Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Identificación de tópicos en dictámenes jurídicos
Aplicación del modelo LDA de clasificación automática de textos

AUTOR: SANTIAGO PRANDINI

TUTOR: SILVIA VIETRI

FEBRERO 2024

Palabras clave

Dictamen jurídico - Procesamiento de Lenguaje Natural – Tópicos - Latent Dirichlet Allocation

Índice

Introducción.....3

1. El Procesamiento de Lenguaje Natural en el ámbito jurídico 5

1.1 El estudio del lenguaje natural desde la perspectiva de la IA 5

1.2 El uso del PLN en el ámbito jurídico 6

1.3 Clasificación automática de textos mediante PLN.....7

2. Detección automática de tópicos mediante el algoritmo de Latent Dirichlet Allocation 11

2.1 El modelado de tópicos: proceso generativo e inferencia estadística 11

2.2 El funcionamiento del algoritmo de Latent Dirichlet Allocation.....13

2.3 Visualización del modelo de tópicos, métricas de evaluación e interpretación del modelo 16

3. La aplicación del algoritmo de LDA sobre un corpus de dictámenes jurídicos 20

3.1 Limpieza y preprocesamiento de datos 20

3.2 La obtención de tópicos sobre un corpus de dictámenes jurídicos 22

3.3 Interpretación de los resultados y evaluación del modelo.....23

Conclusión.....28

Referencias bibliográficas.....30

Introducción

Para una organización dedicada a emitir opinión jurídica sobre temáticas diversas, una tarea fundamental es la búsqueda de jurisprudencia y antecedentes sobre lo ya escrito en la materia, para consultar los criterios vertidos en casos análogos y utilizarlos como referencia para la elaboración del dictamen. Esta búsqueda puede resultar dificultosa si los dictámenes de la organización no se encuentran guardados o categorizados por tópicos o voces que permitan identificar con exactitud los temas sobre los que versan. La tarea de categorizar los dictámenes suscriptos en algunas organizaciones la realiza equipos especializados dedicados a la guarda de estos documentos por voces o temas, no solo para la búsqueda interna sino también para búsquedas académicas.

En el caso analizado la base de datos de la Organización cuenta con aproximadamente un millón de dictámenes jurídicos en formato digital. En su labor diaria los y las agentes requieren la búsqueda de jurisprudencia, para lo cual utilizan un sistema interno que cuenta con un buscador, en el que pueden filtrar por fecha y tema, siendo el tema una etiqueta cargada de forma manual por cada agente según un criterio personal, y no por un equipo especializado. Esta particularidad dificulta en reiteradas ocasiones la efectividad del buscador, haciendo depender el éxito de la búsqueda de la coincidencia de criterios entre quien puso el tema del dictamen y quien lo está buscando, y representa además un obstáculo para la apertura de los dictámenes en un buscador público para investigadores y su consulta por voces. Para solucionar este problema surge la posibilidad de explotar los textos de la base de datos con técnicas de Procesamiento de Lenguaje Natural (NPL o PLN) y analizarlos aplicando metodologías de aprendizaje automático que permitan identificar los tópicos o temas que componen cada documento sin depender del criterio de los y las agentes, posibilitando además en un solo proceso categorizar todos los documentos de la base de datos.

Las áreas de investigación del PLN incluyen: recuperación y extracción de información, minería de datos, traducción automática, sistemas de búsquedas de respuestas, generación de resúmenes automáticos, clasificación automática de textos, análisis de sentimientos, entre otras.

La solución del problema identificado se enmarca en el área de clasificación automática de textos que es la operación de asignar un tópico a un determinado texto, con el fin de agrupar, estructurar y categorizar automáticamente cualquier tipo de documento, permitiendo identificar grupos de documentos que pueden categorizarse dentro de una misma clase.

De los sistemas desarrollados en clasificación automática de texto, se abordarán específicamente aquellos basados en aprendizaje automático no supervisado que puedan

asignar los tópicos basándose en las observaciones pasadas a través del entrenamiento y puedan reconocer las asociaciones entre los fragmentos de texto y proponer una determinada categoría para cada texto. Existen diversas técnicas que pueden ser usadas para obtener estos tópicos, el algoritmo que será utilizado en este trabajo es el modelo latent dirichlet allocation (LDA).

- Objetivo general: Identificar de manera automática los tópicos presentes en los dictámenes jurídicos emitidos por la organización estudiada aplicando el algoritmo de LDA, para luego clasificar los dictámenes de la base de datos de manera de obtener para cada documento una composición de esos tópicos que lo identifique por completo.
- Objetivos específicos: Hallar clusters de dictámenes jurídicos con el fin de identificar el volumen trabajado en cada temática por la organización y representar la composición de cada tópico con las principales palabras que lo componen.

La implementación de estos procesos en un buscador de documentos inteligente significaría un progreso sustancial en la búsqueda de jurisprudencia que los y las agentes realizan diariamente lo que implica un mejor y más veloz servicio para la organización, y una eventual oportunidad de apertura de este sistema de búsqueda para fines académicos.

1. El Procesamiento de Lenguaje Natural en el ámbito jurídico

El objetivo de este primer apartado es realizar un análisis sobre la aplicación del procesamiento de lenguaje natural dentro del ámbito jurídico. Para cumplir con este objetivo se desarrollarán tres subapartados: en el primero se abordará el estudio del lenguaje natural desde la perspectiva de la inteligencia artificial que actúe como marco conceptual desde donde abordar el estudio del procesamiento de lenguaje natural; en el segundo se estudiarán en detalle las aplicaciones ya existentes de PLN en el ámbito jurídico para poder comprender las potencialidades de estas técnicas en el mundo del derecho; y en el tercero se analizarán los métodos que utiliza el PLN para realizar la clasificación automática de documentos con el objetivo de describir el funcionamiento del algoritmo y su forma de identificar los temas de un documento.

1.1 El estudio del lenguaje natural desde la perspectiva de la IA

En la actualidad, el uso de los recursos industriales, naturales o humanos dependen de un manejo eficiente de la información y el conocimiento. En la historia de la humanidad, el conocimiento ha ido incrementándose a pasos agigantados; con la llegada de las computadoras el incremento ha sido aún mayor. Buena parte del saber humano se encuentra en forma digital en distintos tipos de colecciones de datos, según la International Data Corporation para el 2025 los datos en el mundo crecerán un 65% a 175 zettabytes. Sin embargo, lo que es conocimiento para los seres humanos no lo es para las computadoras.

Un lenguaje se puede definir desde el punto de vista funcional lingüístico como una función que expresa pensamientos y comunicaciones entre la gente; desde un punto de vista formal como un conjunto de frases que se forma con combinaciones de elementos tomados de un conjunto llamado alfabeto, respetando un conjunto de reglas de formación y de sentido (Vásquez, Quispe & Huayna 2009). Además de las características fundamentales, debe considerarse que sea funcional, es decir, el lenguaje debe permitirnos expresar nuestras ideas. Podemos distinguir entre dos clases de lenguajes: los lenguajes naturales que son los que han evolucionado con el tiempo para fines de comunicación humana (español, inglés, etcétera.) y los lenguajes formales (matemático, programable etc.) Una de las tareas fundamentales de la Inteligencia Artificial (IA) es la manipulación de lenguajes naturales usando herramientas de computación, los lenguajes de programación juegan un papel importante ya que forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina.

