UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

INTRODUCCIÓN DE LA CIENCIA DE DATOS

Tarea 1

Curso 2025

AUTOR:
Matías Izquierdo - C.I.: 5.288.879-7
Francisco Garchitorena - C.I.: 5.342.952-0

DOCENTES:
María Inés Fariello
Marcelo Fiori
Lorena Etcheverry
Guillermo Moncecchi
Graciana Castro

FECHA: 13 de mayo de 2025

Tabla de contenidos

1. Parte 1

1.1. A

Los datos utilizados para la presente tarea provienen de un archivo .csv, que posee información de los discursos realizados por los candidatos presesidenciales de Estados Unidos, durante la campaña del 2020. La lectura del archivo convierte el set de datos en un DataFrame de Pandas. Las columnas del DataFrame son las que siguen:

- Speaker: Candidato presidencial que dio el discurso.
- Title: Título del discurso.
- Text: Transcripción del discruso.
- Date: Fecha en la que se realizó (ej: Oct 16, 2020).
- Location: Ubicación, en algunos casos el formato es Ciudad, Estado mientras que en otros es únicamente el Estado o la Ciudad.
- Type: El tipo de discurso (ej: discurso de campaña, debate, entrevista, etc.).

Se encuentran datos faltantes tipo NaNs en las columnas "speakers" (3), "location" (18) y "type" (21). Por su parte, en la fila 194 se encuentra un "???" en la columna speakers. A su vez, se verificaron errores de escritura por mayúscula-minúscula y se detectan cero errores.

Realizada esta primera depuración se contabilizan los discursos por candidato y se seleccionan los cinco oradores con mayor cantidad:

Speaker	**Count**
Joe Biden	71
Donald Trump	53
Mike Pence	19
Bernie Sanders	16
Kamala Harris	11

Tabla 1: Número de discursos por orador

1.2. B

Con el objetivo de visualizar el comportamiento de los principales candidatos a lo largo de la campaña se procede a separar los discursos de manera mensual. Para esto se construye un nuevo Dataframe donde las columnas son los distintos oradores y cada fila un mes de la campaña. Con este nuevo Dataframe se realiza un gráfico de barras (figura 1) que ilustra la cantidad de discursos de los candidatos por mes.

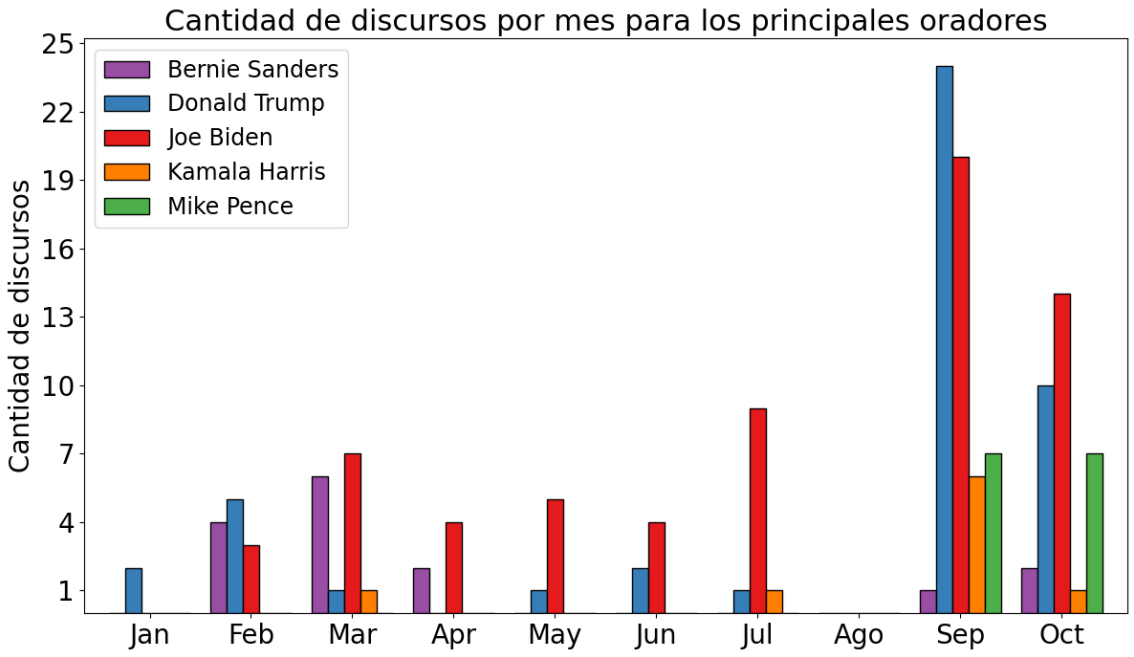

Figura 1: Actividad de los principales candidatos.

