

Uso de la media prescripción en causas de violaciones a DDHH en dictadura: evidencias desde la minería de texto

**Peritaje para CDH-35-2021/011 Caso Vega González y otros Vs.
Chile**

Daniel Giménez

24 de enero de 2023

 danielgimenez.me

 daniel.gimenez@me.com

 +56 9 7580 1985

Índice

Presentación	2
1. Las causas y sentencias analizadas	5
1.1 Panorama general	5
1.2 Procesamiento y organización de los datos de texto para el análisis	6
2. Minería del texto de las sentencias	11
2.1 Primera impresión visual: las nubes de palabras	11
2.2 tf y tf-idf	14
2.2.1 El tf	14
2.2.2 El tf-idf	16
2.3 Estructuras de relaciones	18
Conclusiones	22
Referencias bibliográficas	23

Presentación

El presente documento ha sido elaborado a solicitud de la parte demandante en la causa CDH-35-2021/011 Caso Vega González y otros Vs. Chile. Entrega un diagnóstico realizado con las herramientas y técnicas de la ciencia y analítica de datos sobre la aplicación de la media prescripción (o “prescripción parcial”) contemplada en el artículo 103 del Código Penal de Chile en las sentencias de la Corte Suprema en casos judiciales relacionados con violaciones sistemáticas a los Derechos Humanos cometidas por la dictadura cívico-militar de 1973 a 1990 (de acá en adelante, sólo “la dictadura cívico-militar” o “la dictadura”).

Actualmente hay dos grandes conjuntos de técnicas de ciencia y analítica de datos: el de la minería de datos y el de la minería de texto (también llamada minería de datos de texto). Con el desarrollo y potenciamiento de las redes neuronales, son cada vez más avanzadas y de uso más común distintas técnicas de minería de imágenes, de audios y de video. Para el presente diagnóstico se evaluó la aplicación de las técnicas de todos estos campos. Sin embargo, el material disponible redujo las opciones:

1. No hay material audiovisual o gráfico de las resoluciones de la Corte Suprema. No es posible aplicarles, por tanto, minería de audio, de imágenes o de video.
2. La situación de las estadísticas judiciales generales en Chile es muy deficiente, y con mayor razón aún la de las estadísticas judiciales específicas sobre causas relacionadas con las violaciones a Derechos Humanos durante la dictadura. Para estadísticas judiciales globales existen dos fuentes: la Corporación Administrativa del Poder Judicial (CAPJ) que publica estadísticas menos que básicas en su sitio web: <https://www.pjud.cl/post/estadisticas>¹. La otra fuente son los anuarios de estadísticas judiciales del Instituto Nacional Estadísticas (INE), que, por lo que indican sus textos, se alimentan de las mismas bases de datos de la CAPJ, aunque tabulados con mayores y mejores desagregaciones y aperturas². Los anuarios del INE, sin embargo, se publican después del año concluido y con un gran rezago, a veces de casi un año completo. Por lo tanto, para tener acceso al anuario estadístico con las estadísticas judiciales de 2022 probablemente va a ser necesario esperar hasta el año 2024.
3. Si la situación es así de precaria con las estadísticas judiciales globales, la situación de las estadísticas judiciales específicas sobre causas de violaciones a los Derechos Humanos durante la dictadura es aún peor. En principio, no hay forma de tabular los datos publicados por la CAPJ o por el INE para obtener dichas estadísticas. No se pueden obtener tampoco por Ley de Transparencia solicitándola a alguna de las dos instituciones debido a que, primero, la CAPJ no está regida por ella, y, en el caso del

¹ Adicionalmente, la CAPJ ha puesto a disposición del público una REST API para hacer consultas a su base de datos estadísticos: <https://estadisticaservices.pjud.cl>. Facilita los mismos datos menos que básicos que se pueden obtener visitando el sitio web mencionado.

² Los anuarios del INE se pueden consultar en el siguiente enlace: <https://www.ine.gob.cl/estadisticas/sociales/seguridad-publica-y-justicia/estadisticas-judiciales>

INE, la normativa sólo hace exigible por parte de la ciudadanía la información efectivamente generada por las instituciones públicas; si la información no ha sido generada, la institución pública no está obligada a producirsela especialmente a quien la solicite y sólo porque la solicitó.

4. Las otras dos posibles fuentes, el Instituto Nacional de Derechos Humanos (INDH) y el Programa de Derechos Humanos (PDH) actualmente bajo dependencia del Ministerio de Justicia y Derechos Humanos (antes dependiente del Ministerio del Interior), no tienen ni presumiblemente producen estadística alguna. La página web del PDH le dedica tres párrafos a las causas judiciales³. Y con la página del INDH ocurre lo contrario: mucho texto, poca información.

En resumen, el estado de las estadísticas sobre causas judiciales de violaciones a los Derechos Humanos por parte de la dictadura es peor que precaria: ni siquiera se puede hablar de una situación por el simple hecho de que no existen datos. Esto hace inviable cualquier análisis estadístico o de minería de datos. Al desconocerse el universo total, cualquier inventario de las causas que se encuentre en conocimiento del equipo de la parte demandante carecería de validez estadística: no permitiría saber cuál es su representatividad ni tampoco calcular los niveles de confianza con los que cabría estimar para el universo total los resultados del procesamiento estadístico de las causas conocidas.

La única alternativa de análisis que deja esta situación es la de la minería de texto. Consiste en procesar, organizar y analizar documentos que contienen lenguaje natural. En principio, es indiferente si el contenido de los documentos proviene de la escritura o la transcripción de alocuciones habladas. Su propósito es identificar tendencias y patrones en las relaciones que presentan las palabras (o conjuntos de ellas) contenidas en esos textos específicos. A partir de la identificación de patrones y tendencias es posible construir modelos para clasificar palabras en grupos o tipos de documentos (modelamiento de tópicos). Pero, en principio, la sola identificación de patrones y tendencias permite determinar varias propiedades de los textos y las palabras usadas en ellas. El motor de búsquedas de Google se basa precisamente en estas herramientas de minería de texto: sabe qué páginas mostrar en su página de resultados cuando introducimos palabras para buscar precisamente porque ha procesado el contenido de texto de esos resultados y sabe, gracias a ello, que tienen los puntajes más altos para la búsqueda que realizamos.

En este documento se presentan los resultados de la aplicación de estas técnicas de minería de textos a sentencias judiciales de la Corte Suprema de Chile en causas de Derechos Humanos falladas entre los años 2004 y 2019. Puesto que la tesis de la parte demandante es que la aplicación de la “media prescripción” en las decisiones de la Corte Suprema ha tenido efectos jurídicos concretos, se analizan los textos de las sentencias comparando los de aquellas que mencionan la media prescripción (o sus análogos, como se explica más adelante) con los de las sentencias que no la mencionan. El objetivo de esta comparación es identificar la presencia de diferencias en los fallos finales que pudieran ser significativas.

