

# Tarea 1

## Introducción a la Ciencia de Datos

Del Po, Diego  
O'Flaherty, Julian

Mayo 2025

### 1. Estructura de los datos

Para esta tarea se cuenta con datos tabulares en formato *CSV*. Este archivo colecta información sobre los discursos realizados por candidatos y otras figuras políticas durante las elecciones presidenciales del 2020 en Estados Unidos. Se tienen 269 filas y 6 columnas:

- **Speaker (Orador/es):** Quién dio el discurso. Contiene 3 datos faltantes.
- **Title (Título):** Título del discurso o la jornada donde se dio el discurso. No tiene datos faltantes.
- **Text (Texto):** Texto del discurso. Si hay múltiples oradores, al principio de cada frase se tiene quién la dice y una etiqueta temporal relativa al inicio del evento. No tiene datos faltantes.
- **Date (Fecha):** Día en el cual sucedió el discurso. No tiene datos faltantes.
- **Location (Ubicación):** Dónde se dio el discurso. Pueden ser ciudades, estados o medios de comunicación. Contiene 18 datos faltantes.
- **Type (Tipo):** Qué tipo de discurso fue. Contiene 21 datos faltantes.

Todas las variables se cargan a priori como tipo de dato *object* a modo de facilitar su lectura. Posteriormente, la única variable que sufre un cambio en su tipo de dato es la variable *date*, la cual es convertida al tipo de dato *datetime* con la finalidad de facilitar su manipulación cronológica. A modo de estructurar el informe, vamos a hacer un análisis por columna de datos.

## 2. Speaker (Orador/es)

La columna *Speaker* contiene en texto plano los oradores que estuvieron involucrados en los discursos. En total, sin ningún tipo de procesamiento, tenemos **71 oradores**. Sin embargo, al ver los valores, encontramos dos casos particulares: múltiples oradores y oradores faltantes.

En el caso de los oradores faltantes, podemos identificarlos por tener un valor *Nan*. Además de estas filas, hay un discurso que tiene un orador llamado **???**. Para los casos de múltiples oradores, hay dos casos. El primero y más claro, es el de "Multiple Speakers". Hay 5 instancias con este orador, donde 4 son del "*2020 Republican National Convention (RNC)*" y una del "*2020 Democratic National Convention (DNC)*". El segundo caso es el de "Multiple Speakersz" del cual tenemos 8 instancias. Procesando las transcripciones se pueden estimar<sup>1</sup> los oradores involucrados, que van entre 4 y 225 según el evento.

Es por esto que estos discursos no serán tomados en cuenta, por la cantidad de oradores y por la falta de claridad sobre a quién atribuirle la autoría. Por ejemplo, en *este fragmento de la DNC noche 4*, se utiliza un audio de Donald Trump en forma humorística, el cual es captado por la transcripción como un orador.

En el segundo caso de múltiples oradores, los mismos aparecen como una lista separada por comas. Con un procesamiento similar a la parte anterior, podemos aislar las partes del discurso dichas por cada orador. Luego del procesamiento, obtenemos una distribución de discursos como se observa en la figura 1, a continuación.

Como se podía esperar, dominan el panorama los dos candidatos de los partidos tradicionales, Joe Biden (Demócrata) y Donald Trump (Repúblicano), junto con sus respectivos vices, Kamala Harris y Mike Pence. Bernie Sanders, candidato demócrata, también tuvo una fuerte presencia en la campaña.

## 3. Date (Fecha)

La fecha es el día en el que cada discurso sucedió. En las figuras 2 y 3 se generan visualizaciones temporales de los discursos a través de las semanas, solamente para los cinco candidatos con más participaciones en discursos. Por su parte, la figura 2 muestra la distribución en el tiempo de los discursos para los cinco candidatos con más discursos en el período electoral, dónde se identifican a los

---

<sup>1</sup>La extracción no es perfecta, sino que se asume que cada intervención comienza con el nombre del orador y un *timestamp*. Puede que esta suposición no sea suficiente para extraer todos los involucrados, pero obtiene un estimado razonable.