Esta forma de visualización permite observar fácilmente el aumento de actividad general al final de la campaña, con un pico en el mes de septiembre. Por otro lado se detecta un caso que no sigue esta tendencia como es el de Bernie Sanders, quien comenzó la campaña de forma activa pero luego redujo su participación. También se puede detectar quiénes fueron los candidatos más activos, Joe Biden y Donald Trump, este último con un pico en el mes de septiembre siendo el candidato con más discursos en un mismo mes, mientras que el primero realizó una campaña más pareja en el tiempo.

Un dato tener en cuenta es el año en el que se dieron los discursos. El año 2020 se correspondió con el año de la pandemia del Covid19, lo cual coincide con lo que ilustra la figura 2. Aquí se observa que un 20 % de los discursos fueron virtuales. Si se separa entre antes y después de julio se observa como aumenta la cantidad de discursos virtuales el comienzo de la pandemia, mientras que decrece a medida que se acerca la fecha de la elección (ver figura 3).

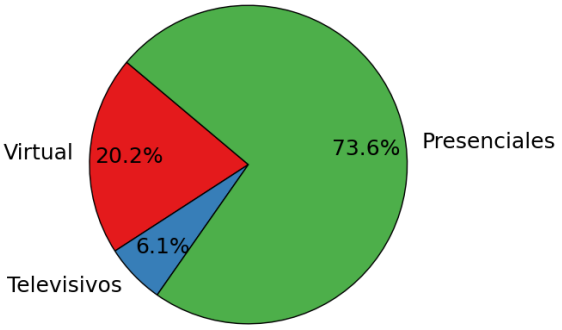

Figura 2: Porcentajes de discursos, virtuales, presenciales y televisivos.

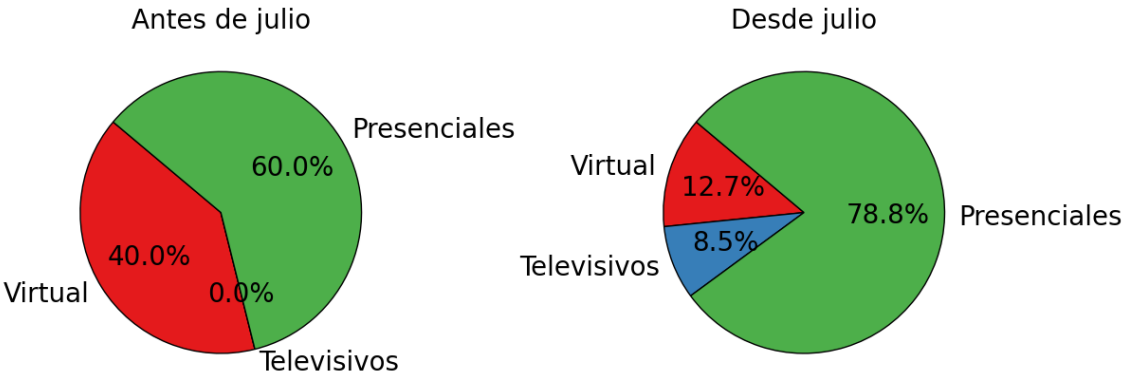

Figura 3: Porcentajes de discursos, virtuales, presenciales y televisivos. Comparación entre las proporciones antes y después de julio.

1.3. C

Es de interés realizar un conteo e identificación de palabras para analizar los discursos de cada candidato. Para esto se debe continuar con la limpieza de datos, analizando el formato de los discursos, realizando una normalización del texto (pasarlo todo a minúsculas) y eliminando los signos de puntuación.

