UNIVERSIDAD DE LA REPÚBLICA

UNIVERSIDAD DE LA REPÚBLICA ORIENTAL DEL URUGUAY
FACULTAD DE INGENIERÍA
INSTITUTO DE COMPUTACIÓN

INTRODUCCIÓN A LA CIENCIA DE DATOS
Mayo de 2025

Entrega 1

GRUPO 11

Karina Cardozo – CI.: 4.135.872-7
Karen García - CI.: 4.945.279-5

UNIVERSIDAD DE LA REPÚBLICA

INDICE

1 Parte 1: Cargado y Limpieza de Datos

A continuación se detalla el proceso de análisis exploratorio, limpieza y normalización de la base de datos de discursos políticos de la campaña presidencial de Estados Unidos de 2020. El objetivo es comprender la calidad de los datos, extraer y normalizar los nombres de los oradores, y preparar la información para análisis posteriores.

1.1 PARTE A

Compruebe que puede correr las primeras dos celdas del notebook, observe el contenido de los dataframes cargados. Reporte si existen datos faltantes en algún campo, o cualquier otro problema de calidad de datos que encuentre. En particular, analice la cantidad de discursos por candidato/a, y a partir de este punto trabaje con los cinco candidatos/as con mayor cantidad de discursos.

1.1.1 Visualización de los datos

Se parte del archivo us_2020_election_speeches.csv, se importan pandas y se lee el CSV, se confirma la carga correcta como data frame *df_speeches* y se inspeccionan los datos.

Dimensiones:

- 269 filas
- 6 columnas que son:
    1. **speaker**: Orador principal o múltiples oradores, separados por coma. A veces aparece como "Multiple Speakers", "Democratic Candidates" o "???".
    2. **title**: Resumen textual del evento / discurso. Suelen incluir el nombre del orador, ubicación, tipo de evento, fecha y medio.
    3. **text**: Transcripción completa, compuesta por bloques: Nombre ORADOR: (mm:ss) seguido del contenido y finalmente salto de línea.
    4. **date**: Fecha en formato Month DD, YYYY (ejemplo, "Oct 16, 2020").
    5. **location**: Lugar (ciudad, estado), medio o etiqueta "Virtual".
    6. **type**: Formato de evento, se identifican principalmente los siguientes,
        - "Campaign Speech": discursos en actos de campaña electoral.
        - "Town Hall": sesiones tipo foro con participación del público.
        - "Debate": enfrentamientos entre candidatos con reglas formales.
        - "Interview": respuestas del candidato en medios o a periodistas.
        - "Press Conference": comparecencias oficiales con preguntas.
        - "Roundtable": mesas de diálogo o discusión con otros actores.
        - "Endorsement": mensajes públicos de apoyo a otro candidato.
        - "Statement": declaraciones oficiales o posicionamientos públicos

Tipos de datos: Todas las columnas aparecen como object.

1.1.2 Análisis de calidad de los datos

Valores faltantes

Speaker: 3 celdas vacías.
Esta información puede recuperarse desde el título y/o el texto, donde se listan claramente los participantes. En este caso, los datos faltantes corresponden con eventos multiorador del DNC (Democratic National Convention):

- DNC Night 1 (Michelle Obama, Bernie Sanders, John Kasich & more)
- DNC Night 2 (Bill Clinton, AOC, Jill Biden & more)
- DNC Night 3 (Barack Obama, Kamala Harris, Hillary Clinton & more)

Location: 18 registros no especifican ubicación ni medio.
Inspeccionando se observa que corresponden a eventos como entrevistas, paneles virtuales, conferencias de prensa nacionales y declaraciones posteriores a resultados electorales. Ejemplos:

- "Joe Biden Speech at the Million Muslim Votes Summit" (evento temático)
- "Bernie Sanders Reacts to Super Tuesday Results" (análisis postelectoral)
- "Nevada Caucus Speech Transcripts" (cobertura múltiple sin lugar único)
- Varios discursos del DNC, que por su carácter virtual/global no tienen una ubicación única.

Type: 21 registros faltantes.
Tras análisis manual, se observa que estos discursos incluyen debates, declaraciones, paneles, entrevistas y eventos como el DNC (Democratic National Convention) que no fueron debidamente etiquetados.

Duplicados
No se identificaron filas duplicadas, por tanto, cada registro de discurso es único.

Fechas parseadas
Convertimos la columna date a datetime con errors='coerce' y ninguna fila falló (0 NaT).
El rango temporal abarca desde el 15 de Enero de 2020 al 16 de Octubre de 2020, abarcando las principales etapas de la campaña electoral.

Observaciones en relación a la columna Speaker
Se observan diferencias de mayúsculas/minúsculas ("Joe Biden" vs "joe biden"), varios oradores por celda (celdas multivariadas) y que se utilizan distintas forma de referirse al mismo orador, ej., "President Trump", "Donald J. Trump", "Donald Trump", "President Donald J. Trump", " Trump", todos refieren a la misma persona.

También hay valores genéricos como "Multiple Speakers" o "Democratic Candidates", o valores desconocidos como “???”, que requieren tratamiento especial.

1.1.3 Limpieza de datos

Aunque no hay errores graves en los datos se han observado áreas a mejorar, a continuación se describe el proceso de limpieza de datos, la extracción precisa de oradores y la unificación de variantes de nombres para garantizar un análisis correcto y consistente de forma tal de construir una base coherente de datos para la comparación entre lo declarado (columna speaker) y lo efectivamente dicho (columna text).

Desde la columna speaker

La columna speaker contiene los oradores declarados por cada discurso, pero, como se ha mencionado puede tener: celdas multivariadas (múltiples oradores separados por comas), mismos nombres escritos de distinta forma, con mayúsculas y/o espacios, y valores genéricos o desconocidos como "Democratic Candidates" o "???".

Pasos aplicados:

- Se dividieron los valores separados por coma en listas individuales.
- Se normalizó el texto (se pasó todo a minúsculas y se eliminaron espacios extra).
- Se eliminó cualquier duplicado por fila.

Resultado: el DataFrame DS_orador_from_speaker contiene por fila la lista de oradores declarados en forma limpia y lista para comparar.

Desde la columna text

La columna text contiene las transcripciones completas de los discursos. Allí, cada intervención sigue un patrón claro:

Salto de línea + Nombre del orador + Marca horaria ": (mm:ss)" (ejemplo: Joe Biden: (00:00)).

Para extraer estos nombres:

- Se aplicó una expresión regular que detecta patrones tipo "Salto + Orador + : (mm:ss)".
- Se generaron listas de nombres por fila.
- Se aplicó normalización (minúsculas, limpieza de espacios).
- Se eliminaron duplicados por fila.

Resultado: el DataFrame DS_orador_from_text contiene por fila los oradores efectivamente detectados en los textos.

Comparación entre Speaker y Text

Para verificar si los oradores declarados en speaker coinciden con los detectados en text para cada discurso, para cada fila:

- Se comparó la lista de DS_orador_from_speaker con DS_orador_from_text.
- Se registraron las diferencias: Oradores que aparecen en speaker pero no en text. Oradores que aparecen en text pero no en speaker.
- Se construyó el DataFrame DF_speaker_vs_text para documentar las discrepancias.

Normalizaciónes realizadas

Durante la comparación se identificaron numerosas variantes de nombres, se observó los casos de mayor ocurrencia para los que se aplicaron reglas de reemplazo y unificación:

Se nombra:	Cuando aparece:	No se modifica cuando aparece:
joe biden	“biden”, “vice president joe biden”	“jill biden”
donald trump	“president trump”, “donald j. trump”, “president donald j. trump”, “trump”	“donald trump jr.”, “trump jr.”
kamala harris	“senator kamala harris”	
mike pence	“vice president mike pence”	
amy klobuchar	“senator amy klobuchar”	

Tabla 1

Reconstrucción Final de Oradores por Discurso

Una vez normalizados los nombres se combinan ambas fuentes de información (from_speaker y from_text) para esto se construyó el DataFrame **DF_speaker_mas_text** de la siguiente forma:

- Utilizando el valor extraído del texto cuando speaker era igual a:
  "democratic candidates"
  "republican candidates"
  "multiple speakers"
  "???"
  “celda vacía”

- En los demás casos se conservaron los oradores de speaker (ya limpiados)

A través de **DF_speaker_mas_text** completo y depurado se puede identificar a los cinco oradores principales y centrar el análisis posterior en ellos. Se registra el total de intervenciones por orador y se extraen las primeras cinco filas ordenadas por número de discursos (mayor a menor) formándose el DataFrame **Top_5.**

Resultado obtenido

Los cinco candidatos con mayor número de discursos acumulados, ordenados de mayor a menor, son:

- Joe Biden,
- Donald Trump,
- Bernie Sanders,
- Mike Pence y
- Kamala Harris.