Se denomina Procesamiento de Lenguaje Natural (de aquí en adelante PLN) al estudio de las interacciones entre las computadoras y el lenguaje natural. Desde la perspectiva de la

inteligencia artificial (IA), el estudio del lenguaje natural tiene dos objetivos: modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas y facilitar la comunicación con la computadora para que puedan acceder a ella usuarios no especializados (Vásquez, Quispe & Huayna 2009). Este trabajo se enmarca en el primer objetivo, donde el lenguaje utilizado dentro de un corpus de dictámenes jurídicos es el objeto de estudio, y se busca comprender los procesos que forman dicho lenguaje para obtener mediante el uso de un lenguaje de programación los tópicos, temas, o voces que conforman el corpus.

El campo del PLN ha tenido un gran crecimiento en los últimos años; sus áreas de investigación incluyen: recuperación y extracción de información, minería de datos, traducción automática, sistemas de búsquedas de respuestas, generación de resúmenes automáticos, clasificación automática de textos, análisis de sentimientos, entre otras. (Hernández, Gómez 2013). El área de investigación en la que se desarrolla este trabajo es la de clasificación automática de textos, por lo que se presentan conceptos y herramientas de PLN con el propósito de extraer información relevante de un conjunto de dictámenes jurídicos que pueda ser usada para diferentes aplicaciones. Por ejemplo, desarrollar un clasificador que permita categorizar esos dictámenes y recomendar etiquetas y a partir de ello crear un buscador de jurisprudencia. En el siguiente subapartado se estudia en detalle las aplicaciones de PLN en el ámbito jurídico.

1.2 El uso del PLN en el ámbito jurídico

Desde una perspectiva jurídica, Goretty Martínez[1] señala que la informática en el ámbito jurídico tiene tres áreas de desarrollo que son: la documentaria, la de control y gestión y la metadocumentaria. Estos desarrollos pueden ser aplicados en cuatro áreas del derecho (Rincón y Martínez, 2021:3): 1. Modelos para la argumentación y toma de decisiones. 2. Clasificación de textos legales. 3. Extracción de información de textos legales. 4. Creación y planificación de un sistema legislativo.

Existen precedentes en el ámbito internacional que muestran la eficacia en el uso de la IA aplicada al campo del derecho y los sistemas de justicia. A manera de ejemplos, en China se cuenta con el primer asistente de IA en los juzgados, llamado Xiao Fa, el cual ofrece

[1] "[…]la informática jurídica, como instrumento del derecho, tiene tres áreas: la informática jurídica documentaria, la informática de control y gestión, y la informática metadocumentaria. La informática metadocumentaria se traduce en los sistemas expertos legales constituidos por una base de conocimiento, mecanismos de inferencia y la interfase entre el usuario y la máquina, pero básicamente estamos hablando de la inteligencia artificial aplicada al mundo del derecho y que tienen vertientes muy educativas porque eventualmente puede coadyuvar en los procesos de educación, de investigación y para la toma de decisiones judiciales" cf. en (Martínez, 2012:828).

asesoramiento legal. Por su parte, en el Reino Unido desde el 2014 está en marcha el programa de reforma de las Cortes y Tribunales de Su Majestad (Rodríguez Carrillo, Barrón González 2022). En Argentina desde noviembre de 2017 funciona Prometea, un software que aplica IA para preparar automáticamente dictámenes judiciales: "Prometea está asociado con un nuevo enfoque de trabajo, que procura aplicar IA en el desempeño de diferentes tareas vinculadas a la práctica del sistema de justicia. En particular, se trata de un sistema de software que tiene como cometido principal la automatización de tareas reiterativas y la aplicación de IA para la elaboración automática de dictámenes jurídicos basándose en casos análogos para cuya solución ya existen precedentes judiciales reiterados. El diseño y la implementación del sistema se enmarcaron dentro del Plan Estratégicos de la Fiscalía, como acciones proactivas para mejorar la eficiencia y la calidad del trabajo de la entidad" (Estevez, Linares & Fillottrani 2020:10).

Estos precedentes mencionados, y el caso argentino en particular, se enmarcan en el área de modelos de argumentación y toma de decisiones definido por Rincón y Martínez entre las posibles para aplicar IA en el ámbito de derecho. El presente trabajo en cambio se desarrolla en el área de clasificación de textos legales y extracción de información, buscando incorporar nuevas técnicas que permitan extraer tópicos de manera automática de un corpus de dictámenes jurídicos. En el siguiente subapartado se desarrollan los métodos de PLN para la clasificación automática de textos.

1.3 Clasificación automática de textos mediante PLN

La clasificación de texto en PLN es la operación de asignar una categoría a un determinado texto, con el fin de agrupar, estructurar y categorizar automáticamente cualquier tipo de documento. En grandes volúmenes de información, la clasificación de texto en PLN permite identificar rápidamente grupos de textos que pueden categorizarse dentro de una misma clase, aun cuando el tema de cada texto sea diferente. Incluso, puede detectar más de una forma de categorización o clasificación de un conjunto de documentos. La clasificación de textos permite analizar, organizar y extraer automáticamente información precisa y relevante para una organización.

Se pueden distinguir tres métodos o enfoques de clasificación automática de documentos:

- Sistemas basados en reglas: realizan su tarea de organización basándose en un grupo de reglas lingüísticas elaboradas a mano, cada una de las cuales proporciona una categoría prefijada y, en conjunto, guían al sistema para identificar las categorías relevantes.

- Sistemas basados en aprendizaje automático: estos sistemas no operan con reglas sino que son algoritmos que aprenden a categorizar basándose en las observaciones pasadas a través del entrenamiento y aprenden a reconocer las asociaciones entre los fragmentos de texto y a arrojar una determinada categoría para un texto de entrada en particular.
- Sistemas híbridos: constan de un clasificador base que puede entrenarse y de un sistema que responde a reglas.

El modelado de tópicos es una técnica no-supervisada de aprendizaje automático que permite inferir temas o tópicos latentes en una colección de documentos considerando la distribución probable de palabras, sin que el investigador introduzca de antemano más información sobre el corpus (Grimmer & Stewart, 2013). Parte de la idea de que los documentos son mezclas de tópicos, en los que un tópico es una distribución de probabilidad sobre palabras. Se basa en reglas de muestreo probabilístico que describen cómo las palabras en los documentos podrían pertenecer a diferentes variables latentes. Durante el ajuste del modelo el objetivo es encontrar el mejor conjunto de variables latentes que pueden explicar los datos observados.

Existen diversas técnicas que pueden ser usadas para obtener estos tópicos. El principal algoritmo y que además será el utilizado en este trabajo, es el modelo latent dirichlet allocation (LDA), propuesto por David Blei, que devuelve por un lado los diferentes tópicos que componen la colección de documentos y por otro lado qué porcentaje de cada tópico está presente en cada documento. Los tópicos consisten en una distribución de probabilidades de aparición de las distintas palabras del vocabulario. El LDA es un modelo probabilístico que parte de la base de que las palabras de un documento pueden asociarse a un tema o tópico dentro del mismo. Calcula la probabilidad de que un tema genere determinadas palabras, así como la frecuencia con la que éstas se distribuyen. Esto, a su vez, le permite determinar las palabras que se asocian con el grupo de temas de un documento y, a continuación, agrupar el documento con otros documentos que contienen un grupo de temas similar.

