Recuperación de texto de audios en la web

Code

Curso Recuperación y análisis de texto con R

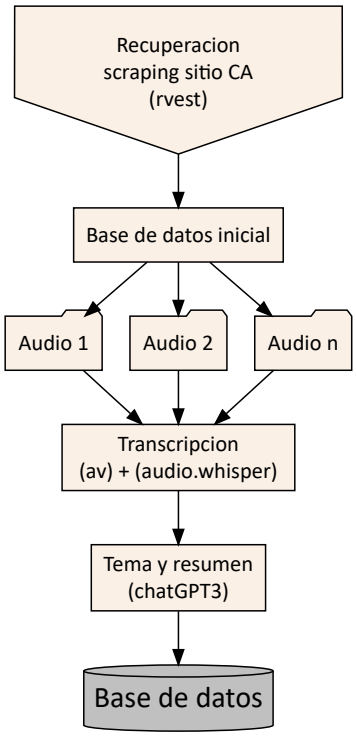
Elina Gómez | Gustavo Méndez

Junio 2023

1. Introducción

En este breve documento quiero compartir algunas potencialidades de R para asistirnos en la investigación en ciencias sociales. Específicamente me concentro en cómo podemos recuperar audios de la web, transcribirlos a texto, identificar el tema principal, obtener un resumen del contenido, y que esa información quede ordenada en una base de datos. En el ejemplo, utilizo un ejercicio con el discurso público del Senador Guido Manini Ríos en los últimos meses. Las principales librerías utilizadas son `rvest` para scrapeo; `audio.whisper` y `av` para transcripción de audio, `rgpt3` para conectar con la API de chatGPT3 y `tidyverse` para manipulación.

Figura 1. Diagrama de contenido



2. Recuperación *(rvest)*

Code

Como el ejercicio supone la recuperación de audios, opté por descargar las audiciones que el partido Cabildo Abierto tiene en Radio Oriental (770AM) y Radio Universal (990AM) y que están subidas a la página web oficial del partido acá (https://cabildoabierto.uy/category/audiciones/). Una alternativa era descargar manualmente todos los audios de Manini Ríos en determinado rango temporal. Sin embargo, eso es bastante largo y tedioso. Además, la idea es que R oficie como asistente de investigación y me ahorre trabajo. Por eso elegí realizar un scrapeo web con `rvest` . Una cosa importante antes del scrapeo es chequear que la página permita realizar ese tipo de operaciones, es decir, que no las prohíba. Esto puede hacerse con el paquete `robotstxt` .

Habitualmente hay dos audiciones por semana y en cada página se incluyen 10 resultados, es decir, un mes y una semana. Cada uno de esos resultados corresponde a una entrada de la página web que refiere a una audición, donde se coloca el reproductor de audio con posibilidad de descarga. Defino scarpear 4 páginas (unos 5 meses, aunque no es exacto porque en verano hubo una interrupción en la frecuencia). Utilizando el paquete `rvest` , hago un scrapeo por capas. Primero recupero las 40 url correspondientes a cada entrada en la cual se encuentran las notas con los audios. Luego scrapeo cada una de esas 40 urls para descargar fecha, título, subtítulo (copete de la nota) y url del audio. En el copete de la nota siempre se pone el nombre del o la columnista, ya que diferentes líderes de CA participan de la audición. A partir de eso filtro y me quedo únicamente con las audiciones de Manini Ríos.

Como resultado de este paso construí un dataframe con 19 observaciones con la fecha, título, subtítulo (copete) y url de los audios del líder de Cabildo Abierto. Finalmente, descargué los audios con la función `downloadfiles()` iterando sobre la variable del dataframe donde están las urls de los audios.

