

Visualización de datos

Trabajo Práctico

Jorge Tenorio Berrio

Mayo 2023

Índice

1. Planteamiento del problema y objetivos de la visualización	3
1.1. Objetivos	3
1.2. Entregables	3
2. Preparación de los datos	3
2.1. Crawler	4
2.2. Limpieza	5
2.3. Extracción de datos	5
2.4. Cruce de datos	5
2.5. Selección de intervenciones	5
3. Procesado y análisis	6
3.1. Análisis 1: Descripción del corpus	6
3.2. Análisis 2: frecuencia de palabras	7
3.3. Análisis 3: selección de características	13
4. Visualización	14
5. Discusión, Conclusiones y posibles mejoras.	17
5.1. Conjunto de datos	17
5.2. Metodología de análisis y resultados	17
5.3. Visualización	17
6. Herramientas utilizadas	18

1. Planteamiento del problema y objetivos de la visualización

El actual panorama político en España se ha visto afectado por dos factores. Por un lado, grupos políticos de carácter extremista han conseguido cierta representación significativa en las cámaras de representantes. Estos partidos suelen caracterizarse por utilizar un lenguaje populista y en ocasiones violento que genera un ambiente de polarización. Por otro lado, desde las últimas elecciones se han dado algunos acontecimientos que han empeorado el clima político. Desde la pandemia, hasta la guerra de Ucrania, pasando por la borrasca Filomena y la erupción del volcán en La Palma, hacen que esta campaña haya resultado especialmente convulsa.

1.1. Objetivos

En este trabajo, se va a tratar de realizar un análisis del lenguaje de los distintos partidos políticos en intervenciones del Congreso de los Diputados de la XIV Legislatura. El objetivo principal es comprobar si existen diferencias en el lenguaje en los discursos de esta cámara. Se pretende comprobar estas diferencias en dos ejes: el primer eje se corresponde con los distintos grupos políticos y posiciones ideológicas-administrativas(izquierda-derecha, gobierno-oposición, regionalistas-nacionales...). Se utilizará el tiempo como segundo eje, permitiendo comprobar si existen variaciones en el discurso en distintos momentos de la legislatura.

De igual manera, el público objetivo de este trabajo pueden ser de dos tipos. Por un lado, el público general con interés en el panorama político actual. Y por otro lado, dar indicios para una investigación dentro del área de las ciencias políticas o de la información.

En este trabajo, se utilizará la palabra como unidad de análisis del lenguaje.

1.2. Entregables

Junto, con esta memoria, se ha creado un repositorio¹ en el que se dispone de los siguientes recursos:

- Datos utilizados en las distintas fases.
- Cuadernos jupyter con cada una de las fases del trabajo. En cada uno se incluye un enlace a *Google Colaboratory*. En cada sección se incluye un enlace a su respectivo cuaderno.
- Documento HTML y PNG con la visualización final.
- Este mismo documento en formato PDF.

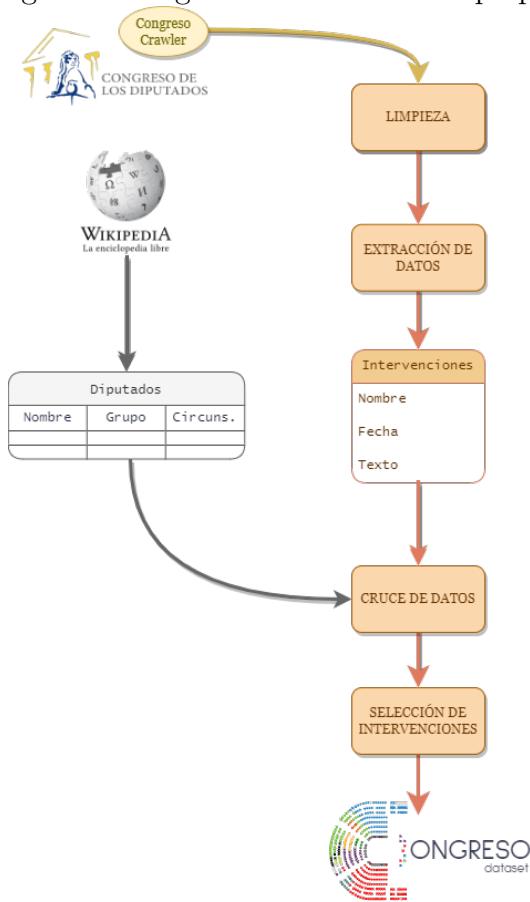
2. Preparación de los datos

El conjunto de textos está formado por intervenciones de parlamentarios del congreso de los diputados. Además del texto, en cada intervención se incluye el parlamentario que lo

¹<https://github.com/Chiriviki/congreso/tree/master>

enuncia, el grupo parlamentario al que pertenece y la fecha en la que se produce la intervención. Para elaborarlo se ha partido de los diarios de sesiones públicos en la web del congreso de los diputados. En esta sección se presentan las distintas fases para obtener el conjunto de documentos que se utilizarán en el análisis posterior. En la figura 1 se muestra un resumen de estas fases.

Figura 1: Diagrama de las fases de preprocesado.



