Tarea 1 - Introducción a la Ciencia de Datos

2025

## 

## Introducción

En las elecciones presidenciales de EE. UU. 2020, los discursos públicos de los candidatos reflejaron las prioridades y estrategias comunicativas en un entorno marcado por la pandemia y la polarización política. Este informe se centra en los cinco aspirantes con mayor número de intervenciones.

El análisis cuantitativo y cualitativo de esos discursos permite identificar patrones discursivos, variaciones temporales en el volumen de intervenciones y diferencias en el vocabulario utilizado por cada aspirante. Además, al centrarse en los cinco candidatos con mayor número de discursos, se obtiene una muestra representativa de las narrativas predominantes en los principales partidos políticos y en los eventos más relevantes de la campaña electoral.

Alguna preguntas clave que intenta responder éste análisis son:

**¿Cómo varía el volumen de discursos a lo largo de la campaña?  
¿Qué vocabulario caracteriza a cada candidato?  
¿Con qué frecuencia se mencionan unos a otros?**

## Datos y Metodología

* **Datos**: se utiliza una base de datos abierta con discursos políticos en el marco de las elecciones de Estados Unidos en 2020.
* **Carga y limpieza de datos**: comprobación de valores faltantes (df\_speeches.info()) conteo de discursos por candidato y filtrado de los cinco con más registros.
* **Análisis temporal**: series de tiempo del número de discursos, señalando picos en debates y convenciones.
* **Normalización de texto**: la función clean\_text() extiende el preprocesamiento (minúsculas, eliminación de puntuación, dígitos y tokens atípicos).
* **Conteo de palabras y frecuencias**: uso de Counter para extraer los top-30 términos por candidato y cálculo del total de palabras.
* **Análisis morfológico**: filtrado de **sustantivos** y **adjetivos** mediante pOs(spaCy) y WordNet, con visualizaciones de **word clouds** para cada categoría.
* **Matriz de menciones mutuas**: construcción de una matriz 5×5 que cuantifica cuántas veces cada candidato menciona a otro, visualizada también como grafo.
* **Diversidad léxica**: aunque no se calculó un ratio tipo-token explícito, el desglose de sustantivos y adjetivos ofrece una aproximación a la variedad léxica empleada.

Este conjunto de análisis ofrece la base necesaria para un informe centrado en los **hallazgos clave**, las **tendencias temporales** y las **dinámicas de vocabulario** entre los candidatos.

## Limpieza de datos

Realizamos la carga de los datos utilizando la librería pandas por medio del comando pd.read\_csv y comprobamos las características de la tabla con el comando df\_speeches.info(). Cuenta con 269 filas y seis columnas:

* *speaker*: exponente del discurso
* *title*: título del discurso
* *text*: texto del discurso
* *date*: fecha
* *location*: lugar
* *type*: tipo del dato

Todas las filas tienen valores en las columnas de *title*, *text* y *date*, pero hay 3 que tienen el *speaker* nulo, 18 filas con *location* nula y 21 con el *type* nulo. Con respecto a los valores nulos de estas dos últimas columnas, decidimos rellenarlos con un valor desconocido (‘unknown’), para que no nos genere ningún problema posterior. Por otro lado, todas las columnas son de tipo “object” y como necesitamos operar luego con la fecha de los discursos, transformamos la columna *date* a tipo fecha con el comando pd.to\_datetime.

Luego analizamos la cantidad de filas por *speaker*, con el fin de contar los discursos que tiene cada exponente. Encontramos que existen exponentes múltiples, por ejemplo, está el debate entre Joe Biden y Donald Trump en conjunto en una misma fila. Nos interesan también este y otros casos, es decir que queremos contabilizarlos como nuevos discursos y conocer qué se dice en ellos, por lo que no los desestimamos. Decidimos tratarlos como discursos nuevos y distintos entre sí, de manera que a este debate de Joe Biden y Donald Trump lo dividimos en dos discursos, dos nuevas filas, uno con lo que dice cada uno. Repetimos este procedimiento para los discursos múltiples que nos interesan puesto que tienen exponentes con posibilidades de ser candidatos con más discursos. Luego de este tratamiento de los datos, la tabla queda con 278 filas.