Tabla 1. Base de datos inicial

fecha	title	subtitle	url_audios
2022-10-11	“Los prestadores de salud tienen que estar obligados a tener cuidados paliativos para evitar el sufrimiento de los pacientes”	En su audición de este martes por AM 770 Radio Oriental, el Senador Guido Manini Ríos se refirió a la votación del proyecto de ley de Eutanasia, el jueves pasado en la Cámara de Diputados, y dijo:	https://www.ivoox.com/_md_93791161_wp_1.mp3 (https://www.ivoox.com/_md_93791161_wp_1.mp3)
2022-10-18	“Como ciudadanos tenemos derecho a criticar a personas que integran la justicia y actúan siguiendo condicionamientos ideológicos”	En su audición de este martes por AM 770 Radio Oriental, el Senador Guido Manini Ríos comenzó refiriéndose al llamado a Comisión General del Ministro del Interior, y dijo:	https://www.ivoox.com/_md_94232440_wp_1.mp3 (https://www.ivoox.com/_md_94232440_wp_1.mp3)
2022-10-25	“Insistiremos con los puntos que presentamos y no fueron contemplados para la reforma de la Seguridad Social”	En su audición de este martes por AM 770 Radio Oriental, el Senador Guido Manini Ríos se refirió al proyecto de ley de Reforma de la Seguridad Social, firmado la semana pasada por el Consejo de Ministros. Y dijo.	https://www.ivoox.com/audicion-del-senador-guido-manini-rios-25-10-2022_md_94846240_wp_1.mp3 (https://www.ivoox.com/audicion-del-senador-guido-manini-rios-25-10-2022_md_94846240_wp_1.mp3)
2022-	“Los cambios que se están	En su audición de este martes por AM 770 Radio	https://www.ivoox.com/_md_95244105_wp_1.mp3

3. Transcripción *(audio.whisper)*

Code

El siguiente paso es transcribir los audios a texto lo que realizo con el paquete `audio.whisper` que utiliza la herramienta *“Whisper” Automatic Speech Recognition model (https://github.com/openai/whisper)* desarrollada por openaAI y cuya documentación puede verse acá (https://github.com/bnosac/audio.whisper). La herramienta tiene diferentes modelos que van desde el menos potente *(tiny)* al más potente *(large)*, aunque no todos están disponibles para español. Los pasos para realizar las transcripciones son sencillos y están bien explicados en el repositorio. Antes de realizar la transcripción utilizo la librería `av` para transformar los audios mp3 a formato de archivo *.wav de 16 bit*, que es el requerido por `audio.whisper` .

Cuanto mayor es la potencia del modelo más demora la transcripción. Como tengo que transcribir varios audios de aproximadamente 15 minutos, para este ejercicio utilizo la versión más liviana *(tiny)*. Luego de la transcripción de cada audio incorporo el contenido en una nueva variable txt en mi base de datos.

Tabla 2. Base de datos con transcripción

fecha	subtitle	title	url_audios	text
2022-10-11	En su audición de este martes por AM 770 Radio Oriental, el Senador Guido Manini Ríos se refirió a la votación del proyecto de ley de Eutanasia, el jueves pasado en la Cámara de Diputados, y dijo:	“Los prestadores de salud tienen que estar obligados a tener cuidados paliativos para evitar el sufrimiento de los pacientes”	https://www.ivoox.com/_md_93791161_wp_1.mp3 (https://www.ivoox.com/_md_93791161_wp_1.mp3)	[Música] Estamos comienzo al espacio del partido cabildo abierto con ustedes el senador Guido, Vanini Rios. Amigos de cabildo abierto de todo el país orientales, bueno, de ella. El jueve pasado en la madrugada se terminó de votar en el Senado de la República el proyecto de Ley de Rendición de Cuentas. Creemos que se atendió a muchos sectores que estaban necesitando un refuerzo económico o rúglos para determinados enprendimientos o cumplir determinados objetivos. Yo quiero destacar que a propuesta de cabildo abierto se logró una mejor asiodiría que significativa en los salarios de aquellos más relegados de la administración pública a administración, trabaja en la administración como son los militares de menor rango. Ellos son atendidos en esta rendición de cuentas todos en general en una escala que va del 7,5% o a los soldados a cerca del 3% a capitanes y algunos mayores. Y por otro lado se votó un concepto