La implementación de las distintas tareas se encuentran en los siguientes enlaces:

- Descarga y limpieza (secciones 2.1 y 2.2): https://github.com/Chiriviki/congreso/blob/master/1.-%20Descarga_corpus.ipynb. Se recomienda no ejecutar este cuaderno por su tiempo de ejecución.
- Extracción de datos, cruce de datos y selección de intervenciones (secciones 2.3 a 2.5): https://github.com/Chiriviki/congreso/blob/master/2.-%20Crea_corpus.ipynb

2.1. Crawler

Se ha desarrollado un *script (crawler)* para descargar el conjunto de documentos con los diarios de sesiones. Los diarios de sesiones se pueden obtener mediante consultas a la url del congreso. Estas webs normalmente están protegidas contra consultas automatizadas por lo

que es necesario simular que somos un usuario. Esto se puede conseguir estableciendo algunos parámetros en la solicitud

Todos los documentos tienen un formato fijo y están numerados en cada legislatura, por lo que es sencillo recorrerlos. Por ejemplo, el documento DSCD-14-PL-140.CODI representa el diario de sesiones del congreso de los diputados del pleno 140 de la XIV legislatura.

Este proceso se realizó el día 6 de marzo de 2023 y se obtuvieron 248 diarios de sesiones.

2.2. Limpieza

En esta fase se procesan los documentos html y se obtiene el texto completo junto con algunos metadatos (especialmente la fecha de realización del pleno). Para ello se ha obtenido el contenido de cierta etiqueta html y eliminado etiquetas con números de página y saltos de línea.

2.3. Extracción de datos

Este es uno de los pasos más importantes del preprocesado ya que se separa el texto plano en intervenciones de cada parlamentario. Todas las intervenciones comienzan con una estructura similar en la que se menciona el nombre del diputado y/o cargo en el gobierno o mesa del congreso. En la figura 2 se muestra un ejemplo. Para llevarlo a cabo se han utilizado varias expresiones regulares.

Figura 2: Ejemplos de intervenciones.

El señor **VICEPRESIDENTE** (Rodríguez Gómez de Celis): Muchas gracias.
A continuación de la palabra señor Baldoví i Roda.

El señor **BALDOVÍ I RODA**: *Moltes gràcies, senyor president.*
Si hay un episodio vergonzoso e ignominioso de la historia de este país es el papel que ha tenido España con el pueblo saharaui. Más de cuarenta años de abandono de un pueblo y algunos que están

2.4. Cruce de datos

Se dispone de un conjunto de textos junto con la fecha y el nombre del parlamentario. En nuestro análisis se pretende tener en cuenta los distintos partidos políticos. En esta fase se enlazan los apellidos de cada parlamentario con el grupo político al que pertenece. Se han utilizado las tablas de wikipedia con información de la legislatura² como origen de datos de partidos.

2.5. Selección de intervenciones

Por último, se ha realizado dos tareas de selección de intervenciones. Por un lado, se han eliminado los errores en la unión con el grupo político. La mayoría de ellos estaban causados por políticos miembros del gobierno o de la mesa del congreso que no iban en la lista de las

²https://es.wikipedia.org/wiki/Anexo:Diputados_de_la_XIV_legislatura_de_España

elecciones, por ejemplo *Nadia Calviño*. Estos errores no se consideran relevantes ya que los partidos a los que pertenecen están sobre-representados en el conjunto de datos.

Por otro lado, se han limitado los grupos políticos considerados en base a dos criterios: que tengan una representación significativa y que existan diferencias ideológicas considerables entre ellos. Estos grupos son:

- Partido Socialista Obrero Español (PSOE)
- Partido Popular (PP)
- Vox
- Unidas Podemos (UP) junto con los grupos de Cataluña (ECP) y Galicia (GeC)
- Esquerra Republicana de Catalunya (ERC)
- Partido Nacionalista Vasco (PNV)

Además, se han eliminado los miembros de la mesa del congreso, ya que sus intervenciones suelen ser cortas y con funciones más administrativas que políticas.

3. Procesado y análisis

A lo largo de esta sección se realizan varios análisis sobre los datos. En primer lugar, se describirá superficialmente el corpus. Posteriormente se realiza un análisis mediante frecuencia de palabras. Por último, se tratará de identificar aquellas palabras que son relevantes para diferenciar los textos. En todos los casos, el análisis se realizará desde dos perspectivas: entre partidos políticos y a lo largo del tiempo.

El código utilizado para esta sección se puede encontrar en:

- Descripción del corpus (sección 3.1): https://github.com/Chiriviki/congreso/blob/master/2.-%20Crea_corpus.ipynb (final)
- Análisis (secciones 3.2 y 3.3): <https://github.com/Chiriviki/congreso/blob/master/3.-%20An%C3%A1lisis.ipynb>

3.1. Análisis 1: Descripción del corpus

El *Congreso Corpus* se compone de 12084 intervenciones de políticos del Congreso de los diputados de seis partidos políticos. Incluye intervenciones únicamente de plenos desde el inicio de la legislatura (diciembre de 2019) hasta febrero de 2023. En total, se compone de casi 6.8 millones de palabras.

Como es lógico, los grupos políticos con mayor representación en el congreso participan más en el congreso (figura 3). Esto da lugar a que el corpus esté desbalanceado entre grupos políticos.



Figura 3: Distribución de intervenciones entre los distintos grupos políticos.

De igual manera, la actividad del congreso no es la misma en las distintas épocas del año. Como se puede ver en la figura 4, en períodos vacaciones hay una disminución en el número de intervenciones. De igual manera, en 2020 la distribución no es la misma que otros años debido a las restricciones de la pandemia. En este caso se han agrupado las intervenciones por meses y diferenciado por años por dos razones. Utilizar el día como unidad de tiempo resultaría en distribuciones muy variables. Comparar el mismo mes en distintos años da una visión del desbalanceado.



Figura 4: Distribución de intervenciones y número de palabras a lo largo del tiempo (meses). No se incluye 2019 al solo disponer datos de diciembre.