1.2 PARTE B

Genere una gráfica que permita visualizar los discursos de los candidatos/as a lo largo del tiempo, con alguna escala temporal adecuada. Comentar si se identifican momentos clave de la campaña. No realizar análisis estadísticos, solamente generar visualizaciones exploratorias.

Procedimiento

En esta fase se parte de Top_5, que contiene los cinco candidatos con mayor número de intervenciones. El objetivo es construir, para cada candidato, su secuencia cronológica de discursos. El procedimiento para esto fue el siguiente:

- Filtrado de registros
  En DF_speaker_mas_text, se seleccionan únicamente las filas donde aparezca el nombre del candidato.
- Recuperación de fechas
  Con los índices resultantes del filtrado, se extraen las fechas correspondientes de df_speeches['date'].
- Construcción del DataFrame
  Se crea un DataFrame temporal, llamado DF_fecha_<candidato> con las fechas de cada uno de sus discursos.

Una vez identificada y ordenada esta información se obtuvo la progresión acumulada de discursos por candidato a lo largo del tiempo aplicando groupby('candidato').cumcount() + 1 para generar la nueva columna count, indicando el número de discursos acumulados hasta cada fecha. Se obtuvo así una serie temporal por orador, que permite visualizar el crecimiento de las intervenciones. Una vez realizado esto se procedió a la representación gráfica de la ***evolución temporal de la cantidad total de discursos por candidato,*** la misma se muestra a continuación.

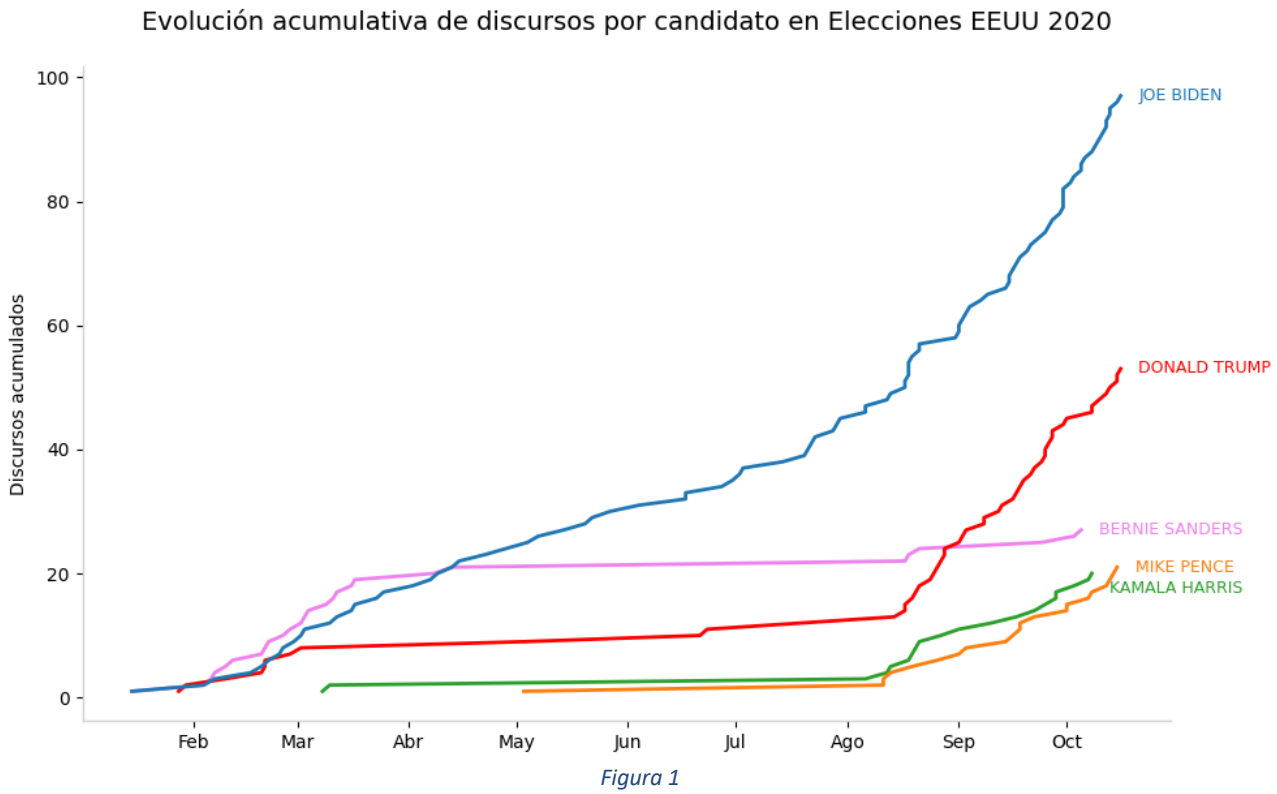

Figura 1

Análisis del gráfico

Al final de las curvas se observa que:

- Kamala Harris, Mike Pence y Bernie Sanders presentan un recuento de intervenciones muy similar (poco más de 20).
- Por su parte, Donald Trump acumula casi el doble de discursos que estos tres candidatos (poco más de 50).
- Joe Biden casi duplica, a su vez, el número de intervenciones de Trump (con unas 100 aprox).

A través de este gráfico se pueden observar los siguientes hitos electorales:

- *7 de abril:* Sanders se retira de la carrera demócrata tras perder las primarias.

- *10 y 11 de agosto:* Mike Pence es anunciado como compañero de formula de Donald Trump y un día después Kamala Harris es anunciada como compañera de fórmula de Biden. Esto coincide con el aumento de sus discursos.

- *Finales de agosto:* Convenciones nacionales demócrata y republicana (evento que usualmente marca el inicio fuerte de campaña). Todas las curvas reflejan una aceleración de discursos a partir de esta fecha.

- *29 de septiembre: Primer debate presidencial.* En este punto, se obseva que, Joe Biden sigue aumentando significativamente sus discursos mientras que Donald Trump presenta una ligera meseta lo cual podría mostrar una posible espera para evaluar la reacción pública y su desempeño en el debate y un probable repliegue y ajuste táctico para reorganizar el mensaje.

En particular, para cada candidato se destaca lo siguiente:

Joe Biden

La curva de Biden (azul) comienza en niveles bajos a fines de enero, e incremena moderadamente hasta julio, con una pendiente constante, lo que indica un ritmo estable de discursos. A partir de agosto la pendiente se eleva gradualmente y se dispara notablemente en septiembre, mostrando un fuerte aumento en la frecuencia de discursos hacia el inicio tradicional de la temporada electoral en Estados Unidos. Biden finaliza con el mayor acumulado de discursos, superando ampliamente al resto de candidatos.

Donald Trump

La curva de Trump (rojo) es casi horizontal desde febrero hasta mediados de agosto, indicando muy pocos discursos acumulados en ese período. A partir de principios de septiembre comienza un incremento abrupto en la pendiente, con una pronunciada subida continua durante septiembre y octubre. Trump intensificó drásticamente sus discursos en el último mes de campaña electoral, pareciendo tomar una estrategia de campaña distinta a la de Biden.

Bernie Sanders

La curva de Sanders (rosa) muestra un crecimiento fuerte en febrero y marzo, con una pendiente pronunciada al inicio de la gráfica, esto refleja su alta actividad durante las primarias demócratas

de 2020. Sanders tuvo un protagonismo discursivo superior al de Joe Biden en la primera parte del año, a pesar de esto, su curva se aplana drásticamente a partir de la primera semana de abril tras perder las primarias demócratas. Su curva se mantiene casi horizontal hasta finales de agosto donde se observa una ligera reactivación con un pequeño aumento, lo cual podría indicar alguna participación en discursos apoyando a los democrátas. Aún así, comparativamente, Sanders acumuló más discursos que Harris y Pence al final de la campaña.