³ Pueden verse en este enlace: <https://pdh.minjusticia.gob.cl/area-juridica/>

Para efectos de transparencia y posibilidad de validación del análisis mediante la reproducción de sus resultados, todos los documentos incluidos en el presente análisis y los códigos de programación (en lenguaje R) desarrollados para procesarlos se encuentran disponibles en un repositorio del github personal del autor: https://github.com/egoipse/peritaje_CDH-35-2021_011. El repositorio contiene el script que reproduce el presente informe, scritp que ejecuta gran parte de los procesamientos de datos necesarios para los resultados acá analizados.

Salvo cuando se indique explícitamente lo contrario en cada caso específico, la fuente de datos para los cuadros y gráficos incluidos en el presente documento es: *Elaboración propia a partir de los archivos de sentencias judiciales de la Corte Suprema contempladas en el presente análisis*.

1. Las causas y sentencias analizadas

1.1 Panorama general

Como se había adelantado en la presentación, para el presente diagnóstico se analizaron sentencias de la Corte Suprema emitidas entre los años 2004 y 2019. En total se consideraron 527 que corresponden a 341 causas distintas; es decir, por cada causa puede haber más de una sentencia analizada.

En cada documento se evalúa si contiene alguna referencia a media prescripción o cualquiera de sus derivados: prescripción gradual, prescripción incompleta, artículo 103. Para efectos del análisis de minería de texto presentado acá, no importa en qué términos o contextos se hace referencia a ellos; podría aparecer una mención para acoger su aplicación o para rechazarla. Para el presente análisis es relevante su mera aparición, sin importar cómo y a propósito de qué.

El Cuadro 1 muestra el resumen de las sentencias analizadas, de las causas que representan y si hacen o no referencia a la media prescripción.

Cuadro 1. Cantidad de sentencias y causas analizadas

Según sentencias que mencionan y no mencionan a la media prescripción o sus derivados

	Sentencias	Causas
No mencionan	187	105
Mencionan	340	236
Total	527	341

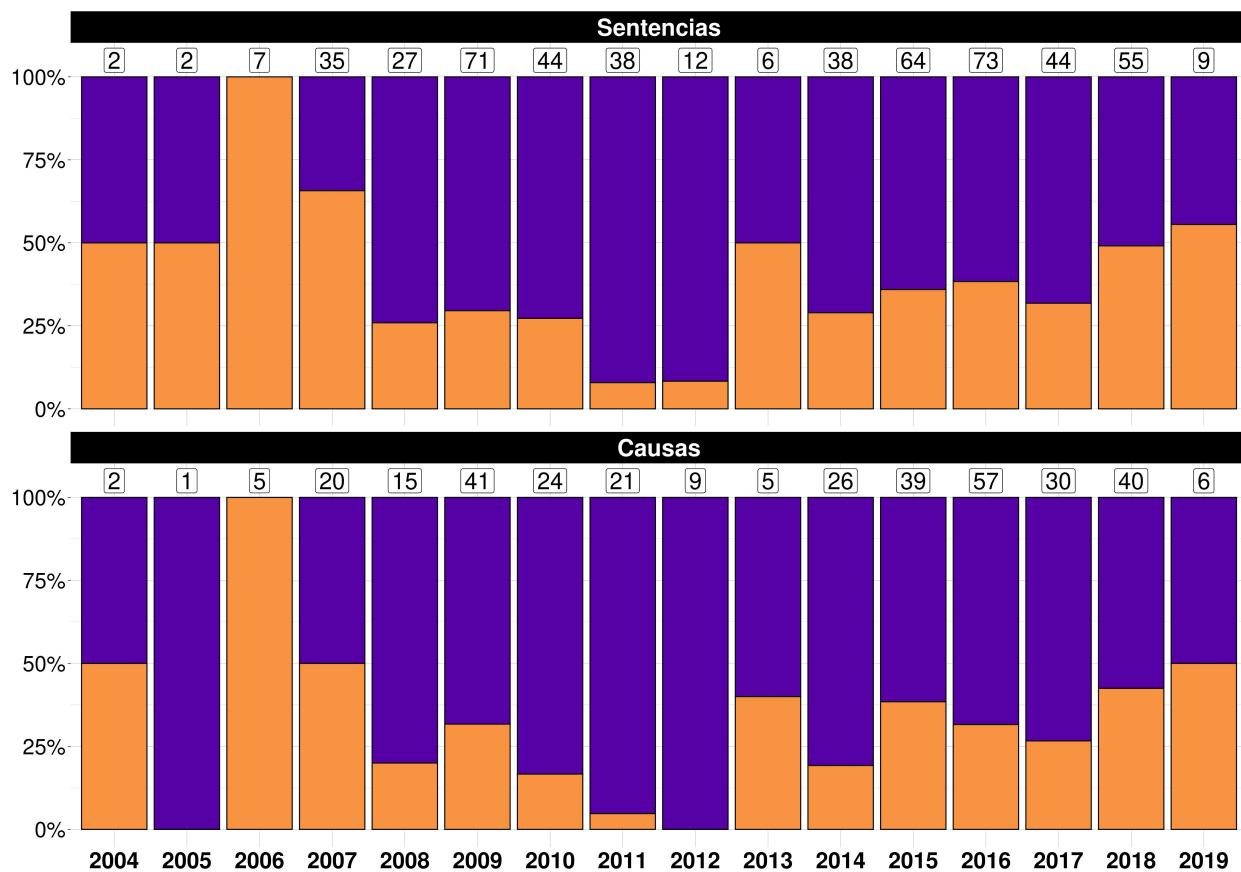
En el Gráfico 1 es posible ver cómo se distribuyen las causas y sentencias analizadas por año. Los números mostrados en recuadros encima de las barras corresponden al total sumado para cada año: en el caso de las sentencias, la suma de las que mencionan y las que no mencionan a la media prescripción y sus derivados. Los años mostrados para las causas corresponden al de la sentencia más reciente de todas las analizadas para cada causa. El número encima de las barras de las causas, por tanto, corresponde al total de causas con sentencias más recientes en cada año.

Como se puede apreciar en el gráfico, casi la mitad de las sentencias se concentran en cuatro años: 2016, 2009, 2015 y 2018 (en ese orden). Ocurre lo mismo con el volumen de causas.

Gráfico 1. Evolución de causas y sentencias analizadas

Por año y según mención de media prescripción o sus derivados

¿Mencionan la media prescripción? Mencionan No mencionan



1.2 Procesamiento y organización de los datos de texto para el análisis

La minería de texto se realizó a través del lenguaje de programación estadística R⁴. Para poder ejecutarla, se siguieron estos procesamientos:

1. **Organizar y homogeneizar archivos.** Todos los textos de las sentencias fueron reunidos por el equipo que demanda de distintas fuentes, algunas probablemente muy antiguas, pues algunos archivos se encontraban en formatos hoy en desuso o que no se usan para documentos, como .doc o .dot. Todos fueron convertidos a formato .pdf.
2. **Importar textos.** Para el análisis, es necesario acceder al contenido de texto de los archivos y almacenarlo en memoria para que el software pueda leerlo.