4. Open AI: Tema y resumen *(rgpt3)*

Code

Una vez obtenido el texto ya podría comenzar a trabajar con procesamiento y análisis de datos. Sin embargo, considero muy útil agregar una etapa más que se puede realizar utilizando la herramienta de openAI *chatGPT3*, cuya API puede conectarse desde R con la líbrería `rgpt3` . En *esta entrada del blog de Elina Gómez (https://www.elinagomez.com/blog/2023-02-21-gpt3-ccss/)* pueden ver los pasos para conectar con la API y otro ejemplo de uso. Un aspecto a tener en cuenta es que la API es de pago, pero para un ejercicio básico de prueba en general alcanza con lo que te permite utilizar gratis.

Lo que le pedí a *chatGPT3*, es que me devuelva para cada uno de los textos transcriptos cuál es el tema principal del que trata y que realice un resumen del contenido. Ambos resultados los agrego como columnas de mi dataframe. Finalizado este proceso tengo una base de datos de las audiciones de Manini Ríos generada por mi asistente de investigación donde consta: fecha, título, subtítulo (copete), tema, resumen y texto completo.

Tabla 3. Base de datos final

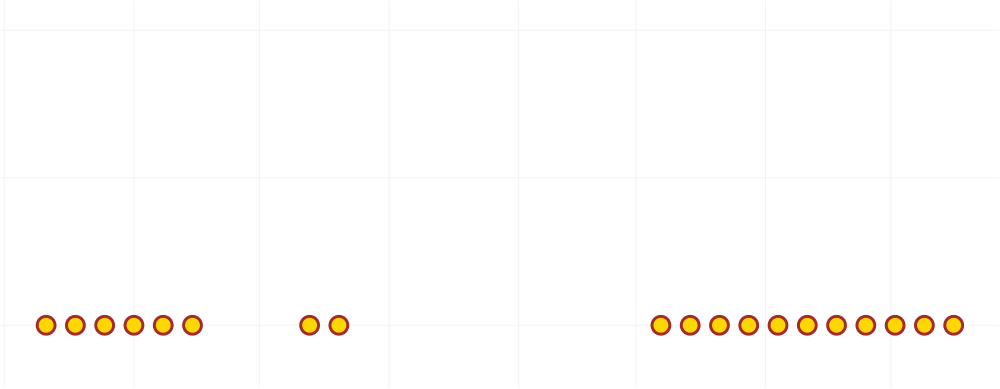
fecha	subtitle	title	url_audios	text	tema	resumen
2022-10-11	En su audición de este martes por AM 770 Radio Oriental, el Senador Guido Manini Ríos se refirió a la votación del proyecto de ley de Eutanasia, el jueves pasado en la Cámara de Diputados, y dijo:	“Los prestadores de salud tienen que estar obligados a tener cuidados paliativos para evitar el sufrimiento de los pacientes”	https://www.ivoox.com/_md_93791161_wp_1.mp3 (https://www.ivoox.com/_md_93791161_wp_1.mp3)	[Música] Estamos comienzo al espacio del partido cabildo abierto con ustedes el senador Guido, Vanini Rios. Amigos de cabildo abierto de todo el país orientales, bueno, de ella. El jueve pasado en la madrugada se terminó de votar en el Senado de la República el proyecto de Ley de Rendición de Cuentas. Creemos que se atendió a muchos sectores	Rendición de cuentas militar.	El Senado d República v proyecto de de Rendició Cuentas, el atiende a m sectores qu necesitan u refuerzo económico. ley incluye u aumento significativo salarios de l militares de menor rang pago por nocturnidad partidas especiales p

