Tercera parte de la entrega

2024-11-29

Objetivos de la Tercera Parte

En esta parte del proyecto, se nos pide implementar un programa llamado `train_analysis.R` que realice el análisis del corpus obtenido en la primera parte, almacenado en el archivo `spanish_train.qcorpus.rds`. El programa debe calcular, para cada documento, lo siguiente:

1. **Frecuencia de verbos del título**: Contar cuántos verbos, convertidos a su forma en infinitivo, presentes en el campo `title` del documento también aparecen en el texto completo correspondiente.
2. **Frecuencia de verbos del resumen**: Contar cuántos verbos, convertidos a su forma en infinitivo, presentes en el campo `summary` del documento también aparecen en el texto completo correspondiente.

Estos resultados se presentan de manera visual mediante dos histogramas: - Un histograma para los verbos del `title`. - Otro histograma para los verbos del `summary`.

En los histogramas obtenidos, se muestra la distribución de las frecuencias de coincidencia y el número de documentos que corresponden a cada frecuencia. Esto permite visualizar de forma clara cómo se relacionan los verbos presentes en los `title` y `summary` con el contenido completo de los textos.

Importación de Código para Tests

Para mantener el documento más organizado y legible, importamos el código de tests desde otro archivo. Esto nos permite ejecutar y mostrar los resultados de las pruebas específicas de las funciones utilizadas en el programa principal sin añadir bloques de código extensos.

La importación se realiza con la función `source()` de R, que permite cargar y ejecutar el contenido de un archivo externo. En este caso, importamos el archivo `test_train_analysis.R`, donde se encuentran definidas las pruebas.

Uso de Archivos RDS para Optimizar el Procesamiento

Dado que el parseo de textos con la función `spacy_parse()` puede ser un proceso lento y costoso, decidimos implementar una estructura `if-else` para gestionar el almacenamiento y la carga de los datos procesados. La idea principal es evitar realizar el parseo en cada ejecución, almacenando los resultados en archivos `.RDS`, que son compactos y rápidos de cargar.

Estructura General

La estructura que utilizamos funciona de la siguiente manera:

- **Si el archivo RDS ya existe**:
  - Cargamos directamente los datos procesados con la función `readRDS()`.
  - Esto reduce significativamente el tiempo de ejecución en futuras sesiones.
- **Si el archivo RDS no existe**:
  - Parseamos los datos correspondientes con `spacy_parse()` y los guardamos en un archivo `.RDS` utilizando `saveRDS()`.
  - Esto asegura que el procesamiento solo se realiza la primera vez.

Función `spacy_parse()`

1. **Tokenización**:
   - Divide el texto en **tokens** (en nuestro caso las palabras de los textos).
2. **Etiquetado gramatical (POS tagging)**:
   - Asocia cada token con su categoría gramatical (por ejemplo, `VERB`, `NOUN`, `ADJ`).
3. **Lematización**:
   - Convierte las palabras a su forma raíz (por ejemplo, "corriendo" → "correr").

El propósito principal de usar `spacy_parse()` en este proyecto es que nos permita trabajar específicamente con los **verbos** en infinitivo de los textos. Esto incluye:

- Identificar todos los verbos en un texto (`pos == "VERB"`).
- Obtener los verbos en su forma **lema** (`lemma`), para evitar inconsistencias debido a conjugaciones.

Resultado de `spacy_parse()`

Tras realizar el `spacy_parse()` sobre las palabras de los campos `title`, `summary` y los textos completos, obtenemos una tabla estructurada que contiene información clave sobre los tokens del texto. En nuestro caso, nos quedamos con las columnas `lemma` (forma en infinitivo de los verbos) y `doc_id` (identificador único de cada documento).

El uso del `doc_id` es bastante importante, ya que nos permite asociar cada verbo con su documento correspondiente.