⁴ Como se adelantaba en la presentación, los códigos y scripts para generar los datos, los indicadores, los cuadros y los gráficos se encuentran disponibles en un repositorio en la cuenta de github del autor. Todas las técnicas aplicadas en este documento se encuentran explicadas detalladamente en Silge y Robinson (2017)

-
3. **Limpieza general.** Puesto que las sentencias provenían de distintas fuentes y épocas, los archivos no presentaban la misma estructura ni patrones de contenido. El criterio general de limpieza para uniformar el contenido fue eliminar mediante código de programación (con la herramienta de las expresiones regulares) todo el texto anterior a la presentación de la sentencia. El patrón general de inicio del contenido es la frase “Vistos:” (en distintas variantes, desde “VISTOS:” hasta “V I S T O S:”). Todo el texto desde esa frase introductoria hasta el final fue considerada para el análisis, incluida la nómina de las autoridades judiciales firmantes de cada sentencia, que generalmente queda registrada al final del documento.
 4. **Limpieza específica para minería de texto.** Para obtener resultados comprensibles y/o relevantes, las técnicas aplicadas requieren que se elimine contenidos muy específicos de los textos incluidos en el análisis:
 - Números;
 - Signos de puntuación;
 - Caracteres especiales; y
 - *Stop words*: aquellas palabras de uso común que aportan gramática pero no significado, como los artículos, los conectivos o los adverbios simples. Existen distintos diccionarios de *stop words* disponibles para el castellano. Para el presente procesamiento se aplicó el diccionario Snowball⁵
 5. **Normalización del texto.** Para una computadora que procesa texto, las palabras “Hola” y “hola” son distintas. Con el objeto de reducir la presencia de distintas versiones de una misma palabra, se suele aplicar alguna normalización de formatos, contenidos y reglas ortográficas. La más común consiste en transformar todas las palabras a mayúsculas o minúsculas. En textos de lenguaje cotidiano, como una conversación casual por redes sociales o chats, además se presuponen elevados “accidentes ortográficos” que van a multiplicar erróneamente las versiones de una misma palabra, como, por ejemplo, las de “normalización” y “normalizacion”. En estos casos también se sustituyen todas las vocales con tilde o diéresis por sus equivalentes puras, al igual que las “ñ” por “n”. Para este análisis, por provenir el texto de fuentes oficiales de elevada formalidad y solemnidad, se espera lenguaje con accidentes ortográficos reducidos. Por lo tanto, la única normalización aplicada fue la conversión de todas las palabras a minúsculas.

Estos son los pasos estandarizados mínimos para preparar los datos para minería de texto. Los siguientes pasos, todos aplicados en el presente análisis, son opcionales o, siendo necesarios u obligatorios, requieren decisiones que se sustentan en los objetivos del análisis o los criterios del investigador:

1. **Limpieza de palabras que no corresponden a la categoría de *stop words* pero que**

⁵ Su listado de *stop words* puede consultarse en este enlace: <http://snowball.tartarus.org/algorithms/spanish/stop.txt>

para los textos analizados en cada caso pueden generar más ruido que luces. Por regla general, la limpieza de estas palabras se realiza *a posteri*, una vez visualizados los primeros resultados de la aplicación de las distintas técnicas. Para el presente análisis, el listado de palabras que se sumaron a las *stop words* para ser eliminadas puede consultarse en el script de generación del documento. Incluyen algunas de las siguientes: “vistos”, “cumplimiento”, “santiago”, “dos”, “mil”, “autos”, “rol”, “corte”, “apelaciones”, “don”, “doña”, “sentencia”, “dictada”, “ministro”, “fuero”, “fj”, “fs”, “fojas”, “jf”, “nº”, “a”, “ed”, “artículo”, “artículos”.

2. **Definición de unidad de análisis.** Por inclinación natural, es común considerar a la palabra como la unidad mínima de un texto. Sin embargo, las palabras sueltas (que en ciencia de datos de texto se llaman “términos” o “unigramas”) suelen carecer de contexto, y el contexto aporta significado. Por ejemplo, el análisis de un texto que tome a la palabra como unidad, puede terminar concluyendo por la aparición recurrente de la palabra “feliz” que se trata de un texto que presenta abundante sentimiento positivo⁶. No obstante, si ese análisis no considera la aparición de dicha palabra en una frecuencia significativa en compañía de otras, que le imprimen el significado directamente contrapuesto (“poco feliz”, “no soy feliz”, “quiero ser feliz y no puedo”, y así sucesivamente) llega a conclusiones equivocadas. Literalmente, los unigramas son palabras sacadas de contexto. La solución para evitar esto es trabajar con unidades de análisis de dos o más palabras: bigramas (dos palabras), trigramas (tres palabras), tetragramas (cuatro palabras), y así sucesivamente para n-gramas. A medida que se aumenta el número de palabras en la unidad de análisis aumenta también el significado de cada unidad de análisis, pero se reduce la probabilidad de aparición frecuente en el texto y, por lo tanto, también la probabilidad de encontrar patrones, tendencias y relaciones relevantes. Para una buena minería de texto la clave está en encontrar la unidad que equilibre significado de palabras con frecuencias de aparición suficientes para arribar a conclusiones relevantes. Por regla general, ese equilibrio se encuentra en el bigrama. Por esta razón, se definió al bigrama como la unidad de análisis para el presente diagnóstico. Como se indicaba, los bigramas son combinaciones de dos palabras que aparecen en un texto. Las combinaciones son de todas las palabras con su antecesora y su sucesora; en la oración “Hola vecino mío” hay dos bigramas: “Hola vecino” y “vecino mío”⁷.
3. **Reducción de variedades.** Muchos n-gramas pueden tener formulaciones distintas, lo que es fuente de importantes distorsiones en el resultado de la aplicación de las técnicas de minería. Por ejemplo, un texto dedicado al estudio del boxeo puede hablar indistintamente de Cassius Clay y Muhammad Ali. Estas múltiples variedades del mismo objeto o sujeto referido pueden dar una idea equivocada de la importancia que el texto en cuestión les da, puesto que las frecuencias de aparición de ambos bigramas se dividirían entre todas esas variedades. Si es importante para el objetivo del análisis

⁶ El “análisis de sentimientos” es una de las técnicas más difundidas de la minería de textos.

