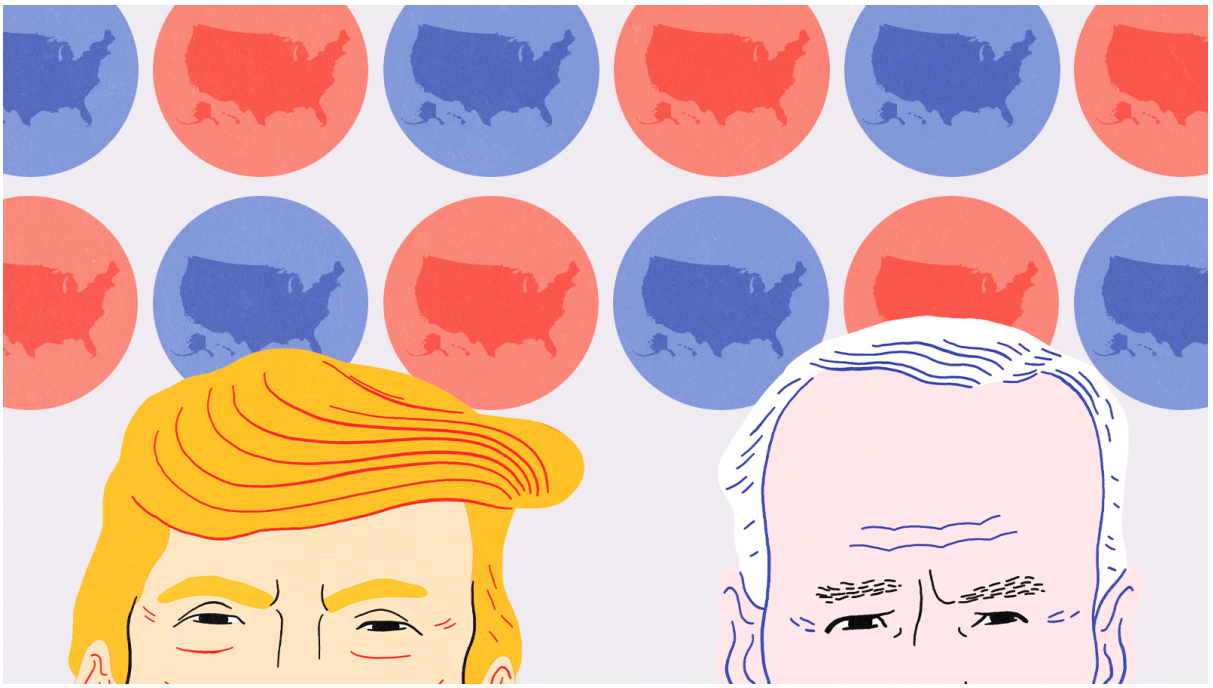
FACULTAD DE INGENIERÍA | UNIVERSIDAD DE LA REPÚBLICA URUGUAY

INTRODUCCIÓN A LA CIENCIA DE DATOS

TAREA 1



Fuente imagen: abcNEWS

GRUPO 23 - Valentina Cornelius, Santiago Quinteros

MAYO 2025

1. Introducción

La Tarea 1 se centra en el estudio de una base de datos abierta con discursos de políticos en el marco de las elecciones de Estados Unidos 2020.

2. Objetivos

Tratamiento de datos de una base de datos abierta. Estudio de discursos en el tiempo, conteo de palabras y menciones cruzadas entre candidatos, para los 5 candidatos con más discursos en la base de datos.

3. Datos

Los datos se extraen del archivo “us_2020_election_speeches.csv”, generando un DataFrame (df) con 269 filas (cantidad de discursos) y 6 columnas (atributos). De estas columnas, las que destacan para alcanzar los objetivos establecidos son: “speaker” (candidato/a, por practicidad se utilizará el término “candidato” aunque también hay candidatas), “date” (fecha) y “text” (texto del discurso).

En una primera explotación de los datos se detecta que hay 3 discursos (1.1 %) donde no hay speaker (datos faltantes en columna “speaker”). Otros problemas detectados en los datos sobre los nombres de los candidatos son los siguientes:

- 14 discursos (5.2 %) tienen candidatos indefinidos: ‘???’, ‘Multiple Speakers’, ‘Democratic Candidates’;
- 14 discursos (5.2 %) tienen candidatos combinados, por ejemplo: ‘Joe Biden, Kamala Harris’.