Los cinco exponentes con más discursos resultantes de esta base de datos trabajada son: los candidatos a presidente Joe Biden y Donald Trump, los candidatos a vicepresidente Mike Pence y Kamala Harris, y Bernie Sanders. Entendemos que los primeros cuatro califican como candidatos, ya sea como candidatos a presidente o como candidatos a vicepresidente. Bernie Sanders, según entendemos, por más que no fue candidato a presidente durante todo el período, sí lo fue en los primeros meses hasta que suspendió su campaña en abril de 2020 y es de nuestro interés analizar también los discursos de un candidato con estas características, considerando que tiene gran cantidad de discursos, incluso más que Kamala Harris en todo el período considerado.

## Resultados y Discusiones

### Discursos de los candidatos a lo largo del tiempo

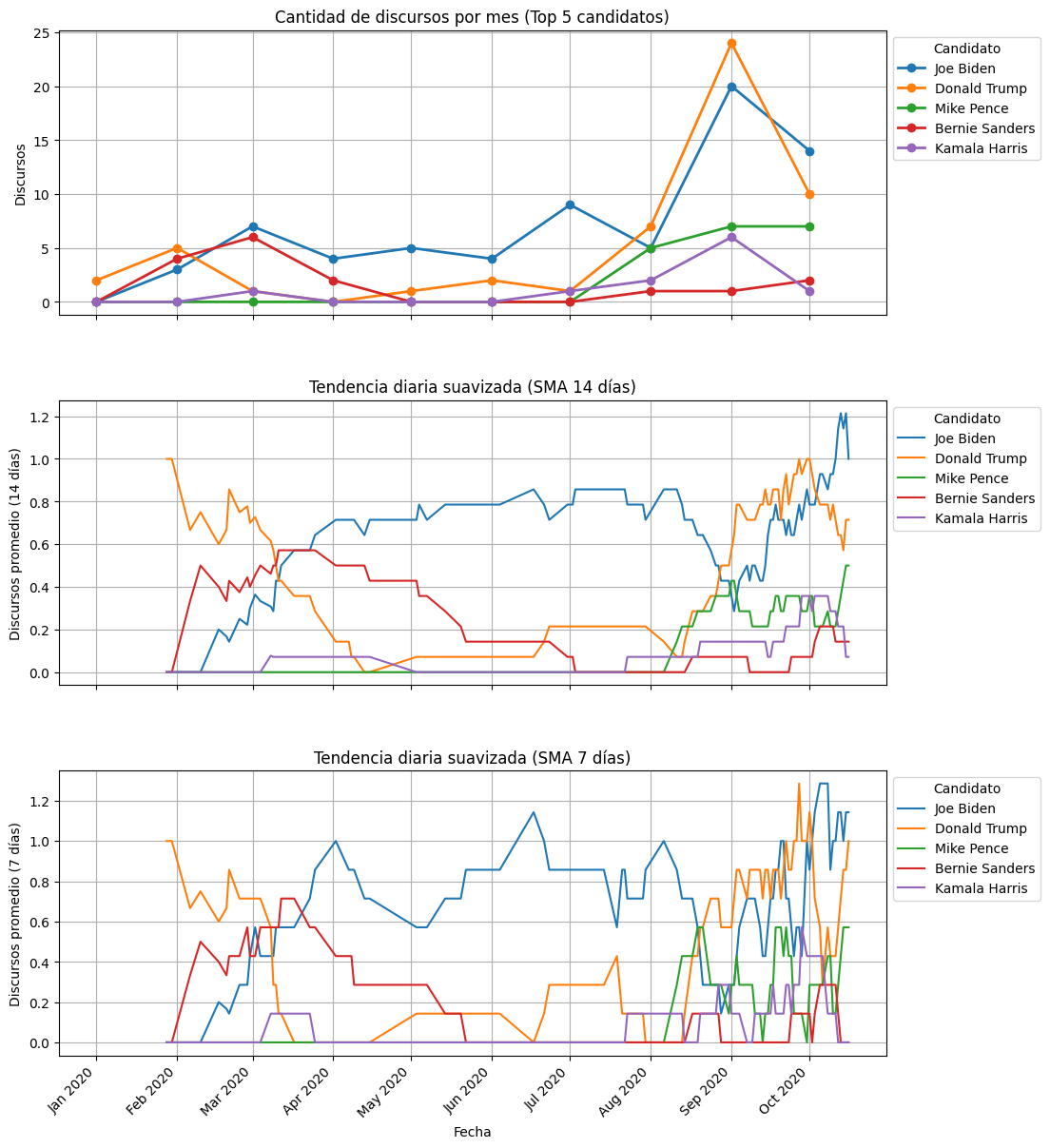
#### **Contexto teórico: Media Móvil Simple(SMA)**