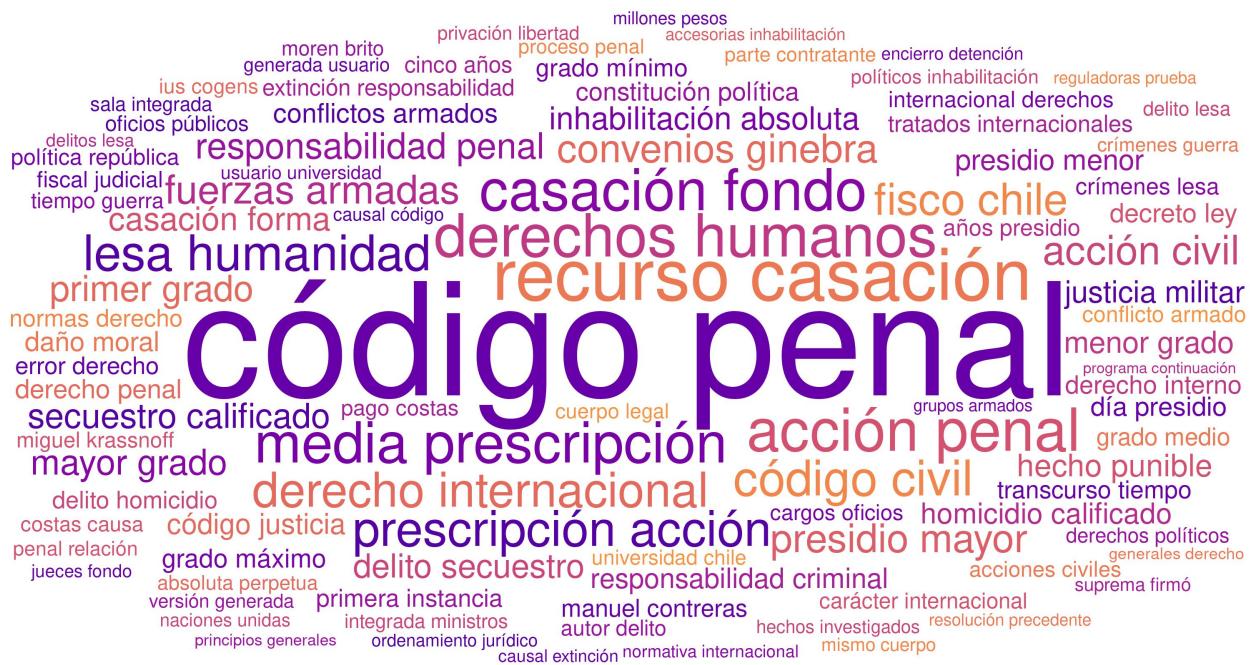
⁷ Como para minería de texto se eliminan las *stop words*, bigramas como “deja desear” deben interpretarse como “deja que desear”.

diagnosticar la importancia del sujeto u objeto referido y no la de los signos que lo refieren, pueden reducirse las distintas variedades transformándolas en una sola. En este caso, se puede cambiar en la base de datos del texto cada “Cassius Clay” por “Muhammad Ali”; así, este último bigrama sumará la frecuencia de aparición de ambos, lo que permitiría ponderar el peso real de la persona “Muhammad Ali” antes que el de los distintos signos que lo identifican en el texto⁸. En el presente análisis se aplicó esta operación para todas las potenciales variedades de “media prescripción”: atendiendo a que se la conoce también como “prescripción gradual” o “prescripción incompleta” (Fernández y Sferrazza, 2009; Parra, 2019), esos bigramas fueron transformados en “media prescripción”. Lo mismo se hizo con el texto “artículo 103”, que es la norma que la establece. También se aplicó esta reducción a los nombres de los imputados: cada bigrama “krassnoff martchenko” fue transformado en “miguel krassnoff”, “contreras sepúlveda” en “manuel contreras”, y así sucesivamente. Y lo mismo para los nombres de las víctimas: “José Tohá” en “Tohá González”; “Hernán Carrasco” en “Carrasco Vasquez”; “Dagoberto San” y “Luis Dagoberto” en “Martin Vergara”; “Recaredo Ignacio” e “Ignacio Valenzuela” en “Valenzuela Pohorecky”, y así sucesivamente.

4. **Limpieza final.** Una vez elegida la unidad de análisis es posible empezar a aplicar las distintas técnicas de minería. Al obtener los primeros resultados puede evaluarse la necesidad de realizar nueva limpieza del texto, excluyendo aquellos n-gramas que, si bien tienen presencia y peso significativos en los documentos, aportan poco al análisis o dificultan la interpretación de las relaciones precisamente porque su peso las distorsiona o impide ver las relaciones entre n-gramas de interés pero menos frecuentes. Eso ocurre con varios bigramas de las sentencias analizadas para este diagnóstico. Obsérvese la nube de palabras del Gráfico 2: la gran frecuencia de apariciones de “código penal” “recurso casación” y “casación fondo” no aporta información relevante, pues, tratándose de sentencias respecto a recursos de casación en su gran mayoría y en referencia a delitos tipificados en el código penal, cabe esperar que registren una frecuencia significativa de menciones; esto, por lo tanto, no nos aporta información nueva o relevante. En este caso, además, sus frecuencias son tan significativas que eclipsan a los otros bigramas y no es posible distinguir por ello si son importantes o no. En vista de esta situación, se excluyeron del cuerpo de texto final todos los bigramas “código penal” y los que contienen la palabra “casación”. Se aplicó el mismo criterio para los nombres de los y las integrantes de la Corte Suprema, que registran altas frecuencias por estar firmados todos los documentos por ellos y ellas.

⁸ Cabe la posibilidad opuesta: que el objetivo del análisis sea determinar con qué universos de signos se asocian “Cassius Clay” o “Muhammad Ali” y si hay diferencias entre ellos. En este caso, el objetivo del análisis impone no reducir las variedades.

Gráfico 2. Nube de palabras de bigramas para todas las sentencias, todos los bigramas



El procesamiento de datos realizado para el presente diagnóstico siguió todos estos pasos. Sus resultados se presentan en el siguiente capítulo.

2. Minería del texto de las sentencias

2.1 Primera impresión visual: las nubes de palabras

Un primer panorama acerca de un texto puede obtenerse de un dato muy sencillo: la frecuencia de aparición de sus n-gramas. Los recursos gráficos habituales, como las barras o las tortas, pueden representar bien y con precisión las distintas frecuencias o sus proporciones. En minería de texto, sin embargo, se usan también otras representaciones gráficas. Una en particular ayuda significativamente a visualizar dicha frecuencia y hacer comparaciones preliminares con facilidad: la nube de palabras.

Una nube de palabras muestra los n-gramas con tamaños y posiciones que representan su frecuencia de aparición en el texto. El n-grama más frecuente se presenta en el centro del gráfico y con el tamaño más grande. A medida que la frecuencia de aparición de los n-gramas va bajando, también disminuye su tamaño y su centralidad en el gráfico. Los n-gramas periféricos y pequeños registran la menor frecuencia de aparición⁹.

La distribución de n-gramas en gráficos de nubes de palabras puede aportar una primera idea visual de las diferencias entre distintos textos. En los casos de las sentencias analizadas, se pueden apreciar sus estructuras en los gráficos 3, 4 y 5, correspondientes al total de las sentencias, a las sentencias que mencionan a la media prescripción o sus derivados y a las sentencias que no las mencionan, respectivamente. Para cada caso, se incluyen en el gráfico los 100 bigramas con frecuencias más altas.

Tratándose de textos referidos a los mismos tópicos (y uno de ellos siendo la suma de los otros dos), sus nubes de palabras no presentan diferencias muy notorias; pero hay algunas diferencias de matices que sí pueden aportar a entender las inclinaciones de las sentencias. Por ejemplo, “derechos humanos” es el bigrama más frecuente si consideramos todas las sentencias o las que no hacen referencia a la media prescripción, pero en las que la mencionan pasa al segundo lugar después de, precisamente, “media prescripción”.

