Informe Final
Visualización y análisis de tópicos de las causas no contenciosas del Tribunal de Defensa de La Libre Competencia

Nombre	Lugar de Trabajo	Cargo en el proyecto
Priscila Rodríguez	Comisión Nacional de Energía	Jefa de Proyecto
Kiumarz Goharriz	Comisión Nacional de Energía	Visualización de Datos
Nicolás Torrealba	FOSIS	Científico y Analista de Datos
Eduardo Jimenez	Ministerio de Economía	Científico y Analista de Datos

1. Introducción

Durante los últimos años, si bien el número de causas no contenciosas[1] que el Tribunal de Defensa de la Libre Competencia (TDLC) ha debido analizar, no ha variado significativamente, si lo ha hecho el impacto mediático de sus sentencias, toda vez que las mismas han afectado de distinta forma la vida de las personas.

Casos emblemáticos como la colusión de las tres principales cadenas de farmacias en el precio de 220 medicamentos (2009), la colusión en el precio de los compresores de los refrigeradores entre Whirpool y su competidora Tecumseh (2012), la colusión de seis importantes navieras en el proceso de contratación de transporte marítimo de automóviles (aún en análisis) o bien, los famosos casos de colusión del mercado de los pollos, supermercados, confort y pañales, dan luces de la relevancia del impacto de las sentencias de este organismo en la sociedad.

No obstante, a pesar de la importancia de las decisiones del TDLC, no existen dentro de la página web del organismo módulos de visualización, que permitan de manera rápida tener una primera mirada de los mercados que son analizados con mayor frecuencia, las materias más relevantes, ni el número de veces en que el TDLC acoge o rechaza las solicitudes de los denunciantes en las causas no contenciosas, entre otros.

Todo lo anterior, sirvió de motivación para el trabajo que se explica a continuación.

[1] De acuerdo al artículo 817 del Código de Procedimiento Civil, *"son actos judiciales no contenciosos aquellos que según la ley requieren la intervención del juez y en que no se promueve contienda alguna entre partes*", por otro lado, de acuerdo al artículo 818, "*aunque los tribunales hayan de proceder en algunos de estos actos con conocimiento de causa, no es necesario que se les suministre este conocimiento con las solemnidades ordinarias de las pruebas judiciales. Así, pueden acreditarse los hechos pertinentes por medio de informaciones sumarias. Se entiende por información sumaria la prueba de cualquiera especie, rendida sin notificación ni NOTA intervención de contradictor y sin previo señalamiento de término probatorio*".

2. Tema o Problema a resolver

Dado el contexto anterior, el problema que se propone resolver en este proyecto, es el poder visualizar de mejor forma la información pública de los actos resolutorios de las causas no contenciosas del TDLC, para lo cual, se definen los siguientes objetivos:

- **Objetivo 1:** Lograr que todos aquellos interesados en la información que emite este organismo regulador, por ejemplo: la Fiscalía Nacional Económica; los organismos reguladores; académicos y algunos estudios de abogados; entre otros, puedan visualizar de forma rápida e ilustrativa las estadísticas más relevantes relacionadas con las causas no contenciosas.
- **Objetivo 2**: Con la ayuda de lenguaje para *text mining*, analizar en profundidad los actos resolutorios de las causas no contenciosas y ver si es posible que las herramientas de *machine learning* entregadas en el Diplomado, nos permitan clasificar esta información en tópicos que tengan sentido de tal forma que, a través de un *dashboard* programado en *Shiny* y con diferentes visualizaciones, un interesado pueda entender rápidamente de que se trata un determinado documento, junto con tener un buen y rápido análisis del contenido.

Las acciones que se tomarán para el logro de los anteriores objetivos son:

- la creación de una plataforma en Shiny para la visualización de las estadísticas más relevantes (objetivo 1) y mostrar la nube de palabras que permita inferir rápidamente de que se trata un determinado documento (objetivo 2)
- el análisis de *text mining* y el uso de *Latent Dirichlet Allocation* (LDA), para realizar la clasificación de los tópicos por documento.