Para el procesamiento de los datos se decide no considerar los discursos donde no se definen los nombres de los candidatos (datos faltantes e indefinidos, que representan un 6.3 % del total). Los otros casos mencionados se procesan, replicando los discursos y dejando un sólo nombre. El contenido de los discursos se corrige para que coincida con el candidato según se explicará a continuación.

Explorando los discursos (columna “text”), se observan los siguientes problemas:

- los datos no contienen solamente el texto de un discurso, sino que incluyen, para cada intervención, el nombre de la persona (o una referencia genérica) y el minuto en el que se expresa;
- el texto del discurso asignado a un candidato particular puede presentar intervenciones de otras personas.

Que se repita el nombre del candidato no es un problema si se tiene en cuenta al interpretar los resultados del conteo de palabras. El problema es que aparezcan discursos de otras personas: se considerarían palabras que no corresponden. Para solucionarlo, se filtran los diálogos manteniendo solamente los que se corresponden con el “speaker”. Algo a destacar en este punto, es que los candidatos a veces aparecen con otros nombres en los textos de los discursos. Por lo tanto, antes del filtrado, se uniformiza la forma en la que se hace referencia al “speaker” en los discursos, identificando “sinónimos”. La Tabla 1 presenta las alternativas consideradas para los candidatos con más discursos (ver Top 5).

Tabla 1. Alternativas de denominación de candidatos/as

Nombre de candidato/a	Alternativas al nombre
Donald Trump	President Trump, President Donald J. Trump, Donald J. Trump
Joe Biden	Vice President Biden
Kamala Harris	Senator Harris
Mike Pence	Vice President Mike Pence

Luego de aplicado el tratamiento de los datos descrito, resulta una base final de 284 filas (discursos), con una nueva columna “Phrases”, en la cual solamente hay discursos del candidato identificado en dicha fila.

4. Análisis Descriptivo

Se realiza un análisis descriptivo de la base de datos, considerando los 5 candidatos con más cantidad de discursos.

4.1. Top 5 candidatos

La Figura 1 muestra los 5 candidatos con más discursos.

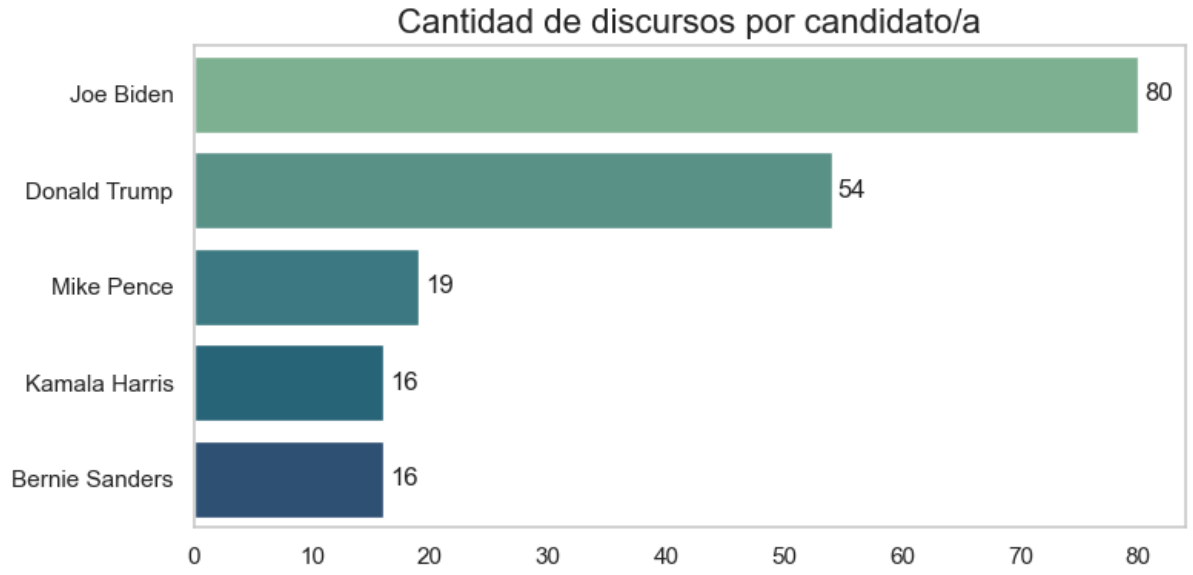


Figura 1. Top 5 de candidatos con más discursos en la base de datos analizada

Los dos candidatos con mayor cantidad de discursos son Joe Biden y Donald Trump, ambos candidatos por la presidencia de Estados Unidos en el año 2020. Luego, le siguen Mike Pence, candidato a vicepresidencia por Trump, y Kamala Harris, candidata a la vicepresidencia por Biden. Finalmente, Bernie Sanders es otro de los 5 candidatos con más discursos, quien también participó por la presidencia en la parte temprana de las elecciones por el partido demócrata.