3.2. Análisis 2: frecuencia de palabras

Antes de comprobar las diferencias entre los distintos textos es necesario tener algo que comparar. Se ha representado cada intervención de forma numérica. Cada una está definida por un vector tan largo como el vocabulario, con el valor de la frecuencia de la palabra w_i en la posición i del vector. Utilizar la frecuencia absoluta tiene problemas con palabras muy frecuentes, por lo que se ha utilizado TF-IDF³ como medida. Esta métrica castiga el peso de aquellas palabras que son más frecuentes en el corpus.

³<https://es.wikipedia.org/wiki/Tf-idf>

Para realizar este proceso se ha utilizado `sklearn.feature_extraction.text.TfidfVectorizer`, el vocabulario se ha limitado a las 5000 palabras más frecuentes, se han desechado las *stop-words* definidas en `nltk.corpus.stopwords` (español) y se ha utilizado normalización l_1 , esta normalización hace que la suma de todos los valores de un vector sea 1. Normalizarlos permite lidiar con distintas longitudes de textos. Además, permite ser interpretado como una proporción. Sin embargo, para agrupar distintas intervenciones será necesario utilizar el promedio para mantener la normalización.

Realizar un análisis y realizar visualizaciones de datos con 5000 variables es realmente complicado. De una forma u otra, es necesario seleccionar una parte. En la figura 5 se muestra un diagrama de barras de las 20 palabras más frecuentes del corpus. Se prefiere una disposición horizontal para facilitar la lectura. Se puede ver que la mayoría son términos comunes en el lenguaje parlamentario. Hay un hecho interesante: la aparición de *aplausos* como palabra con un frecuencia muy alta. Esto se debe a que cada vez que alguna de las bancadas aplaude, esto queda registrado en el diario de sesiones con el término entre paréntesis.

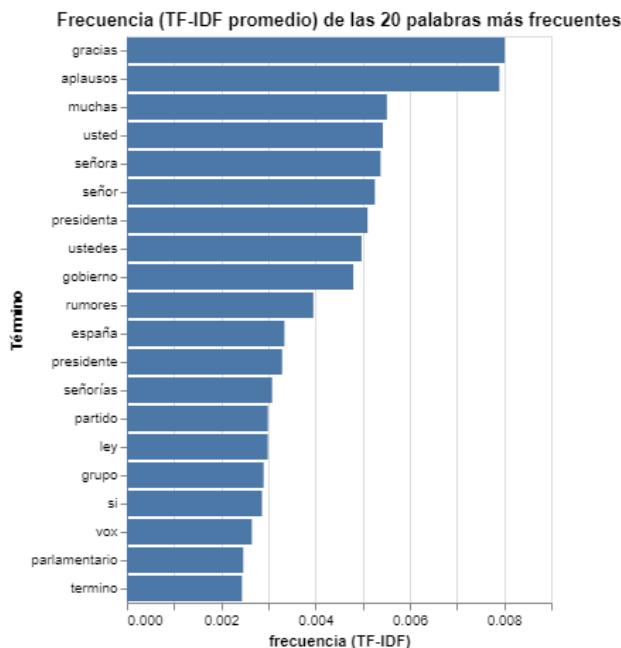


Figura 5: Diagrama de barras de las 20 palabras más frecuentes junto con el valor de su frecuencia.

Una alternativa de visualización son las nubes de palabras (no disponible en Altair). En esta visualización, se disponen las palabras aleatoriamente en el espacio con un tamaño proporcional a su frecuencia. En la figura 6 se muestra la nube de palabras de todo el corpus. Los colores han sido asignados aleatoriamente.



Figura 6: Nube de palabras del corpus completo. El tamaño de las palabras está definido por su frecuencia (TF-IDF promedio).

La nube de palabras no aporta información exacta del valor de la frecuencia, pero es capaz de recoger muchas más palabras y dar una visión mucho más rápida que un gráfico preciso como el de barras. Además, el valor de TF-IDF no es fácilmente interpretable, por lo que no resulta muy útil dar su valor exacto. En ambos casos, es necesario limitar el número de variables de algún modo.

Los puntos anterior nos ofrecen un vistazo global de los textos. Pero si los agrupamos de cierta manera, podemos reorientarnos a los objetivos del trabajo. Podemos empezar agrupando las **intervenciones por fechas** y de este modo comprobar la evolución de los discursos en el tiempo. Para este caso se utilizado el año-mes como unidad de tiempo, de esta forma es posible agrupar un número considerable de intervenciones y compararlas a lo largo de la legislatura. Se han considerado dos tipos de gráficos.

En primer lugar se considera representar los 20 términos más comunes en el corpus (los del primer análisis) mediante un *streamgraph*. Este tipo es muy común a la hora de representar variables categóricas a lo largo del tiempo. En nuestro caso tiene varios inconvenientes. En este gráfico, hemos representado términos con frecuencias muy altas, pero puede haber ocultación entre valores grandes y pequeños. Una solución posible sería utilizar escalas no lineales pero este tipo de gráfico no lo permite. Por otro lado, no nos interesa comparar los valores absolutos entre términos (uno de las ventajas de este gráfico), si no la progresión dentro cada término.