Ejemplo de los Resultados

Verbos del campo `title`:

```
   doc_id     lemma
2   text1     decir
4   text1  peligrir
28  text4 reconocer
40  text5  inspirar
57  text6     parar
70  text7     tener
```

Verbos del campo `summary`:

```
   doc_id    lemma
13  text1 advertir
17  text1 invertir
26  text1    tener
28  text1  detener
48  text2   lanzar
52  text2 disparar
```

Verbos de los textos completos:

```
   doc_id    lemma
3   text1  recibir
15  text1 terminar
32  text1    decir
47  text1 recortar
50  text1   llegar
61  text1    hacer
```

Función `verbs_create_list`

La función `verbs_create_list` organiza los verbos extraídos del corpus en una estructura más manejable, agrupándolos por documento. Su objetivo principal es asociar cada verbo a su documento correspondiente, utilizando el identificador único de cada documento (`doc_id`).

La función toma los verbos extraídos de un campo del corpus, como `title` o `summary`, y los organiza en una lista, donde cada elemento corresponde a un documento específico. Si un documento no contiene verbos en un campo determinado, se incluye en la lista con un vector vacío, asegurando que todos los documentos estén representados de manera consistente.

El resultado de esta función es una lista con nombres que corresponden a los `doc_id` del corpus y cuyo contenido son los verbos en su forma lematizada. Esta estructura facilita el acceso y análisis de los datos, permitiéndonos comparar los verbos presentes en los títulos o resúmenes con los del texto completo de cada documento.

Resultados tras aplicar la función sobre los dataframes obtenidos anteriormente:

Aplicando para `verbs_titles`:

```
$text1
[1] "decir"    "peligrir"

$text2
character(0)

$text3
character(0)

$text4
[1] "reconocer"
```

```
$text5
[1] "inspirar"

$text6
[1] "parar"
```

Para `verbs_summary`:

```
$text1
[1] "advertir" "invertir" "tener"    "detener"

$text2
[1] "lanzar"    "disparar"  "dispersar"

$text3
character(0)
```

Lo mismo con `verbs_text`:

```
$text1
[1] "recibir"  "terminar" "decir"    "recortar" "llegar"   "hacer"    "juntar"
[8] "proteger"

$text2
 [1] "hindúes"   "gritar"    "tener"     "decir"     "producir"  "volver"
 [7] "empezar"   "lanzar él" "resultar"  "arrestar"  "evitar"    "liderarar"

$text3
[1] "ver"     "mostrar" "buscar"  "perder"  "ganar"   "crear"
```

Eliminación de Verbos Duplicados

Como indica el enunciado, necesitamos trabajar con los **verbos únicos** de los campos `title` y `summary`, ya que no queremos contar verbos repetidos dentro de un mismo documento. Para lograr esto, definimos la función `make_unique()`, que elimina duplicados en cada elemento de una lista. Su objetivo es garantizar que cada sublista contenga solo valores **únicos**.

La función recorre cada elemento de la lista (correspondiente a un documento) y, si este no está vacío, utiliza la función `unique()` para filtrar los valores duplicados. De esta forma, el resultado es una lista en la que cada sublista contiene únicamente los verbos únicos asociados a su documento.

Ejemplo de Uso

Si observamos el contenido original de `list_verbs_summary`, en el documento **text__5** aparece el verbo `tener` dos veces.

```
$text5
[1] "infectar" "tener"    "tener"    "ver"
```

Después de aplicar `make_unique()`, el resultado muestra este verbo solo una vez.

```
$text5
[1] "infectar" "tener"    "ver"
```

Cálculo de Frecuencias de Verbos

Llegamos al núcleo del proyecto: **contar cuántas veces los verbos presentes en los campos `title` y `summary` aparecen en el texto completo de cada documento**. Para ello, definimos la función `freq_verbs`, que calcula la frecuencia total de los verbos en el texto correspondiente.

Explicación de la Función `freq_verbs`

La función `freq_verbs` toma tres argumentos:

1. **`text_list`**: Lista de verbos en los textos completos.
2. **`list_comp`**: Lista de verbos a comparar (pueden ser de `title` o `summary`).
3. **`corpus_ids`**: Identificadores únicos de los documentos.

El objetivo de esta función es recorrer los textos y contar cuántas veces los verbos presentes en `list_comp` aparecen en el texto completo correspondiente. Para ello:

- Se inicializa un vector vacío llamado `freq`, cuya longitud corresponde al número de documentos, y se asignan los `corpus_ids` como nombres.
- Para cada documento, se recorren los verbos de `list_comp` y se cuenta cuántas veces aparecen en el texto completo (`text_list`) utilizando `sum()`.
- Finalmente, el vector `freq` almacena la frecuencia total de los verbos por documento y se devuelve como resultado.