Esta situación se puede verificar numéricamente con un indicador (tf) que se presenta en la siguiente sección. Pero de la inspección visual de estos gráficos y en base a los antecedentes sobre el peso de cada tipo de sentencia en el total (de acuerdo al Cuadro 1, las que mencionan la media prescripción representan el doble de las que no) se puede inferir que la mención a Derechos Humanos en las sentencias que mencionan la media prescripción cae de forma importante.

⁹ El algoritmo que construye el gráfico puede intentar acomodar los n-gramas con frecuencias menores en los pequeños espacios que quedan entre los de mayor tamaño; por lo tanto, podrían quedar en una posición muy central. Se sobreentiende por su tamaño que se encuentran entre los menos frecuentes, y no cabe interpretar su importancia por la posición central que ocupan en el gráfico.

Gráfico 3. Nube de palabras de bigramas para todas las sentencias



Comparando directamente los gráficos que mencionan y no mencionan a la media prescripción (4 y 5, respectivamente) se pueden apreciar bigramas que en unos casos son más centrales y en otros más periféricos, y viceversa. Por ejemplo, “lesa humanidad”, “código civil” y “derecho internacional” son más periféricos en las sentencias que mencionan a la media prescripción, mientras que son absolutamente centrales en los que no la mencionan. Ocupan el lugar de, precisamente, los bigramas “media prescripción”, que, como es obvio, no aparece en estas últimas sentencias, y “acción penal”, muy central en el primer grupo.

Otro matiz importante es la mayor centralidad de bigramas ligados a las acciones civiles en las sentencias que no mencionan a la media prescripción: “código civil”, “daño moral”, “acción civil”, “fisco chile”. En las que sí la mencionan, adquieren mayor centralidad los bigramas que refieren a acciones penales: “prescripción acción”, “acción penal”, “responsabilidad penal”. Todos ellos están presentes en el primer grupo de sentencias, pero en posiciones más periféricas.

En tercer lugar, en las sentencias que mencionan a la media prescripción es posible apreciar con claridad los nombres de tres oficiales con alto rango en los aparatos represivos de la dictadura: “miguel krassnoff”, “manuel contreras” y “moren brito”. No ocupan un lugar central, no tienen un tamaño muy visible, pero ahí están. En las sentencias que no mencionan a la media prescripción se puede apreciar con toda claridad los bigramas “miguel krassnoff” y “manuel contreras”, pero no “moren brito”.

Gráfico 4. Nube de palabras de bigramas para sentencias que mencionan la media prescripción



Finalmente, las periferias en los tres gráficos son muy similares si se evalúan los bigramas que aparecen y sus frecuencias. Debido a su menor importancia visual, por esta vía no es posible determinar si hay cambios significativos en posiciones. Y no cabe esperar que tales cambios afecten significativamente a las relaciones entre bigramas (que se analizan en la sección subsiguiente). La imposibilidad de detectar en las zonas periféricas de los gráficos patrones relevantes que marquen diferencias visualmente significativas entre los distintos grupos de textos indica que el punto de corte (los 100 bigramas más frecuentes por grupo) no es analíticamente inadecuado. Birgramas con menores apariciones en los textos no aportarían mejores pistas sobre la composición de los distintos tipos de texto y las diferencias entre ellos.

Gráfico 5. Nube de palabras de bigramas para sentencias que NO mencionan la media prescripción



2.2 tf y tf-idf

La nube de palabras de los bigramas aporta una primera perspectiva gráfica de las sentencias analizadas. Pero esta perspectiva se puede ampliar, completar y precisar con otras representaciones gráficas comparativas y, sobre todo, con distintos indicadores sobre la estructura que forman los n-gramas. Existen dos indicadores que ilustran algunas propiedades de un texto y que van a completar numéricamente el análisis gráfico presentado hasta ahora: el tf y el tf-idf.

2.2.1 El tf

El “tf” (“term-frequency”) es la proporción de veces que aparece un término (en el caso del presente diagnóstico, un bigrama) en relación a la frecuencia total de términos aparecidos en el texto¹⁰. Por ejemplo, el minicuento “El dinosaurio” de Augusto Monterroso está compuesto por un total de siete palabras o unigramas: “Cuando despertó, el dinosaurio todavía estaba allí.” El tf de la palabra “Cuando” se obtiene dividiendo el total de veces que aparece (1 vez) por el total de palabras que tiene el texto (7 palabras): 0,14. En este caso, como todas las palabras son únicas, todas tienen el mismo tf. Pero cambia con la variación de las frecuencias. Supóngase que el cuento en realidad dijera: “Cuando despertó allí, el dinosaurio todavía

¹⁰ En motores de búsqueda, especialmente el de Google, el tf es una medida alternativa a la simple distribución de frecuencia de palabras, que se conoce como “keyword density”.

estaba allí”. Ahora el texto tiene 8 palabras en total; “allí” alcanzaría un tf de 0,25 (2 apariciones en relación a un total de 8 palabras) y todas las otras un tf de 0,125.

Puesto que se trata de un indicador neto, que se calcula en relación a los totales de cada documento o grupo de documentos, es apto para comparaciones: un tf mayor indica mayor presencia de un n-grama en un texto que un tf menor (y viceversa). Además, es fácilmente interpretable (puede asumir valores entre 0 y 1) y no requiere transformaciones ulteriores (como normalizaciones, estandarizaciones o logarítmicas) para hacer posible la comparación.

Cuadro 2. Term Frequency (TF) y ranking de bigramas en función a su TF

Según sentencias que mencionan y no mencionan a la media prescripción o sus derivados

	Mencionan la media prescripción		No mencionan la media prescripción	
	Ranking	TF	Ranking	TF
media prescripción	1	0,002407	-	-
derechos humanos	2	0,001870	1	0,002140
acción penal	3	0,001825	5	0,001239
prescripción acción	4	0,001639	7	0,001055
derecho internacional	5	0,001494	4	0,001514
lesa humanidad	6	0,001366	2	0,001604
responsabilidad penal	7	0,001200	34	0,000539
convenios ginebra	8	0,001154	12	0,000881
código civil	9	0,001151	3	0,001527
Fuerzas armadas	10	0,001142	15	0,000800
presidio mayor	11	0,001051	11	0,000891
fisco chile	12	0,001000	6	0,001072
secuestro calificado	13	0,000966	14	0,000834
acción civil	14	0,000943	13	0,000837
responsabilidad criminal	15	0,000907	-	-
inhabilitación absoluta	16	0,000848	10	0,000894
mayor grado	17	0,000831	17	0,000774
primer grado	18	0,000817	9	0,000941
justicia militar	19	0,000763	-	-
delito secuestro	20	0,000754	8	0,001005
hecho punible	21	0,000754	30	0,000569
código justicia	22	0,000719	-	-
homicidio calificado	23	0,000703	16	0,000784
conflictos armados	24	0,000697	45	0,000479
menor grado	25	0,000692	33	0,000542
Promedio	—	0,001108	—	0,000996

El cálculo del tf para los bigramas de las sentencias nos aporta una primera medida cuantitativa de las diferencias entre el grupo que menciona a la “media prescripción” y el que no la menciona. El Cuadro 3 muestra un listado de los 25 bigramas más frecuentes en todos los textos, sus tf respectivos y la posición que ocupan en el ranking de frecuencias en cada grupo.