Kamala Harris

La curva de Harris (naranja) permanece prácticamente horizontal desde enero hasta mediados de agosto, indicando muy escasa actividad en discursos acumulados. A finales de agosto / principios de septiembre la pendiente empieza a aumentar moderadamente como resultado de haber sido elegida como vicepresidenta en la formula presidencial de los democratas.

Mike Pence

La curva de Pence, candidato a la vicepresidencia republicano (verde) es similar a la de Harris: prácticamente plana hasta agosto, seguida de un pronunciado cambio a principios de septiembre. A partir de esa fecha, Pence incrementa su pendiente y acumula discursos de forma creciente durante septiembre y octubre.

Observaciones finales

Joe Biden (azul) no solo lidera en cantidad total, sino que nunca deja de subir. Su curva no presenta tramos planos, lo cual no sucede en los demás candidatos. Esto refuerza la idea de una estrategia basada en presencia continua y escalonada, posiblemente pensada para construir visibilidad sostenida, incluso antes de la campaña formal.

Donald Trump tiene una curva con un tramo casi plano por más de 6 meses (febrero-agosto), seguido de un cambio drástico, lo que evidencia una campaña más concentrada en la recta final, probablemente para generar impacto inmediato.

Finalmente, resulta interesante notar que, si se compara la pendiente de las curvas (cantidad de discursos por unidad de tiempo) entre fines de agosto y octubre se tiene que:

- la pendiente de los presidenciables es bastante similar entre sí,
- de igual forma la pendiente de los vice-presidenciables es bastante similar entre sí,
- en ambos casos, la pendiente de los presidenciables en mucho mayor que la de los vice-presidenciables (visualmente, casi el doble de discursos por unidad de tiempo).

1.3 PARTE C

Una de las funciones básicas que se desea realizar, es el conteo de palabras: cuántas veces aparece cada palabra agrupando por distintos criterios. Para ello, primero es necesario normalizar el texto (i.e: pasarlo todo a minúsculas) y eliminar los signos de puntuación. De no hacerlo, las secuencias "You", "you." y "you," se contarían como palabras distintas. La función clean_text(...) realiza parte de esta tarea, pero se debe completar agregando algunos signos de puntuación y cualquier otra normalización que considere oportuna. Comprobar el resultado observando el contenido de df_speeches_top_5, algunas celdas más abajo. Comente todas las transformaciones de texto que haya agregado y justifique.

1.3.1 Preparación del Texto

En esta etapa, el foco se centró en preparar los discursos para su análisis textual, asegurando que cada palabra, orador y estructura de texto estuvieran correctamente representados. Dado que los discursos contenían intervenciones combinadas de varios candidatos y múltiples inconsistencias en los nombres, el proceso se organizó en tres grandes pasos: **normalización de nombres**, **extracción de intervenciones por orador** y **limpieza estructural del texto**.

Normalización de nombres
Como se explicó anteriormente, se desarrolló la función normalize_names para unificar las diferentes formas en que los candidatos eran mencionados en el texto (por ejemplo, "President Donald J. Trump", "Trump", etc.). Esta función fue aplicada sobre la columna text para asegurar la correcta identificación de los cinco oradores principales y facilitar su posterior separación.

Separación de intervenciones por candidato/a
Un mismo discurso podía contener fragmentos correspondientes a varios oradores. Para evitar asignaciones incorrectas o dobles conteos, fue necesario aislar las intervenciones específicas de cada uno.
Trabajamos sobre el DataFrame filtrado previamente para discursos que involucraban a los cinco candidatos más activos (Top 5), y aplicamos una función personalizada extraer_intervenciones_por_orador, basada en una expresión regular que identifica bloques de texto en formato: ***Salto de línea + Nombre del orador + (mm:ss)***
Esto permitió separar los discursos en fragmentos y registrar, para cada fila del DataFrame, cinco nuevas columnas (una por orador), con el contenido específico dicho por cada uno.

Limpieza de encabezados y texto para análisis
Si bien los encabezados que indicaban el nombre y tiempo del orador eran útiles para segmentar el discurso, no aportaban información semántica y podían distorsionar el conteo de palabras al repetir nombres propios.
Para resolver esto, se aplicó la función limpiar_encabezados, que eliminó estos encabezados y los reemplazó por un separador neutro (//), manteniendo la estructura del discurso sin introducir ruido léxico.
El resultado de esta limpieza se almacenó en un nuevo DataFrame df_speeches_sin_encabezados, una versión depurada de los discursos, ya segmentada y estructurada por orador.

Consolidación del discurso por candidato/a
A partir de los textos individuales ya limpios y separados, se creó el DataFrame DISCURSOS_TOTALES, que consolida el contenido de cada candidato en una única celda. Este DataFrame tiene cinco filas (una por orador) y permite observar el discurso unificado de cada figura política.
Antes de continuar con los análisis, se aplicó la función clean_text, encargada de preparar el texto para un tratamiento lingüístico más riguroso. Esta función

- Convierte todo a minúsculas,
- Elimina tildes y caracteres especiales,
- Sustituye signos de puntuación por espacios para unificar formas (“hope,” → “hope”),
- Homogeniza los espacios en blanco.

A partir del texto limpio, se genera una nueva columna WordList, que contiene la lista de palabras tokenizadas (una por celda), base para todos los análisis posteriores.

El resultado de este proceso es una estructura limpia, normalizada y organizada por candidato, que constituye la base para los análisis de frecuencia de palabras, emociones, menciones cruzadas y estilo discursivo que se presentan en las siguientes secciones.

1.3.2 Conteo de Palabras según distintos criterio

En primer lugar cabe preguntarse que se entiende por “distinto criterio”.

Para responder esto se parte de la base de que el lenguaje en la política no es neutral: cada término seleccionado construye realidades, moviliza emociones y define agendas. Analizar palabras clave en discursos permite desentrañar no solo prioridades temáticas, sino también **estrategias retóricas**, **sesgos cognitivos** y **marcos ideológicos**.

Como ha sido mencionado, se partió de una estructura normalizada (DISCURSOS_TOTALES), donde cada registro representa un discurso unificado por candidato/a, con palabras tokenizadas y limpias. Esta estandarización garantiza comparabilidad al eliminar ruido (stopwords, caracteres especiales) y homogenizar formatos.

1.3.2.1 Criterios Utilizados

1.3.2.1.1 Palabras Positivas/Negativas

Procedimiento

Se considera interesante analizar una serie de 10 palabras positivas y negativas ya que:
- Las palabras positivas activan el sistema de recompensa cerebral, asociando al candidato con emociones gratificantes, términos como *"prosperity"* o *"opportunity"* proyectan una visión compartida de futuro, esencial en campañas y palabras como *"peace"* o *"justice"* suavizan polarización, atrayendo a votantes moderados.
- Las negativas operan estableciendo problemas urgentes que justifican la necesidad de liderazgo (ej.: *"poverty"* como llamado a acción), términos como *"chaos"* o *"messy"* deslegitiman adversarios o gobiernos anteriores, *"coronavirus"* o *"terror"* vinculan discursos a eventos traumáticos.

Las palabras seleccionadas fueron:

Positivas:	*Negativas:*
hope (esperanza)	**crisis** (crisis)
change (cambio)	**chaos** (caos)
unity (unidad)	**broken** (quebrado/roto)
freedom (libertad)	**dysfunctional** (disfuncional)
justice (justicia)	**corruption** (corrupción)
peace (paz)	**poverty** (pobreza)
opportunity (oportunidad)	**unemployment** (desempleo)
prosperity (prosperidad)	**coronavirus** (coronavirus)
kindness (bondad)	**terror** (terror)
love (amor)	**messy** (desordenado)

Para el filtrado y conteo se explotó la columna WordList (lista de palabras por discurso) mediante df_exploded para desagregar palabras en filas individuales. Se filtraron las palabras pertenecientes a cada lista (df_positivas, df_negativas). Mediante groupby se contaron las ocurrencias por palabra y candidato.