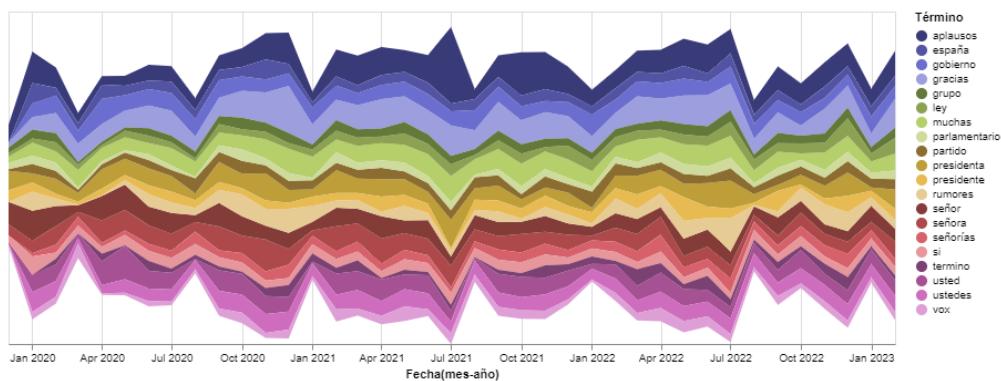


Figura 7: *Streamgraph* de la frecuencia (TF-IDF promedio) de los 20 términos más frecuentes del corpus.

En la figura 8 se muestra un gráfico de área de la progresión de la frecuencia de los

20 términos más frecuentes separados. Representarlos de este modo permite utilizar escalas independientes en el eje Y para evitar la ocultación. Se ha utilizado un gráfico de área en lugar del típico gráfico de líneas de las series temporales para diferenciar mejor entre las distintas filas del gráfico.

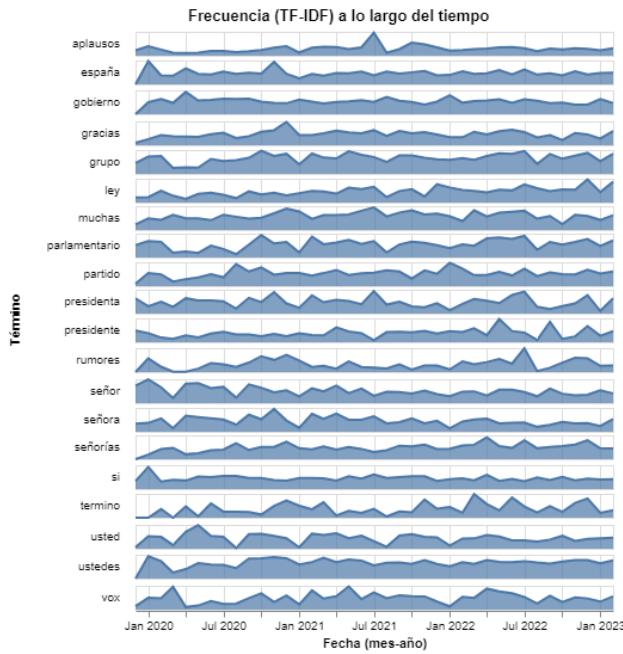


Figura 8: Frecuencia (TF-IDF promedio) a lo largo del tiempo de los 20 términos más frecuentes del corpus.

Los dos gráficos anteriores no aportan mucha información para el análisis. Principalmente debido a la selección de términos realizada. Que sean términos muy frecuentes no significa que sean relevantes para el análisis. Por otro lado, se puede apreciar existen bajadas de frecuencia generales en los meses de enero y agosto. Un posible indicio de que se han trasladado los sesgos del desbalanceado.

Otra forma de agrupar la intervención es por **partido político**. Para llevarlo a cabo, hemos agrupado las intervenciones utilizando de nuevo el promedio. Para representarlos gráficamente podríamos utilizar gráficos de barras al igual que en el análisis inicial, utilizando el color como elemento diferenciador entre partidos. Sin embargo, este tipo de gráficos tiene varios inconvenientes.

- Si se utiliza un barra distinta para cada partido, aumenta considerablemente el número de elementos a representar en un mismo eje, ya que necesitamos una barra distinta dentro de cada partido y término.
- Si se utilizan barras apiladas perdemos la coherencia con los datos. Recordemos que se ha utilizado el promedio para agruparlos, y que el resultado de sumar todos los términos es uno. Si apilamos las barras sería equivalente a agrupar utilizando la suma.

Un ejemplo de ambos gráficos se muestran en la figura 9.