Cabe señalar que un tercer objetivo, mucho más ambicioso, por cierto, y que va más allá del alncance de este proyecto, es predecir en base a los expedientes de las causas no contenciosas de los último diez años, la probabilidad de que una nueva causa se acepte o se rechace. Este proyecto queda en carpeta para realizarlo en una fase posterior a la finalización del diplomado.

3. Características del Dataset

La "base de datos" utilizada en este proyecto está compuesta por dos partes. En primer lugar, tenemos aquella que contiene la identificación y caracterización de 66 Causas no Contenciosas del TDLC, con las siguientes variables de caracterización:

(a) **Categoría**: Identifica el tipo de causa entre: (i) Resolución, (ii) Informe; y (iii) Instrucciones de carácter general.
(b) **Link**: Dirección URL con la ubicación de todas las causas dentro de la página del TDLC. Esta información permite descargar los documentos de las causas, que representan la información principal para el análisis de minería de texto.
(c) **Nombre**: Identificador único con el tipo de causa con su año y numeración.
(d) **Causa**: Identificador único con numeración administrativa de cada causa.

(e) **Conclusión**: Variable categórica construida a partir del análisis de cada causa. Identifica el tipo de conclusión resultante del trabajo del TDLC. . La variable tiene 11 categorías.
(f) **Mercado**: Identifica el tipo de mercado que es tratado en cada causa. La variable tiene 17 categorías, incluyendo una definida como "Otros".
(g) **Materia**: Identifica el tipo de materia tratada en cada causa. La variable tiene 11 categorías.

En segundo lugar, se tiene los documentos finales con la información de cada causa no contenciosa del TDLC. Los documentos contienen distintas secciones dependiendo del tipo la categoría del documento:
(a) **Resolución**: Incluye una parte expositiva, una parte considerativa y una parte resolutiva.
(b) **Informe**: Incluye una parte expositiva, una parte considerativa y las conclusiones del Tribunal.
(c) **Instrucción de Carácter General**: Incluye información de los antecedentes vistos durante el proceso, información de las consideraciones tenidas en relación con los antecedentes, y el detalle de las instrucciones de carácter general impartidas por el Tribunal.

En este proyecto, la información de los documentos es analizada en forma integral como un único texto por cada causa no contenciosa.

4. Análisis Exploratorio y Transformaciones

4.1. Análisis Exploratorio

El análisis exploratorio se enfocó en una revisión preliminar de los textos de las causas no contenciosas del TDLC. Utilizando las direcciones URL se descargan archivos .PDF con los documentos. Los que luego son cargados en R bajo distintos formatos, para luego realizar procesos de limpieza y análisis de la información contenida en los documentos como un grupo y en forma individual.

Se utilizaron dos métodos para aplicar técnicas de limpieza de los textos: en primer lugar, se armó un "*Corpus*" con todos los documentos y se utilizaron funciones de limpieza al interior de este corpus. Luego se optó por individualizar cada documento como un "*character vector*", y se utilizaron funciones generales de edición de textos. En este último caso se construye finalmente un "*data frame*" para poder gestionar los documentos como un conjunto.

Las técnicas de limpieza consideradas son las utilizadas de manera estándar para análisis de texto y son las siguientes:

(a) Se eliminan puntuaciones
(b) Se eliminan números

(c) Se deja el texto en letras minúsculas
(d) Se eliminan "*stopwords*" (palabras irrelevantes)
(e) Se aplica "*stemming*" (dejar las raíces de las palabras)
(f) Se reducen los espacios en blanco
(g) Se eliminan los acentos.

Luego, con los documentos editados, se construyen una matriz de Documentos-Términos para hacer análisis de "Bolsas de palabras". Se construyen matrices con monogramas y bi-gramas. Se construyen "Nubes de palabras" del Corpus como un todo, además de cada documento en forma individual.

