre categorías temáticas. TF-IDF, en particular, resaltó palabras altamente representativas de cada tipo de cuento.

- Word Embeddings: Útiles para capturar similitud semántica entre documentos, aunque en este caso BoW/TF-IDF fue más claro para la diferenciación temática específica.

- Análisis POS: Reveló que los cuentos para adultos tienen más sustantivos y adjetivos, mientras que las fábulas tienen más verbos de acción y sustantivos concretos