Como señala Blei en el artículo donde presenta el modelo, los investigadores en el campo de la recuperación de información (RI) habían logrado avances en la modelación de corpus de documentos. La metodología propuesta por estos reduce cada documento del corpus a un vector de números reales, cada uno de los cuales representa frecuencias de conteo. En el esquema término-frecuencia/frecuencia inversa del documento representado como *tf-idf* (Salton y McGill, 1983), un vocabulario básico de “palabras” o “términos” se elige y se forma un recuento de los números de ocurrencias de cada palabra en cada documento. Tras una normalización, este recuento de términos en el documento se compara con un conteo de frecuencia del término en todo el corpus. Al final se obtiene una matriz cuyas columnas contiene

los valores tf-idf para cada uno de los documentos en el corpus, reduciendo documentos de longitud arbitraria a listas de números de longitud fija. Blei señala que si bien la reducción tf-idf tiene algunas características atractivas, especialmente en su identificación básica de conjuntos de palabras, el enfoque también proporciona una cantidad relativamente pequeña de reducción en la longitud de la descripción y revela poco en cuanto a la estructura estadística entre documentos o dentro del mismo documento.

Para abordar estas deficiencias, los investigadores de RI propusieron otras técnicas de reducción de dimensionalidad, la más notable es la *indexación semántica latente* (LSI) (Deerwester et al., 1990). LSI utiliza una descomposición de valores singulares de la matriz para identificar un subespacio lineal en el espacio de características tf-idf que capture la mayor parte de la variación en la colección. Este enfoque puede lograr una compresión significativa en grandes colecciones.

Luego Hofmann (1999) presentó el modelo probabilístico de Latent Semantic Analysis (de ahora en adelante pLSA) que modela cada palabra en un documento como una muestra de un modelo de mezcla, donde los componentes de la mezcla son variables aleatorias multinomiales que pueden ser vistas como representaciones de temas. De esta manera cada palabra es generada por un solo tema, y cada documento se representa como una lista de proporciones de esa mezcla (o temas) y por lo tanto es reducido a una probabilidad de distribución de un conjunto fijo de temas.

Blei señala que si bien el trabajo de Hofmann es un paso útil hacia el modelado probabilístico del texto, está incompleto ya que no proporciona un modelo probabilístico a nivel de documentos. En pLSA, cada documento es representado como una lista de números (las proporciones de mezcla para los temas), y no hay un modelo generativo probabilístico para estos números., lo que lleva a dos problemas: el número de parámetros en el modelo crece linealmente con el tamaño del corpus generando problemas de sobreajuste, y no está claro cómo asignar probabilidad a un documento fuera del conjunto de entrenamiento.

Los métodos de reducción de dimensionalidad se basan en la suposición de ‘bolsa de palabras’ en la cual se puede descuidar el orden de las palabras en un documento, en la teoría de la probabilidad esta es una suposición de intercambiabilidad de palabras. Además suponen también que los documentos son intercambiables, es decir el orden específico de los documentos del corpus también se puede descuidar. Siguiendo el teorema de representación de Finetti (1990) que establece que cualquier colección de variables aleatorias intercambiables se puede representar como una distribución de mezcla, Blei entiende que para trabajar con representaciones intercambiables de palabras y documentos es necesario considerar modelos mixtos que capturen esa intercambiabilidad tanto de palabras como de documentos. Esta línea

de razonamiento es la que lo lleva al desarrollo del modelo LDA, donde se demuestra que a través del teorema de Finetti se puede obtener una estructura estadística significativa intradocumento mediante la distribución de la mezcla.

Con el objetivo de analizar las aplicaciones de procesamiento de lenguaje natural en el ámbito jurídico, en este apartado se comenzó repasando a grandes rasgos las características de los diferentes tipos de lenguaje para introducir luego al PLN como el estudio de la interacción entre las computadoras y el lenguaje natural, y presentar sus diferentes campos de aplicación. Como paso siguiente se identificaron las aplicaciones de PLN en el ámbito jurídico, y se destacaron algunos casos que ya se encuentran en funcionamiento. Por último y ya en el ámbito de clasificación de textos legales y extracción de información, se destacaron los diferentes métodos dentro del PLN que permiten realizar esas tareas, y se profundizó sobre la técnica no supervisada de aprendizaje automático de modelado de tópicos, específicamente sobre el algoritmo LDA. En el siguiente apartado se desarrollará la forma en que trabaja el algoritmo para obtener tópicos sobre un corpus de documentos.

2. Detección automática de tópicos mediante el algoritmo de Latent Dirichlet Allocation

El objetivo de este segundo apartado es explicar el funcionamiento del algoritmo de Latent Dirichlet Allocation para comprender la forma en que detecta los tópicos presentes en un corpus de documentos. Para cumplir con este objetivo se desarrollarán tres subapartados: en el primero se abordará la técnica de modelado de tópicos resaltando la diferencia entre un proceso generativo y un proceso de inferencia estadística; en el segundo se presentará el fundamento teórico del algoritmo de LDA y su funcionamiento; y en el tercero se detallarán las formas de visualización, las métricas de evaluación y las formas de interpretación del modelo.

2.1 El modelado de tópicos: proceso generativo e inferencia estadística

Dentro del PNL, el modelado de tópicos basa su funcionamiento en la representación de documentos mediante un Modelo de Espacio Vectorial o VSM (Salton et al, 1975). VSM utiliza un enfoque léxico que analiza las palabras de un documento individualmente, y no establece relaciones semánticas entre palabras. Por lo tanto no importa su orden de aparición en el documento.

Como primer paso se determina el conjunto de palabras que componen el vocabulario. Cada documento de texto queda definido como una secuencia de palabras (bolsa de palabras). Cada documento es una bolsa de palabras o un subconjunto del vocabulario. Entonces, dado un vocabulario V de tamaño N, se representa cada documento como un vector de tamaño N, donde en cada posición aparece la frecuencia de término de esa palabra para ese documento. El valor numérico que se emplea para indicar la importancia de la palabra es la frecuencia de término. El corpus queda representado como una matriz de tamaño DxN, donde D es el número de documentos del corpus. Cada fila de esta matriz es un documento representado en VSM.

Como se mencionó en el apartado anterior, Blei destacó que en el modelo de pLSA cada documento es representado como una lista de números (las proporciones de mezcla para los temas), y no hay un modelo generativo probabilístico para estos números lo que genera dos problemas: el número de parámetros en el modelo crece linealmente con el tamaño del corpus, lo que lleva a serios problemas con sobreajuste; y no queda claro cómo asignar probabilidad a un documento fuera del conjunto de entrenamiento. Un modelo de tópicos es, en cambio, un modelo generativo de documentos. Los modelos de tópicos se basan en la idea de que los documentos son mezclas de tópicos, en los que un tópico es una distribución de probabilidad de palabras. Para hacer un documento, se opta por una distribución de tópicos, posteriormente cada palabra en el documento se elige de un tópico. Técnicas estadísticas se pueden utilizar para

invertir este proceso: inferir el conjunto de tópicos que son responsables de la generación de una colección de documentos.