Los tf y sus rankings confirman lo que habían mostrado las nubes de palabras: algunos bigramas centrales en los textos de las sentencias que mencionan a la media prescripción pierden centralidad en las que no la mencionan. Y viceversa. Es significativo el caso de “lesa humanidad”, que presenta el sexto mayor TF de todos los bigramas del primer grupo de sentencias, pero está en el segundo lugar entre los bigramas del segundo grupo, el que no menciona a la “media prescripción” o sus derivados.

También se confirma la mayor relevancia de los bigramas relacionados con el proceso penal en el primer grupo que en el segundo: “acción penal”, “prescripción acción”, “responsabilidad penal”, “responsabilidad criminal”; y, al mismo tiempo, la menor relevancia de bigramas relacionados con la parte civil de las demandas: “código civil”, “fisco chile” y no alcanza a ingresar en el ranking de los 25 bigramas más mencionados el “daño moral”, muy relevante, como se pudo apreciar en las nubes de palabras, en las sentencias que no mencionan la media prescripción.

Finalmente, es destacable el hecho de que entre los 25 primeros bigramas de las sentencias que sí mencionan la media prescripción se encuentran algunos completamente marginales en el grupo de las que no la mencionan. Por ejemplo, “conflictos armados” se encuentra en el lugar 24 en el primer grupo, pero cae al 45 en el segundo; “hecho punible” cae del lugar 21 en el primer grupo al 30 en el segundo; y “responsabilidad penal”, que cae del 7 al 34.

2.2.2 El tf-idf

Un segundo indicador que puede mostrar diferencias entre ambos grupos de sentencias es el que se conoce como tf-idf. Consiste en la multiplicación del tf, ya descrito, por el *inverse document frequency* (idf). El idf, al igual que el tf, es un indicador sobre un término o n-grama en una colección de documentos. Se calcula como el logaritmo natural del cociente entre el número total de documentos del corpus o colección y el número de documentos del corpus o colección en que está presente el término o n-grama evaluado:

$$idf_{(n-grama)} = \ln \left(\frac{N^{\circ} \text{ documentos}}{N^{\circ} \text{ documentos con } n\text{-grama}} \right)$$

A diferencia del tf, que se puede calcular para un n-grama dentro de un documento con independencia del corpus de documentos en que se encuentre, el idf es un indicador que se mide en relación a dicho corpus. En otras palabras, el tf de un n-grama para un documento no va a variar si se lo está comparando con un grupo de documentos u otro: va a ser siempre el mismo. En cambio, el idf es uno si se evalúa en relación a un grupo de documentos y es otro si se lo evalúa en relación a un grupo distinto de documentos. Tómese como ejemplo hipotético el bigrama “espacio tiempo” para el libro *Sobre la Teoría de la Relatividad Especial y General* de Einstein. El tf de ese bigrama se calcula en relación al total de bigramas contenidos en ese libro. Por lo tanto, va a ser el mismo si se evalúa el libro en el marco de una colección de libros de física o en una colección de literatura latinoamericana. En cambio, el idf del bigrama va a ser distinto en relación a cada grupo. En relación al primer grupo va a ser significativamente más bajo que en relación al segundo grupo, porque una mayor cantidad de

libros de física va a contener el bigrama en comparación a la cantidad de libros de literatura latinoamericana que lo mencionarían.

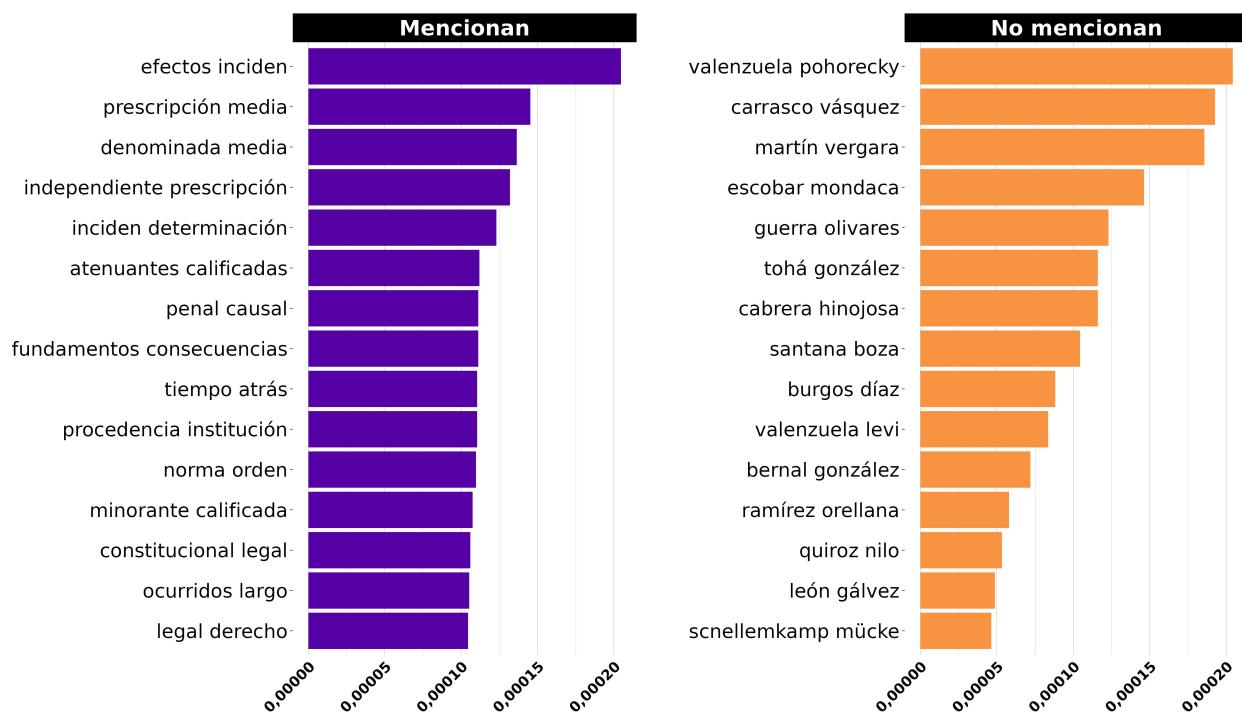
En este sentido, el idf identifica no tanto cuáles son los temas relevantes de un documento o texto, sino cuáles son los n-gramas que hacen más único a dicho documento en el marco de una colección o un corpus de varios documentos, qué términos son los que lo identifican (Silge y Robinson, 2017, Capítulo 3)¹¹.

En cualquier caso, el indicador aplicado al análisis de un texto no es el idf, aislado. Es el producto del idf y el tf, el tf-idf. Este producto pondera la frecuencia de aparición de un n-grama por su nivel de exclusividad en el corpus de documentos. Eso es lo que permite evaluar el aporte específico de términos de parte de cada documento (o grupo de documentos) al corpus completo. En otras palabras, el tf-idf entrega la importancia del documento para un término antes que la importancia del término para el documento, como hace el tf puro.