Aplicamos esta función y obtenemos las frecuencias de los verbos de cada `title` en su texto y de cada `summary` en su texto.

Veamos las frecuencias de los `titles` 50 - 70:

```
freq_verbs_titles[50:70]
```

```
text50 text51 text52 text53 text54 text55 text56 text57 text58 text59 text60
    14      0      0      0      4      0      1      6      2     13      4
text61 text62 text63 text64 text65 text66 text67 text68 text69 text70
     0      0      1      0      0      0      0      0      0      0
```

Veamos las frecuencias de los `summary` 50 - 70:

```
freq_verbs_summary[50:70]
```

```
text50 text51 text52 text53 text54 text55 text56 text57 text58 text59 text60
     0      0      0      0      8      0      0      5      2     19     11
text61 text62 text63 text64 text65 text66 text67 text68 text69 text70
     9      6      6      1     10      2      4      0     12      0
```

Testeos de la Función `freq_verbs`

Para confirmar que los resultados de la función `freq_verbs` son correctos, definimos una nueva función llamada `test_freq_verbs`. Esta función evalúa un subconjunto de documentos y verifica:

- La frecuencia de verbos encontrados en los títulos (`title`) y resúmenes (`summary`) comparados con el texto completo.
- Los verbos coincidentes entre los diferentes campos y las frecuencias correspondientes.

El objetivo de esta prueba es asegurar que los verbos se están contando correctamente y que los resultados coinciden con las expectativas.

Hacemos un test de los textos 56 - 58:

```
test_freq_verbs(list_verbs_text, list_verbs_title_unique, list_verbs_summary_unique, corpus_ids)
```

—— Testing freq_verbs Function ——

Document ID: text56
Text Verbs: saber desafeír responder animar realizar seguir

Title Verbs: saber
Matched Title Verbs: saber
Matched Count (Titles): 1

Matched Verbs:

Matched Count (Summaries): 0

Document ID: text57
Text Verbs: desconocer dar recetar mandar recibir enfermar deber hacer esparcir aumentar temer interesar admitir fracasar detectar confirmar necesitar requerir decir pedir hacer solicitar monitorear detectar apoyar desarrollar designar dirigir combatir anunciar adelantar intentar calmar estabilizar desconocer intentar lidiar creer estar decir causar pasar saber comentar reportar considerar creer presentar monitorear llevar dar contar dar reflejar señalar desconocer mostrar indicar distribuir saber agregar coincidir tener detectar permitir aislar demorar considerar recordar comenzar decidir fabricarir examinar comenzar enviar encontrar funcionar tener cambiar él demorar comenzar hacer hacer señalar colocar ayudar prevenir enviar realizar contar realizar limitar contar hacer ultramar contar ofrecer enviar confirmar confirmar tener enviar confirmar esperar recordar obtener realizar limitar acceder creer comiencir extender él conducir opinar comentar terminar tener existir cumplir someter detectar indicar hacer especificar adoptar abordar anunciar contagiar realizar él corregir acudir hacer él prescribir mostrar presentar encontrar aceptar tener añadir detectar practicar contagiar presentar contagiar tener crear hacer hacer él hacer creer creer estar él afirmar convertir él contener tener manifestar transmitir hacer existir servir detener considerar desconocer someter detectar hacer dejar publicar rondar transcurrir reportar él publicar comenzar hacer ascender indicar tomar tener reportar realizar realizar analizar reportar evaluar realizar intentar contactar conocer realizar dejar ofrecer obtener detectar separar requeír tener llevar presentar requeír acudar tener desembolsar cobrar hacer desestimir ir encontrar definir cubrir presentar tener indicar indicar posponer dejar comprar responder preocupar llevar él vivir entrar limitar utilizar creer tomar contagiar acceder autoaislar él evitar opinar registrar requerir hacer proteger trabajar elegir hacer él autoaislar él indicar coincidir sugerir alentar quedar él trabajar tomar presentar tener tener hacer tener tener escuchar autoaislar él dejar trabajar tener hacer él quedarir mantener él comentar continuar trabajar experimentar agravar contagiar manejar ocurririr opinar pedir tomar garantizar indicar anunciar garantizar quedar él tomar tener proporcionar dirigir afirmar decir requerir contar hacer hacer pasar durar dar señalar visitar recibir descargar activa él perdertir