4.2. Discursos en el tiempo

Para poder estudiar cómo se distribuyen los discursos en el tiempo, se realizó la conversión de los datos de fecha (columna “date”) a un formato adecuado.

La Figura 2 muestra la distribución de los discursos de los candidatos a lo largo del tiempo. Se observa que todos los discursos se dan en el año 2020. La mayoría de los discursos se concentran hacia los meses de setiembre y octubre, que corresponden a los meses de debates presidenciales. A excepción del candidato Berni Sanders, cuyos discursos se concentran hacia marzo. Los resultados son coherentes, ya que Sanders retiró su candidatura en el mes de abril[1].

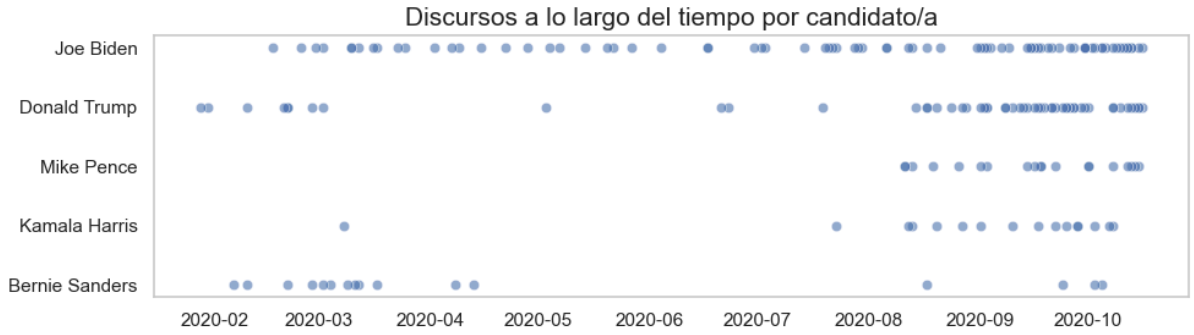


Figura 2. Distribución de los discursos de los candidatos a lo largo del tiempo

La Figura 3 muestra la cantidad de discursos por candidato por mes. Se observan picos claros en el mes de setiembre. Esto es coherente ya que ese mes se realizó el primer debate de las elecciones generales[2]. A esa altura de la campaña, Biden y Trump eran los candidatos a la presidencia, por lo que es razonable que sean los candidatos con picos más altos en cantidad de discursos ese mes. Los siguen Harris y Pence, quienes eran candidatos a la vicepresidencia.

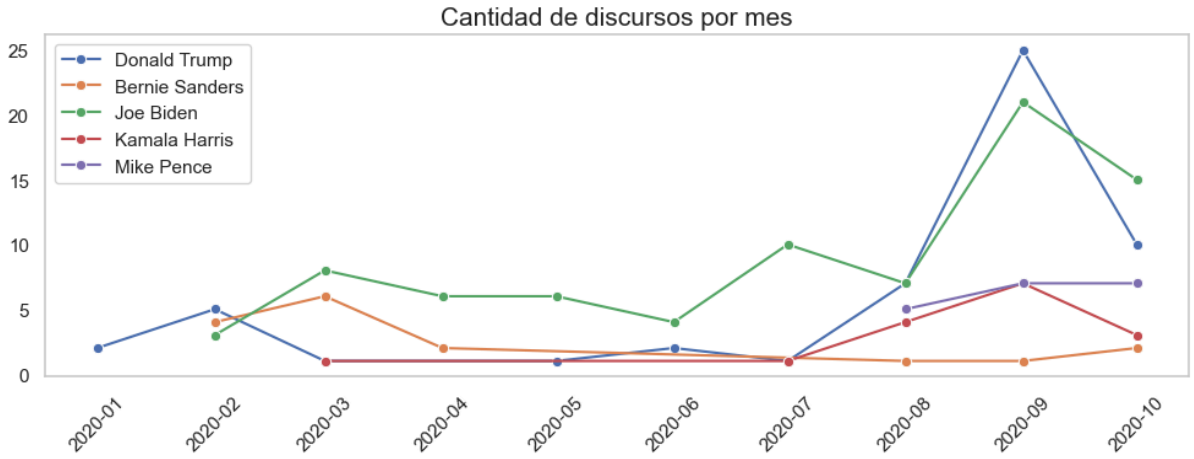


Figura 3. Variación de la cantidad de discursos por mes por candidato/a

[1] Fuente: https://edition.cnn.com/election/2020/results/president
[2] Fuente: https://apps.npr.org/elections20-primaries/

Los discursos finalizan en octubre porque las elecciones fueron el 3 de noviembre. En este gráfico se ve cómo Joe Biden fue el candidato más constante dando discursos, mientras que Donald Trump centró su estrategia en tener un pico máximo de discursos en setiembre.