Resultados obtenidos

A continuación se muestran los gráficos obtenidos

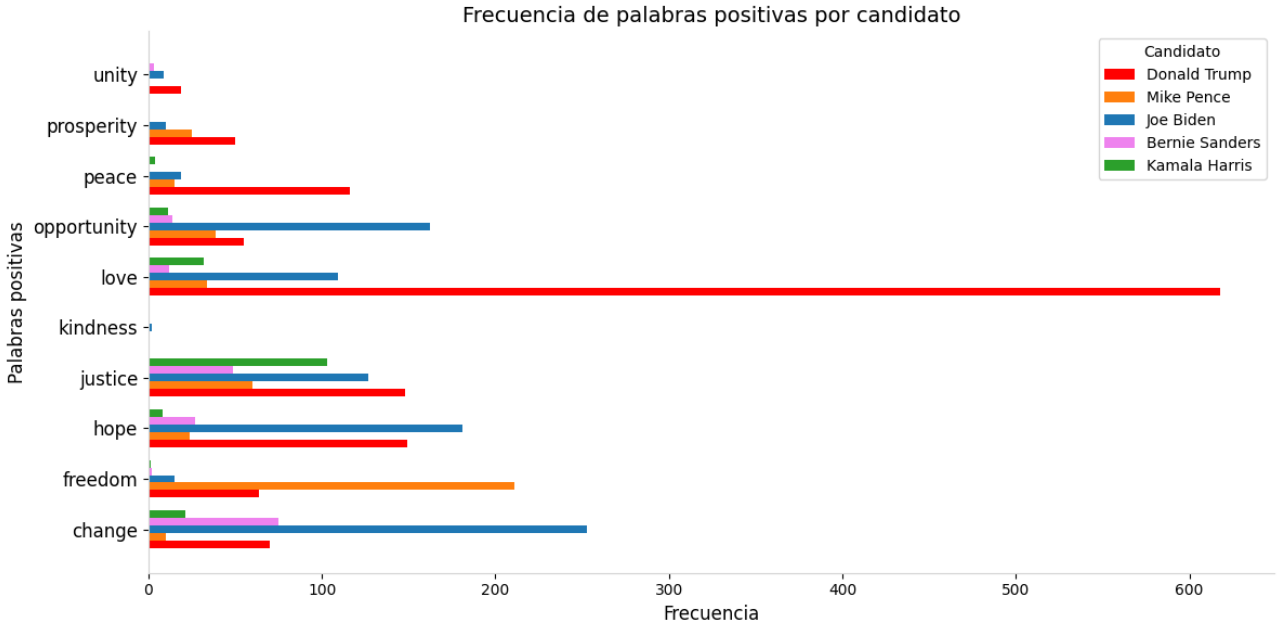

Figura 2

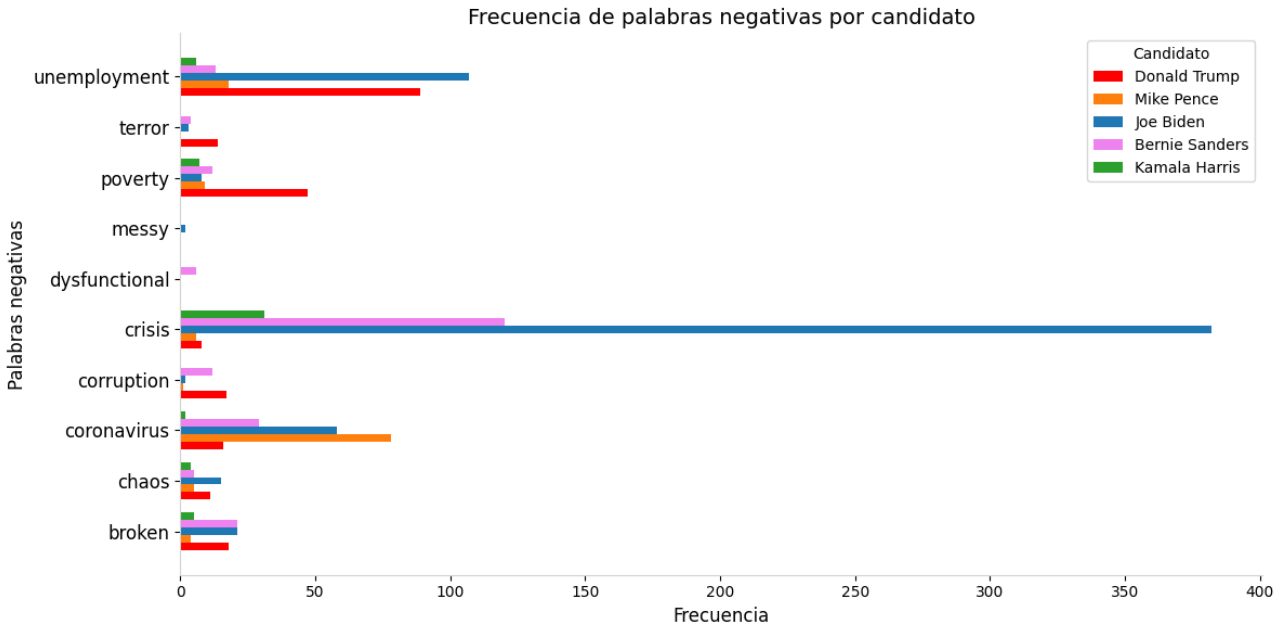

Figura 3

Análisis de resultados

Las palabras positivas alcanzan hasta **600** unidades, mientras que las negativas llegan a **400**. Esto sugiere que, en general, los candidatos utilizan más palabras positivas**,** lo que es coherente con la tendencia política de enfatizar soluciones, esperanza y unidad.

En relación a las palabras negativas destacan términos asociados a problemas estructurales, *"unemployment"* y *"poverty"*, y otros más coyunturales *"crisis" y "coronavirus".*

Por otro lado, si se visualiza particularmente las palabras más utilizadas de los presidenciables se observa que Biden hace un mayor uso de palabras negativas que Trump lo que podría explicarse por la posición de cada candidato (presidente vs. retador):
Trump como presidente que busca la reelección:

- podría evitar términos negativos como *"crisis"* para no asociarse con problemas
- en su lugar, usaba *"prosperity"* o *"victory"* para destacar logros económicos pre-pandemia.

Biden como retador:

- su estrategia podría depender en criticar la gestión de Trump, usando *"crisis"* para movilizar a votantes descontentos y *"change"* para incentivar al votante al cambio.

1.3.2.1.2 Diversidad Léxica

Procedimiento

El análisis de la diversidad léxica se realiza midiendo la proporción de palabras únicas/totales, esto es interesante pues es un indicador que ayuda a entender las estrategias comunicativas, el perfil ideológico y la efectividad retórica de los oradores políticos. Un léxico reducido (baja diversidad) facilita la memorización (consignas como "Make America Great Again"), mientras que alta diversidad sugiere expertise (ej.: Sanders al detallar políticas públicas).

Para realizar esto se realizó lo siguiente:

- Limpieza Adicional: Se excluyeron palabras con ≤3 caracteres para evitar monosílabos no significativos.
- Cálculo de Métricas: Para cada candidato, se calculó la proporción entre palabras únicas y total de palabras en sus discursos. Se utilizó un diccionario (lexico) para acumular conteos únicos y totales, derivando en un DataFrame (df_lexico) con la razón de diversidad.

Resultados obtenidos

A continuación se muestra el gráfico obtenido.

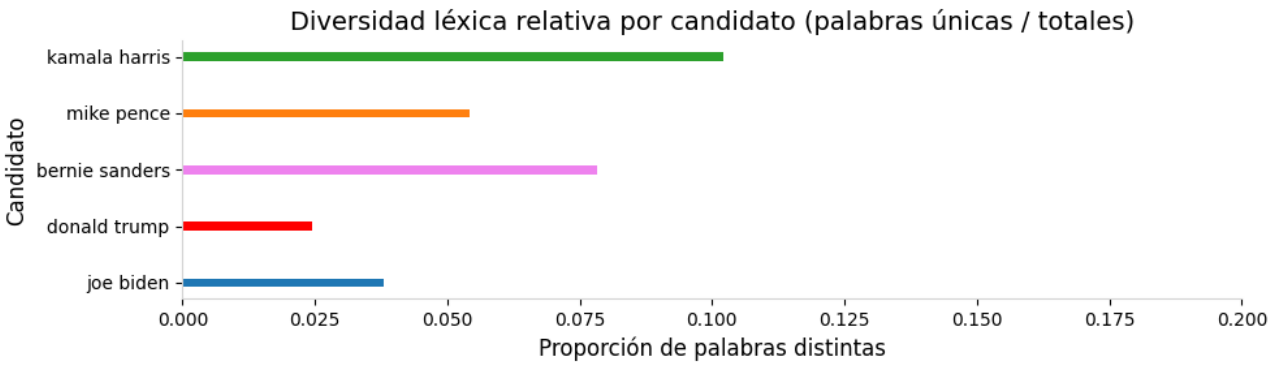

Figura 4

Se observa que Kamala Harris y Bernie Sanders son los dos oradores con mayor diversidad léxica mientras que los dos presidenciables son los dos oradores con menor diversidad léxica. Esto podría ser un reflejo de estrategias comunicativas

- **Baja diversidad léxica**: Sugiere repetición deliberada de eslóganes o palabras clave (ej.: "Make America Great Again" de Trump). Esto facilita la memorización y cohesión del mensaje, ideal para movilizar bases con mensajes claros.
- **Alta diversidad léxica**: Indica un discurso técnico o detallado (ej.: Bernie Sanders hablando de reformas estructurales), útil para proyectar autoridad y sofisticación ante audiencias especializadas.
- **Adaptación al público**: Candidatos que buscan atraer a votantes jóvenes o urbanos (como Kamala Harris) suelen emplear un léxico más diverso para abordar temas como justicia social o cambio climático. Líderes populistas (como Trump) priorizan un lenguaje sencillo y repetitivo para conectar con audiencias menos académicas.