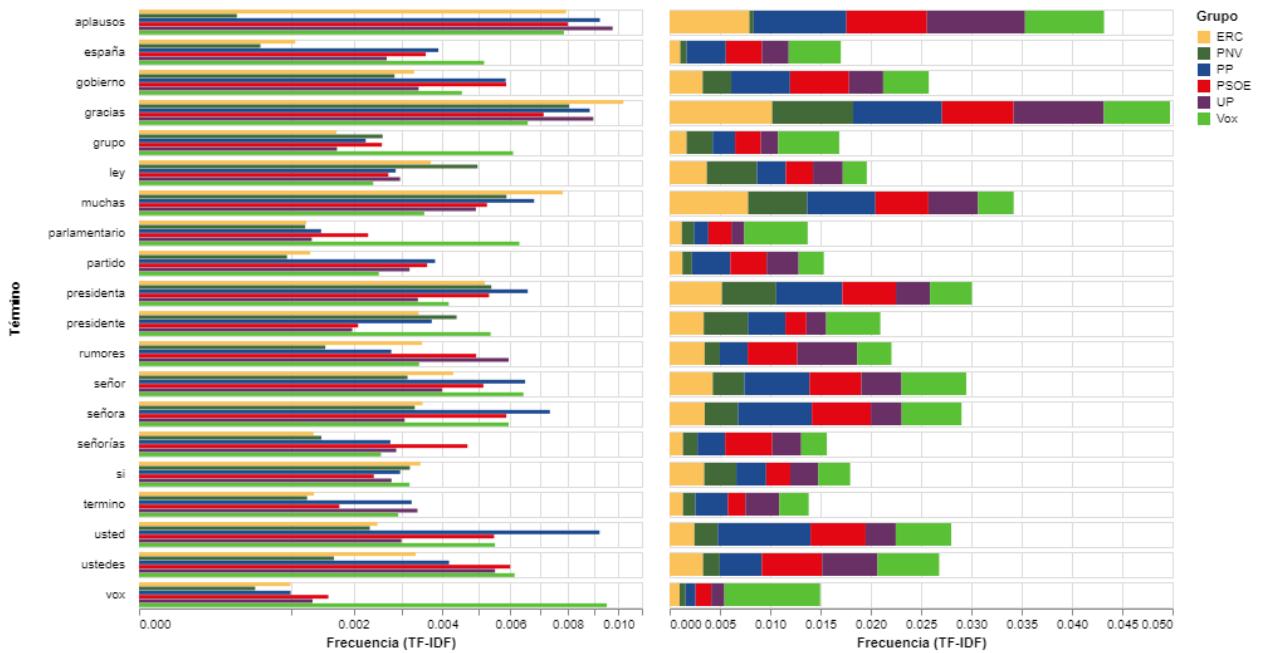


Figura 9: Gráficos descartados para mostrar la frecuencia de cada término en los distintos partidos políticos.

Para solucionarlo, se ha utilizado un gráfico de dispersión de una sola dimensión. Para cada término, se representa el valor de su frecuencia dentro cada partido como un punto en un recta. De esta forma es posible comparar el uso de ese término en particular en cada uno de los discursos. En la figura 10 se muestra un ejemplo con los 20 términos más frecuentes. Se ha utilizado la raíz cuadrada en la escala para evitar ocultamiento de valores más pequeños. Este gráfico tenía un problema de solapamiento que se ha solucionado introducción de un selector interactivo en la leyenda que resalta los valores del partido utilizando la transparencia.

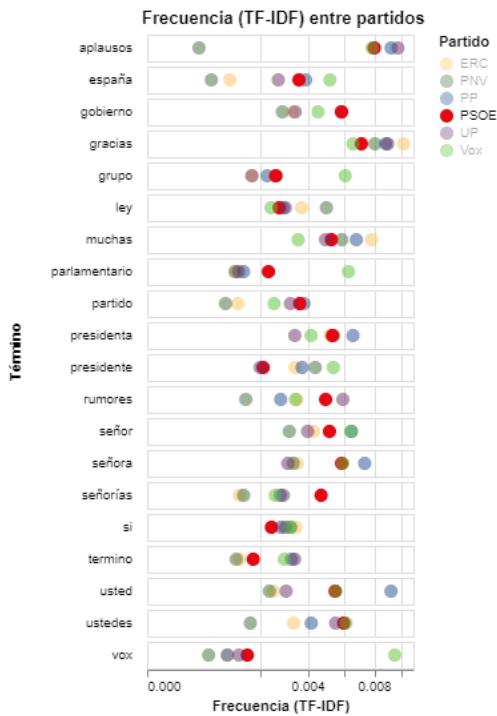


Figura 10: Frecuencia (TF-IDF promedio) de los 20 términos más frecuentes del corpus en cada partido político.

En este gráfico podemos ver diferencias significativas. Por ejemplo, el término aplausos tiene una frecuencia similar en la mayoría de partidos excepto en el PNV. Esto podría ser un indicio de que este partido no utiliza la técnica del *autoaplauso* típica en el congreso. También podemos ver que palabras como *grupo*, *parlamentario* y *vox* tienen una frecuencia mayor en Vox. Si vemos algunas de sus intervenciones, nos daremos cuenta que en este partido es muy común utilizar la expresión *grupo parlamentario Vox* para referirse al propio partido.

Esta forma de agruparlos nos permite utilizar también nubes de palabras. En la figura 11 se muestra nubes de palabras de los 100 términos más frecuentes en el discurso de cada uno de los partidos, representados cada uno con su respectivo color. Se puede ver que se replican, en cierta medida, las observaciones realizadas en el gráfico anterior.

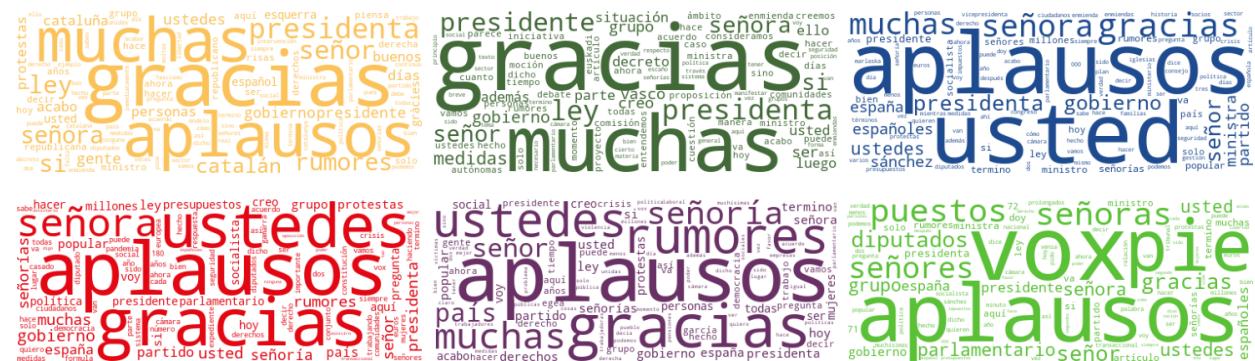


Figura 11: Nubes de palabras (TF-IDF promedio) de los 100 términos más frecuentes de cada partido. De izquierda a derecha: ERC, PNV, PP (arriba), PSOE, UP, Vox (abajo).