Title Verbs: fracasar combatir dejar
Matched Title Verbs: fracasar, combatir, dejar
Matched Count (Titles): 6

Matched Verbs: contagiar
Matched Count (Summaries): 5

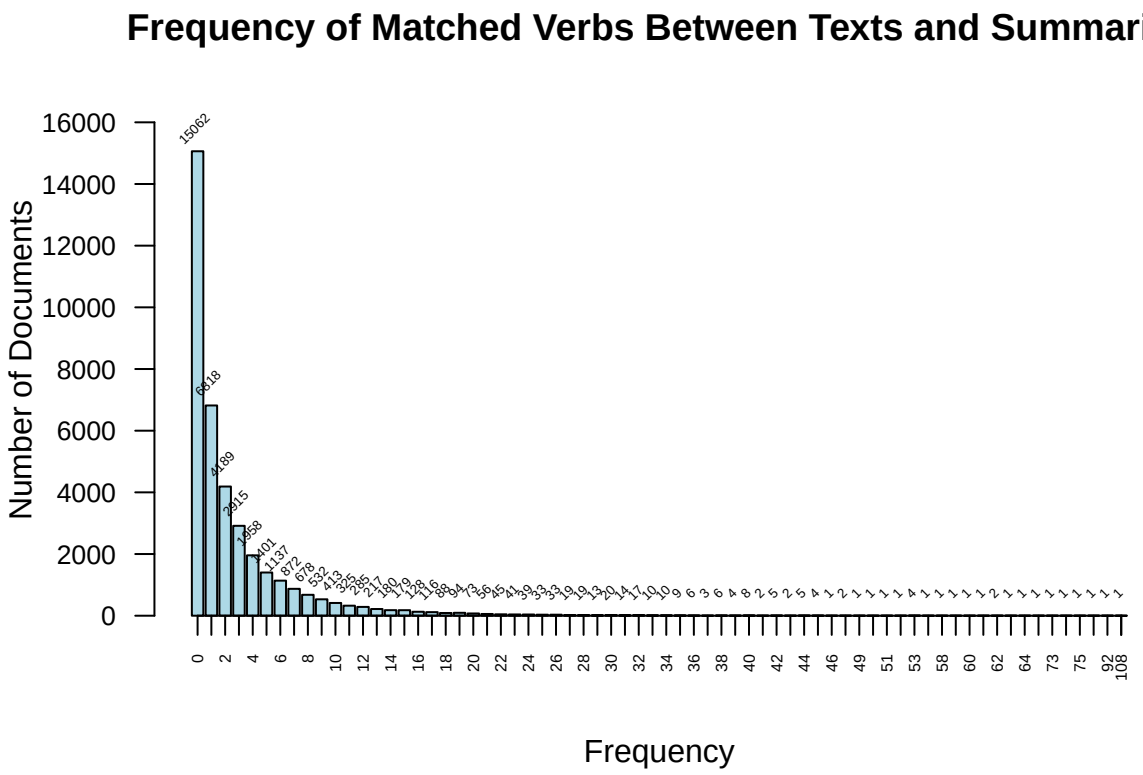
Document ID: text58
Text Verbs: estar participar participar esperar producir estar alcanzar él decir decidir viajar pasar tratar resolver interesar viajar asistir