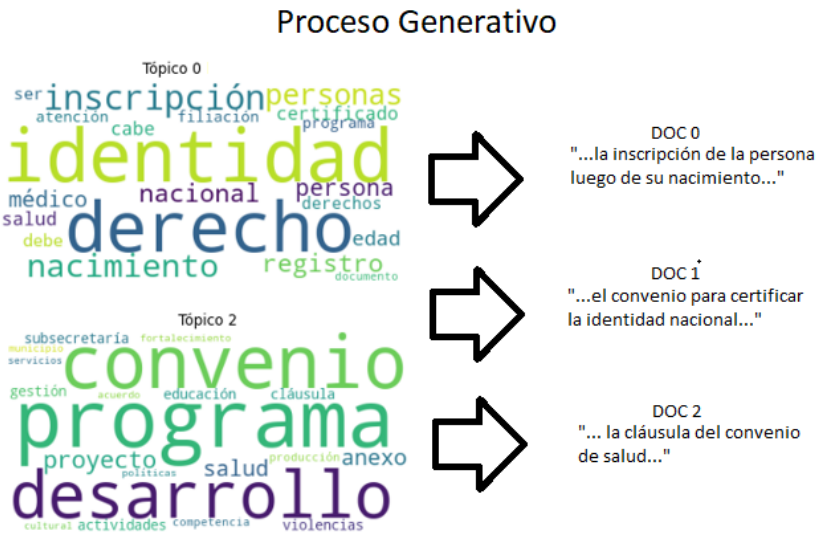
Un modelo generativo de documentos se basa en reglas de muestreo probabilístico que describen cómo las palabras en los documentos podrían ser generadas sobre la base de variables latentes (escondidas). Durante el ajuste de un modelo generativo, el objetivo es encontrar el mejor conjunto de variables latentes que pueden explicar los datos observados (las palabras en los documentos)

En las imágenes que siguen, se presenta el enfoque de modelado de tópicos de dos formas distintas: como un modelo generativo y como un problema de inferencia estadística. La **Imagen 1** muestra el proceso generativo que se ilustra con los tópicos 0 y 2, temáticamente relacionados con "registro de personas" y "convenios", que se muestran como bolsas que contienen distribuciones sobre palabras. Diferentes documentos pueden ser producidos escogiendo las palabras de un tópico en función del peso dado al tópico. Por ejemplo, los documentos 0 y 2 se generaron mediante el muestreo sólo de los tópicos 0 y 2 respectivamente, mientras que el documento 1 se generó por una mezcla de los dos tópicos. No hay exclusividad de que las palabras formen parte de un solo tópico, esto permite que los modelos de tópicos sirvan para captar la polisemia, donde la misma palabra tiene varios significados.

Como se mencionó con anterioridad, el proceso generativo no hace ninguna suposición sobre el orden de las palabras que aparecen en los documentos. La única información relevante para el modelo es el número de veces que se producen las palabras.

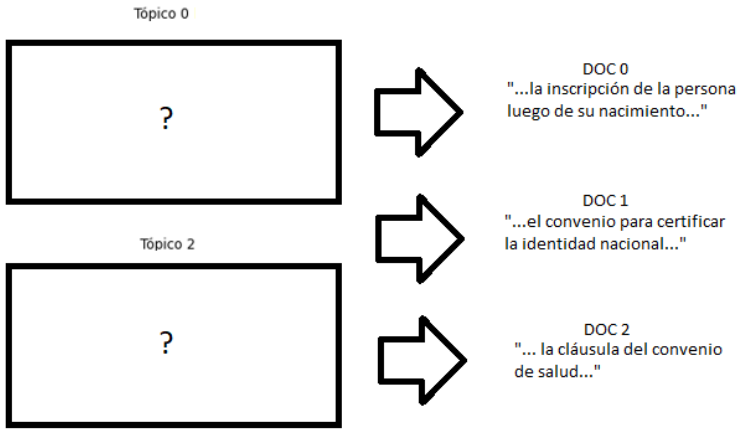
En la **Imagen 2** se presenta el problema de la inferencia estadística. Teniendo en cuenta las palabras observadas en un conjunto de documentos, se desea saber los tópicos más probables que hayan generado los datos. Esto implica inferir la distribución de probabilidad de palabras asociadas con cada tópico y la distribución de los tópicos para cada documento.

Imagen 1: Modelado de tópicos-Proceso generativo



Fuente: Elaboración propia

Imagen 2: Modelado de tópicos-Inferencia estadística



Fuente: Elaboración propia

En el presente trabajo se presenta una aplicación de inferencia estadística mediante el algoritmo LDA, el objetivo será inferir el conjunto de tópicos que son responsables de la generación de una colección de dictámenes. En el siguiente subapartado se explica en detalle el funcionamiento del algoritmo de Latent Dirichlet Allocation

2.2 El funcionamiento del algoritmo de Latent Dirichlet Allocation

Una variedad de modelos probabilísticos de tópicos se ha utilizado para analizar el contenido de los documentos y el significado de las palabras. Todos estos modelos utilizan la misma idea

fundamental, que un documento es una mezcla de tópicos, pero hacen ligeramente diferentes supuestos estadísticos. En LDA si las observaciones son palabras en documentos, cada documento puede verse como una mezcla de varios tópicos. Esto es similar a pLSA excepto que en LDA se asume que la distribución de tópicos tiene una distribución de Dirichlet. En teoría de la probabilidad y estadística, la distribución de Dirichlet (en honor a Peter Gustav Lejeune Dirichlet), generalmente denotada Dir(α), es una familia de distribuciones de probabilidad continuas multivariada, parametrizadas por un vector alfa (α) perteneciente al conjunto de los números reales positivos. Las distribuciones de Dirichlet frecuentemente se utilizan como distribuciones a priori en estadística Bayesiana.

Para explicar el funcionamiento de este modelo se presenta la notación de placas de la **Imagen 3**. Este sistema de placas se utiliza a menudo para representar modelos gráficos probabilísticos lo que permite que las dependencias entre las muchas variables se puedan capturar de manera concisa. La placa exterior representa los documentos, mientras que la placa interior representa las posiciones de palabras repetidas en un documento determinado; cada posición está asociada con una elección de tópico y palabra. Los nombres de las variables se definen según lo indica la Tabla 1.

Tabla 1: Variables del modelo LDA

Variable	Significado
K	Número de tópicos
M	Número de documentos
N	Número de palabras en cada documento (el documento i tiene Ni palabras)
α	Parámetro de la distribución de Dirichlet en las distribuciones de tópicos por documento
β	Parámetro de la distribución de Dirichlet en las distribuciones de palabras por tópico
θi	Es la distribución de tópicos para el documento i
φk	Es la distribución de palabras para el tópico k
Zij	Es el tópico para la j-ésima palabra en el documento i
Wij	Es la palabra específica

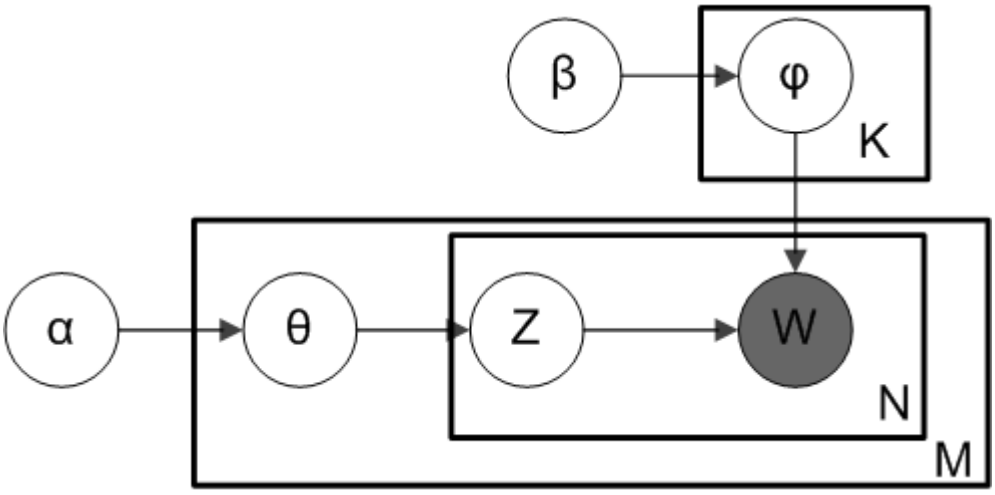
Fuente: Elaboración propia

En la **Imagen 3** el hecho de que W esté sombreado significa que las palabras Wij son las únicas variables observables y las otras variables son variables latentes. Como propone Blei en el artículo original, se puede utilizar la distribución de Dirichlet a priori para modelar la distribución tópico-palabra, siguiendo la intuición de que la distribución de probabilidad sobre las palabras en un tópico está sesgada, de modo que sólo un pequeño conjunto de palabras tiene alta probabilidad. El modelo resultante es la variante de LDA más aplicada en la actualidad, donde K es el número de tópicos y los φ son vectores V-dimensionales que almacenan los