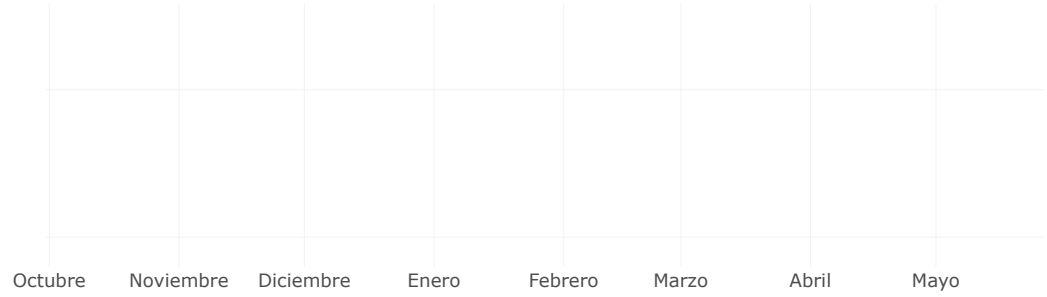
5. Visualización

Hay diversas formas de visulización de la información recogida en la base. Las más interesantes son las que surgen de procesamientos de datos con las distintas técnicas vistas a lo largo del curso (nubes de palabras, redes de co ocurrencias, frecuencias, etc.). En este caso, haremos algo bien sencillo y cuyo objetivo es solamente darnos un panorama sintético de la evolución del discurso de Manini Ríos a partir de la ferecuencia de audiciones y de los temas que trató en cada una

Code

Gráfico 1. Tema principal en audiciones de Manini Ríos





Esta representación gráfica sería más interesante si abarcara períodos temporales más largos e integrara a más figuras políticas, para que pueda compararse la evolución del discurso público y la agenda de cada uno. A modo de ejemplo, en los siguientes gráficos se integran audiciones hipotéticas de Julio María Sanguinetti, José Mujica y Laura Raffo, colocando además una línea que representa un evento X de interés, en este caso, la presentación en el parlamento de las modificaciones a la reforma de la seguridad social por parte del Poder Ejecutivo.

Si el gráfico tuviese información real (que no es el caso!), nos permitiría visualizar información de las temáticas que trataron los líderes políticos antes y después de un evento de interés. Además podríamos hacer operaciones y cálculos de frecuencia absoluta y relativa de audiciones, desagregado por tema, entre otras posibilidades.

Esta primera aproximación visual a los datos no ingresa al análisis textual más profundo, sin embargo,nos permite conocer más acerca de nuestros datos. Es importante utilizar las visualizaciones durante todo el proceso de análisis y no sólo como herramienta de presentación de resultados!

Gráfico 2. Simulación de audiciones de Manini Ríos, Mujica, Raffo y Sanguinetti (pre y post reforma SS)