Title Verbs: viajar
Matched Title Verbs: viajar
Matched Count (Titles): 2

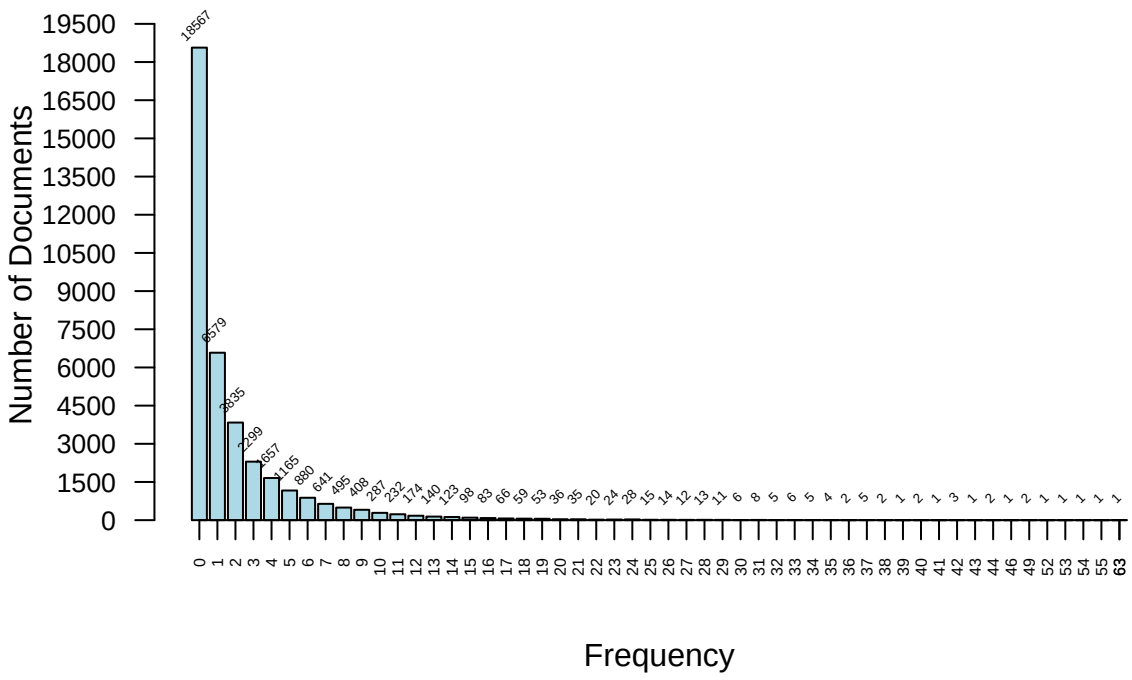
Matched Verbs: participar
Matched Count (Summaries): 2
—

Visualización de Frecuencias de Verbos

Para analizar la cantidad de documentos que tienen cierta frecuencia de coincidencia entre los verbos de los resúmenes (`summary`) y los textos completos, generamos un **histograma**. Este histograma nos permite observar cómo están distribuidas las frecuencias de coincidencia entre los documentos.



Histograma para ver las frecuencias de los verbos del campo **títulos**:



Como se puede observar hay gran cantidad de documentos que tienen frecuencia 0. Al principio esto nos pareció muy raro, pero tras realizar numerosas pruebas y verificar nuestros cálculos, hemos llegado a la conclusión de que estos son correctos. Nuestra hipótesis es que el problema radica en las limitaciones del modelo `es_core_news_sm` utilizado por `spacy_parse()`. Este modelo, aunque eficiente y rápido, no identifica correctamente ciertos verbos. Por ejemplo, no reconoce el verbo "ser", lo cual es especialmente significativo dada su alta frecuencia en el idioma español. Además, tampoco maneja bien los verbos pronominales, lo que afecta aún más la precisión de los resultados.

Creemos que si hubieramos usando un modelo de lenguaje español más grande que `es_core_news_sm` los resultados habrían sido más representativos y precisos.

Testeo para Frecuencias Específicas

Una vez calculadas las frecuencias de coincidencias de verbos entre los resúmenes (`summary`) y los textos completos, podemos realizar testeos para analizar documentos que tienen una frecuencia específica. Esto nos permite observar en detalle los verbos coincidentes en esos casos particulares y asegurarnos de que el cálculo se realiza correctamente.

Función `test_specific_frequency_summary`

La función `test_specific_frequency_summary` busca documentos que tienen una frecuencia específica de coincidencias en los resúmenes. Para cada documento encontrado: 1. Muestra los verbos del texto completo y del resumen. 2. Indica los verbos coincidentes y cuántas veces se encontraron en el texto.

Si no se encuentran documentos con esa frecuencia, la función informa que no hay coincidencias.

Hagamos un test, por ejemplo, con frecuencia **46**, solo debe aparecer 1 documento con la suma de frecuencias de cada verbo encontrado equivalente a 46.