Estos dos tipos de visualizaciones ofrecen dos perspectivas de los mismos datos. Por un lado, el gráfico de puntos permite ver como se distribuyen los partidos dentro de cada término. Por el otro, la nube de palabras muestra la distribución de las palabras dentro de cada partido. En ambos casos, se requieren métodos de selección de los términos a mostrar.

3.3. Análisis 3: selección de características

Hemos visto, que este tipo de análisis y visualizaciones están limitados por el número de términos y el criterio para seleccionarlos. Para solucionarlo, se va a ponderar cada término con un valor de **importancia o relevancia** (a partir de aquí se hará uso de estos términos indiferentemente). Este valor estará relacionado con la capacidad de discriminación del término. Aplicado a los partidos (o fechas), representaría la capacidad de discriminar entre partidos político (o fechas) dado el valor de frecuencia del término.

Para llevarlo cabo se utiliza la información mutua⁴ como medida de importancia de cada término. Esta métrica mide la reducción de incertidumbre de una variable, cuando otra es conocida. Un valor de 0 significa que la variable no aporta información. Valores más altos equivalen a la cantidad de información aportada. Uno de las ventajas de esta medida es que no se limita a relaciones no lineales.

Este proceso se va a utilizar para calcular dos valores de información mutua. Uno para la relación entre la frecuencia de cada término con el partido político de la intervención, y otro para la frecuencia de cada término con la fecha de celebración del pleno. Para la implementación se han usado `sklearn.feature_selection.mutual_info_classif` y `sklearn.feature_selection.mutual_info_regression` respectivamente. La fecha se ha asumido como un valor numérico.

Un simple gráfico de barras es suficiente para mostrar las palabras con mayor importancia. En la figura 12 se puede ver que *aplausos* es la palabra más relevante discriminando partidos políticos. También se puede ver que aparecen palabras propias de partidos de carácter nacionalista como *euskadi* o *gràcies*. Por otro lado, *presupuestos* es la palabra más relevante para las fechas. También aparecen términos propios de acontecimientos como *ucrania*, *alarma*, *pandemia*, *guerra* o *crisis*.

⁴https://es.wikipedia.org/wiki/Informaci%C3%B3n_mutua

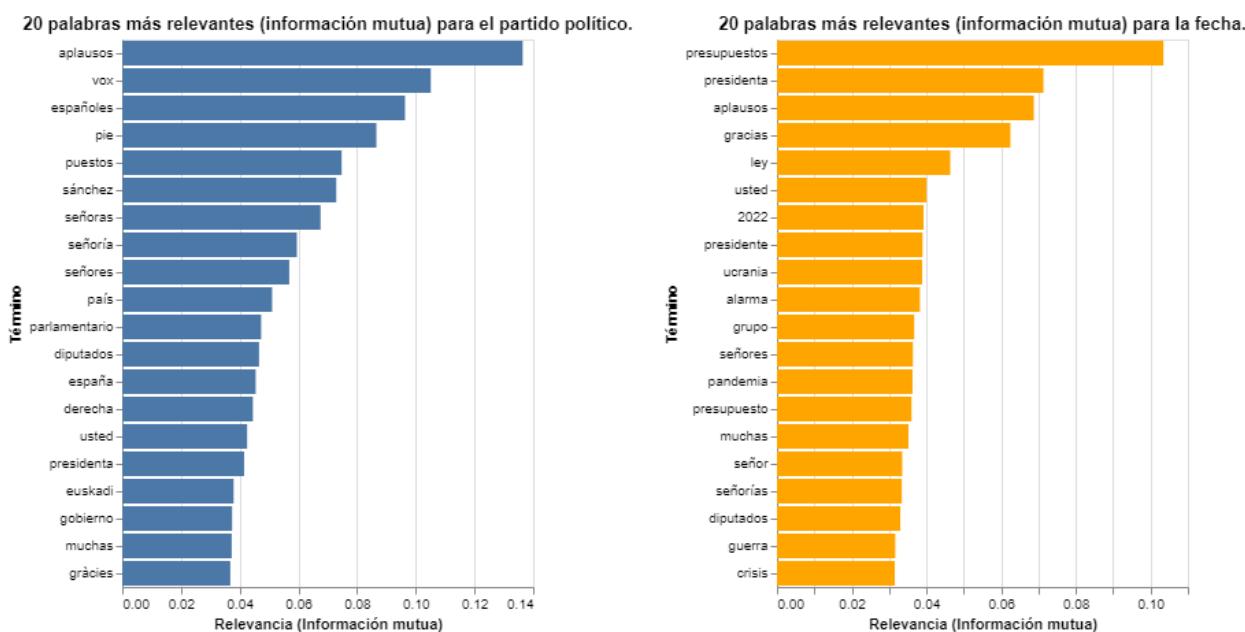


Figura 12: Valor de relevancia (información mutua) de los 20 términos más relevantes en ambas tareas (partido y fecha).

4. Visualización

Llegados a este punto, se ha planteado la pregunta: ¿Qué queremos mostrar? La respuesta es, aquellas palabras que diferencian a los partidos políticos y momentos de la legislatura, y el porqué: los valores cuantitativos de esas diferencias.

Los pasos para crear esta visualización se pueden ver en <https://github.com/Chiriviki/congreso/blob/master/4.-%20Final.ipynb>.

Para llevar a cabo la primera parte se han predisposto dos gráficas: una para las diferencias temporales y otra para las diferencias entre partidos. En ambos casos, se han seleccionado los 20 términos más relevantes y predisposto cada uno en una fila (tal y como se hizo en los gráficos de barras anteriores).