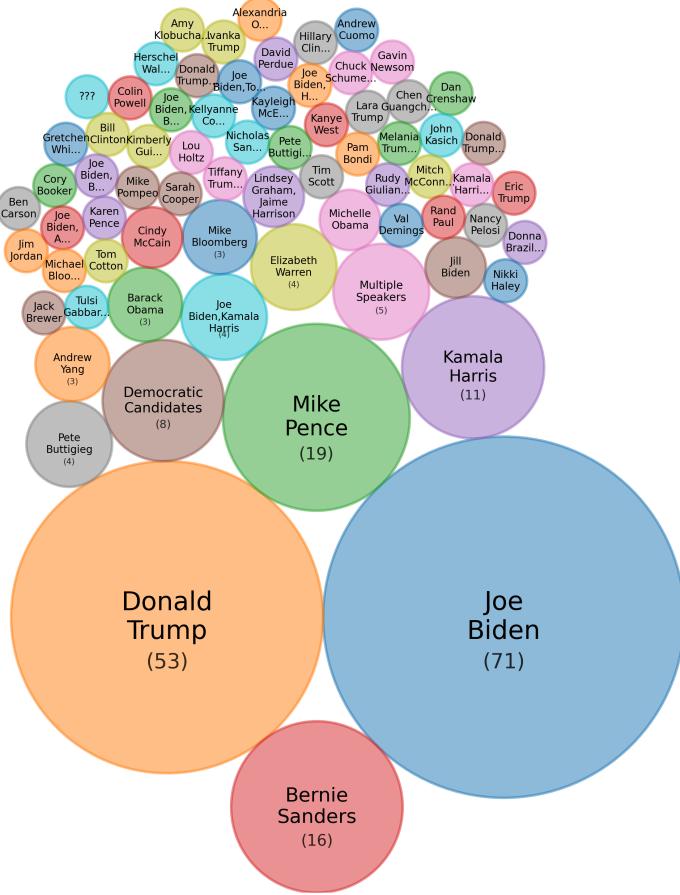


Figura 1: *Circle packing* plot donde el radio del círculo representa la cantidad de discursos en los que el *speaker* participó.

candidatos demócratas en tonalidades de azul y a los candidatos republicanos en tonalidades de rojo.

Uno de los primeros aspectos que se nos vienen a la vista es como el Partido Demócrata fue más constante en los discursos que el Partido Republicano. Si bien parecen tener una intensidad similar a lo largo de la campaña (aumenta sobre el final), el partido liderado por el candidato Joe Biden realizó al menos un discurso por semana en el período de tiempo que cuenta nuestro set de datos.

Otro detalle importante a observar es la baja de discursos luego de Marzo, que coincide con el arranque de la pandemia provocada por el virus SARS-CoV-2, donde se pusieron medidas de cuarentena, evitando que sucedan eventos

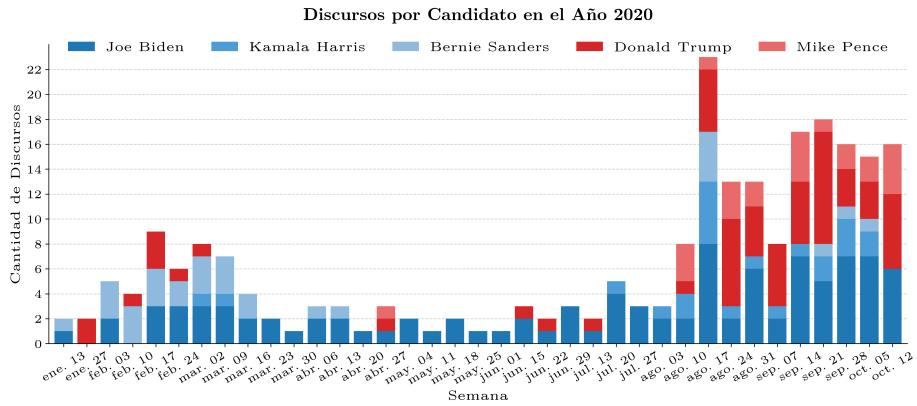


Figura 2: Discursos por semana del año 2020 dados por los cinco candidatos con más discursos.

multitudinarios. Más adelante haremos un análisis de la locación de los discursos, donde esto también será relevante.