Una **Media Móvil Simple** (Simple Moving Average **SMA** inglés) es un promedio aritmético que suaviza una serie temporal para revelar su tendencia subyacente. Por ejemplo, en vez de quedarnos con el valor de un solo día, calculamos el promedio de ese día y de los N días anteriores para cada punto en el tiempo. Así, cada nuevo valor de la serie resulta de condensar lo que ocurrió en ese pequeño tramo de días, atenuando los picos y valles aislados y mostrando con más claridad la tendencia general. Si en el arranque no hay suficientes días, basta con promediar los que sí estén disponibles para no dejar espacios vacíos. Finalmente, si quisiéramos centrar ese promedio, en lugar de mirar solo hacia atrás, podríamos repartir la ventana a ambos lados del día en cuestión, obteniendo una curva aún más “alineada” con los máximos y mínimos reales.

¿Para qué sirve?

* Reducir el ruido de corto plazo y destacar la dirección general.
* Comparar horizontes: **SMA** corta (7 días) muestra reacciones inmediatas; SMA más larga (14 días) revela la tendencia a mediano plazo.
* Alineación de picos: usando center=True alinear visualmente picos y valles, pero con desfase, mientras que con center=False mantienen fidelidad a la secuencia temporal real.

### Tendencias: Discursos



**Análisis mensual**

En la primera gráfica de esta figurase presenta la cantidad de discursos en cada mes. Vemos queJoe Biden comienza casi inactivo en enero–febrero, sube con fuerza en marzo (7 discursos) y mantiene un flujo moderado (4–5/mes) hasta julio, cuando da un salto a 9. El pico máximo es en septiembre (20) antes de bajar ligeramente en octubre (14).

Donald Trump tiene 2 discursos en enero, tiene un pequeño repunte en febrero (5), luego cae a 1–2/mes hasta agosto, y en septiembre alcanza su máximo (24) —el mes más “movido” globalmente— antes de descender a 10 en octubre.

Mike Pence prácticamente estuvo ausente hasta julio, luego se incorporó con 5 discursos en agosto, 7 en septiembre y 7 en octubre, mostrando su actividad creciente en el último tercio del período.

Bernie Sanders concentró casi toda su actividad en marzo (6) y abril (2), para casi desaparecer de mayo en adelante (1–2/mes o cero). Refleja que su campaña principal fue pre-primarias.

Kamala Harris tuvo una presencia prácticamente nula hasta agosto (2 discursos), luego un pico en septiembre (6) y se reduce en octubre (1). Muestra una entrada tardía, ligada a su anuncio de candidatura y debates de Vicepresidentes.

Lo más relevante es el mes de setiembre, recta final antes de las elecciones, donde Biden y Trump concentran sus máximos de cantidad de discursos.

Los candidatos con menor perfil (Sanders, Harris, Pence) tienen ratos muy focalizados: Sanders en marzo, Harris y Pence a partir de agosto.

**SMA 14 días**

Con una media 14 días eliminamos casi todo el ruido semanal, revelando tendencias a mediano plazo. Esto se presenta en la segunda gráfica de la figura.

Vemos que Biden ya estabiliza su promedio en ~0.7–0.8 discursos/día desde abril hasta poco antes de la campaña final, cuando sube hacia ~1.0.

Trump muestra claramente tres fases: alta actividad inicial (enero-febrero), declive prolongado y recuperación lenta desde julio hasta ~0.9 en septiembre.

Sanders queda casi “aplanado” después de marzo, y Harris y Pence sólo emergen con una rampa suave en el último tercio.

A grandes rasgos se elimina casi todo el ruido semanal, y se aprecian las tendencias de mediano plazo.

**SMA 7 días**

En la tercera gráfica de la figura se aprecia una rápida subida de Sanders en marzo, que alcanza ~0.7 discursos/día (equivalente a 5–6 en una semana) y luego cae.

Biden muestra un ascenso constante desde marzo hasta un pico de ~1.2 discursos/día en octubre.

Trump cae fuerte en abril–junio (~0.1–0.2) y luego sube de mayo en adelante, recuperando ~1.3 en septiembre.