parámetros de las distribuciones palabras-tópico donde V es el número de palabras. Es útil pensar en las entidades representadas por Θ y φ como matrices creadas al descomponer la matriz original palabra-documento que representa el corpus de documentos que está siendo modelado. En esta imagen Θ consta de filas definidas por documentos y columnas definidas por tópicos, mientras que φ consta de filas definidas por tópicos y columnas definidas por palabras. De esta forma cada φ se refiere a un conjunto de filas cada una de los cuales es una distribución de palabras y cada Θ se refiera a un conjunto de filas cada una de las cuales es una distribución de tópicos.

De acuerdo con lo expuesto por Blei y Jordan en el artículo citado, los parámetros α y β de Dirichlet suelen ser parámetros definidos a priori que pueden tomar valores dentro del intervalo abierto (0,1). Si se presupone que el documento está categorizado, se recomienda valores de α pequeños; mientras que si se presupone que un documento va a tratar sobre varios temas, se recomienda poner un valor de α grande. El valor β depende de la cantidad de palabras que definen un tópico, se recomienda que si un tópico necesita de pocas palabras el valor de β sea próximo a 0; por el contrario si un tópico requiere muchas palabras, se recomienda que β tome un valor próximo a 1.

Imagen 3: Notación de placas modelo LDA



Fuente: Ghahramani, Zoubin (August 2007). Graphical models

El proceso generativo que realiza el modelo de LDA para un corpus D compuesto de M documentos cada uno de Ni longitud, puede resumirse en los siguientes pasos:

1- Elige Θi~Dir(α) donde i ∈ {1,…., M} y Dir(α) es una distribución de Dirichlet con un parámetro simétrico α (α<1)

2- Elige φk~Dir(β) donde k ∈ {1,…., K}

3- Para cada una de las posiciones de las palabras i, j, donde i ∈ {1,…., M} y j ∈ {1,…., N}
   a- Elige un tópico Zij ~ Multinomial (Θ)
   b- Elige una palabra Wij ~ Multinomial (φ)

Para el caso de inferencia estadística la implementación de un modelo LDA transforma los datos de entrada y establece la siguiente representación:

- Tópicos: LDA devuelve una matriz de tamaño KxV, es decir, se define cada tópico ki: {i ∈ (0, K-1)} como una composición del aporte de cada palabra del vocabulario V en el tópico.
- Documentos: LDA devuelve una matriz de tamaño DxK, es decir, se define cada documento di: {i ∈ (0, D-1)} como una composición del aporte de cada tópico del conjunto K en el documento.

Detallado el mecanismo mediante el cual el modelo LDA infiere los tópicos latentes en un conjunto de documentos, en el siguiente subapartado se introduce una técnica que permite representar visualmente los resultados obtenidos del modelo, y colabora con la interpretación de estos.

2.3 Visualización del modelo de tópicos, métricas de evaluación e interpretación del modelo

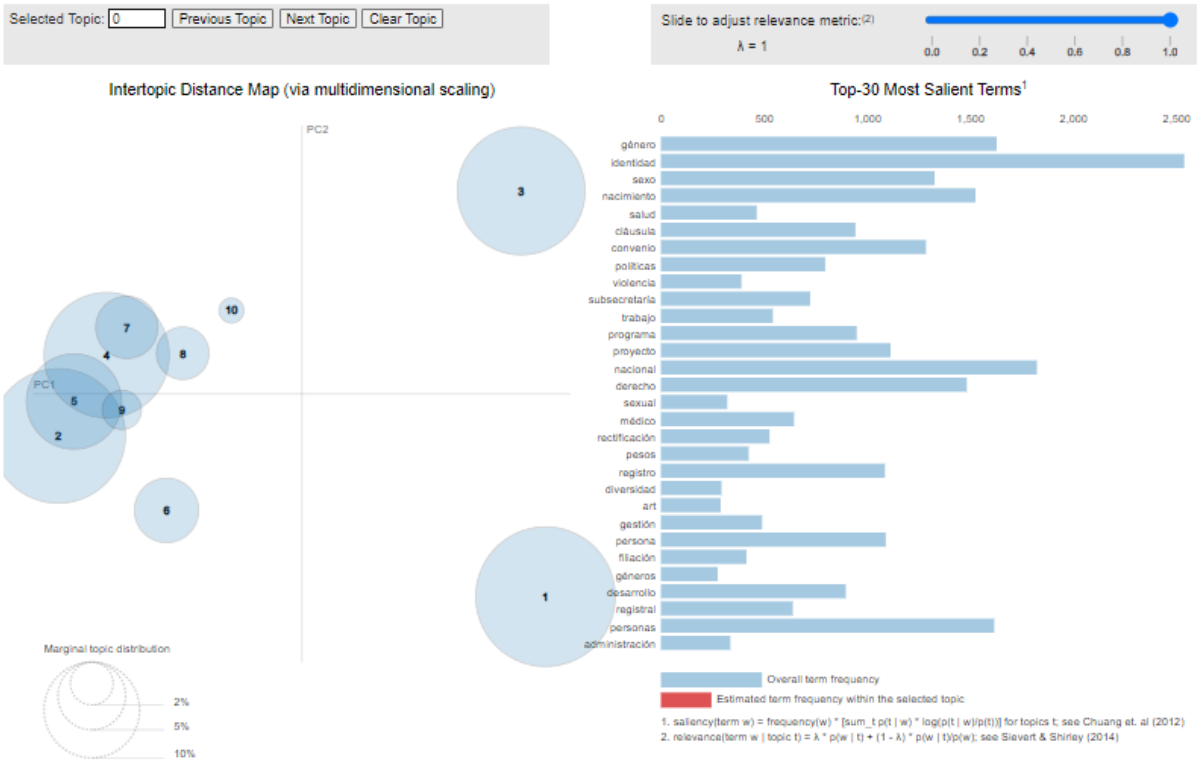
Para representar los datos de salida del modelo y poder interpretarlos se utilizan técnicas de visualización de tópicos. Chaney y Blei en su trabajo sobre visualización de datos destacan que los objetivos de este proceso deben ser: resumir el corpus estudiado visualizando la composición de los tópicos, presentar las relaciones entre los documentos evaluados y los tópicos descubiertos y evidenciar relaciones entre los documentos evaluados. Estos objetivos establecen las bases de la extracción de información de una colección de documentos empleando un modelo de tópicos.