Como mencionamos anteriormente, es notorio como a medida que se acercó el momento de las urnas, la cantidad de discursos por partido se hizo más elevada. Como suele suceder en muchas partes del mundo, los distintos movimientos políticos intensifican su campaña en el cierre para poder consolidar a su electorado y captar a la mayor cantidad de indecisos posible. Algo interesante es que este aumento en la intensidad de la campaña se dio tres meses antes del momento de la elección, lo cual podría indicar que debido a la pandemia los partidos políticos tenían incertidumbre sobre qué proporción del electorado iba a votar de forma anticipada, por lo que las acciones para captar votantes comenzaron con algo de anticipación. Otra cuestión, podría decirse, es que los partidos políticos necesitaban promover el voto de las personas, dado que en Estados Unidos el voto no es obligatorio, y estas eran unas elecciones donde la definición de quién iba a gobernar el país iba a ser por un acotado margen.

Si miramos a nivel individual, vemos como Bernie Sanders tiene apariciones al principio de la campaña, cuando competía con Joe Biden por la interna demócrata, para luego parar y re-aparecer sobre el final de la campaña para apoyar a Biden. No es casualidad que Sanders sea uno de los candidatos con más discursos ya que él contaba con una sólida base de votantes que el Partido Demócrata no podía perderse de cara a la competencia con Donald Trump.

En el caso de Mike Pence, observamos que comienza a hablar públicamente sobre el final de la campaña. Probablemente la designación de Pence como presidente de la Comisión Especial de la Casa Blanca sobre el Coronavirus, le quitó presencia en los discursos durante una buena parte del año.

Kamala Harris, por su parte, tiene presencias más acotadas en cuanto a discursos se refiere con una concentración de estos en los meses de Agosto y Setiembre. Sin embargo, su rol fue no menor en las elecciones ya que ella representaba al futuro generacional del partido y reforzó temas como justicia racial, salud pública y unidad nacional.

Como era de esperarse, los candidatos a la presidencia Joe Biden y Donald Trump son las personas con más discursos a lo largo de la semana. El primero fue el candidato más constante del año, donde realizó al menos un discurso por semana. Trump por su parte, tuvo una participación más leve al inicio y la intensificó en los últimos tres meses anteriores a las elecciones, lo cual es coherente considerando que en ese momento ocupaba el cargo de presidente.

En la figura 3, que agrega a la gráfica anterior los datos de los demás oradores en la campaña, aparece un pico de discursos en Agosto. En esta fechas, sucedieron tanto la *Democratic National Convention* (del 17 al 20 de Agosto), como la *Republican National Convention* (del 24 al 27 de Agosto). Estos eventos, originalmente planificados para Julio pero aplazados por la pandemia, tienen como objetivo la elección del candidato a presidente y vicepresidente. Durante los mismos, exponentes del partido ofrecen discursos y participan de debates informativos, resultando en el pico que observamos en la gráfica.

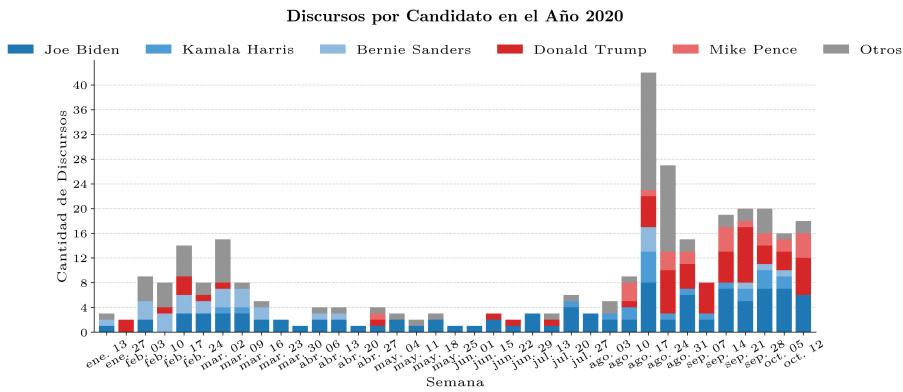


Figura 3: Discursos por semana del año 2020 dados por los cinco candidatos con más discursos, más la categoría Otros.