El Gráfico 6 muestra los 15 bigramas con mayor tf-idf para las sentencias que mencionan y las que no mencionan a la media prescripción.

Gráfico 6. TF-IDF de bigramas en sentencias de la Corte Suprema

Según mención y no mención de la media prescripción



Como se puede apreciar con toda claridad, las sentencias del segundo grupo, las que no mencionan la media prescripción, son relevantes y se distinguen por su referencia a los

¹¹ Por esta capacidad, el idf se usa en ocasiones para explorar un texto sin necesidad de eliminar las *stopwords*, que se suponen frecuentes en todos los textos y, por ello, registrarán un bajo idf, generalmente equivalente a 0. Mientras la limpieza “manual” de stopwords supone aplicar fuerza bruta, la minería que se aplica a los n-gramas con tf-idf mayor a 0 tiene un fundamento matemático.

nombres de las víctimas de violaciones a los Derechos Humanos en la dictadura. A excepción del bigrama “schnellenkamp mücke”, todos los otros corresponden a nombres de víctimas de la dictadura. Las sentencias del primer grupo, en cambio, son relevantes para bigramas del proceso criminal y muy en particular para tres tipos de términos:

1. Los relacionados con el artículo 103 del Código Penal: “prescripción media” o “denominada media”;
2. Los que tienen relación con beneficios para imputados o culpables, como “atenuantes calificadas” o “minorante cualificada”;
3. Y los que hacen referencia al derecho y el orden jurídico: “norma orden”, “constitución legal”, “legal derecho”.

2.3 Estructuras de relaciones

Otra forma de determinar si existen diferencias entre dos o más grupos de textos es analizando las relaciones que presentan los términos en los documentos analizados. Existen varias formas de evaluar estas relaciones. Sólo para fines ilustrativos, en este diagnóstico se utilizará la más intuitiva: los grafos.

Un gráfico de grafos muestra las conexiones (edges) que existen entre distintas entidades (nodes). En este caso, las entidades serán los términos que componen los bigramas y las conexiones son establecidas en función al peso o proporción de aparición de los términos de los bigramas en conjunto. Si, por ejemplo, en los textos existen los bigramas “derecho civil” y “código civil”, el gráfico mostrará el término “civil” conectado mediante líneas y flechas a “derecho” y “código”.

El algoritmo que crea los grafos agrupa entre sí los términos que registran mayor frecuencia de co-aparición. Se presentan en distintas zonas del gráfico los grupos de términos que presentan relación entre sí.

Para el presente análisis, además, el algoritmo clasificará los términos en grupos o clústers. El algoritmo evalúa las frecuencias y pesos de las relaciones para incluir cada término en un clúster. Su criterio matemático es maximizar las similitudes entre integrantes del mismo clúster y maximizar las diferencias entre los integrantes de los distintos clústers. Cada clúster está representado con colores: los términos que aparecen con el mismo color pueden considerarse con frecuencias y co-apariciones en el texto muy similares entre sí y distintas a las de los términos en otros colores que, por ello, quedan agrupados en otros clústers.

En resumen, los grafos nos entregan representación gráfica de tres propiedades de las relaciones entre términos en los textos:

1. La existencia misma de relaciones, en este caso, de co-aparición de los términos: si forman parte de un bigrama que aparece al menos una vez en los textos, entonces tienen alguna relación, que se representa con una línea y una flecha de conexión. La dirección de la flecha indica la posición en los bigramas: el término desde el que parte es el primero y el término al que apunta es el segundo del bigrama.

-
2. La distribución general de los términos: un racimo de términos conectados indica la presencia de relaciones directas o indirectas entre ellas. Siguiendo con el ejemplo mencionado, “derecho” y “código” podrán no aparecer conectados entre sí porque no existe un bigrama compuesto por ellos, pero aparecen en la misma red de conexiones, en el mismo racimo, por estar conectados ambos al término “civil”. Si en el texto estuviera el bigrama “demanda civil”, pertenecería al mismo racimo; lo mismo “derecho constitucional” (unido por el término “derecho”), y así sucesivamente. No habiendo relación directa entre términos (no habiendo conectores que los unan en los grafos), la pertenencia a un mismo racimo es indicio de proximidad semántica o de coincidencia de tópico en un texto dado. Si los mismos términos aparecen en racimos distintos en los grafos de otro texto, entonces hay una diferencia.
 3. La proximidad o distancia entre términos según su agrupación: la sola pertenencia a un mismo racimo de dos más términos podría dar una señal imprecisa de relaciones entre ellos. Por ello se configuró el algoritmo para que hiciera una clasificación o agrupación en clústers, identificados en los grafos siguientes por los distintos colores. Estos clústers permiten apreciar si todos los términos de un racimo pertenecen efectivamente a los mismos grupos o si pueden a su vez clasificarse en grupos distintos. Comparando dos documentos o grupos de documentos, por lo tanto, es posible establecer similitudes y diferencias entre ellos revisando no sólo las conexiones entre términos o los racimos a los que pertenecen, sino también los clústers en que han sido clasificados sus términos.

En los Gráficos 7 y 8 se puede ver la aplicación de estos tres criterios. Muestran, respectivamente, los grafos de los textos de las sentencias que mencionan a la media prescripción (o sus derivados) y los textos de las sentencias que no los mencionan. Se incluyen únicamente los términos de los 50 bigramas con mayor frecuencia en cada grupo de textos.

En términos generales, y pese a las diferencias en las formas que adquieren las distribuciones de los términos, se puede apreciar similitud en la estructura de relaciones entre ambos grupos de sentencias, similitud esperada tratándose de textos que refieren a lo mismo: sentencias judiciales en materias criminales por violaciones a los Derechos Humanos en dictadura.

No obstante, y como se mencionó a propósito de esto mismo al analizar las nubes de palabras, los matices indican algunas diferencias a destacar. Por ejemplo, el término “guerra” pertenece a distintos racimos en ambos grafos. En el grupo 1, de sentencias que mencionan a la media prescripción, está en el racimo que comparte con “transcurso” y “tiempo”, pero con la flecha apuntando desde “tiempo” hacia “guerra”. En el segundo grupo de textos, en cambio, está en un racimo que comparte con “crímenes”, “lesa” y “humanidad”. ¿Cómo se interpreta esto? En las sentencias del primer grupo hay alusión a “tiempos de guerra”, que es una de las tesis esgrimidas por la defensa de los responsables de las violaciones a los Derechos Humanos para justificarlas y solicitar que sean exculpados. En el segundo grupo, en cambio, es receptor de la flecha que lo conecta con “crímenes”, que da cuenta de la mención en varios de los textos de las sentencias a que en tiempos de guerra rige el derecho internacional humanitario, que tipifica y establece sanciones a los crímenes de guerra. Esta diferencia

muestra que las sentencias que mencionan a la media prescripción refieren a la tesis de las defensas de “los tiempos de guerra” mientras que las sentencias que no la mencionan son proclives a referir la tesis del juicio contra los crímenes de guerra. Es decir, el mismo término, por sus relaciones en los grafos, nos indica una diferencia en las inclinaciones de los textos.