Del análisis exploratorio realizado se pudieron sacar las siguientes conclusiones:

- Los documentos considerados en el proyecto presentan una estructura común y relativamente ordenada, que permitiría realizar análisis de minería de texto en mayor profundidad para extraer información relevante.
- Debido a que todos los documentos son de carácter legal, y uno de los objetivos del proyecto es poder caracterizar cada documento para diferenciarlo del resto, es que se reconoce la necesidad de profundizar en la identificación y limpieza de "palabras irrelevantes", para mejorar la calidad de los documentos.
- Se reconoce la necesidad de profundizar en la aplicación de técnicas de "*stemming*" y/o "*lemmatization*", para poder agrupar palabras comunes y así poder reflejar de manera más adecuada su frecuencia.
- Desde un punto de vista técnico, se opta por eliminar los acentos de las palabras para mejor visualización de la información con técnicas como la "nube de palabras".

4.2. Transformaciones

El proceso de transformar los textos del TDLC tiene como objetivo obtener y procesar estos documentos para posteriormente hacer análisis de minería de texto. Este proceso cuenta con dos secciones. En la primera se descargan los archivos desde direcciones URL y se leen en R para su procesamiento. En la segunda sección se realiza edición, limpieza y almacenamiento de la información, para su posterior uso en procesos de **minería de texto** y **visualización** de la información.

4.2.1. Carga de documentos en R

En esta sección, se crea un "*data frame"* con la información de 66 documentos desde un archivo Excel, con una variable que contiene las direcciones URL de cada documento considerado y otra variable con el nombre de cada documento. Se utiliza esta información para descargar archivos .PDF de cada documento. Luego se leen estos archivos como "*character vector"* dentro de R.

4.2.2. Edición y limpieza de textos

En esta sección se editan los textos de los documentos para que puedan ser utilizados en minería de texto. Para esto se crea un nuevo "*data frame"* que será utilizado como fuente para armar "*Corpus*" para el análisis. La edición y limpieza de los textos se hace sobre los "*character vector"* individuales, y luego el resultado se va almacenando en el "*data frame"*. Por último, se guardan resultados en archivos .rds, para su uso posterior en minería de texto y visualización.

- La primera edición consiste en convertir el "*character vector"* que originalmente se lee como una variable lista, en un solo "*string*" con el texto compilado. Luego se aplican las siguientes técnicas de limpieza de textos: (1) Se eliminan puntaciones, (2) Se eliminan números, (3) Se deja el texto en letras minúsculas, (4) Se eliminan "*stopwords*" (palabras irrelevantes), (5) Se reducen los espacios en blanco, y (6) Se eliminan los acentos.

- Luego se crea un "*Corpus*" con los documentos editados hasta este punto, para construir finalmente una matriz de Documentos-Términos. Con esta matriz se identifican dos nuevos grupos de "*stopwords*", que luego son aplicados junto con un criterio de truncamiento, para terminar con la limpieza de los documentos.
  A partir de la matriz se identifican las palabras más frecuentes a lo largo de todos los documentos ("*sparsity*"), y se aplica un criterio cualitativo para generar una lista de "*stopwords*" a partir de esta información. Se opta por utilizar un criterio de excluir palabras que aparecen en al menos un 95% por de los documentos considerados. Luego de excluir de este grupo aquellas palabras que se consideran relevantes, se obtiene una lista de "*stopwords*" de 41 palabras.
  Después, se vuelve a utilizar la matriz para identificar las palabras más frecuentes en el "*Corpus*" como un todo ("*frequency*"), y se vuelve a aplicar un criterio cualitativo para generar una lista de "*stopwords*". Al analizar las palabras más frecuentes en el 'corpus' se observa que existe una mayor cantidad de palabras 'relevantes' que se deberían mantener. Por esta razón, se opta por utilizar como criterio excluir las primeras 50 palabras más frecuentes, excluyendo de este grupo las palabras consideradas relevantes. De esto se obtiene una lista de "*stopwords*" de 36 palabras.

- Finalmente se aplican las siguientes técnicas de limpieza de textos: (i) Se eliminan "*stopwords*" definidas por "*sparsity*", (2) Se eliminan "*stopwords*" definidas por "*frequency*", (3) Se truncan las palabras hasta 5 caracteres, y (4) Se truncan las palabras hasta 6 caracteres. Cabe señalar que el criterio de truncamiento se determinó en base recomendaciones recibidas del equipo orientador del Diplomado. Los dos últimos criterios se aplican y se guardan por separado para el análisis posterior. Los resultados se almacenan en formato "*data frame"*.