En primer lugar, se analizaron los discursos para visualizar el formato que presentaban. Dado que son transcripciones, se observó que muchos de ellos correspondían a conversaciones entre el candidato de interés y otros agentes (entrevistadores, multitudes, presentadores, etc.). El formato de estos diálogos se da de la forma:

```
Crowd: (05:51)
We love you.
Mike Pence: (05:52)
I know this is Packer's Country.
Crowd: (05:53)
Yes it is.
Mike Pence: (05:55)
...
```

Por este motivo, dado que el objetivo es analizar exclusivamente lo dicho por el candidato, se procede a eliminar todas las intervenciones realizadas por agentes externos (por ejemplo, el público, identificado en las transcripciones como "Crowd"). El texto preprocesado conserva únicamente las palabras pronunciadas por el candidato, eliminando también estructuras del tipo "Nombre del hablante (HH:MM:SS)", que son comunes en este tipo de registros. Para realizar esta limpieza se utiliza la biblioteca de Python re, que permite identificar dichos patrones mediante expresiones regulares. Asimismo, dado que los candidatos no siempre son mencionados por su nombre completo, sino que pueden aparecer referidos por títulos o apodos (por ejemplo, "President Trump" en lugar de "Donald Trump"), se implementó un sistema de alias para asegurar que se consideren todas las formas bajo las cuales cada candidato es mencionado.

Con el texto preprocesado para incluir únicamente lo que expresa el candidato, se

procede a limpiar el texto, eliminando signos de puntuación y pasando todo el texto a minúsculas. Al código base presentado en la tarea se le agregan los siguientes signos de puntuación: (,), !, ., ;, ..., ¿, ¡, “, ”, ”, ’, ‘, ’, {, }, $, 0, 1, -, /. También se detecta que los discursos están escritos con contracciones típicas del inglés (ej: you’re, don’t, etc.) que, si no se corrigen, harán que la contabilización no sea correcta. Por esta razón, se seleccionan ciertas contracciones usuales y se las cambia por palabras completas (ej: you are, do not, etc.), nuevamente utilizando la librería re.

2. Parte 2

2.1. A

Realizadas estas modificaciones se procede con la contabilización. Para lograrlo se toman los discursos normalizados y corregidos y se los transforma en listas donde cada elemento es una palabra. A esta columna del DataFrame se la denomina “WordList”. Luego se “explota” estas listas utilizando la función de pandas *explode*, generando un nuevo Dataframe con una fila por palabra perteneciente al discurso del orador X. Como hay palabras que se repiten dentro del mismo orador es necesario agruparlas utilizando la función *groupby*. Además, se agrega una columna con la contabilización. Finalmente, se cuenta con un Dataframe de 3 columnas: orador, palabra y cantidad de veces que la dijo.

La información procesada se presenta en forma de gráfica de barras (figura 4), donde se muestran cinco grupos de palabras (uno por orador) con las cinco palabras más repetidas por el orador correspondiente. Los datos se normalizaron, considerando la cantidad de palabras total expresada por cada candidato, con el fin de visualizar de forma más clara la frecuencia con la que cada palabra es expresada, sin importar la cantidad de discursos que haya dado cada candidato.

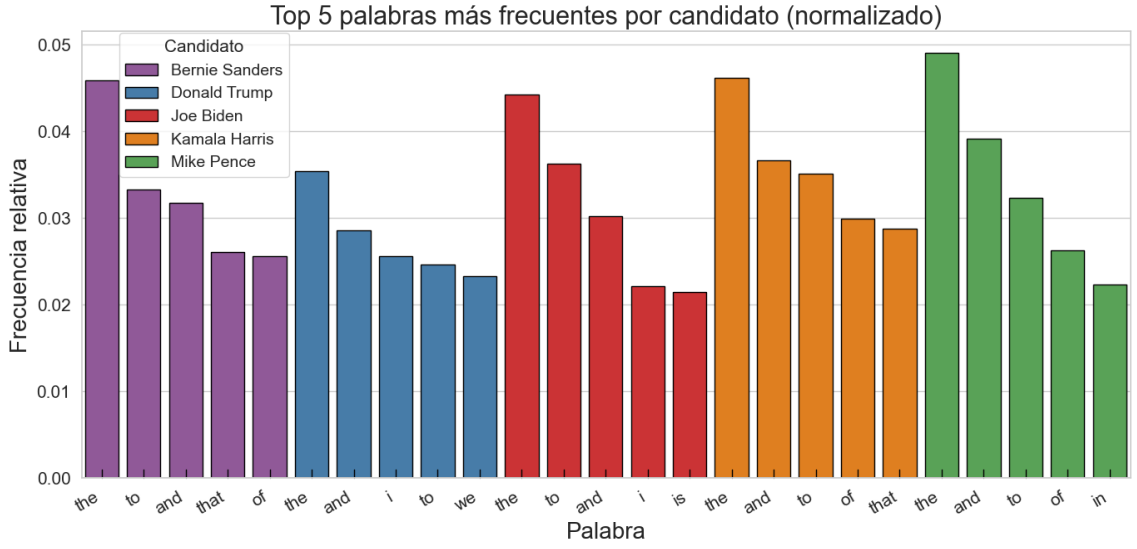

Figura 4: Cinco palabras más frecuentes por candidato.