4.3. Conteo de palabras

Para poder estudiar la cantidad de veces que aparece cada palabra en los discursos, es necesario normalizar el texto y eliminar los signos de puntuación.

Los cambios realizados fueron:

- Conversión del texto a minúsculas.
- Eliminación de signos de puntuación: ".", ",", “;”, ":", "?", “!” , "(", ")", "[", "]", "...", "“", "”"
- Eliminación de otros elementos: "\n", "1","2","3","4","5","6","7","8","9","0".
- Eliminación de “stop words”, por ejemplo “we’re”, “it’s”, “i’m”.

La denominación “stop words” hace referencia a palabras que se usan frecuentemente pero que no ofrecen información significativa por sí solas. Existen recursos que permiten extraer estas palabras según el idioma. En Anexo se presenta una lista de las palabras eliminadas.

Estas transformaciones se realizaron para poder obtener conclusiones más ricas de las palabras más mencionadas por los candidatos. Por ejemplo, que la palabra más repetida de un candidato sea “the” es entendible, porque es una palabra que se usa mucho, pero ese resultado no permite concluir nada en el contexto de las elecciones.

Una vez generado el listado “limpio” de palabras por candidato/a, se realizó el recuento de las diferentes palabras. La Figura 4 muestra las palabras más mencionadas por candidato/a.

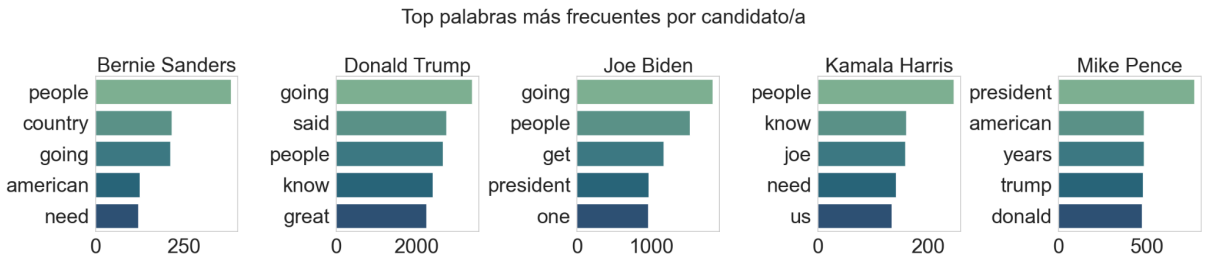


Figura 4. Palabras más mencionadas por cada candidato/a

Se observa como hay palabras comunes que repiten varios candidatos, como es el caso de “people”, “going” y “president”. Viendo el top 5 de palabras no se nota ninguna que permita identificar claramente un estilo o ideología marcada en el discurso. Tal vez los candidatos Sanders y Pence, son los que parecen tener un discurso más centrado en los estadounidenses al mencionar “american” (y “country” en caso de Sanders).

Algunas ideas para modificar esta visualización con el fin de encontrar diferencias entre partidos políticos, fechas o lugares serían:

- Agrupar los discursos según el partido político del candidato y ver las palabras más mencionadas según ese criterio.

- Agrupar los discursos según el mes, y ver las palabras que más usadas en cada uno. Se podría hacer una visualización global, o discriminando también entre partidos, para ver si hay palabras que empiezan a aparecer a partir de cierto momento en uno o ambos grupos.
- Asociar una geolocalización al discurso según el lugar donde fue realizado (en caso de ser posible, ya que hay discursos virtuales). Se podrían ver las palabras más usadas en diferentes regiones o Estados. Estas palabras podrían representarse en un mapa, o generar nubes de palabras con la forma de algún estado en particular.

En cuanto al total de palabras, Donald Trump no fue el candidato con más discursos, pero fue el que más palabras dijo, como se muestra en la Figura 5. Se entiende que es razonable que los dos candidatos a la presidencia sean los que más hayan hablado.