Una de las técnicas utilizadas en este trabajo es el algoritmo de nube de palabras, una técnica visual que manifiesta la importancia del conjunto de palabras de cada tópico. Para mostrar la nube de palabras se usa el módulo Matplotlib.pyplot y WordCloud de Python que posibilita el diseño de la gráfica.

Por otro lado, Sievert y Shirley proponen el método LDAvis para visualizar la composición de los tópicos. Este método propone una visualización interactiva que permite observar los tópicos representados mediante círculos en un plano de dos dimensiones, donde el área de cada círculo

representa qué tan significativo es cada tópico en el corpus. La posición de cada tópico se calcula mediante la distancia entre tópicos proyectadas al plano de dos dimensiones. A su vez se presenta mediante un gráfico de barras las palabras más relevantes para el tópico seleccionado, superponiéndose una barra azul y una roja, que indican el número de veces que se repite la palabra en el corpus y la preponderancia en el tópico seleccionado.

Imagen 4: Vista general de la herramienta LDAvis



Fuente: Elaboración propia mediante Python

El diagrama de barras de la **Imagen 4**, en la que no se ha seleccionado ningún tópico, se representa las 30 palabras con mayor prominencia en el corpus. La medida de prominencia se corresponde con la propuesta por Chuang et al. y se calcula de acuerdo con la siguiente fórmula:

$$Prominencia(w) = Tf_{corpus(w)} * \sum_{i=0}^{K-1} p(k_i|w) * \log\left(\frac{p(k_i|w)}{p(k_i)}\right)$$

Como se observa en la **Imagen 5**, cuando se selecciona un tópico se presentan sobre el diagrama de barras las medidas antes explicadas. La **Imagen 6** ilustra que si seleccionamos una palabra concreta, podemos ver gráficamente a que tópicos pertenece, el área de los círculos pasa a indicar el grado de pertenencia de la palabra al tópico. Por último cabe mencionar que, como

se observa en la parte superior izquierda de la herramienta de visualización LDAvis, hay un parámetro libre $\lambda \in (0,1)$. Este parámetro influye a la hora de ordenar las palabras más relevantes de cada tópico y se extrae de un aporte de Sievert y Shirley, que propone una medida de la relevancia de una palabra en un tópico k de acuerdo a la siguiente fórmula:

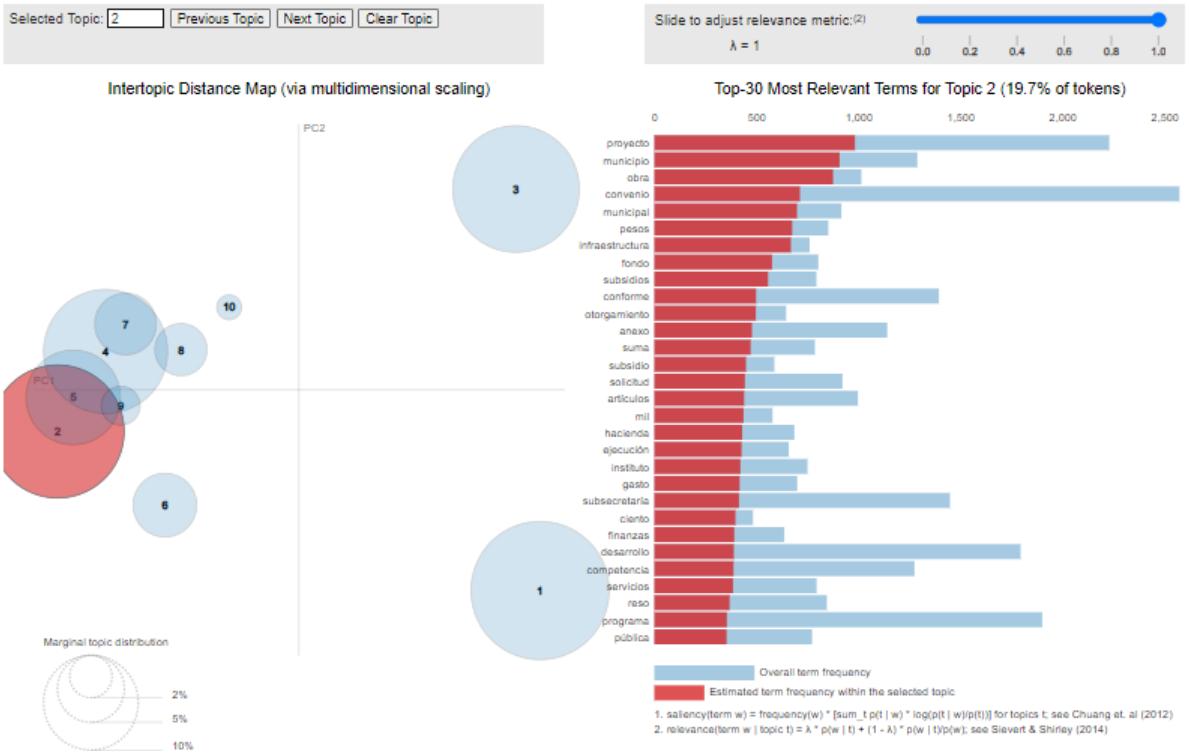
$$Relevancia(w|k) = \lambda * \log [p(w|k)] + (1 - \lambda) * \log [\frac{p(w|k)}{p(w)}]$$

Como se puede observar el parámetro λ determina el peso dado a la probabilidad del término w bajo el tópico k respecto a su *lift* (característica de interés que mide hasta qué punto ocurre conjuntamente w y k si fuesen independientes). Si $\lambda=1$ la relevancia queda definida por la probabilidad de la palabra w dado el tópico k.

En la evaluación de los modelos de tópicos se suele asumir que el espacio latente de tópicos es semánticamente significativo. Bajo este supuesto se puede implementar una medida de evaluación llamada 'Perplejidad del modelo probabilístico'. Sin embargo como establece Chang et al. estas medidas no evalúan lo significativo semánticamente que es el modelo. Si bien evalúan lo bueno que es el modelo predictivo probabilístico, no abordan objetivos de evaluación humana. El Modelado de Tópicos debe evaluarse también desde la perspectiva del entendimiento humano, los tópicos deben poder expresar información entendible y coherente, por lo que el análisis del modelo debe ser cualitativo y cuantitativo. En cuanto a lo relativo a la presente investigación los documentos a evaluar son dictámenes jurídicos, la heterogeneidad de este tipo de documentos en cuanto a su temática hace que sea complejo evaluar cuantitativamente el número óptimo de tópicos. Por este motivo, se emplearán valoraciones cualitativas, presentando la composición de los tópicos con LDAvis a los profesionales encargados de elaborar los dictámenes jurídicos quienes determinarán si los tópicos identificados por el modelo tienen sentido para los dictámenes evaluados, y si la cantidad de tópicos es la adecuada para representar el total de dictámenes.