Un factor que se observa en los resultados es que las palabras más frecuentes son del tipo "stopwords". Estas últimas se refieren a palabras muy comunes y frecuentes en un idioma, que por sí solas aportan poco significado al análisis textual. Dado que en sí no aportan ningún tipo de información sobre los discursos, interesaría eliminar estas palabras, para así obtener más información que permita diferenciar a los candidatos por partido político, fecha o lugar donde da el discurso.

2.2. B

La figura 5 ilustra la cantidad de palabras expresadas por cada candidato en el total de sus discursos. Para obtenerla, se sumó la cantidad de palabras de cada fila de la columna "WordList" (mencionada previamente), para cada candidato.

Se observa, que Donald Trump es el candidato con mayor cantidad de palabras expresadas, mientras que Joe Biden lo sigue. Esto no sigue el mismo orden observado en la tabla 1, donde se veía que Joe Biden es el candidato con mayor cantidad de discursos.

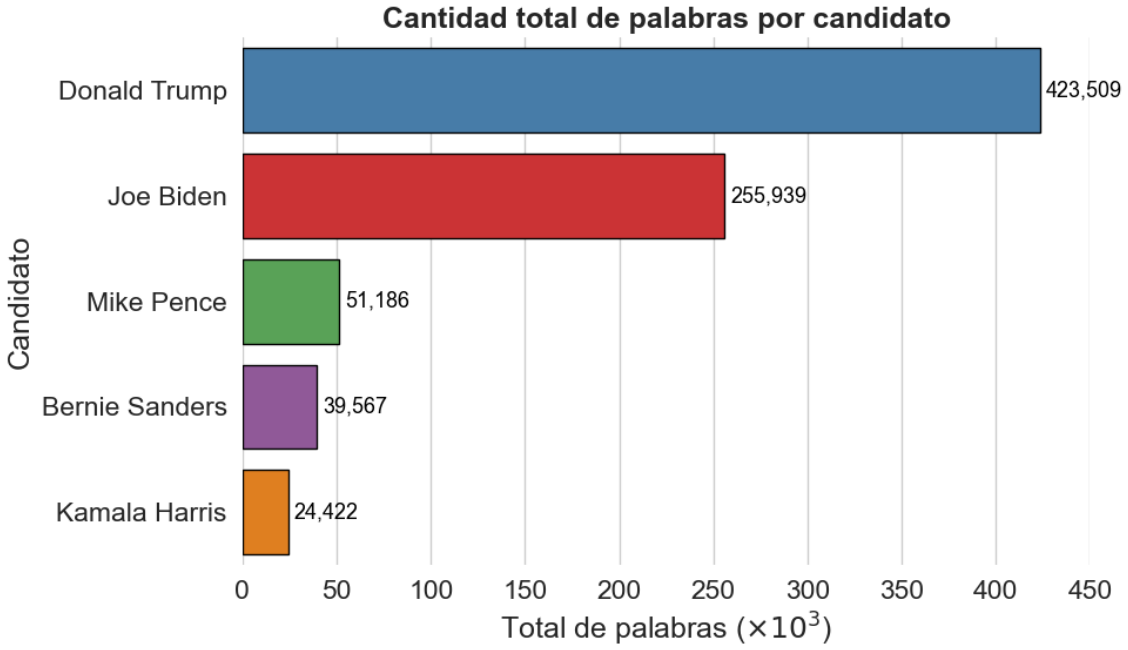

Figura 5: Comparación de cantidad de palabras expresadas por cada candidato.

Por otro lado, interesa observar cómo se diferencian los candidatos en términos de vocabulario, analizando la cantidad de palabras diferentes entre sí expresadas por cada candidato. Para ello, se utiliza la columna "WordList" del DataFrame, y se agrupa todas las palabras de las distintas filas utilizando la función *groupby*. Luego, con la función *set*, se eliminan las palabras repetidas, para así contar todas las palabras distintas expresadas por cada candidato.