Pence y Harris comienzan a marcar tendencia apenas entran en campaña: Harris ronda 0.3–0.5 en septiembre y Pence 0.5–0.6.

O sea se detectan subidas súbitas, ya que la ventana de 7 días conserva más variabilidad. A grandes rasgos se revela cómo reaccionan los candidatos a eventos concretos (debates, convenciones) con subidas y caídas que no se aprecian en el agregado mensual.

**Análisis integrado**

En septiembre, cuando el recuento mensual de discursos alcanza su punto más alto para Biden y Trump, vemos en la curva de 14 días que ese repunte no es un simple pico aislado sino una tendencia sostenida: la línea se inclina claramente hacia arriba desde julio, confirmando que ambos construyeron un gran impulso en esa fase de campaña. Al mismo tiempo, la versión de 7 días resalta las respuestas más inmediatas a eventos puntuales.

Por ejemplo, pequeños saltos en la actividad de Biden tras debates clave o fluctuaciones más abruptas de Trump alrededor de anuncios específicos que el promedio de dos semanas suaviza. Mientras Sanders dispara su actividad en marzo y luego cae bruscamente (algo que la SMA-14 atenúa pero la SMA-7 evidencia con claridad en sus subidas repentinas).

Harris y Pence apenas emergen en agosto, con una pendiente suave en la media de 14 días y picos semanales que delatan su entrada en escena. Así, al superponer recuento mensual y medias móviles de 14 y 7 días obtenemos un relato continuo: los hitos generales, la confirmación de su persistencia y la intensidad de las reacciones inmediatas.

### Conteo de palabras

#### Con el objetivo de realizar el análisis cuantitativo y cualitativo de los discursos, necesitamos hacer el conteo de palabras, es decir contar las veces que aparecen las distintas palabras en los discursos de cada candidato según distintos criterios. Para esto normalizamos el texto, pasando todas las palabras a minúsculas, y eliminamos los signos de puntuación (los más comunes incluido el punto y también el signo de exclamación, el de interrogación, los paréntesis, comillas y porcentaje). De esta manera nos aseguramos de que contamos la misma palabra las veces que sea así aparezca de diferentes formas. Un detalle a tener en cuenta es que a las palabras que son compuestas (con apóstrofe) las contamos como tales, por más que puedan escribirse como dos palabras distintas, ya que en el idioma inglés bajo consenso general de la lingüística pueden ser consideradas como palabras.

### 2.1 Cantidad de palabras por candidato

Con el fin de encontrar los candidatos con mayor cantidad de palabras, contamos la cantidad de palabras de cada uno, lo que se visualiza en la siguiente tabla:

| Candidato | Cantidad de palabras total |
| --- | --- |
| Donald Trump | 533.016 |
| Joe Biden | 414.607 |
| Mike Pence | 108.661 |
| Kamala Harris | 67.966 |
| Bernie Sanders | 55.539 |

### Vemos que el candidato con más palabras es Donald Trump, seguido por Joe Biden. Esto indica la cantidad de palabras totales en todos los discursos. Claramente era más probable que los candidatos con más discursos tuvieran más palabras pronunciadas, aunque en este caso particular se invierte el orden entre estos dos candidatos, indicando que Trump pronuncia más palabras por discurso. Si se quisiera saber esto, deberíamos realizar un promedio. Por otro lado, contamos las palabras cada vez que aparecen así estén repetidas, si se quisiera indagar en la calidad del vocabulario usado por cada candidato deberíamos contar la cantidad de palabras únicas.