Alta diversidad léxica puede generar percepción de competencia técnica, pero también de distancia emocional si el lenguaje es demasiado técnico, (por ejemplo: Hillary Clinton fue criticada por usar un léxico complejo, percibido como "elitista"). Baja diversidad léxica promueve identificación emocional (mensajes simples = más accesibles), pero arriesga parecer superficial o carente de profundidad.
En este caso, parece tener sentido que los presidenciables apelen a un lenguaje más sencillo para estar más cerca del pueblo y generar mayor identificacion emocional.

2 Parte 2: Conteo de Palabras y Visualizaciones

2.1 PARTE A

Realice una visualización que permita comparar las palabras más frecuentes de cada uno de los cinco candidatos/as. Sin necesidad de implementarlo, proponga ideas para modificar esta visualización con el fin de encontrar diferencias entre partidos políticos, fechas, o lugares.

2.1.1 Procedimiento

Una vez consolidado el texto limpio y normalizado de discursos totales por candidato (DISCURSOS_TOTALES), el siguiente paso consistió en analizar qué términos aparecen con mayor frecuencia en el lenguaje de cada candidato/a. Para esto, se utilizó la columna WordList, que contiene listas de palabras.

El objetivo fue construir una representación sintética del vocabulario predominante por orador. En lugar de observar cada discurso por separado, se unificaron todos los textos por candidato/a para formar un solo gran bloque por persona, y sobre ese conjunto se aplicó un conteo de palabras. Así, se identificaron las cinco palabras más repetidas por cada uno.

Los resultados se almacenaron en un nuevo DataFrame llamado df_frecuencias, que vincula cada palabra frecuente con su respectiva frecuencia y el candidato/a que la utilizó.

2.1.2 Visualización

Para interpretar estos resultados se construyó una visualización de barras. En ella, el eje X representa las palabras más frecuentes, el eje Y indica la cantidad de veces que fueron pronunciadas, y los colores diferencian a los candidatos.

Esta representación resulta útil no solo por su claridad visual, sino porque permite detectar de inmediato si ciertos términos son comunes a varios oradores o si, por el contrario, hay palabras exclusivas que caracterizan a un solo discurso político.

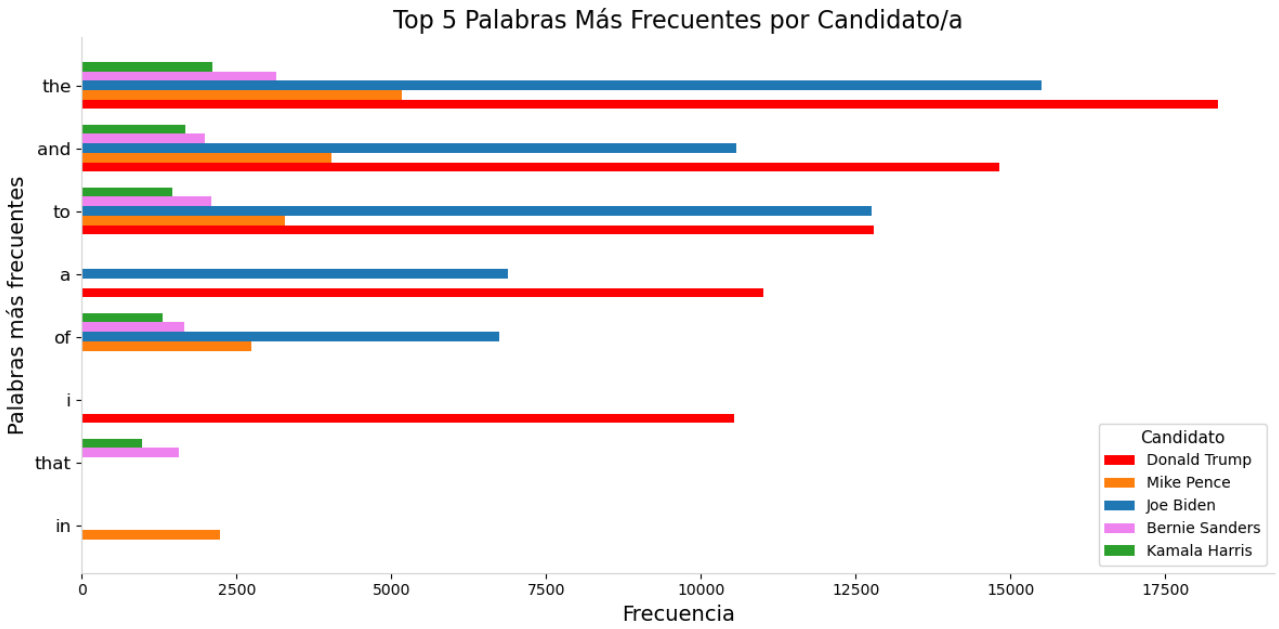

Figura 5

Complementariamente, se diseñó una segunda visualización donde cada candidato/a es representado en gráficos individuales. Esta alternativa permite observar en detalle las particularidades del lenguaje de cada uno, sin superposiciones ni solapamientos visuales.

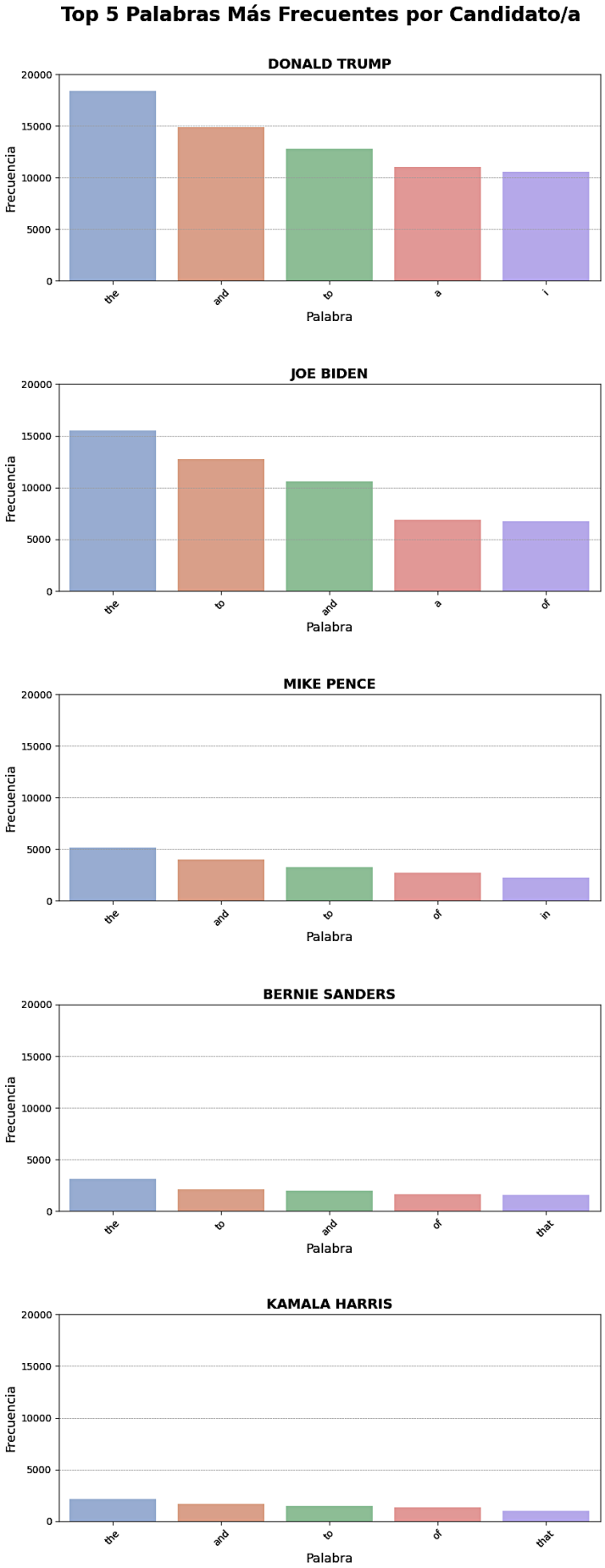
Top 5 Palabras Más Frecuentes por Candidato/a
DONALD TRUMP
Frecuencia
Palabra
the
and
to
a
i
JOE BIDEN
Frecuencia
Palabra
the
to
and
a
of
MIKE PENCE
Frecuencia
Palabra
the
and
to
of
in
BERNIE SANDERS
Frecuencia
Palabra
the
to
and
of
that
KAMALA HARRIS
Frecuencia
Palabra
the
and
to
of
that