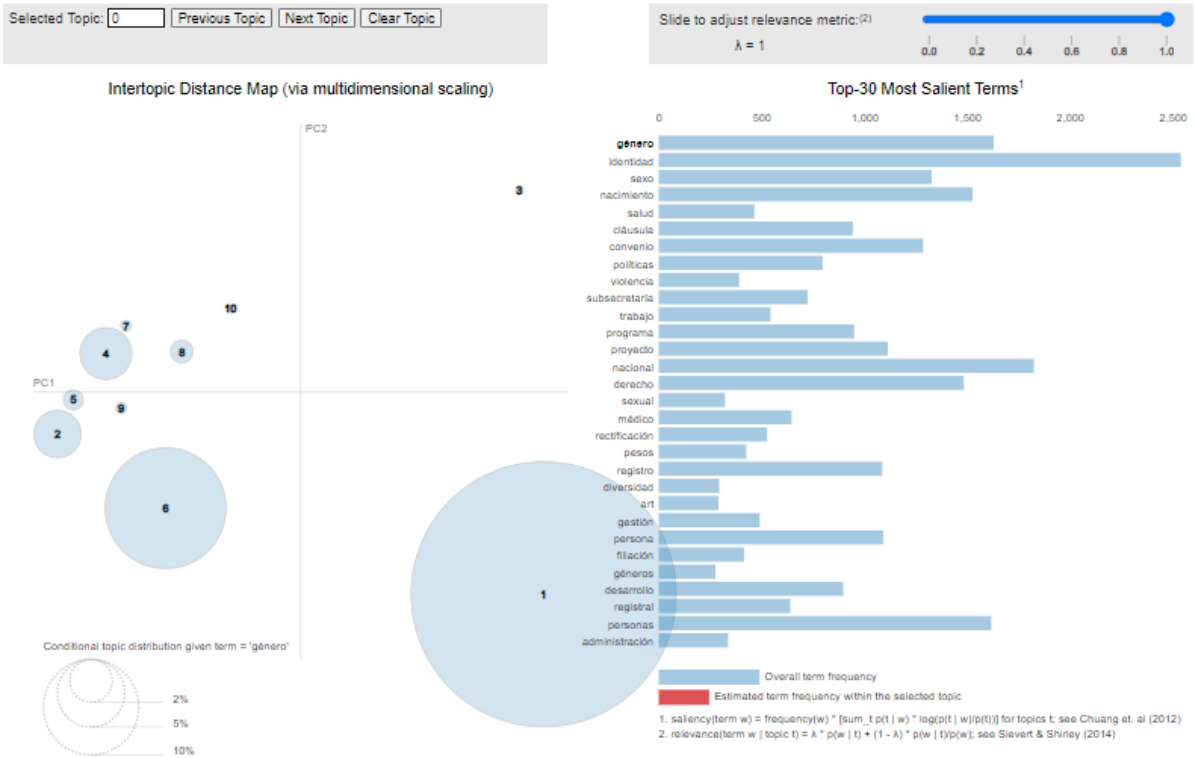
Con el objetivo de explicar el funcionamiento del algoritmo de Latent Dirichlet Allocation para comprender la forma en que detecta los tópicos presentes en un corpus de documentos, en este apartado se comenzó introduciendo la técnica de modelado de tópicos diferenciando entre un modelo generativo y uno de inferencia estadística. Como paso siguiente se desarrolló el fundamento teórico del algoritmo de LDA para echar luz sobre el funcionamiento del mismo. Por último se presentaron técnicas de visualización que permiten interpretar los resultados del modelo. En el siguiente apartado se desarrollará la aplicación de este algoritmo sobre el corpus de dictámenes jurídicos seleccionado para este trabajo.

Imagen 5: Vista de selección de un tópico con la herramienta LDAvis



Fuente: Elaboración propia mediante Python

Imagen 6: Vista de selección de una palabra con la herramienta LDAvis



Fuente: Elaboración propia mediante Python

3 La aplicación del algoritmo de LDA sobre un corpus de dictámenes jurídicos

El objetivo de este tercer apartado es desarrollar la aplicación del algoritmo de LDA sobre el corpus de dictámenes jurídicos estudiados en el trabajo y presentar los resultados obtenidos. Para cumplir con este objetivo se desarrollarán tres subapartados: en el primero se explicarán las tareas llevadas a cabo en limpieza y preprocesamiento de datos, fundamentales para el éxito del modelo; en el segundo se presentarán los resultados obtenidos detallando los pasos que llevaron a la elección de los tópicos óptimos para el caso estudiado; y en el tercero se analizarán los resultados obtenidos en detalle y se realizará una interpretación de estos.

3.1 Limpieza y preprocesamiento de datos

Para llevar a cabo el trabajo el primer paso consiste en la recopilación de los datos de una manera que sea posible explotarlos en un entorno utilizando lenguajes de programación. En la organización analizada cada dictamen que se suscribe acompañando un expediente queda almacenado en su versión Word en el sistema interno del Organismo. Mediante un proceso coordinado con el equipo de informática se exporta a un archivo de extensión csv una muestra significativa de dictámenes, de manera que en la primera columna queda establecido el índice mientras que en la segunda se extiende en cada celda el dictamen completo en formato texto. Cabe destacar que este trabajo se realiza en una primera instancia sobre una muestra de 400 dictámenes tomados de una de las Secretarías de la Organización ya que corresponde a un período de prueba. La Secretaría en cuestión tiene una delegación de temáticas amplias pero específica, que no abarca el total de temas tratados por la Organización.

Cargado el archivo en un entorno de programación se comienzan las tareas de modeling sobre los dictámenes que van a entrenar el modelo de LDA. El primer paso es la limpieza y el preprocesado de los dictámenes, este paso requiere no solo de herramientas de programación sino de conocimiento sobre la temática y trabajo en conjunto con quienes elaboran los documentos, para asegurar que las palabras que se conserven en la limpieza sean las adecuadas para interpretar el contenido del dictamen y evitar que palabras importantes para la interpretación sean eliminadas en la limpieza previa.

La limpieza del corpus comienza por aplicar una función a todo el conjunto. Esta función le aplica a la columna que contiene el texto del dictamen cuatro modificaciones: elimina los

caracteres especiales, elimina las palabras de un solo carácter (generalmente conectores), elimina los espacios en blanco dobles, y lleva todo el texto a minúscula. Finalmente se guarda el texto resultante de esta función en una nueva columna, a la que se le aplica la función 'tokenizer' la cual lleva a cabo la división del texto obtenido en 'tokens' o términos individuales.

Una vez obtenida la columna con el texto limpio y organizado en tokens, se le aplica la función 'stopwords' o palabras vacías de la librería NLTK, que elimina aquellas palabras que carecen de sentido cuando se las escribe solas o sin la palabra clave. Las palabras que pueden ser consideradas 'stopwords' dependen del idioma, para el idioma español esta función considera un poco más de 300 palabras. Son, básicamente, conjunciones, artículos, preposiciones y adverbios. Esta lista fue revisada por los profesionales encargados de la elaboración de los dictámenes quienes no identificaron ninguna palabra para quitar de la lista.

Filtradas las 'stopwords' se aplica una función de lematización sobre los tokens. Esta técnica sirve para reducir variantes morfológicas de la forma de una palabra a raíces comunes o lexemas; con el fin de, por ejemplo, mejorar la habilidad de los motores de búsqueda para mejorar las consultas en documentos. Básicamente consiste en remover el plural, el tiempo, o los atributos finales de la palabra.

Realizada la lematización se lleva a cabo la confección de un diccionario que contenga todas las palabras que no fueron filtradas una única vez. Los diccionarios son objetos que asignan un identificador numérico a cada palabra única y se pueden usar para obtener el identificador a partir de la palabra y viceversa. Para esto se utiliza la librería Gensim que crea un objeto de tipo Dictionary pasándole como argumento las listas de palabras de la columna modificada y tokenizada en los pasos previos.