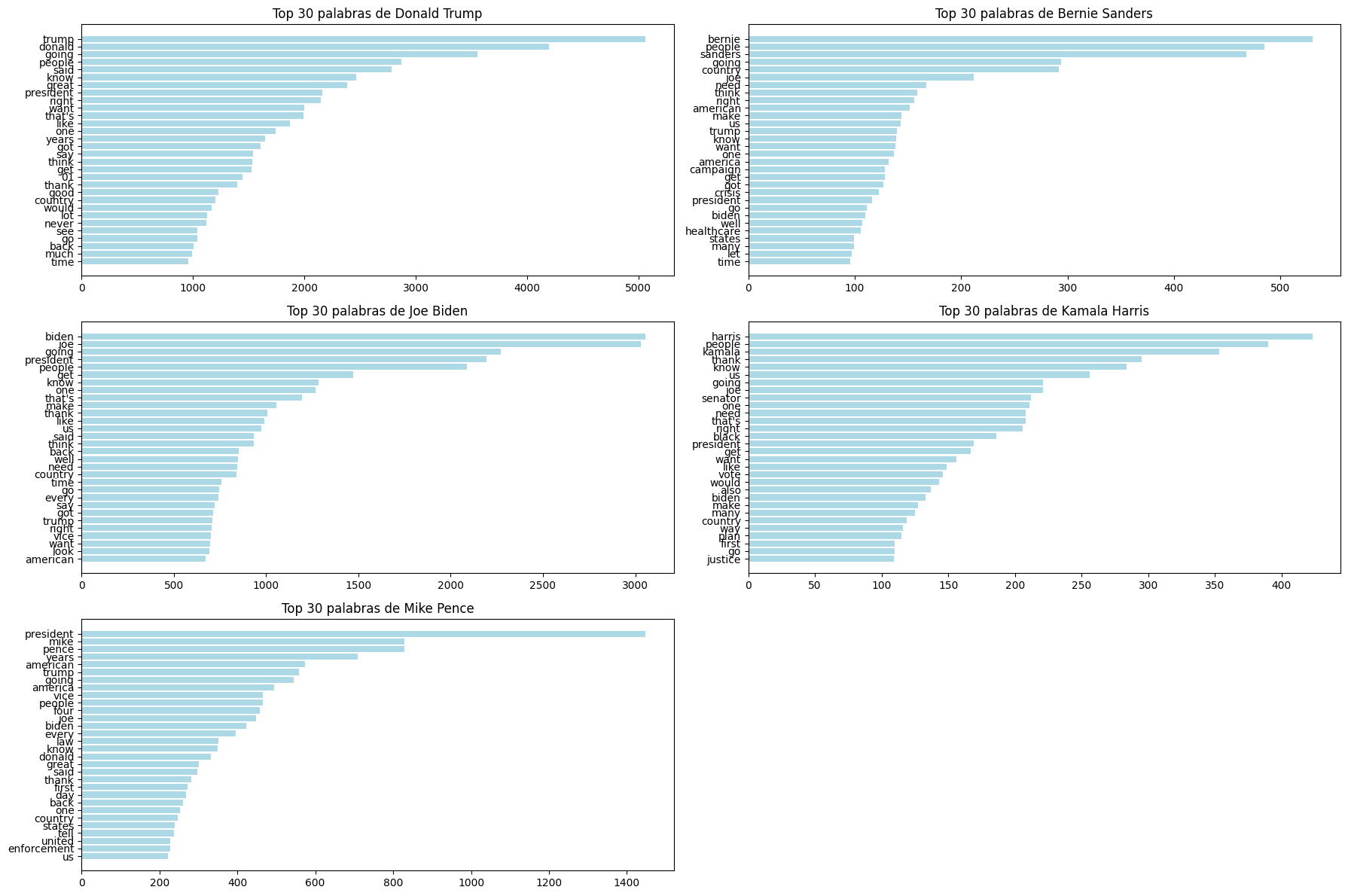
### 2.2 Palabras más frecuentes por candidato

##### **Barcharts**

Se hizo en primer lugar un conteo inicial de las 30 palabras más usadas por los candidatos. Al visualizar el Top-30 completo, mediante *barcharts*, vimos que las palabras más frecuentes usadas por Biden y Trump eran “the” (~17500 menciones) y “to” (~12000). Para los casos de Sanders, Pence y Harris, estas mismas palabras, apenas superan las 3 000–5 000. Por ello se procedió a la extracción de palabras con un contenido relevante. De asta forma aplicamos la lista de *stopwords* de NLTK, conjunto estándar de palabras vacías en inglés, y filtramos los artículos (“the”, “a”), preposiciones (“of”, “in”) y conjunciones (“and”, “but”) antes de contar.

El efecto fue inmediato como puede verse en la siguiente figura, en los *barcharts* filtrados, Biden pasa a tener en cabeza “biden” (~3100), “joe” (~3000) y “people” (~2200); Trump presenta “trump” (~5200), “donald” (~4000) y “said” (~2200).

En los demás candidatos se aprecian con más peso terminos como “campaign”, “president” o “senator” con recuentos de varios cientos.