Los resultados se visualizan en la figura 6. Nuevamente, Donald Trump presenta el vocabulario más diverso.

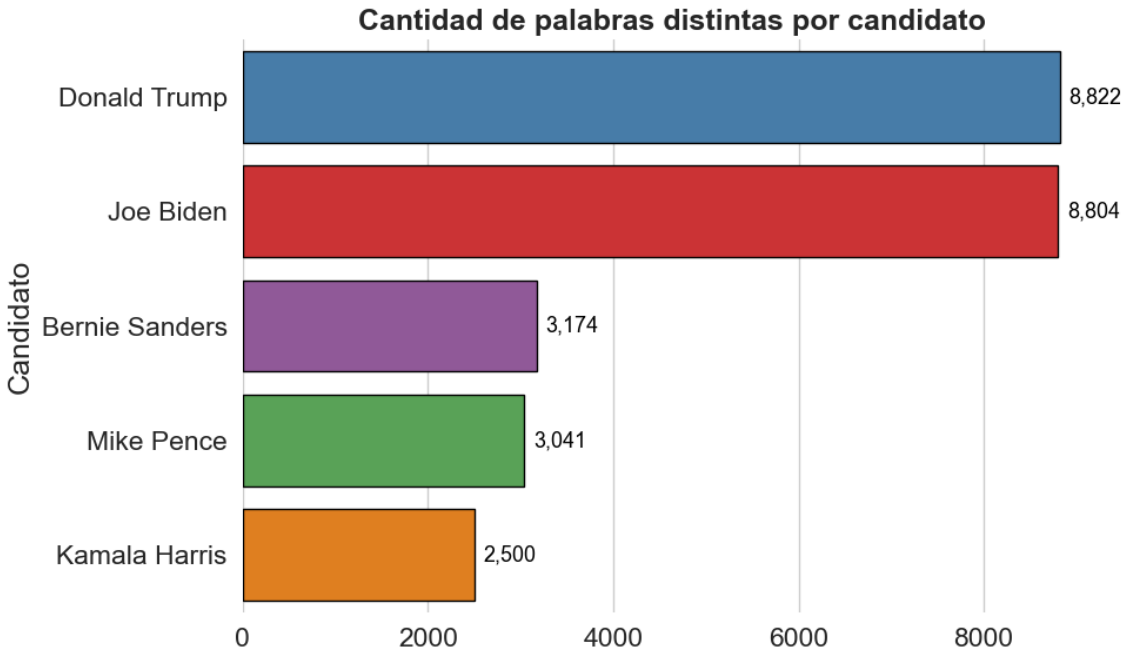

Figura 6: Comparación de cantidad de palabras distintas entre sí expresadas por cada candidato.

2.3. C

Previo a analizar las menciones entre candidatos, se tuvo en cuenta las distitnas formas de mencionar a cada candidato. Para ello, se normalizó la forma en la que cada candidato es nombrado, colocando su nombre y apellido. En la tabla 2 se ilustra el nombre de cada candidato con sus respectivos alias.

Candidato	**Alias**
Joe Biden	vice president biden, vicepresident biden, senator biden, biden, joe biden
Donald Trump	president trump, trump, donald trump
Mike Pence	mike pence, vice president pence
Bernie Sanders	bernie sanders, senator sanders, mr sanders
Kamala Harris	kamala harris, senator harris, mrs harris

Tabla 2: Lista de candidatos y sus alias utilizados para detectar menciones en los discursos.

La tabla 3 ilustra la cantidad de veces que cada candidato menciona a los otros. Para visualizar más fácilmente los resultados, se ilustra la figura 7. Se observa que Donald Trump es quien menciona más a los restantes candidatos, siendo Joe Biden su objetivo más frecuente. Es interesante observar que tanto Joe Biden como Donald Trump poseen una gran cantidad de menciones a sí mismos, factor no observado en el resto de los candidatos.