=== Testing for Frequency: 46 ===

Document ID: text2042
Text Verbs: hablar, pedir, decir, tener, llevar, sumir él, contar, ocurrir, ayudar, reconciliar él, entender, pedir, ver, visitar, modificar, preservar, creer, aprovecher, hacer, estresar, aprobar, pedir, quedar yo, tener, vivir, compartir, hablar, llegar, ir, llevar, intimidar, querer, admitir, hablar, otorgar, quedar, compartir, saber, compartir, evaluar yo, hacer, preguntar, creer, pensar, proponer, saber, reaccionar, pensar, querer, comenzar, invitar, tomar, ir, empezar, mirar, sentar, entender, pasar, mirar, mirar, mirar, decir, analizar, saber, acercar, decir, hablar, decir, hablemos, comenzar, oscurecer, decir, ir, hablar, saber, llegar, sentar, sacar, decir, tomar, decir, tener, tomar, pasar, tomar, ocurrir, encontrar, empezar, hablar, saber, entrar, querer, empezar, preguntar, querer, querer, decir, estudiar, hacer, entrar, hacer, pensar, pensar, pensar, hablé, pensar, pensar, pensar, hacer, prir, creer, tener, decir, pensar, estar, explorar, creer, querer, decir, atraer, pasar, decir, pasar, -es, locar, estar, -hablamos, sentir, tener, atraer, parecer, ir, llamar, hablar, hacer, -si, estar, hacer, masturbo, -cuar, masturbar, acariciar, hacer, ver, esperar, recordar, decir, ir, levantar, coger, tener, creer, pensar, decir, rechazar, hablar, sacar, preguntar, hacer él, tener, conocí, dar, hacer, decir, -¿perir, hacer, acostándome, masturbo, tener, masturbo, -¿y, consentir, acariciar él, consentir, hacer, consentir, pasar, hacer, preguntar, creer, pensar, proponer, saber, reaccionar, pensar, querer, pensar, pasar, decir, entender, ver, querer, saber, saber, permitir, tener, agradar él, llegar, oscurecer, llevábar, hablar, llevábamos, tomar, encargar, seguir, preguntar, tocar, tocar, recordar, poner, decir, hacer, incomodar, -¿y, hacer, subir, saber, sentir, invadir, costar, abrazar, tocar, explicar, ocurrir, seguir, seguir, seguir, sentir, -nar, -sí, recordar, decir, ir, levantar, coger, recordar, decir, jurar, llegabas, estar, sentir, sentar, temblar, seguir, consentir, recordar, venir, saber, besar, besar, besar, recordar, desvestir, recordar, recordar, estar, querer, escoger, hacer él, decir, conocer, hablar, decidir, hacer, saber, hacer, querer, hacer, hacer, saber, hacer, liderar, acostar, recordar, pensar, contar él, recordar, tratar, jugar, hacer, saber, hacer, saber, pasar, perder, hacer, creer, aburrir, tener, decir, ir, masturbar, saber, esperar, hacer, saber, ir, masturbar, vestí, creer, visitar, ir, salir, caer, salir, ir, querer, encontrar, llegar, hacer, echar, sentir, creer, comenzar, comenzar, doler, doler, doler, querer, estar, decir, decir, abusar, querer, curar, querer, escoger, hacer él, decir, conocer, hablar, decidir, hacer, querer, hacer él, atraer, tener, saber, volver, venir, volver, hacer, bañé, sentir, creer, llorer, acostar, dormir, escuchar, asustar, recordar, hacer, bañé, ocurrir, perder, querer, hablar él, hablar él, hablar, rezar, meditar, recordar, pasar, ver, querer, ver él, querer, estar, sentir, mirar, acercar, hablar yo, llorar, llorar, llorar, llamar, decir él, decir, decir, decir, abusar, querer, curar, tener, decir, querer, confesar, sentir, confesarme, decir, cuento, hablar, pasar, confeser, comulgar, comulgar, tener, fui, confesar, confesarme, recir, decir, robar, parecer, pensar, hacer, seguir, llorar, llorar, volví, llamar, llamar, preguntar yo, estar, esconder, contestar, ver, poner, contestar, querer, hablar él, trasladar, hablar, decir, buscar, decir, pasar, romper, jodir, recordar, recibir él, volviéndomir, empezar, hablar, sacar, sentir, pasar, verter, excitar, mirar, sentar, seguir, tener, empezar, mostrar yo, creer, aceptar, entrar, haber él, decir, decir, pasar, sentir, parar él, cuestionar, decir él, hacer, saber, tener, saber, terminir, salir, volver, confesar, decir, hacer, hacer, decir, venir, confesarme, acostar, decir, usar, tener, decir él, quedar, volver, abrir, abrir, vivir, seguir, tener, querer, abusar, conocer, compartir, compartir, empezar, salir, sentíar, pasar, veíar, decir, creer, llorar, llorar, llorar, llamar, decir él, dejar, hablar, generar, pasar, empezar, pasar, jesuitar, querer, empezar, hablar, dar, pasar, dar, hacer, contar, decir, sentir, pedir, hacer, decir, valar, hacer, estar, tener, hacer, mostrar, obrar, creer, creer, aceptar, entrar, haber él, decir, decir, abusar, aprovechar, significar, hacer, pintar, hacer, tener, tener, dirigir, sucedierar, tocar, poner, tomar, cumplir, utilizar, lograr, querer, importar, quiera, creer, configurar, pasar, tener, querer, tener, sentir, bañé, pasar, querer, acostar yo, sospechar, pasar, pasar, pintar, decir, crear, hacer, importar, tener, sentir yo, juzgar, parecer, preguntar, tener, tener, vivir, destacar, ayudar, salir, vivir, tener, repetir, pasar, abrir, compartí, saber, ir, encontrar, juzgar, encontrar, aceptar, odiar, hacer, ver, aproximar, vivir, reconocer, gustar, casar yo, hacer, servir, creer, creer, tener, decidir, tener, terminar, decir, querer, hacer, decidir, querer, saber, saber, querer, exigir, tener, querer, tener, célibir, durar, masturbar yo, creer, decidir, querer, vivir, impedir, querer, tener, decidir, vivir, reconocer, jesuitar, salir, decir, salir, acostar, incumplir, decir, salgar, vivir, tomar, decir, tuvistar, gustar, tuvierar, tomes, importar, pasar, descartar, ver, tener, hablar, decir, tener, saber, manejar, tener, juzgar, aceptar, vivir, jesuitar, hablar, masturbar, querar, pasar,