- Con los "*data frame"* de los textos editados se construyen "*Document-Term Matrix*" (DTM), que posteriormente son utilizadas para hacer minería de texto. Se construyen 4 DTM, dos para cada nivel de truncamiento considerado. Un grupo de DTM donde los términos se consideran individualmente, y otro grupo de DTM donde se construyen bi-gramas, esto es, pares de palabras agrupadas.

- Por último, se almacena la información procesada en archivos .RDS, para su posterior uso en análisis de minería de texto y visualización.

5. Análisis de Data Science / Modelo / Evaluación

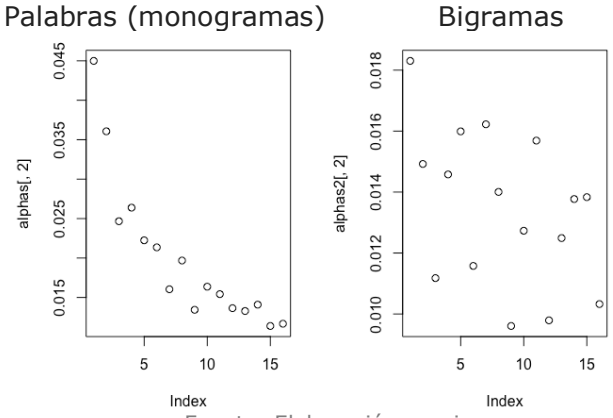
Para este proyecto, era necesario encontrar un modelo que permitiera encontrar los tópicos que subyacen a las causas no contenciosas, con el fin de poder ordenar la información de forma rápida y útil, junto con intentar comprender patrones de conducta subyacentes del TDLC. Inicialmente se pensó que para esto se podría usar LSA, dado que era una de las metodologías cubiertas en el Diplomado. Sin embargo, este algoritmo entrega resultados que son de difícil interpretación, dismuyendo su aporte para el público objetivo. Es así, que ampliamos nuestra investigación y descubrimos los modelos ***Latent Dirichlet Allocation*** **(LDA)**.

LDA es un modelo no supervisado generativo probabilístico, la idea detrás del modelo es que los documentos están representados por una mezcla aleatoria de temas subyacentes (o tópicos latentes). LDA asume que cada documento puede ser representado por una distribución probabilística sobre tópicos latentes y que esta distribución de tópicos en todos los documentos comparte una distribución de *Dirichlet a priori* uniforme. Cada tópico en el modelo LDA también esta representado como una distribución probabilística sobre palabras y la distribución de palabras en tópicos que comparten también una distribución *Dirichlet a priori* uniforme. Esto, dando como resultado una lista de tópicos donde las palabras con mayor relevancia, o aporte, tienden a representar de mejor forma el contenido del tópico.

La estimación de estos modelos se realizó a través de dos metodologías. La primera corresponde al algoritmo ***Vectorial Expectation Maximization*** **(VEM)**, el cual descansa en descubrir la máxima verosimilitud estimada de los parámetros cuando la data del modelo depende de ciertas variables latentes. El algoritmo tiene dos etapas, calculo de expectativas y maximización. La segunda corresponde a ***Gibbs sampling***, el cual es un algoritmo basado en Monte Carlo vía cadena de Markov, método que genera muestras de una distribución conjunta, cuando las distribuciones condicionales de cada variable pueden ser eficientemente calculadas.

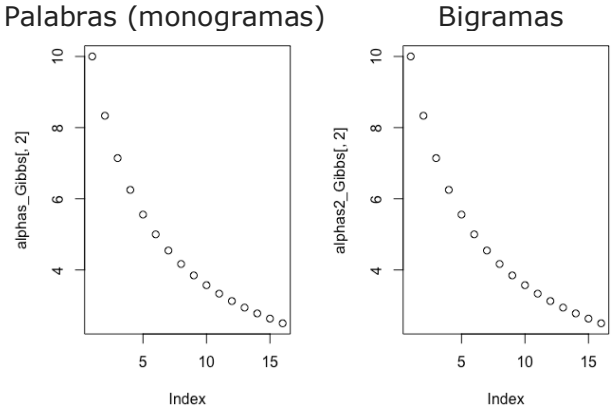
En concreto, se estimaron modelos que tienen entre 5 y 20 tópicos, bajo ambas metodologías antes mencionadas, para palabras (monogramas) y bigramas. A continuación, se muestra el *alpha*, medida de concentración de la distribución de los tópicos sobre los documentos, donde un *alpha* alto implica que un documento esta representado de forma más homogénea entre los tópicos.