#### 4. Text (Texto)

En la columna *text* nos encontramos con la transcripción de los discursos. Lo que resulta importante es que a pesar de que cada fila está identificada con un orador, en general, en la transcripción notamos que varias personas participan

de ese mismo discurso. Por esta razón, se intentó separar estas intervenciones de distintos oradores con la ayuda de expresiones regulares de forma de saber con la mayor exactitud posible quién dijo qué. Al separar cada intervención, nuestro set de datos ahora pasa a tener una fila por intervención (en vez de una fila por discurso), identificando en *speaker* quién fue la persona que dijo ese fragmento del discurso.

El procesamiento anterior nos permite analizar de forma más eficiente los discursos de los cinco candidatos. En concreto, se realizó una nube de palabras para cada uno de los cinco candidatos a modo de analizar cuáles fueron las 100 palabras más mencionadas por éstos. Antes de realizar la visualización, fue necesario normalizar el texto con la finalidad de que una misma palabra escrita de formas distintas no aparezca más de una vez.

Para lo anterior fue necesario realizar las siguientes transformaciones sobre la columna *text*:

- Se eliminó todo lo que está entre paréntesis rectos, como por ejemplo: [crosstalk...], [inaudible...], etc.
- Se eliminó el texto específico *Commercial: (48:14)* ya que rompía la separación de las intervenciones dentro del discurso *Donald Trump ABC News Town Hall Transcript with George Stephanopoulos in Philadelphia*.
- Se eliminaron los patrones : (*mm:ss*) y : (*hh:mm:ss*).
- Se separaron con expresiones regulares las intervenciones dentro de un mismo discurso, generando como resultado una lista de tuplas donde el primer elemento de la tupla es el orador y el segundo lo que dijo.
- Cada par orador-discurso se convirtió en una fila del set de datos, utilizando el primer elemento de la tupla como *speaker* y el segundo como *text*.
- Se homogeneizaron los nombres puesto que por ejemplo Donald Trump aparecía como: President Trump, President Donald J. Trump, President Donald Trump, etc.
- Se crea la función `search_punctuation`, para encontrar todos los signos de puntuación presentes en los discursos.
- Se utiliza la función `clean_text` para normalizar el texto, agregando los signos de puntuación obtenidos en el punto anterior.
- Se convierte cada discurso en una lista de palabras.

En la figura 4, tenemos una nube por candidato con las 100 palabras más mencionadas por éstos y de ellas se pueden notar algunos perfiles en sus discursos.

Por ejemplo, se nota como Donald Trump hace muchas veces mención a la palabra *right*, así como también hace muchas veces mención de la palabra *great* (parte de su lema *Make America Great Again*), reflejan una visión conservadora y nacionalista centrada en la restauración de valores tradicionales, el fortalecimiento de la soberanía nacional y una crítica al declive percibido del país en comparación a una grandeza pasada del mismo. Por su parte, el candidato republicano a la vicepresidencia hace muchas veces mención a los candidatos a la presidencia, probablemente en forma de comparación y en pos de resaltar los logros del gobierno. También se ve su enfoque relacionado a una fuerte seguridad nacional ya que muchas veces menciona a las fuerzas de seguridad del país (*law enforcement*).

Por su parte, los candidatos demócratas parecen hacer mención a la necesidad de un cambio con la palabra *now* que aparece mucho en los discursos de Biden y Sanders. Particularmente, Biden nombra mucho la palabra *president* lo cual podía referir a críticas al actual gobierno encabezado por Trump. Harris y Sanders nombran mucho a Biden en sus discursos seguramente en forma de respaldo al candidato demócrata y en pos de incitar el voto de las personas al mismo, viéndose esto último reflejado en la aparición de la palabra *"vote"* entre las más frecuentes. También en los discursos de estos dos candidatos se hace mención a la pandemia del COVID-19, probablemente cuestionando las acciones del gobierno y proponiendo otro enfoque de acción ante la crisis sanitaria.



Figura 4: Las 100 palabras más dichas por los candidatos en todos sus discursos.

Para hacer un análisis más granular de la campaña se pueden hacer más nubes de palabras con distintos filtros. Por ejemplo, se puede hacer una nube de palabras por mes, buscando encontrar los tópicos de interés en cada momento de la campaña. Se puede hacer un análisis en base a la localización, la cual estudiaremos en la siguiente sección, dividiendo por ejemplo entre las instancias virtuales y las presenciales. Hacer un análisis geográfico a nivel estatal no sería particularmente útil, dado que no se cuenta con volumen de datos en cada estado para sacar conclusiones.