viví, podiar, masturbar, volví, comenzar, explotar, ver, decir, escribir, oír, ver, estar, pasar, contar, hacer, ver, hacer, sentir, invitar, hacer, tomar, contar, ver, tener, juzgar él, terminar, juzgar, creer, pedir él, abuso, permitir, suponer, querer, tener, decidir, vivir, creer, profesar, abusar, ver, cambiar, hacer, jesuitar, educar, creer, educar, gustar, recibir, sentir, juzgar, acepté, evitar, llegar, tener, entender, explicar, decir, estar, quedar, tirar, llevar, creer, intentar, hacer, saber, tirar, creer, creer, querer, cambiar, creer, llegar, hacer, ir, creer, hacer, querer, dedicar, descubrir, querer, dedicarme, creer, ir, dar, volver, intentar, pasar, pensar, llegar, pensar, estar, tener, tener, tener, hacer, dejar, construir, creer, conocer, entrar, jesuitas, descubrir, creer, interesar

Summary Verbs: hablar, pasar, buscar, construir
Matched Verbs (Summaries) with Counts:
hablar (19), pasar (25), buscar (1), construir (1)
Matched Count (Summaries): 4

Podemos hacer lo mismo con los **títulos**, por ejemplo, con los que tengan frecuencia **63**. Según el histograma, debe aparecer 1 documento con frecuencia 63.