Figura 1: ***Alpha*** **para palabras y bigramas para documentos que tienen entre 5 y 16 tópicos** *(Metodología VEM)*



Fuente: Elaboración propia

Figura 2: ***Alpha*** **para palabras y bigramas para documentos que tienen entre 5 y 16 tópicos** *(Metodología Gibbs)*



Fuente: Elaboración propia

Adicionalmente, se analizó el resultado de las palabras de cada tópico para los 64 modelos calculados. Como resultado, se considera que la mejor opción corresponde a palabras con 10 tópicos y bajo la metodología de *Gibbs sampling*. Esto, debido a que el *alpha* en los modelos *Gibbs* muestra un buen nivel de dispersión entre los documentos, de hecho, como se puede ver en los gráficos, es un punto de inflexión, antes de observar que los documentos están determinados principalmente por un par, o tan solo un tópico. Además, dada la naturaleza de los tópicos y los documentos, no parece razonable esperar que un documento sea determinado por un solo tópico. A continuación, se muestran la lista de las principales diez palabras para cada tópico, ordenadas según relevancia.

Tabla 1: Principales diez palabras por tópico

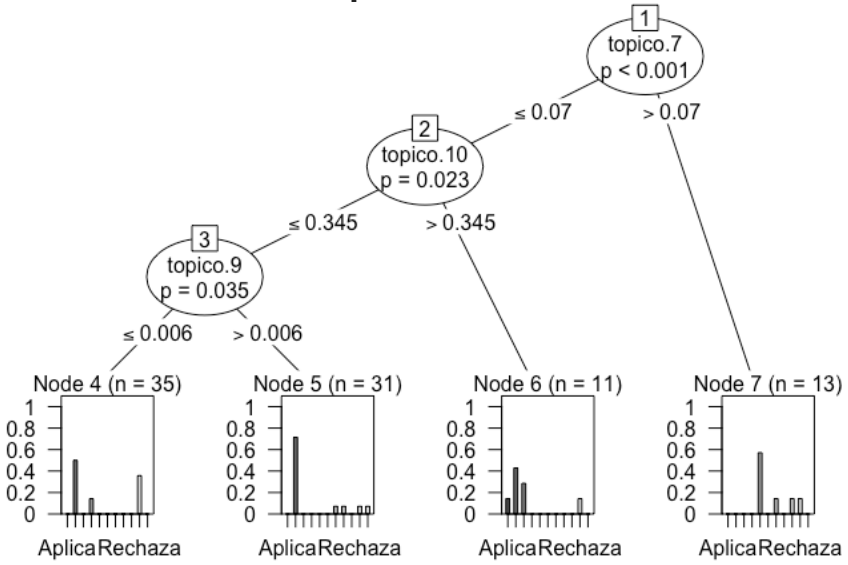
Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Retail	*Servicios públicos*	*Hidrocarburos*	*Afectación a la libre competencia*	*Puertos/ Aduanas*	*Competencia*	*Telecomunicaciones*	*Tópicos variados*	*Aeroportuario*	*Prensa*
fusion	derecho	terpel	Servicio	Concesión	Existencia	Telefonia	bases	Lan	Publicidad
Product	aguas	gas	Distribucion	Frente	Presencia	Servicio	Hotels	Aerolineas	Concesión
Consulta	proyecto	Combustible	Comercio	Puerto	Consideración	Movil	Estacionamiento	aeropuerto	radial
supermercado	generación	distribución	Prestamo	Carga	Competencia	Operador	Municipio	rutas	glr
Operación	solicitud	copec	Producto	Portuario	Efecto	Telecomunicaciones	Administración	pasaje	radios
importación	central	propiedad	Electricidad	Servicio	información	Espectro	Consulta	transporte	radio
consumidor	endesa	glp	solicitud	Atraque	Operación	Banda	Afp	Operación	medios
comercio	aprovechar	metrogas	información	establecimiento	participación	Mhz	Servicio	Acuerdo	radiodifusión
proveedor	consulta	consulta	Precios	tarifa	empresas	Tecnología	Publicidad	consulta	consulta
fertil	transmición	gnl	asociación	integración	podria	red	licitación	ruta	grupo