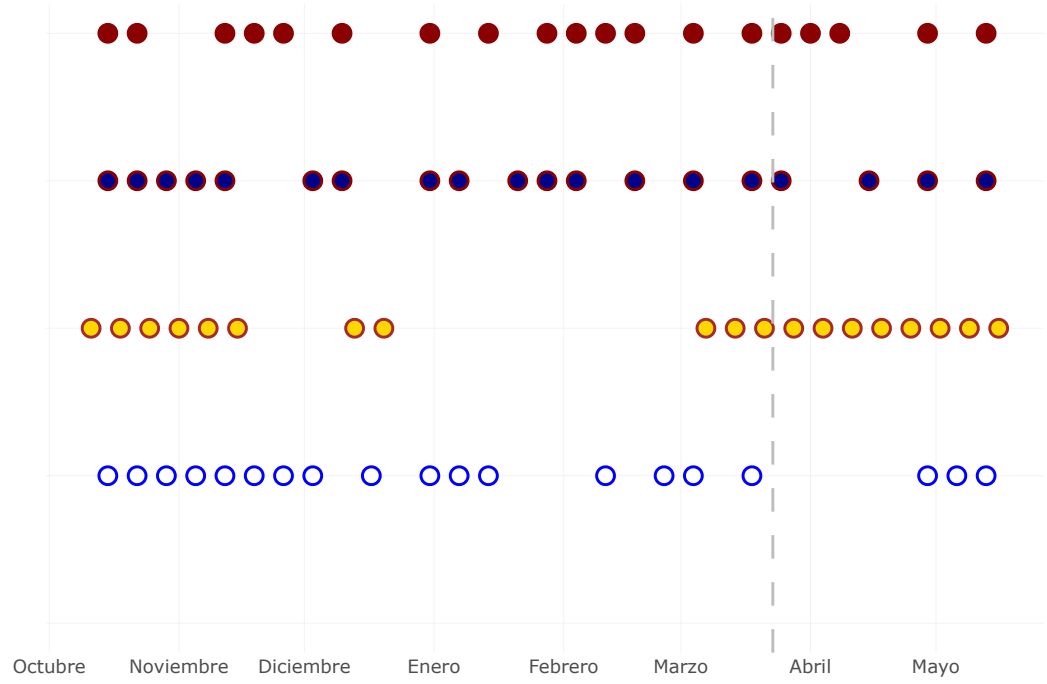
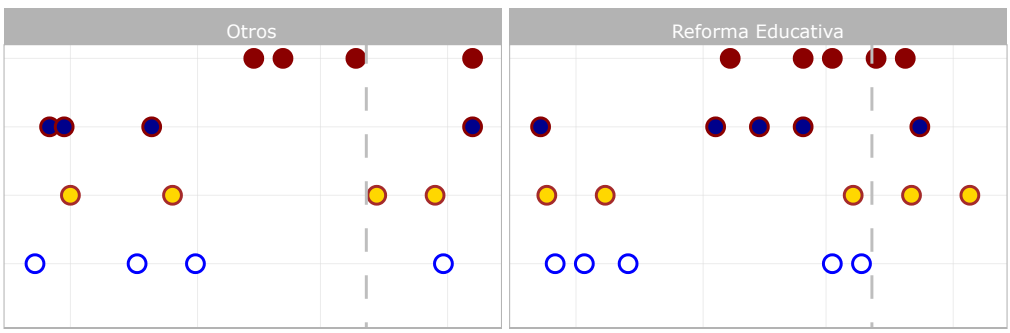
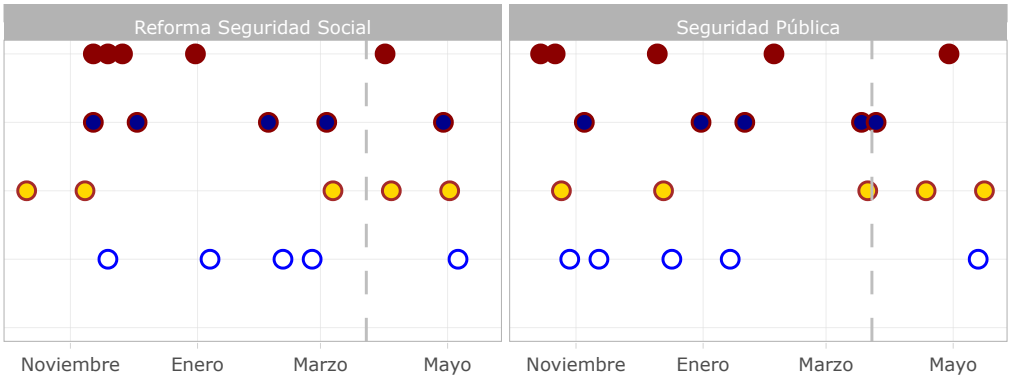


Gráfico 3. Simulación de temas en audiciones de Manini Ríos, Mujica, Raffo y Sanguinetti (pre y post reforma SS)





6. Posibles aplicaciones

¿Para qué se puede usar esto de la transcripción? ¿qué nos aporta?

Ahorra tiempo: transcripción de entrevistas es el mejor ejemplo

Trabajar con fuentes no tan habituales y pasibles de comparación: discursos no escritos de políticos/as, entrevistas, programas de radio, fragmentos de noticieros, etc.