#### Esta visualización podría modificarse con el fin de encontrar diferencias entre partidos políticos por ejemplo indicando con distintos colores las barras de los gráficos que corresponden a candidatos del partido demócrata (Joe Biden, Kamala Harris y Bernie Sanders) y las barras de los gráficos que corresponden a candidatos del partido republicano (Donald Trump y Mike Pence). También se podrían multiplicar estas gráficas por diferentes períodos, viendo por ejemplo las palabras más pronunciadas por cada candidato en el primer mes considerado y por otro lado en el último mes considerado. Por último se podría hacer un análisis discriminando los lugares donde se pronuncian los discursos, por ejemplo visualizar las palabras más frecuentes de cada candidato en discursos virtuales.

#### **Elección de Biblioteca: identificación y conteo de sustantivo, verbos y adjetivos**

En un primer acercamiento empleamos WordNet por su extenso inventario léxico, pero pronto comprobamos que ese filtro puramente estático solo verifica la existencia de una palabra sin analizar su uso real: WordNet te dirá que “record” puede ser sustantivo o verbo, pero no sabe en qué contexto aparece en tu discurso. En cambio, spaCy POS-tagging tokeniza cada texto, o sea, separa palabras y puntuación y asigna a cada token una categoría gramatical (sustantivo, verbo, adjetivo) basándose en el contexto, y por tanto, generando una desambiguación de los casos, como por ejemplo: “record the speech” vs. “set a record”.

Además, spaCy aplica lematización agrupando palabras relacionadas a su significado fundamental, por ejemplo agrupa “running” con “run”, “children” con “child”), y gracias a su implementación en Cython, que compila el código crítico a C, y a su arquitectura modular hacen que sea rápido y eficiente. Por estas razones, spaCy se convirtió en la base de nuestro análisis gramatical.

Extraer simultáneamente **sustantivos**, **verbos** y **adjetivos** nos brinda una visión muy completa de los discursos. Los sustantivos señalan entidades y conceptos clave; los verbos revelan acciones y compromisos (“prometer”, “construir”) y los adjetivos aportan matices valorativos (“importante”, “nuevo”).

##### **Wordclouds**

Para representar estas tres categorías, generamos word clouds que capturan tanto de qué hablan, así de cómo presentan sus principales ideas. Este enfoque filtra el ruido y destaca los términos de verdadero peso. De esta manera vemos en la siguiente figura que “country” y “people” están dentro de los sustantivos más utilizados por los cinco candidatos y además que “president” está entre los más utilizados de todos excepto en Sanders, lo que se puede explicar por la suspensión temprana de su candidatura. En cambio, menciona mucho las palabras “campaign”, “workers” y “healthcare”. En la nube de palabras de Trump, podemos ver las palabras “job”, “wall”, “hell”, etcétera, y en la nube de palabras de Biden vemos “crisis”, “vice”, “way”, etcétera.



### 2.3 Menciones mutuas

##### **Normalización de los nombres**

En los textos originales las referencias a un candidato aparecen de formas muy variadas (“Joe”, “Biden”, “President Biden”, etc.), lo que dificulta un conteo confiable. Para unificar todas esas variantes en una sola cadena consistente (“Joe Biden”, “Donald Trump”, etc.) aplicamos un proceso de tres pasos:

**Eliminar encabezados**(columna text): quitamos todo lo anterior al primer salto de línea para descartar timestamps o metadatos.

**Unificar en apellidos**: reemplazamos cada “Nombre Apellido” por su apellido (“Joe Biden” → “Biden”), asegurando que cualquier mención completa quede registrada como una única unidad.

**Reconstruir nombres completos**: transformamos cada apellido de vuelta a “Nombre Apellido” (“Biden” → “Joe Biden”), de modo que todas las menciones finales sean idénticas y contables sin ambigüedades.

##### **Matriz de menciones**

Con los nombres normalizados, construimos una **matriz de menciones** 5×5 —donde la fila i y la columna j indican cuántas veces el candidato i mencionó al candidato j, sumando para cada discurso, las ocurrencias exactas de “Nombre Apellido”.