Tabla 3: Menciones entre candidatos

	Joe Biden	Donald Trump	Mike Pence	Bernie Sanders	Kamala Harris
Joe Biden	73	471	0	8	6
Donald Trump	830	348	31	39	10
Mike Pence	249	307	0	5	37
Bernie Sanders	60	124	0	1	1
Kamala Harris	36	31	0	0	1

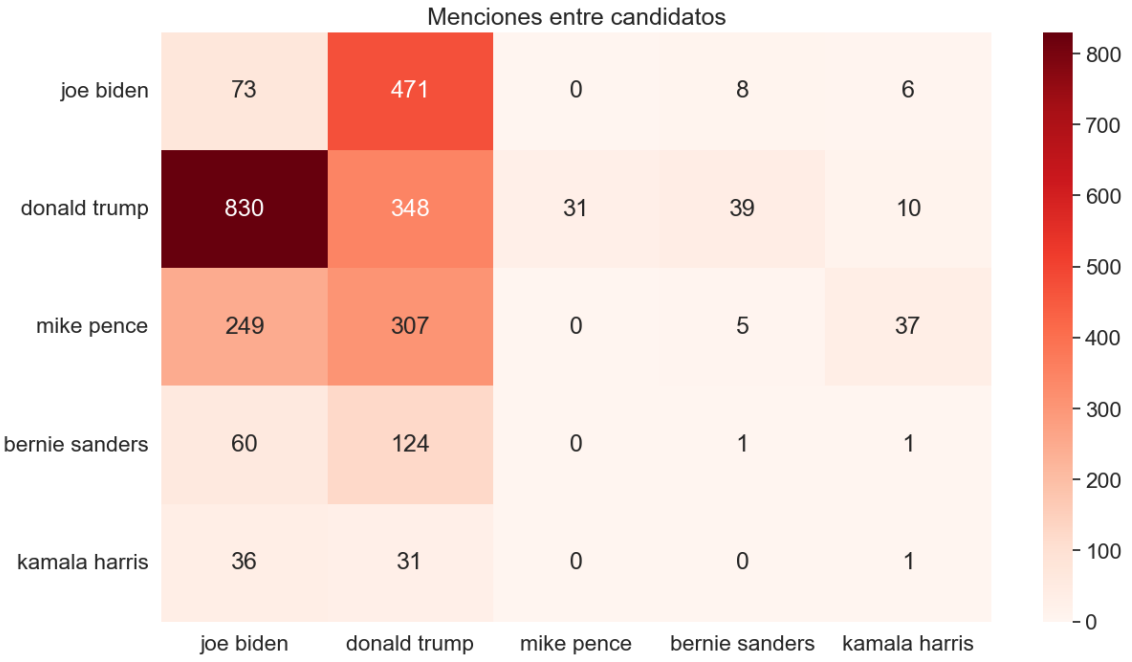

Figura 7: Menciones entre candidatos.

2.4. D

A partir del análisis desarrollado se pueden contestar ciertas preguntas, referentes a la información que poseen los discursos de cada candidato.

Una pregunta que surge es: ¿Qué diferencias existen en los temas más tratados entre los candidatos? Para responderla, se podría utilizar la lista de palabras más frecuentes por candidato (después de eliminar las stopwords), analizando sus frecuencias relativas para ver qué términos son más distintivos de cada uno. Esto permitiría observar cuál es el foco temático de cada discurso. Un análisis más exhaustivo consistiría en realizar este mismo estudio, pero segmentando por locación geográfica del discurso. Por ejemplo, se podrían comparar las palabras más frecuentes en discursos dados en el norte vs el sur de Estados Unidos, y visualizar sus diferencias.

Por otro lado, se podría contestar: ¿Qué tan centrados están los discursos en atacar o mencionar a los oponentes? Este análisis podría realizarse observando la cantidad de menciones explícitas a los nombres de los otros candidatos, tal como se visualiza en la figura 7. Además, se podría extender este análisis para clasificar el tono de esas menciones (positivo, neutral o negativo) mediante técnicas de análisis de sentimiento.

Otra pregunta interesante es: ¿Cómo fue evolucionando el lenguaje de los discursos a lo largo del tiempo? Esto podría abordarse observando cambios en el vocabulario de cada candidato, su complejidad, o la aparición de ciertos términos asociados a eventos históricos específicos (como por ejemplo el asesinato de George Floyd).

También se podría preguntar: ¿Qué emociones predominan en los discursos de cada presidente? Para ello, se podría aplicar un análisis de sentimientos o detección de emociones, clasificando fragmentos del discurso en categorías como ira, esperanza, miedo o entusiasmo.

Finalmente, otra línea de análisis podría abordar: ¿Qué estructura narrativa sigue cada presidente en sus discursos? Aquí se podría examinar la proporción de oraciones que hacen promesas, apelan al patriotismo, señalan problemas o llaman a la acción, usando etiquetas gramaticales o análisis temático.