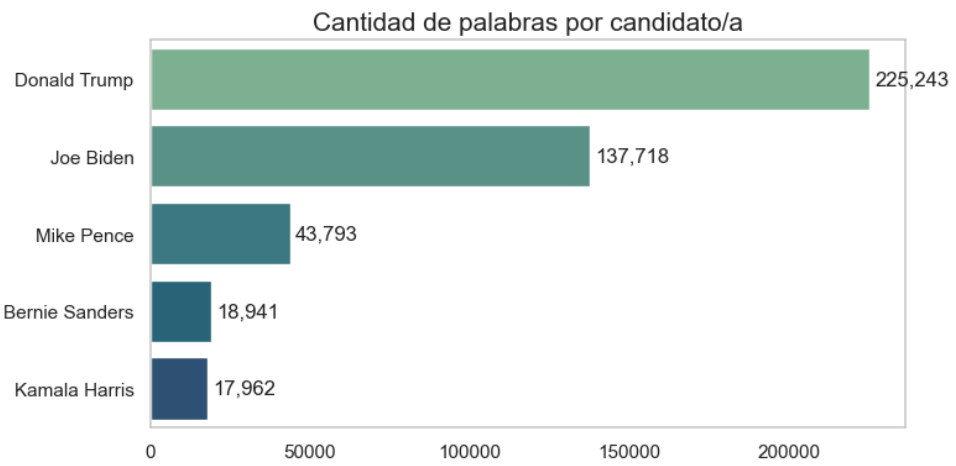


Figura 5. Cantidad de palabras por candidato/a

Se observa que, si bien cualitativamente se espera obtener resultados similares a los de otros grupos, el número de palabras obtenido seguramente difiera dadas las hipótesis asumidas al procesar los datos.

4.4. Menciones cruzadas

Para analizar las menciones cruzadas se propone la Figura 6, que presenta un mapa de calor, donde figura la cantidad de veces que se mencionaron entre candidatos.

La metodología utilizada para llegar al conteo de veces que un candidato fue nombrado, se basó en el siguiente cálculo:

- Las veces que un candidato nombra a otro por su nombre (p. ej., “Joe”), más las veces que lo nombra por su apellido (p. ej., “Biden”), menos las veces que lo nombra por su nombre completo (p. ej., “Joe Biden”).

Esta nos pareció la mejor aproximación a las menciones, debido a que dentro de los discursos hay varias formas de nombrar a otro candidato.

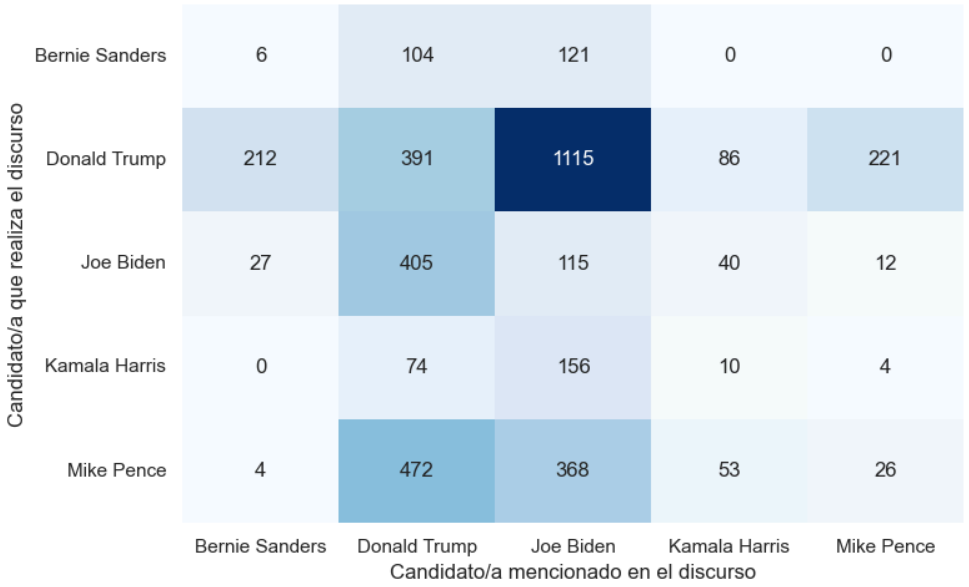


Figura 6. Cantidad de menciones cruzadas entre candidatos

Revisando los candidatos que se mencionaron entre sí, el que más menciono al resto fue Donald Trump, con un máximo de 1115 menciona a Joe Biden. Por otro lado, el candidato más mencionado fue Joe Biden, que tanto en su oposición, como miembros del partido demócrata fue uno de los más mencionados. Ambos vicepresidentes mencionaron principalmente al candidato a la presidencia de su partido, con menciones significativas a la oposición también. Bernie Sanders, solamente se mencionó al mismo, y a los candidatos a la presidencia, sin nombrar nunca a los candidatos a la vicepresidencia. Por último, Kamala Harris fue la candidata con menos menciones, siendo más nombrada por la oposición, Donald Trump, que su propio partido.