=== Testing for Frequency: 63 ===

Document ID: text8911
Text Verbs: hacer, almorzar, disfrutar, tener, fletar, existir, contar, tener, conceder, representar, conocer, vivir, representar, tener, influir, añadir, interesar, recibir, utilizar, acompañar, tener, poseer, explicar, servir, llevar él, tener, pertenecer, pedir, tener, necesitar él, decir, contar, subir él, pagar, equivaler, ganar, tener, solicitar, trabajar, tener, dar, comprar, acompañar, costar, tener, recibir, establecer, tomar, constar, llegar, ver él, convertir él, deber, volver él, tener, aumentar él, determinar, soler, nombrar, dar yo, afirmar, explicar, reúnir, significar, subir, matizar, evaluar, aumentar, hacer, incluir, reunamos, evaluar, subir él, darer, contar, impugnar, tener, someter él, tener, saber, quedar, llamar, pedir, quejar él, asegurar, decidir, tener, tener, poseer, tener, vivir, tener, tener, constar, tener, administrar, tener, parecer, recordar, tener, llenar, tener, hacer, colocar, decir, acompañar, tener, hacer él, contar, planchar, cocinar, limpian, hacer, pagar, tener, vivir, pernoctar, pagar, pasar, tener, pagar, decidir, vivir, tener, hacer él, pagar, vivir, asegurar, dejar, tener, tener, pagar, explicar, compartir, tener, hacer, tener, vivir, vivir, alquilar, recibir, reembolsar, considerar, vivir, pedir, tener, pagar, explicar, hacer, compartir, tener, pagar, asegurar, existir, dormir, lavar, tener, llenar, llegar, tener, reservar, tener, tener, suscribir él, llevar yo, leer, devolver él, decir, leer, ofrecer, leer él, añadir, tener, secretacer, pagar, pagar, comer, llevar, recoger, contar, traer, calentar, lavar, tener, tener, contratar, explicar, recibir, contratar, formar, atender, compartir, preparar, encarguir, proporcionar, trabajar, soler, ocupar, preparar, organizar, decir, tener, compartir, atribuir él, opinar, dar, tener, necesitar, significar, aumentar, necesitar, ayudar, ampliar, proporcionar, explicar, recibir, informar, contar, estar, utilizar, trabajar, hacer, decir, controlar, ostentar, recibir, elegir, incluir, gastar, presentar, someter, recibir, depender, recibir, deducir, almorzar, abonar, explicar, tratar, compartir, tener, tener, pagar, añadir, soler, volar, pagar, tener, recibir, cumplir, ofrecer, llamar, decir, proporcionar, tener, garantizar, trabajar, tener, recibir, demostrar, buscar, ganar él, entender, tener, tener, demostrar, buscar, acabar, explicar, regular, pasar, ejercer, cancelar, considerar, ejercer, tener, recibir, trabajar, tener, tener, tener, secretacer, -trabajar, ganar, participar, realizar, opinar, recibir, descargar, activa él, perdertir

Title Verbs: tener, pagar
Matched Verbs (Titles) with Counts:
tener (51), pagar (12)
Matched Count (Titles): 2

Retos y Soluciones

Problema con el Cálculo de Frecuencias

Uno de los desafíos más importantes que enfrentamos fue calcular correctamente las frecuencias de coincidencia de verbos entre los textos y los campos title y summary. Inicialmente, los histogramas mostraban frecuencias máximas de solo 10, lo cual resultaba poco realista dado que muchos textos contienen cientos de

verbos. Además, las frecuencias estaban incorrectamente distribuidas; por ejemplo, el histograma indicaba que había 4 documentos con una frecuencia de coincidencia de 8, pero al verificar manualmente, no había ninguno.

Identificación del Problema Después de analizar el código y realizar múltiples pruebas, descubrimos que el error se debía a que solo estábamos considerando los verbos únicos en lugar de sumar todas las veces que estos aparecían en los textos completos. Esto generaba resultados inconsistentes y poco representativos.

Solución Implementada La solución fue realizar una pequeña modificación en el cálculo de las frecuencias. En lugar de simplemente contar si un verbo estaba presente, sumamos todas las apariciones de cada verbo en el texto correspondiente. Esta corrección, aunque sencilla, tuvo un impacto significativo, permitiéndonos obtener histogramas precisos que reflejan correctamente la distribución de frecuencias.

Validación con Tests Para asegurarnos de que los cálculos eran correctos, utilizamos los tests desarrollados específicamente para esta tarea. Estos tests nos permitieron verificar las coincidencias verbo por verbo, documentando los resultados y confirmando que las frecuencias calculadas eran exactas. Gracias a estas pruebas, podemos afirmar con confianza que los resultados ahora son consistentes y precisos.