#### 4.1. Menciones cruzadas entre candidatos

Además de la frecuencia de las palabras, es de particular interés hacer un análisis de las menciones que hace un candidato de otro. Para este análisis, se parte de los textos anteriormente procesados, y se buscan menciones a los apellidos de los candidatos. Esto último introduce la hipótesis de que cuando se usa el apellido refiere al candidato, lo cual no es cierto estrictamente, por ejemplo Ivanka Trump, pero argumentamos que una mención a familiares directos de un candidato tiene como objetivo comentar sobre el candidato.

Grafo Dirigido de Menciones entre Candidatos

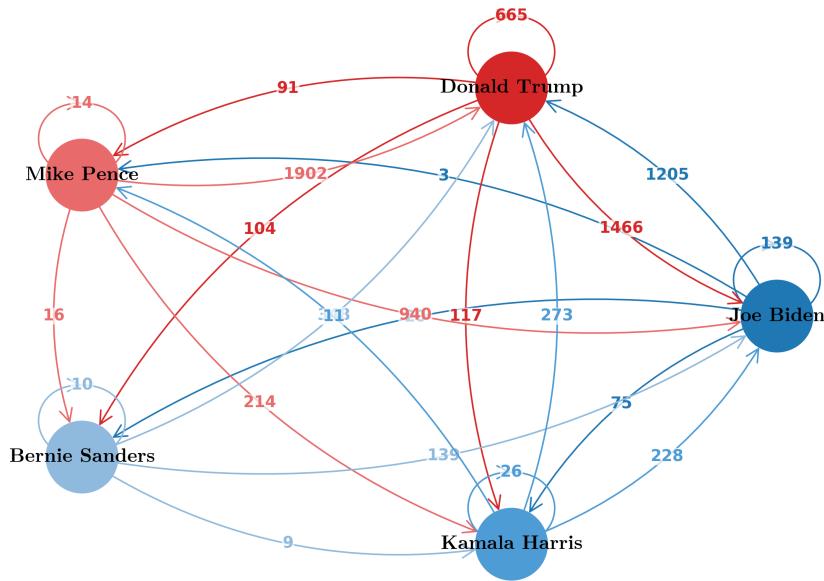


Figura 5: Grafo de menciones entre candidatos. El grafo es dirigido, con la cantidad de menciones incluida en la flecha.

En la figura 5 vemos el grafo de menciones resultante. Lo primero a notar, es que los candidatos más referenciados son Trump y Biden, candidatos a presidencia de cada partido. Algo interesante que surge de este gráfico, es que los republicanos tienen un discurso mucho más personal, haciendo muchas más menciones a otros candidatos comparando con sus contrapartes demócratas.

## 5. Location (Ubicación)

La columna de *location* contiene la ubicación donde sucedió el discurso. Como ya se ha mencionado, la campaña se dio en el marco del COVID-19, lo cuál puso una limitante en la realización de eventos multitudinarios. Es por esto que esta elección se caracterizó por tener una gran cantidad de eventos virtuales.

Los datos de la columna son texto libre, por lo que es necesario tratarlos para poder utilizarlo. Lo primero es tratar de dividir entre los discursos virtuales y los presenciales. Se observa que los discursos que fueron presenciales tiene el formato *Ciudad, Estado* o simplemente *Estado*, lo que nos permite asignarles uno de los 50 estados. Todas aquellas ubicaciones que no pudieron ser asignadas, serán tratadas como virtuales<sup>2</sup>. Luego de este procesado, para los 5 candidatos con más apariciones obtenemos **46 apariciones virtuales y 132 apariciones presenciales.**, con 10 discursos sin localizar. Estos 9 de los 10 discursos tiene *NaN* como valor para la localización, y el restante sucede en la Casa Blanca, siendo ambiguo si fue presencial o virtual.