A continuación se presenta dicha matriz, la cual revela quiénes se citan a sí mismos más veces y cuánto se citan entre pares de candidatos.

| menciona a: | **Joe Biden** | **Donald Trump** | **Mike Pence** | **Bernie Sanders** | **Kamala Harris** |
| --- | --- | --- | --- | --- | --- |
| **Joe Biden** | 3298 | 724 | 29 | 59 | 95 |
| **Donald Trump** | 1391 | 5551 | 256 | 244 | 107 |
| **Mike Pence** | 462 | 588 | 840 | 11 | 82 |
| **Bernie Sanders** | 241 | 144 | 0 | 530 | 1 |
| **Kamala Harris** | 281 | 122 | 8 | 5 | 438 |

##### A partir de esta matriz encontramos en primer lugar que existe una tendencia a autopromocionarse. En la diagonal se destacan los números de auto-mención de Trump y Biden, con valores mucho más grandes que los de sus compañeros de fórmula o rivales secundarios. Esto es en parte porque también son quienes tienen ayor cantidad de discursos, pero comparando con las menciones al resto de los candidatos vemos que ambos concentraron gran parte de su discurso en sí mismos. Mientras tanto, Pence, Harris y Sanders o bien se centraron más en hablar de otros o lo hicieron con menor volumen personal.

##### También en cuanto a los enfrentamientos principales, los valores más altos se encuentran entre Trump y Biden. Trump menciona a Biden una mayor cantidad de veces que a la inversa. Esto refleja que la campaña giró, sobre todo, alrededor de su rivalidad directa. En cambio, las referencias de los vicepresidentes a los candidatos del otro bando son mucho más moderadas.

##### 

##### **Posibles preguntas complementarias al análisis**

*¿Quiénes son los centros de la conversación tanto por la atención que concentran como por la que generan en el debate entre candidatos?*

Para responder a esta pregunta, imaginamos a cada candidato como un punto en un plano donde el eje X refleja cuántas veces él mismo alude a otros (usando el método G.out\_degree(weight='weight') de NetworkX) y el eje Y muestra cuántas veces los demás lo citan (con G.in\_degree(weight='weight')). Aquellos situados en la esquina superior derecha son los auténticos “centros de gravedad” del discurso, pues no solo lanzan el debate hacia sus rivales con frecuencia, sino que también concentran la atención del resto.

*¿Cómo se distribuyen geográficamente los discursos de campaña, y en qué estados se concentraron más intervenciones de los candidatos?*

Para responder a esta cuestión, se podría partir de la columna “location” de df\_speeches que contiene cadenas del tipo “estado, ciudad” en dónde se brindaron los discursos, y posiblemente, eliminando las versiones en cadenas de television. A partir de allí:

Se podrían extraer los valores, limpiando espacios y unificando mayúsculas, de modo que cada registro cuente con dos campos separados. A continuación, emplearíamos una librería como GeoPandas junto con Geopy para traducir esos nombres de lugares a coordenadas espaciales, de modo que cada discurso se convierta en un punto cartográfico. Con esa información, integraríamos los recuentos de intervenciones en un mapa base de los estados de EE. UU., usando herramientas de visualización como Folium o Plotly para colorear cada polígono estatal según la cantidad de eventos allí celebrados. El resultado es un mapa intuitivo que destaca de un vistazo los “puntos calientes” de la campaña.

*¿Qué cadenas televisivas cubrieron con mayor frecuencia los encuentros de cada candidato y, en conjunto, cuál fue la red que más los entrevistó?*

Antes de nada, revisaríamos todos los valores de la columna location para muestrear cada cadena mencionada y compararlos con una lista de referencia de emisoras de EE. UU. (ABC, NBC, CNN, CBS, Fox, etc.). De esta forma, podríamos detectar variantes (“ABC News”, “ABC-US”, “American Broadcasting Co.”) y asegurarnos de que ninguna se nos escape.

A continuación, unificaríamos cada una de esas variantes bajo el nombre oficial de la red (por ejemplo, todo “ABC News” o “ABC-US” pasaría a “ABC”), de modo que al agrupar realmente sumemos todas las entrevistas para cada estación. Luego, con esos valores ya limpios, agruparíamos por candidato y por cadena para contar cuántas veces cada aspirante participó en ABC, NBC, CNN, etc., y al mismo tiempo totalizaríamos las apariciones de cada red para ver cuál dominó la cobertura de debates y town halls. Finalmente, lo presentaríamos en un gráfico claro —barras apiladas o heatmap— que muestre de un vistazo qué cadena entrevistó más a cada candidato y cuál fue la más activa en toda la campaña.