Como valores cuantitativos se han usado:

- Relevancia (información mutua): muestra el grado en que ese término diferencia los distintos valores. Se ha utilizado un gráfico de barras sencillo.
- Frecuencia (TF-IDF): muestra la distribución del término. Se ha utilizado un gráfico de área para el gráfico tiempo, y dispersión 1D para los partidos políticos.

Todos estos gráficos se han mostrado en la sección anterior. Desde el punto de vista estético, se pueden destacar las siguientes acciones.

Se ha añadido un título con el logo del congreso para hacer referencia a la fuente de los datos. Se han seleccionado los colores del logo para diferenciar los dos gráficos de datos. Existe cierto solapamiento con los colores de algunos partidos políticos que se ha solucionado añadiendo cierto grado de transparencia.

Se ha añadido un punto artístico al título mediante una nube de palabras compuesta por las palabras más relevantes y su frecuencia en cada uno de los partidos políticos. Se ha seleccionado una forma de gráfico (semi)circular simulando los gráficos de resultados de encuestas (estos gráficos, a su vez, hacen alusión a la forma del congreso). El número de intervenciones de cada partido en los datos define el ángulo de cada porción del gráfico. El orden de las porciones está definido por su posición ideológica, al igual que en los gráficos de encuestas (retocado para evitar solapamiento de colores). En el gráfico final se muestra en baja resolución, por lo que no se pueden observar los detalles. En la figura 13 se muestra este gráfico en mejor resolución.



Figura 13: Nube de palabras de las palabras más relevantes del corpus. La frecuencia en cada partido define su tamaño. El ángulo de cada porción está definido por el número de intervenciones tratadas.

En la figura 14 se muestra la visualización final. Respecto a los datos, si nos fijamos en la gráfica temporal podemos observar que hay términos asociados a tendencias ligadas a la legislatura, como *presupuestos* o *presupuesto*, que se repiten anualmente. Y otras ligadas a eventos como puede ser *pandemia* o *crisis*, que inician en marzo de 2020; *guerra y ucraña* en febrero de 2022; o *diputados* en el inicio de la legislatura.

Por otro lado, si nos fijamos en la gráfica por partidos, podemos ver que hay el grupo parlamentario Vox tiene un lenguaje más distinto al resto. En la mayoría de términos con una gran relevancia es el partido que más se distancia. Estos términos son *vox*, *españoles*, *pie*, *puestos*, *señoras*, *señores*, *parlamentario* y *diputados*. También se puede ver que los partidos nacionalistas tienen un mensaje distinto, con una frecuencia menor en términos como *españa* o *españoles*, y mayor en palabras propias de su propia lengua o región, como *gràcies* o *euskadi*. Por último, volver a mencionar el uso casi nulo de *aplausos* por parte del PNV, hecho explicado en el análisis parcial.