Fuente: Elaboración propia

Luego, para intentar comprender mejor como afectan los tópicos encontrados a las resoluciones del tribunal, se agregaron las composiciones de pesos de cada tópico por documento al resumen de información de resoluciones del TDLC.

Para entender mejor la relación de los tópicos encontrados y los documentos, se decidió usar la metodología ***Random Forest*** para encontrar el mejor ajuste posible a la determinación del Tribunal, usando como input, la información de pesos por documento de cada tópico, sumado a las variables de caracterización, usando como variable dependiente (o a describir), a la conclusión de las causas. Inicialmente, se considereo como *benchmark* un modelo *one-rule*, donde asumir la aprobación de todas las resoluciones lleva a una tasa de error de 47,7%. En contraste, se estiman tres modelos, el primero con un nivel de error levemente mejor (42,4%), un segundo modelo con un error de 37,9% y un tercer modelo con un error de 36,4%. Se consideró que el modelo con error de 36,4% era un ajuste adecuado. Este modelo considera cada uno de los tópicos encontrados y la variable mercado, que indica el mercado al que pertenece la causa. El resultado se muestra en la figura 3, donde las cajas de frecuencia indican en el eje X a cada una de las opciones que puede tomar el tribunal (aplica, aprueba, confirma, emite opinión favorable, fija condiciones, impone medidas, entrega instrucciones, modifica anterior instrucción, no a lugar, rechaza y traslada).

Figura 3: Árbol de decisión para modelo con tasa de error 36,4%



Fuente: Elaboración propia

Finalmente, es posible concluir que los tópicos revelan los principales temas de las causas, ya sean sus mercados o algún otro componente relevante, como afectación de la libre competencia. Adicionalmente, el árbol de decisión presenta resultados razonables en la jerarquización de tópicos al momento de evaluar la resolución del tribunal, mostrando solides en los resultados del modelo LDA. Sin embargo, los modelos de árbol decisión analizados aún carecen de robustez suficiente, por un lado, a causa de contar únicamente con 66 causas y, por otro lado, debido a que se requiere mayor profundización en que otras variables se debiera contar, con el fin de obtener resultados más robustos.

6. Dificultades

Se observaron las siguientes dificultades durante el desarrollo del proyecto:

(a) El proceso de limpieza de los textos presentó un desafío mayor en relación con la aplicación de las técnicas de "*stemming*", "*lemmatization*" y "*stem-completion*". Principalmente porque estas metodologías están más desarrolladas para el idioma inglés, que para el español. Como se señaló en la sección de Transformaciones, se optó por utilizar un criterio de truncamiento de las palabras, y sigue siendo una meta pendiente implementar técnicas gramaticales más adecuadas, y luego completar las palabras para una interpretación más amable de los resultados.
(b) Durante el proceso de visualización de la información, se presentaron varias dificultades en la implementación de la "Nube de palabras", de acuerdo con los objetivos propuestos para el proyecto. En particular, hubo dificultad para implementar en formato *Shiny* las funcionalidades de "*tokenize*" y "*weighting*" de una "*Document-Term Matrix*". Además de observarse que la visualización de la nube trunca las palabras en los bordes.
(c) El objetivo propuesto de lograr hacer predicciones con la información utilizada en este proyecto sigue siendo un desafío pendiente. En particular, se observa que la cantidad de casos (66 documentos) puede ser relativamente bajo para aplicar técnicas de predicción, independiente de la riqueza de información en cada caso. Junto el desafio de encontrar otras variables relevantes que pudiesen aportar a tener un mejor ajuste de los modelos de árbol de decisión.
(d) Capacidad de procesamiento. Aunque trabajamos solo con 66 documentos, la información contenida es inmensa, implicando un costo de procesamiento mayor al esperado, lo cual llevo a que no pudiéramos generar todos los modelos deseados.