Ejemplos: - comparar las temáticas tratadas por Mujica y Manini en sus audiciones radiales durante un año - comparar el resumen de noticias entre diferentes noticieros radiales durante x tiempo

Extra: Transcripción modelo medium

Code

La transcripción con el modelo *tiny* de `audio.whisper` es buena pero no perfecta como puede verse en la columna text de la base de datos. Por ejemplo, tiene bastantes problemas con los nombres propios (“Manini”, “Vanini”, “Vanivi”, etc.). Si se quiere obtener una transcripción fidedigna y precisa es mejor optar por modelos de `audio.whisper` más potentes como *medium* o *large*. Lo que hay que tener en cuenta es que demoran mucho tiempo más en ejecturase: *velocidad vs precisión*. Al hacer la consulta con chatGPT3, devuelve prácticamente el mismo resultado cuando se le pide tema y resumen! Así que la elección del modelo dependerá del objetivo de la transcripción y con qué corpus textual se vaya a trabajar. Si se trabaja con el texto completo, es mejor un modelo potente. Si se va a trabajar con resumenes creados por AI, es más razonable usar modelos más livianos y rápidos.

Tabla 4. Tiny vs Medium

text_tiny	text_medium
[Música] Vamos con mi enzo al espacio del partido Cabildo Abierto con ustedes el senador Guido Vanimi Rios A mi go de Cabildo Habierto de todo el país Orientales Bueno, dia En estos últimos días si lugar a dudas Acaparó los titulares, los grandes titulares en la actividad política nacional Todo lo referido Al intercambio de propuestas y de ideas en torno Al proyecto de reforma de la seguridad social Que desde el medio cubre Ingressó el Parlamento Que a finales del mes de diciembre Fue aprobado en el Senado Y que ahora están en la etapa final de su confideración En la cámara de representantes Cuando se votó En el mes de diciembre en el Senado nuestro partido cabido abierto Dejo bien en claro Que lo hacía para cumplir con el crono grama trasado Para no retrasar más una agenda O uno plazo Para aprobar una ley de este tipo Que ya en el cuarto año de gobierno Es bastante tardillo por las características de una ley Que siempre contendrá artículos polémicos Se votó en el Senado y se dijo Que en la cámara de representantes se iba a continuar Buscando mejorar algunos aspectos del proyecto Claroamente en octubre Cuando se presentó el proyecto al Parlamento Digimos en aquello oportunidad Que en el Parlamento Y vamos a buscar mejorar el proyecto Y eso exactamente lo que hicimos Lo que dijimos en octubre Fue lo que hicimos en todo este período Hasta el día de hoy Que en esquieren sembrar la idea de que cabildo Faltó a una palabra empeniada Están realmente faltando a la verdad Taclaramente	Damos comienzo al espacio del partido Cabildo Abierto, con ustedes el senador Guido Manini Ríos. Amigos de Cabildo Abierto de todo el país orientales, buenos días. En estos últimos días sin lugar a dudas acaparó los titulares, los grandes titulares en la actividad política nacional, todo lo referido al intercambio de propuestas y de ideas en torno al proyecto de reforma de la seguridad social que desde el mes de octubre ingresó al parlamento, que a finales del mes de diciembre fue aprobado en el Senado y que ahora está en la etapa final de su consideración en la Cámara de Representantes. Cuando se votó en el mes de diciembre en el Senado, nuestro partido Cabildo Abierto dejó bien en claro que lo hacía para cumplir con el cronograma trazado, para no retrasar más una agenda o unos plazos para aprobar una ley de este tipo que ya en el cuarto año de gobierno es bastante tardío por las características de una ley que siempre contendrá artículos polémicos. Se votó en el Senado y se dijo que en la Cámara de Representantes se iba a continuar buscando mejorar algunos aspectos del proyecto. Claramente en octubre cuando se presentó el proyecto al parlamento, dijimos en aquella oportunidad que en el parlamento íbamos a buscar mejorar el proyecto y eso es exactamente lo que hicimos, lo que dijimos en octubre fue lo que hicimos en todo este período hasta el día de hoy. Quienes quieren sembrar la idea de que Cabildo faltó a una palabra empeñada están realmente