Figura 6

2.1.3 Palabras más frecuentes

En general y en particular para cada candidato: ¿Encuentra algún problema en los resultados?

Los primeros gráficos construidos permitieron visualizar las palabras más utilizadas por cada uno de los cinco candidatos. Sin embargo, al analizar los resultados, surgió una limitación importante: las palabras más frecuentes eran, en su mayoría, palabras vacías del idioma (stopwords), como "the", "to", "and" o "I". Estas palabras, si bien forman parte esencial del lenguaje, no aportan contenido ni permiten diferenciar el estilo discursivo ni las prioridades temáticas de cada orador.

Este hallazgo evidenció la necesidad de refinar el análisis. Se decidió, entonces, repetir el procedimiento excluyendo las stopwords, lo que permitió filtrar el “ruido” lingüístico y enfocarse en términos más relevantes.

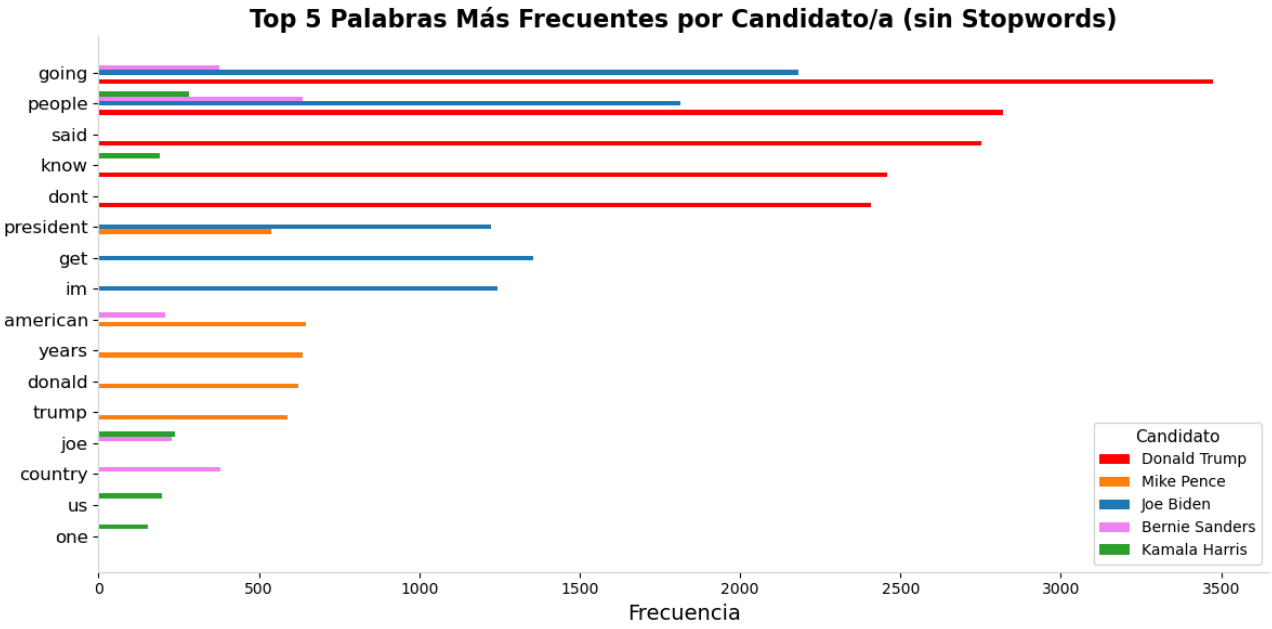

Figura 7

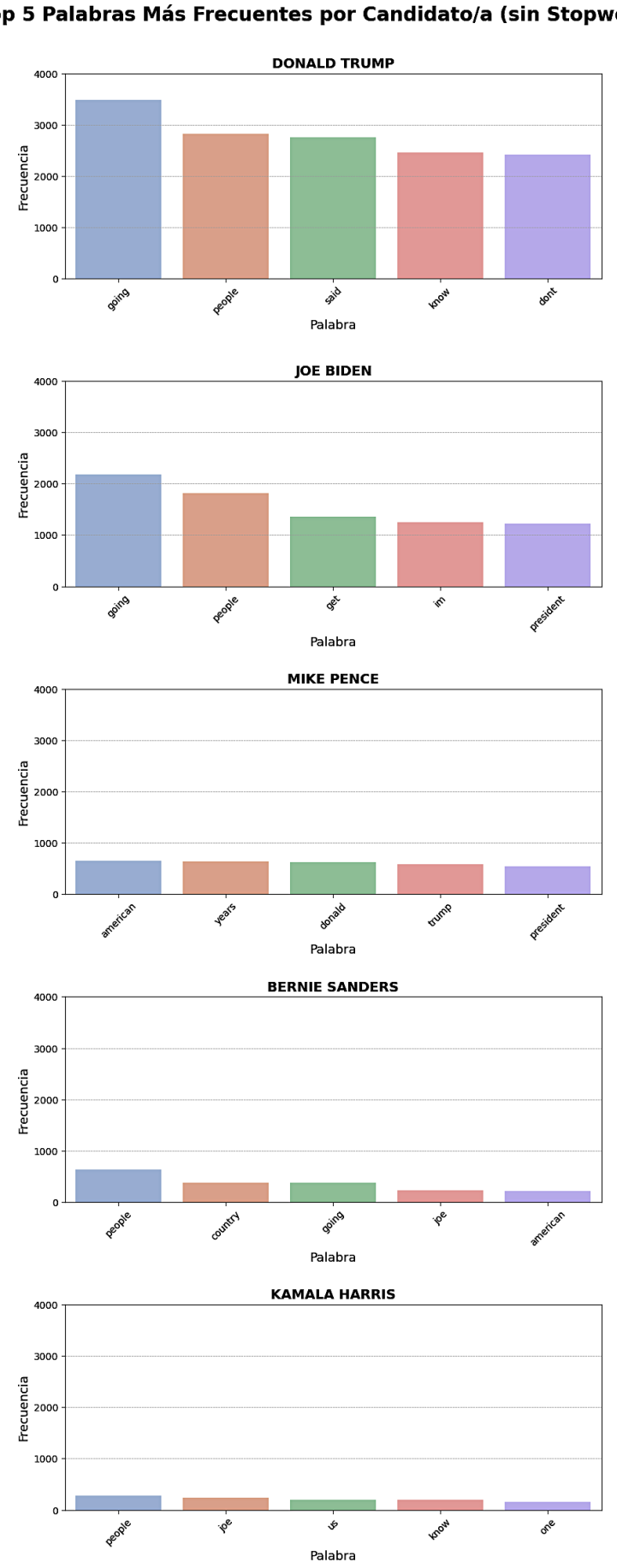
Top 5 Palabras Más Frecuentes por Candidato/a (sin Stopwords)
DONALD TRUMP
Frecuencia
Palabra
going
people
said
know
dont
JOE BIDEN
Frecuencia
Palabra
going
people
get
im
president
MIKE PENCE
Frecuencia
Palabra
american
years
donald
trump
president
BERNIE SANDERS
Frecuencia
Palabra
people
country
going
joe
american
KAMALA HARRIS
Frecuencia
Palabra
people
joe
us
know
one

Figura 8

2.1.4 Palabras más representativas
(sin stopwords)

Con el nuevo gráfico fue posible identificar con mayor precisión qué ideas, personajes o conceptos son centrales en la narrativa de cada candidato.