5. Preguntas

Se plantean tres preguntas que se podrían responder utilizando la base de datos analizada.

1. ¿El tema “Cambio Climático” está presente en la campaña?
   Para estudiarlo, se podría buscar cuántas veces aparece “climate change” en los discursos de los diferentes candidatos. Se podrían incluir eventualmente otras palabras que pudieran hacer referencia al mismo tema (se debería utilizar el discurso procesado antes de separar las palabras, ya que se podrían buscar frases además de palabras).

2. ¿Cuando un/a candidato/a se nombra a sí mismo/a, es porque habla de él/ella en tercera persona, o está citando una frase donde alguien más lo/la nombró?
   Para estudiarlo, se podría considerar que si el candidato está citando algo que dijo otra persona, el texto estará entre comillas. Se podría generar un nuevo discurso eliminando los textos que estén entre comillas; de esta forma, al contar en ese nuevo discurso las veces que aparece su nombre, se estarían contando las “autorreferencias”. La resta entre la cantidad original y las “autorreferencias” serían citas de otras personas.

3. ¿Qué sentimientos aparecen en los discursos políticos de cada candidato/a?
   Análisis de sentimientos en los discursos políticos: los discursos políticos pretenden generar emociones en la población, que pueden ser positivas (p. ej., “esperanza”) o negativas (p. ej., “resentimiento”) . Utilizando algoritmos de machine learning como el procesamiento de lenguaje natural (NPL), se podría estudiar la emociones contenidas dentro de los discursos de cada candidato/a.

ANEXO

Palabras eliminadas del texto mediante NLTK[3]

{'other', 'more', "she's", 'haven', 'only', "i'm", "didn't", 'very', 'have', 'he', "i'll", 'his', 'shan', 'me', 'ours', "weren't", 'don', "mightn't", 'll', 'its', 'than', 'with', 'wouldn', 'which', 'yourselves', 'does', "he'd", 'hasn', 'nor', 'between', 'a', 'your', 'itself', 'few', 'further', 'we', 'him', 'themselves', 'were', 'i', 'this', 'after', "she'll", 'why', 'has', "you're", 'their', 'ma', "isn't", 'wasn', 'as', 'where', 'once', "we've", "he'll", 'of', 'if', 'each', 'they', "they've", 'didn', 'her', 'mustn', 'up', 'at', 'yourself', "should've", 'm', "you've", "you'll", 'then', 'doesn', 'theirs', "we're", 'yours', 'below', 'ourselves', "haven't", 'no', "hadn't", 'needn', 'herself', 'own', "we'll", "won't", 'aren', "doesn't", 'that', 'here', 'she', 'couldn', 'can', "hasn't", 'just', 're', 'about', 'against', 'off', 'too', 'my', 'o', 'hadn', 'during', "he's", 'an', 'these', 'is', 'weren', "i've", 'some', "aren't", 'will', 'to', 'had', 'all', 'are', "they'd", 'again', "wasn't", 'mightn', "couldn't", 'same', 'it', 'y', "she'd", 'now', 'or', "we'd", 'over', 'above', 'd', 'shouldn', 't', "they'll", 'when', 'them', 'from', "they're", "shouldn't", 'there', 'did', 'any', 'our', 'not', 'such', 'being', 'what', 'those', 'in', 'myself', 'won', "wouldn't", 's', 'on', 'been', "i'd", "shan't", 'but', 'while', 'and', 'into', 'having', 'ain', 'you', 'most', 'through', "that'll", 'himself', 'am', 'under', 'until', 'who', 'both', 'hers', 'do', 'for', "you'd", 'down', "don't", 'was', 'should', 'out', 've', 'the', "needn't", 'so', 'because', 'by', 'whom', 'isn', "mustn't", 'before', 'be', "it'll", "it'd", "it's", 'doing', 'how'}

[3] Fuente: https://pythonspot.com/nltk-stop-words/