Gráfico 7. Grafo de relaciones para sentencias que mencionan la media prescripción



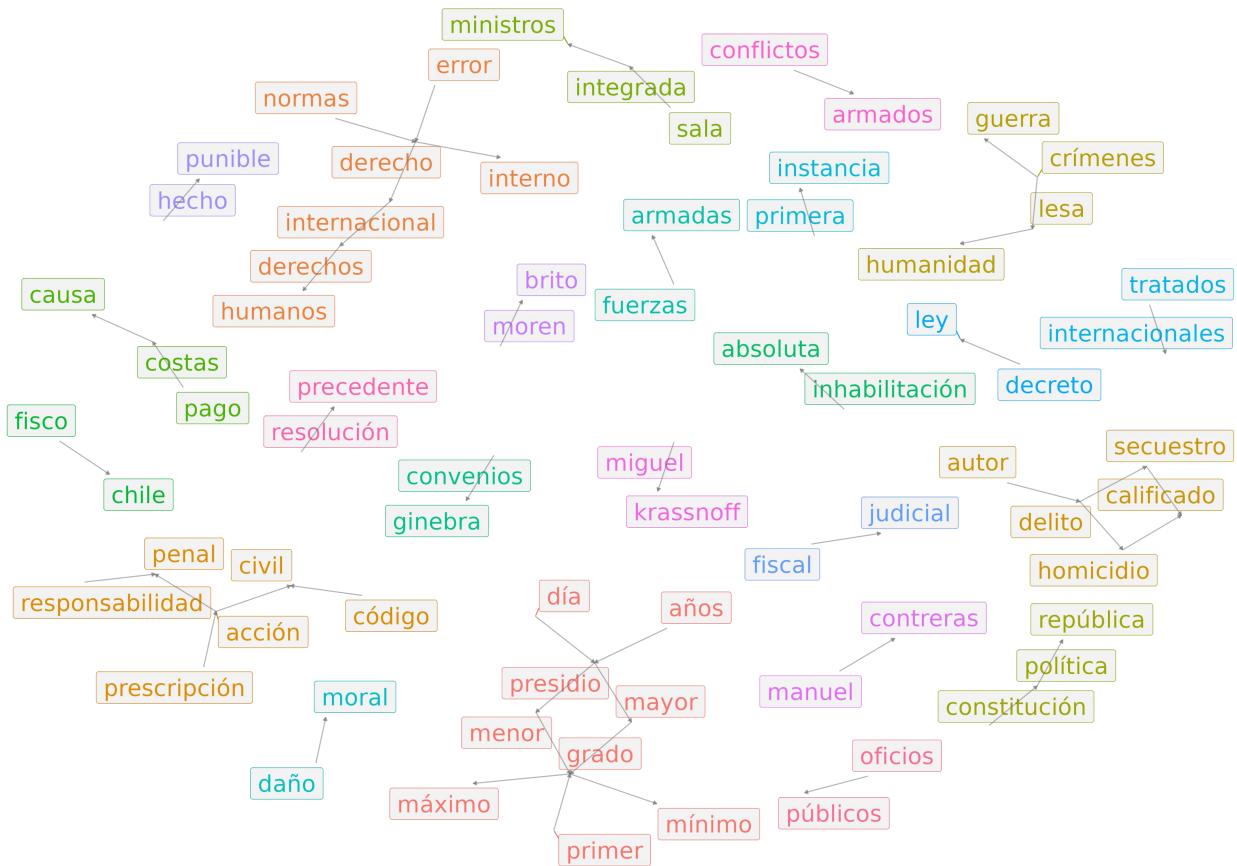
La misma situación aplica con varios otros términos:

- como se había adelantado, los términos “lesa” y “humanidad” en el segundo grupo de sentencias se encuentran en el racimo de “crímenes” y “guerra”; en el primer grupo se encuentran aislados;
- “derecho” y “penal”: están en el mismo racimo en el primer grupo de sentencias, pero en racimos distintos en el segundo grupo;
- En el primer grupo de sentencias, el término “responsabilidad” comparte racimo con “extinción”, “criminal” y “derecho”; en el segundo grupo está en el racimo de “penal”, “prescripción” y “acción”;

Finalmente, una importante diferencia a destacar entre ambos grafos es la existencia de un racimo muy grande en el primer grupo de sentencias, que une a “media prescripción”, “justicia militar”, “derechos humanos”, “extinción responsabilidad”, “derecho internacional”, “normas derecho” y otros términos menores más. No se encuentra un racimo así de grande en

el segundo grupo de sentencias. Varios de esos mismo términos están distribuidos en varios racimos. Curiosamente, esos distintos racimos en el segundo grupo de sentencias coinciden con distintos clústers dentro del racismo grande del primer grupo. Eso es indicio de que algún término puede haber generado la unión de ese gran racimo de forma más accidental y azarosa y no ser una relación estructural. Por lo tanto, debe tomarse con precaución el hecho de que tantos tópicos tan diversos hayan quedado conectados en una misma red.

Gráfico 8. Grafo de relaciones para sentencias que NO mencionan la media prescripción



Conclusiones

Pese a que no cabe esperar diferencias estructurales entre ambos grupos de sentencias por, como se adelantaba, tratar sus textos de un mismo tópico (causas judiciales sobre violaciones a Derechos Humanos en dictadura), la minería de texto que se aplicó en el presente diagnóstico ha permitido identificar algunas diferencias importantes:

1. En el grupo de sentencias que hacen referencia explícita a la media prescripción tienen mayor centralidad los bigramas que refieren al proceso penal; en las sentencias del segundo grupo, la centralidad la tienen “lesa humanidad” y los bigramas que refieren a proceso civil.
2. En relación a las sentencias del primer grupo, las del segundo grupo son especialmente relevantes para las víctimas; es decir, aquello que aporta este grupo al corpus total de texto y que no aporta el otro grupo son los bigramas correspondientes a sus nombres y/o apellidos. Por el otro lado, las sentencias del primer grupo son relevantes para tres tipos de bigramas:
 - Los relacionados con la media prescripción;
 - Los relacionados con beneficios para imputados o culpables;
 - los que hacen referencia al derecho y el orden jurídico.
3. En la comparación más estructural de las relaciones entre los distintos términos se aprecian diferencias que podrían ser más que simples matriciales. En un primer análisis, la más significativa es el papel que juega el término “guerra” en ambos grupos de sentencias.

Todas estas diferencias permiten concluir que la apelación a la media prescripción marca diferencias en las composiciones y estructuras de los textos de las sentencias.

Referencias bibliográficas

- Fernández, K. y Sferrazza, P. (2009). La aplicación de la prescripción gradual en casos de violaciones de derechos humanos. *Estudios constitucionales*, 7(1), 299-330.
- Parra, F. (2019). Los efectos de la media prescripción penal. *Revista de derecho (Concepción)*, 87(246), 247-285.
- Silge, J. y Robinson, D. (2017). *Text mining with R: a tidy approach*. O'Reilly.