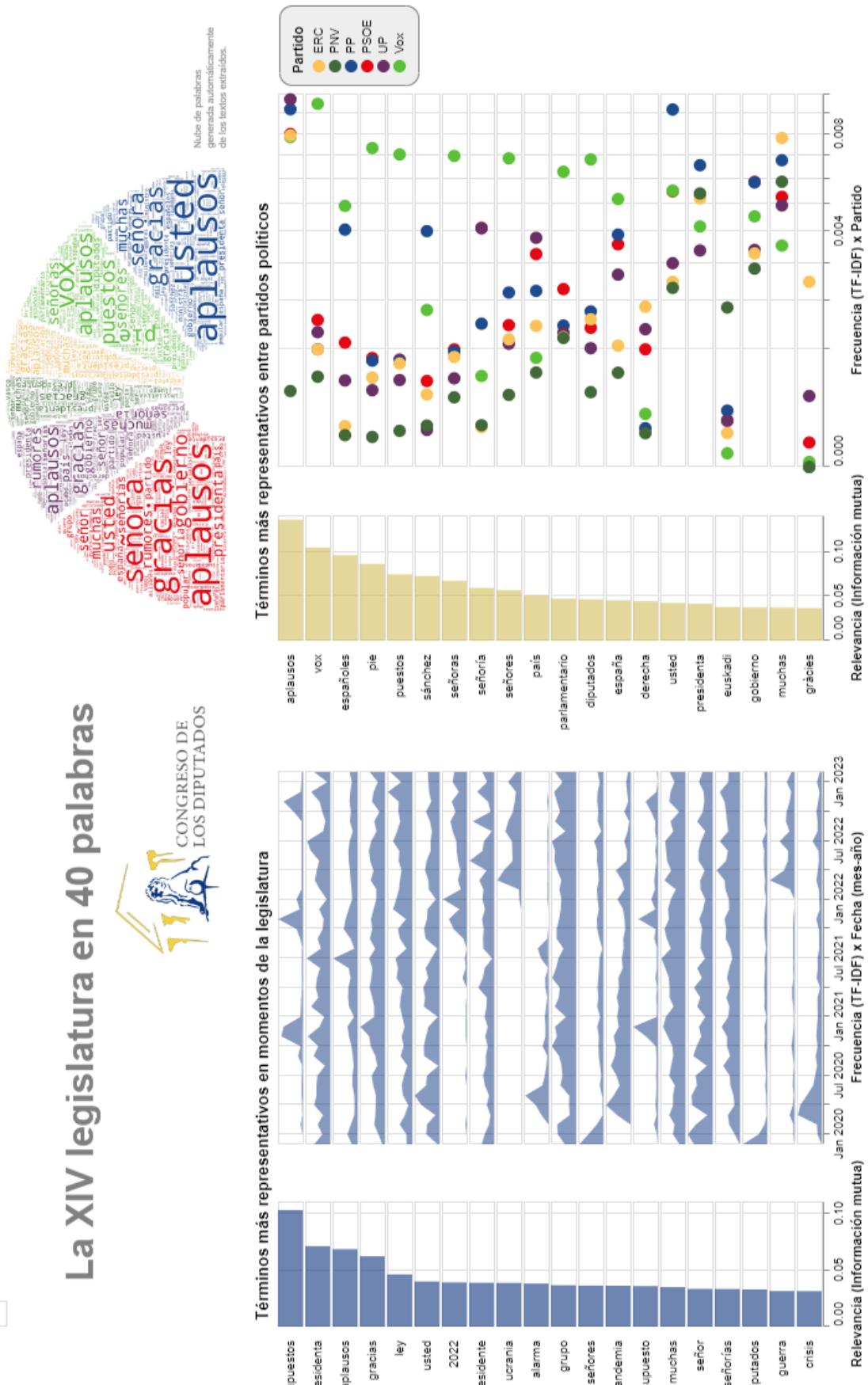


Figura 14: Visualización final

5. Discusión, Conclusiones y posibles mejoras.

En este trabajo se ha elaborado un corpus de intervenciones de políticos españoles en el congreso de los diputados, se ha realizado un análisis del lenguaje a nivel de palabra y elaborado una visualización para mostrar los resultados.

A continuación, se van a tratar distintos puntos del trabajo de forma independiente.

5.1. Conjunto de datos

Pese a que el conjunto de datos es público, la labor de extraer y estructurar los textos ha supuesto un gran esfuerzo, principalmente debido a su gran volumen. Este conjunto de datos junto con la metodología para extraerlos se considera de gran valor desde el punto de vista de análisis político y puede suponer un punto de partida para análisis más extensos. Se podría ampliar por varias vías: incluyendo más partidos políticos, otras legislaturas, otras órganos como comisiones, etc.

5.2. Metodología de análisis y resultados

Analizar los textos a nivel de palabra mediante TF-IDF e Información Mutua es una técnica bien asentada. Se han conseguido indicios de diferencias en el discurso. Los resultados más interesantes se han observado en el análisis entre partidos, mientras que el análisis por fechas ha mostrado resultados más triviales.

A pesar de que puede resultar tentador realizar el análisis con técnicas más sofisticadas, no lo plantearía para futuros trabajos inmediatamente. Utilizar técnicas complejas complica la labor de comunicación de resultados, objetivo principal de este trabajo. Este hecho se comprobado con el concepto de relevancia por ejemplo. Simplemente analizando la desviación en cada uno de los grupos, hubiese dado resultados más sencillos de explicar y probablemente mejores.

En los objetivos del trabajo se puso el foco en realizar el análisis en dos ejes: fechas y partidos. Este planteamiento ha añadido complejidad al análisis que no se ha sabido lidiar, concluyendo en un análisis por separado. Creo fundamental en posteriores trabajos redefinir los objetivos y focalizarse en alguno de los dos ejes.

Por último, hay indicios que el desbalanceado del dataset ha sesgado profundamente el análisis. En futuros análisis hay que buscar técnicas para lidiar con esto.

5.3. Visualización

Esta es la parte más relevante de este trabajo y a su vez en la que más dificultades me he encontrado. Primero comentaré la herramienta utilizada para la visualización (Vega Altair), para posteriormente centrarme en la visualización. Su enfoque declarativo considero que es muy útil, permitiendo muy buenos resultados con muy pocas líneas de código. Pero a lo largo de esta práctica he tenido varios inconvenientes que deben ser tenidos en cuenta para el futuro.

En primer lugar puedo destacar su mala escalabilidad, con grandes conjuntos de datos la transformaciones son muy lentas y en muchas ocasiones el navegador directamente deja de funcionar sin previo aviso. Es de vital importancia transformar correctamente los datos para

minimizar el numero de filas del dataset. Si sobrepasamos el límite por defecto empezaremos a encontrar los problemas planteados.

Por otro lado hay que tratar su documentación. Desde un punto de vista introductorio la documentación oficial es muy buena. Pero si queremos realizar alguna tarea muy específica encontrar cómo realizarla puede suponer un horror. Considero que la documentación a nivel de clase no es la más adecuada.

Por último, la librería tiene algunos comportamientos impredecibles y bugs. Animo al lector a tratar de concatenar horizontalmente múltiples objetos `Row` complejos y tratar de que queden alineados.

En cuanto a la visualización realizada, a pesar de que los objetivos se han cumplido parcialmente (muestra indicios de diferencias en el lenguaje y da una visión superficial de estas diferencias), desde mi punto de vista no cumple la labor informativa que pretendía. Principalmente, este hecho ha sido provocado por un análisis complejo y con gran cantidad de datos que no se ha conseguido simplificar. Sin embargo, estoy contento con la mejora desde la primera iteración. Reducir complejidad, tratar de mejorar algunos aspectos estéticos y solucionar defectos de diseño creo que ha aportado cierta profesionalidad al gráfico. En futuras mejoras, para mejorar la visualización habría que replantear el enfoque, y redefinir objetivos y técnicas de análisis

6. Herramientas utilizadas

Este trabajo se ha realizado completamente en python en un entorno de cuadernos Jupyter. Además de Pandas y Numpy (típicas en las labores de análisis de datos) las librerías utilizadas en cada fase del trabajo han sido:

- requests y BeautifulSoup: descarga y limpieza de documentos HTML.
- NLTK: preprocessado (stopwords).
- Scikit-Learn: análisis (TF-IDF, Información mutua).
- Altair: visualización (excepto nubes de palabras).
- wordcloud: nubes de palabras.
- cv2, PIL, y matplotlib: recortado, extracción de máscaras, gestión e impresión de imágenes para las nubes de palabras.