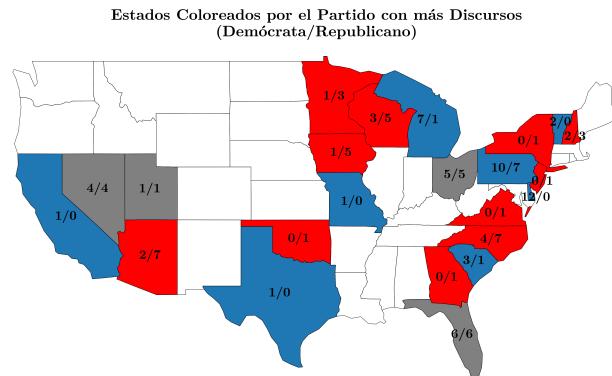


Figura 6: Mapa con los discursos por estado de los dos partidos principales. El color del estado indica cual fue el partido con más presencia, estando en gris cuando tuvieron el mismo número de apariciones.

Comencemos analizando las apariciones públicas. En la figura 6 se observa un mapa de los Estados Unidos con los estados resaltados en base a las apariciones presenciales. El color del estado es acorde al partido con más presencia, pintado de gris cuando hay un empate y en blanco cuando no hay apariciones registradas. Lo primero que se observa es que las actividades presenciales fueron pocas debido

<sup>2</sup>Si se usa toda la data, existen algunos casos donde solo aparece la ciudad. Para simplificar el procesado y las gráficas posteriores, analizamos las apariciones de los 5 candidatos con más apariciones.

a la crisis del COVID-19, con poco menos de la mitad de los estados sin actividad de los principales candidatos durante la campaña.

Según [1], la elección del 2020 se definiría por los estados Florida, Pennsylvania, Michigan, North Carolina, Arizona y Wisconsin. Viendo el mapa de apariciones, se ve un claro foco de los candidatos por hacer campaña en esos estados, con Florida con 6 discursos cada uno y Pennsylvania con 10 discursos demócratas y 7 republicanos.

Viendo las apariciones virtuales, la gran mayoría de instancias están bajo la categoría "Virtual", sin detalles de la instancia. Además de esas instancias, tenemos un número de apariciones en canales de comunicación tradicionales, específicamente en *CNN*, *ABC*, *NBC* y *Fox News*.

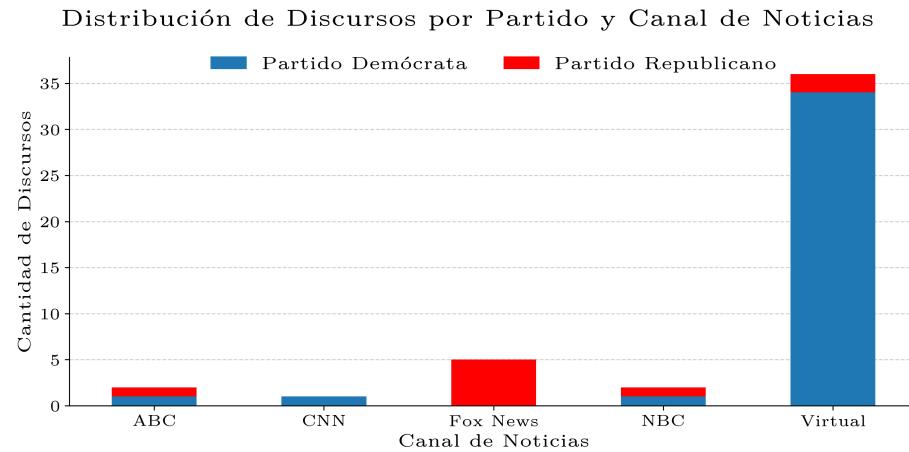


Figura 7: Distribución de apariciones virtuales entre canales de televisión y otros medios.

En la figura 7 vemos la distribución entre canales. Una primera conclusión rápida es el gran uso que se hizo, especialmente por parte de los demócratas, de medios de comunicación no tradicionales para dar discursos, con más de 30 instancias de este tipo por parte de sus candidatos. Respecto a los medios tradicionales, se pueden observar sesgos de los que usualmente se los acusa, con *CNN* teniendo solo apariciones demócratas y *Fox News*, aunque la cantidad de apariciones no permite sacar conclusiones fuertes.

## Referencias

- [1] Patsy Widakuswara. These us 'swing' states may decide 2020 election, October 2020. Accessed: 2025-05-11.