7. Potencial aspecto ético

A diferencia de otros proyectos donde la información recolectada podría permitir identificar a un determinado individuo, ya sea directa o indirectamente (datos personales) o bien, la información de este individuo permitiría hacer inferencia sobre sus características físicas o morales o incluso, obtener alguna conclusión respecto a elementos de su vida privada (datos sensibles), el *dataset* de este proyecto pareciera, a primera vista, no vulnerar elementos de privacidad. No obstante, se perciben algunos elementos de opacidad analfabeta, puesto que el público objetivo detrás (economistas

y principalmente abogados), debido a lo complejo de la programación que se utilizó para alcanzar los objetivos, no será capaz de entender los algoritmos de una forma sencilla. Por otro lado, tampoco se percibe que el algoritmo desarrollado tenga elementos de discriminación, puesto que no se está clasificando personas. En este sentido, el potencial aspecto ético de este proyecto es mínimo, en comparación con lo que se podría observar en aquellos que trabajan con información específica de personas y/o empresas.

8. Conclusiones

A lo largo de este desarrollo hemos aprendido diferentes aristas de los proyectos relacionados al manejo masivo de datos.

En primer lugar, algunos mencionados en el punto 6, donde la recolección de información supuso un gran desafío, y nos vimos obligados a construir matrices aledañas que nos permitieran acceder de la forma que necesitábamos a la información.

Por otra parte, una vez obtenida la información aprendimos que, para obtener buenos resultados en un procesamiento masivo de textos, debemos seguir una lógica similar a la siguiente: "construir el *character vector*" → "Crear el Corpus" → "Aplicar las técnicas de limpieza de textos aprendidas en el diplomado" → "*Document-Term Matrix*" → "Exportar archivo .rds". Con esto, logramos crear la base de lo que sería después el desarrollo.

Para el caso del análisis de datos, vimos que existen muchas barreras con las que no contábamos, tan básicas como recursos informáticos (si no hubiéramos tenido computadores con buenas capacidades de hardware no hubiéramos podido correr los modelos) hasta temas más cualitativos, como ser conscientes de lo importante que es tener claras las hipótesis que buscamos responder durante los análisis.

Por último, una vez obtenidas las bases de datos en el formato que buscábamos, teníamos que preguntarnos: "¿Qué buscamos responder con estos datos? ¿A quién va a ir dirigido nuestro producto?" En este sentido, reconocemos que nos hemos visto más desafiados, ya que nuestra respuesta puede ser muy amplia ("Cualquiera que busque temas o estadística relacionada al TDLC") o muy específica ("Para el mismo tribunal, poder entender internamente su operación y funcionamiento").

Finalmente, es muy satisfactorio haber obtenido tópicos que reflejaran correctamente los contenidos de los documentos y que a su vez fueran los más relevantes. Sin embargo, esto no es suficiente para poder modelar el comportamiento del TDLC, debido a la poca cantidad de causas existentes. Por otro lado, con el tiempo se podra tener más información, permitiendo estimar con mayor precisión los comportamientos subyacentes del tribunal.

Dado que queremos seguir mejorando el producto en términos de procesamiento de los datos, limpieza, visualización y análisis; tenemos la intención de dar una vuelta de ideas más, para entender mejor hasta dónde podríamos llegar y qué preguntas de negocio / mercado / sociales, podemos responder con un desarrollo como el que logramos.

En cualquier caso, al respecto del trabajo en equipo, creemos que hemos superado satisfactoriamente diferentes desafíos que nos han surgido, de forma coordinada y confiando siempre en el trabajo colaborativo.

Anexo

A continuación se entrega al link a los documentos de trabajo (rmd y bases de datos utilizadas). Una particularidad de este trabajo fue la necesidad de cortar en cuatro partes el proceso, dada la intensidad de uso de recursos computacionales.

https://github.com/kgoharriz/TDLC-UAI