Algunas observaciones relevantes emergieron de esta visualización:

Referencias cruzadas: se evidencian menciones frecuentes entre candidatos, lo que sugiere una estrategia de confrontación o alusión directa. Por ejemplo, Kamala Harris y Bernie Sanders mencionan a Joe Biden, mientras que Mike Pence alude tanto a Donald Trump como a sus opositores.

Temas compartidos: términos como *"people"*, *"american"*, *"president"* o *"going"* se repiten entre varios candidatos, lo que puede indicar un consenso temático en torno a ciertos ejes de campaña o una coincidencia en el enfoque de los discursos.

Límites del enfoque: dado que se muestra únicamente el top 5 de palabras por candidato, este tipo de visualización deja fuera términos que, aunque importantes, aparecen con menor frecuencia. Esto podría llevar a perder matices discursivos que sí son relevantes en el análisis cualitativo.

2.1.5 Ideas para modificar esta visualización
Con el fin de encontrar diferencias por partido político, fecha o lugar

Una vez obtenidas las palabras más frecuentes por candidato, se abre la posibilidad de enriquecer el análisis agregando nuevas dimensiones que permitan comprender mejor cómo, cuándo y dónde se utilizan ciertos términos. En lugar de contar únicamente apariciones absolutas, una alternativa más informativa sería calcular **frecuencias relativas** como en el ejercicio anterior, de forma de comparar candidatos con diferentes volúmenes de discurso en forma más equitativa.

A continuación proponemos nuevas ideas para los tres ejes de análisis:

1. **Comparación por partido político**
   Analizar el lenguaje no solo por individuo, sino por **partido**, de forma de identificar patrones discursivos comunes dentro de cada coalición:
   - Usar colores diferenciados por partido ayudaría a resaltar bloques ideológicos en las visualizaciones.
   - Se podrían **promediar las frecuencias relativas** por partido, para detectar temas que actúan como "marco común" dentro de un grupo político.
   - También sería posible construir **nubes de palabras por partido**, en lugar de por candidato, para comparar las temáticas entre demócratas y republicanos.

2. **Evolución en el tiempo: eventos y coyuntura**
   Como se mostró anteriormente, la cantidad de discursos varió antes y después de ciertos eventos, lo cual abre la puerta a explorar si también hubo cambios en el tipo de lenguaje utilizado:
   - Comparar discursos en torno a **debates presidenciales**, anuncios importantes o resultados electorales.

- Identificar **cambios en el vocabulario dominante** a lo largo de la campaña: aparecen nuevos temas o desaparecen otros.
- Detectar **cambios estratégicos** en el mensaje de cada candidato ante nuevas circunstancias políticas.

3. **Adaptación al territorio: Evaluar si cambia el vocabulario utilizado según el lugar**

Integrar la **variable geográfica**:

- Utilizando la columna location, es posible agrupar discursos por región, estado o tipo de localidad (urbana, rural, virtual).
- Se podría investigar si los candidatos **ajustan su discurso** según la audiencia local. Por ejemplo, si aparecen más menciones a "empleo" o "industria" en zonas económicamente más complicadas.
- Esto permitiría explorar si existe una **estrategia discursiva adaptativa**, en lugar de un mensaje uniforme para todo el país.

2.2 PARTE B

Corra el código que permite encontrar los candidatos/as con mayor cantidad de palabras. En caso de encontrar algún problema luego de realizar la visualización, comente a qué se debe y proponga formas de resolverlo.

A partir del texto previamente limpiado, segmentado por orador y consolidado en el DataFrame DISCURSOS_TOTALES, se contabilizó el número total de palabras pronunciadas por cada uno de los cinco candidatos principales.

Dado que las magnitudes superaban las 100.000 palabras en algunos casos, se optó por representar la información con una escala ajustada (división por 10.000) y mediante un gráfico de barras horizontales, lo que permitió una comparación visual clara y ordenada.

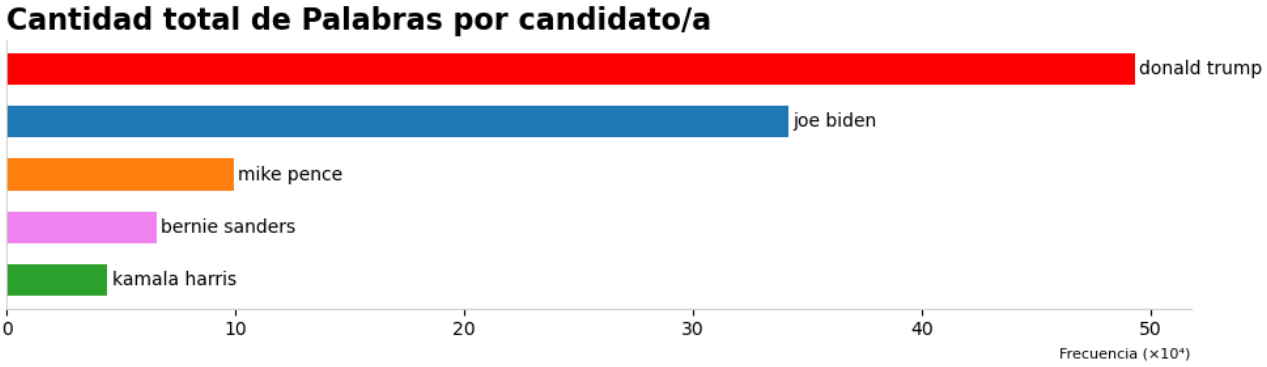

Figura 9

Conteo total de palabras por candidato/a

El resultado del análisis muestra que **Donald Trump** y **Joe Biden** lideran ampliamente en cantidad de palabras pronunciadas, seguidos por **Mike Pence**, **Bernie Sanders** y **Kamala Harris**. Este patrón refleja de forma coherente los niveles de exposición pública que se espera de los candidatos presidenciales frente a los vicepresidenciables o a quienes se retiraron anticipadamente.

Resulta especialmente llamativo que, si bien Joe Biden fue el candidato con más discursos, Donald Trump fue quien pronunció más palabras en total, lo que sugiere diferencias en el estilo discursivo y en la extensión promedio de sus intervenciones.

¿Se presentó algún problema en el conteo? ¿Cómo se evitó?
No detectamos problemas significativos durante el conteo de palabras, entendemos que fue resultado de el preprocesamiento del texto previo a realizar el conteo.
En particular:

1. **Segmentación de intervenciones**:
   Se extrajeron únicamente los fragmentos de texto pertenecientes a cada orador, lo cual fue crucial porque muchos discursos incluían a más de una persona o a fragmentos de oradores secundarios como por ejemplo "crowd". Sin esta separación, el total de palabras de un candidato podría haber incluido fragmentos de otro.
2. **Eliminación de encabezados**:
   Los encabezados del tipo "Trump: (00:30)" fueron eliminados para evitar que el nombre del orador se contabilizara como palabra cada vez que introducía una intervención.
3. **Normalización**:
   La normalización de nombres ("Trump" como "Donald Trump") no solo afecta al nombre del orador en el encabezado sino también al cuerpo del discurso, alterando la cantidad final de palabras por orador.
   Se aceptó este posible desvío como un costo menor frente al beneficio de obtener una base coherente y más comparable, y tener nombres normalizados para más adelante realizar el conteo de menciones cruzadas.

2.3 PARTE C

Construya una matriz de 5x5, donde cada fila y columna corresponden a un candiato/a, y la entrada (i,j) contiene la cantidad de veces que el candiato/a "i" menciona al candiato/a "j". Opcional: genere un grafo dirigido con esa matriz de adyacencia para visualizar las menciones.

Se construyó una matriz de 5x5 denominada mentions_matrix, donde:

- **Las filas representan al candidato que habla.**
- **Las columnas representan al candidato que es mencionado.**
- **Cada celda (i,j) contiene la cantidad de veces que el candidato/a i menciona explícitamente al candidato/a j en el conjunto total de sus discursos.**

Con esta matriz podemos estudiar las relaciones estratégicas dentro del juego político: quién ataca a quién, quién refuerza alianzas o quién evita referirse al otro.

El insumo utilizado fue el texto consolidado y normalizado por candidato (DISCURSOS_TOTALES), garantizando consistencia en la forma de los nombres y evitando errores de conteo.

¿Qué observamos en la matriz?

- **Trump y Biden** son los **más mencionados** por otros**, y se mencionan mutuamente con mucha frecuencia**, reflejando su protagonismo y su confrontación directa como principales rivales. La mención de Trump a Biden supera las mil apariciones.
- **Mike Pence**, a pesar de su menor cantidad de palabras, **hace numerosas referencias a Trump y Biden**, lo que se alinea con su rol de vicepresidente y figura de apoyo en la campaña republicana.
- **Kamala Harris y Bernie Sanders aparecen menos integrados en estas dinámicas de mención**, probablemente por haber tenido un protagonismo limitado en los tramos

centrales de la campaña o por enfocarse en discursos más temáticos y no tan confrontativos.

- **Joe Biden también menciona a Pence y a Harris**, probablemente para reforzar el contraste entre su fórmula y la fórmula opositora, destacando tanto a su compañera como a su principal rival, el binomio Trump/Pence.

	EL CANDIDATO DE LA PRIMERA COLUMNA MENCIONA A:				
	joe biden	**donald trump**	**bernie sanders**	**mike pence**	**kamala harris**
joe biden	-	623	6	0	13
donald trump	1071	-	51	38	12
bernie sanders	84	192	-	0	2
mike pence	477	620	8	-	60
kamala harris	111	110	0	4	-

Tabla 2

Visualización como grafo dirigido

Para facilitar la interpretación visual de estas relaciones, se construyó un **grafo dirigido**, donde:

- Cada nodo representa un candidato.
- Cada flecha indica que un candidato menciona a otro.
- El **color de las flechas distingue las interacciones entre distintos pares de nodos**
- **El grosor de las flechas (cantidad de menciones)**, permite identificar visualmente vínculos fuertes o más débiles o inexistentes.
- **El tamaño de nodo (menciones totales de todos los otros candidatos),** permite identificar de quien hablaron más el resto de los candidatos

Esta representación convierte la matriz numérica en un **mapa discursivo de las relaciones entre candidatos**. Permite ver con claridad quién posiciona a quién en el centro del debate político. Refuerza visualmente las apreciaciones realizadas respecto a la matriz.

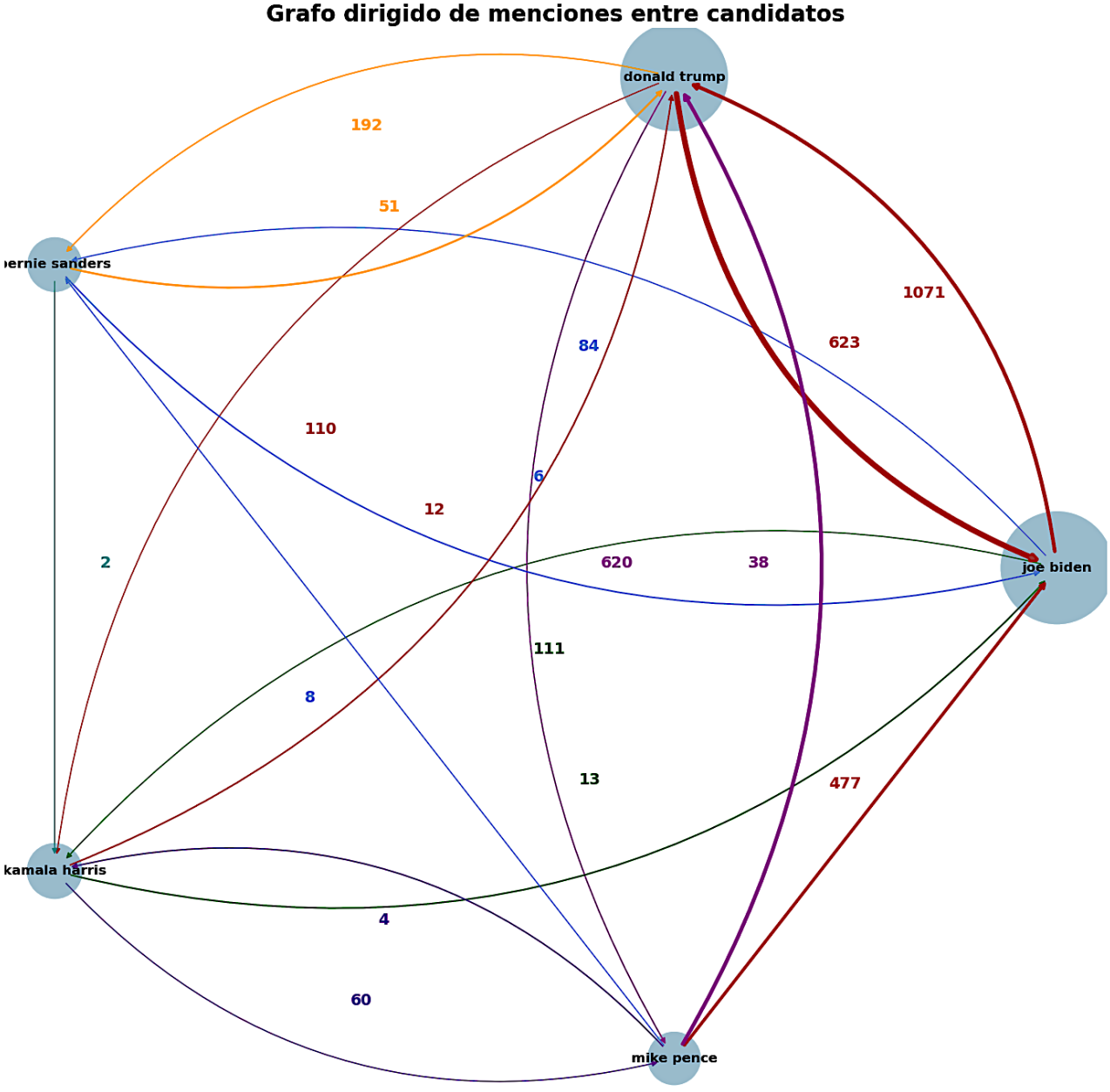
Grafo dirigido de menciones entre candidatos
donald trump
ernie sanders
joe biden
kamala harris
mike pence
192
51
1071
623
84
110
6
12
2
620
38
111
8
13
477
4
60

Figura 10

2.4 PARTE D

Proponga al menos tres preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas (sin implementar nada).

1. **Palabras diferentes según la ubicación geográfica**
   - Como mencionamos antes, usando la columna location, podríamos comparar si los candidatos adaptan su lenguaje según la audiencia.
   - Evaluar si usan más palabras frecuentes como "trabajo" o "familia" en ciertos lugares o si cambian el tono.
   - Esto ayuda a ver si realmente intentan conectar con el público o si repiten lo mismo en todos lados.

2. **¿Qué tan fácil o difícil es entender lo que dicen?**
   - Se puede medir cuán complejo es el lenguaje de cada candidato: evaluar si usan frases largas o cortas, también evaluar si usan palabras comunes o más técnicas.
   - Hay fórmulas para medir legibilidad, que sirven para saber si están hablando "para todos" o si su discurso es más académico o técnico.
   - Esto ayuda a entender a quién están tratando de llegar.

3. **Tipo de palabras usan**
   - Se puede analizar si cada candidato/a usa más verbos (acción), sustantivos (cosas, ideas) o adjetivos (descripciones).
   - Es una forma de ver si están "haciendo promesas", "describiendo problemas" o "contando historias".

4. **¿Qué emociones predominan en los discursos?**
   - Aplicar análisis de sentimientos (positivo, negativo, neutro).
   - Usar léxicos emocionales para identificar emociones específicas (miedo, alegría, enojo...).
   - Comparar perfiles emocionales entre discursos, etapas o candidatos.

3 Conclusiones

En resumen, este informe muestra cómo, partiendo de la limpieza y normalización de los datos, podemos extraer y comparar las estrategias de discurso de los candidatos. El análisis de discursos a lo largo del tiempo, los conteos de palabras y la diversidad léxica, entre otros, revelan diferencias claras en el tono y la intensidad del mensaje, mostrando estrategias de campaña distintas entre los candidatos. Estos resultados sientan las bases para futuras indagaciones, por ejemplo, el análisis emocional o en la evolución temática regional.