Con el diccionario creado se limpian las palabras más raras o demasiado frecuentes. Para esto se aplica la función 'filter_extremes', que mantiene únicamente en el diccionario aquellos tokens que se encuentran en al menos n dictámenes (no_below) y los que están contenidos en no más del n% de dictámenes (no_above). Este paso consiste en un ensayo de prueba y error crucial para el éxito del modelo, en el cual los integrantes de la Organización tienen un rol fundamental ya que son los que conocen la temática de los dictámenes que alimentan el modelo. Realizadas las pruebas se decide conservar aquellas palabras que se registren en al menos 20 dictámenes y en no más del 60% del total de documentos.

Tras la limpieza y el preprocesamiento de datos el conjunto quedó en condiciones para ser utilizado por el modelo. En el siguiente subapartado se detalla el proceso llevado a cabo para la ejecución del modelo de LDA y la obtención de tópicos.

3.2 La obtención de tópicos sobre un corpus de dictámenes jurídicos

Finalizada la limpieza del corpus y obtenidos los tokens para cada dictamen, se crea un modelo LdaModel de la librería Gensim al que se le pasan como argumento el corpus creado con la bolsa de palabras de cada dictamen y el diccionario con todas las palabras del corpus, y se le indica adicionalmente los siguientes parámetros: número de tópicos, número de documentos que será utilizado en cada pasada de entrenamiento, número de pasadas por el corpus durante el entrenamiento y la densidad de tópicos por documento (un mayor valor de este parámetro implica que los documentos estén compuestos de más tópicos). Se entrena el modelo y se obtiene como resultados los n tópicos indicados.

Como primer paso se corre el modelo indicando un solo tópico para todo el corpus, con el objetivo de identificar las palabras con mayor relevancia para el total de dictámenes. Se aplica la técnica mencionada de nube de palabras para este único tópico, obteniendo los resultados que se muestran en la **Imagen 7**. A mayor tamaño de la palabra, mayor relevancia en el corpus.

Imagen 7: Nube de palabras del corpus

Fuente: Elaboración propia mediante Python.

De la figura se destacan palabras como “identidad”, “nacional” y “nacimiento” relacionadas a temas registrales tratados por la organización que son los que mayor volumen implican, por lo que cobra sentido que al tomar sólo un tópico sean las que predominen. Con este tópico general como parámetro se realizó un proceso iterativo, donde se ejecutó el modelo

aumentando de a un tópico por paso, y se consultó con los profesionales encargados de elaborar el dictamen la coherencia de los resultados obtenidos. Se identificó que a partir de los 5 tópicos las palabras que componen cada tópico comienzan a repetirse sin cambiar su significado. Por esto se tomó 4 tópicos como el óptimo para el corpus de dictámenes analizados.

Definido el número de tópicos en 4 para el caso de análisis, se definieron los parámetros específicos del modelo y se lo ejecutó. Acto seguido se imprimieron los tópicos obtenidos, integrados por palabras claves que tienen ponderaciones conforme su importancia dentro del tópico, lo que posibilita etiquetarlos. Los resultados obtenidos se muestran en la tabla 2.

Tabla 2: Tokens ponderados para los tópicos resultantes del modelo

Ítem	Tokens Ponderados
1	0.026 "identidad" + 0.016 "nacimiento" + 0.016 "nacional" + 0.014 "sexo" + 0.014 "derecho"
2	0.014 "convenio" + 0.010 "cláusula" + 0.008 "municipio" + 0.008 "obra" + 0.006 "infraestructura"
3	0.016 "programa" + 0.013 "políticas" + 0.012 "violencia" + 0.012 "género" + 0.011 "convenio"
4	0.012 "comunidad" + 0.011 "programa" + 0.009 "subvenciones" + 0.009 "instituto" + 0.007 "cláusula"

Fuente: Elaboración propia

Observando la tabla se puede identificar cada tópico según las palabras que mayor preponderancia tienen en él. Así los dictámenes que estén compuestos en mayor medida por el tópico 4 tratarán temáticas de programas y subvenciones para la comunidad, los que estén representados en mayor medida por el tópico 3 tratarán cuestiones de género, los del tópico 2 convenios con municipios para obras e infraestructura y los del tópico 1 cuestiones registrales. Sin perjuicio de ello, cada dictamen puede versar sobre más de un tópico.

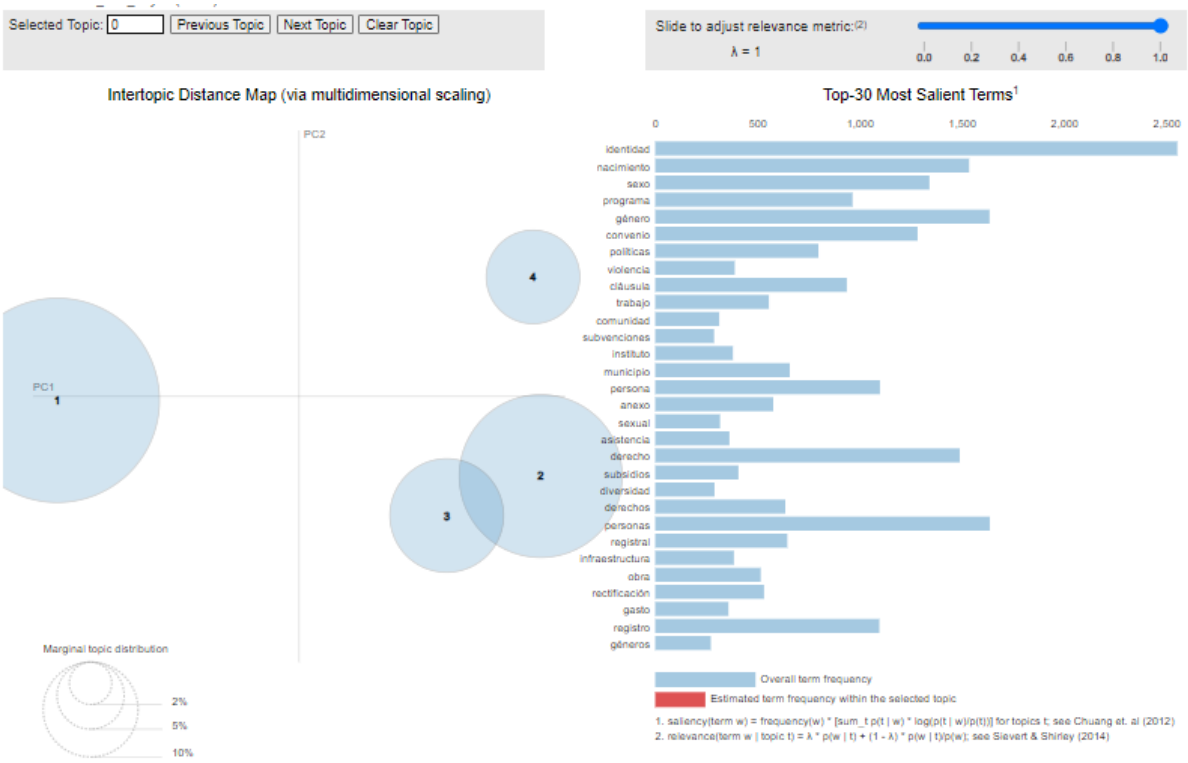
En el siguiente subapartado, con el objetivo de mejorar la interpretación del algoritmo, se complementa el análisis con la herramienta visual que propone el módulo mencionado LDAvis en Python, que permite el diseño de un gráfico interactivo. Como evaluación del modelo se presentan además algunas pruebas realizadas con el equipo de profesionales de la Secretaría mencionada.

3.3 Interpretación de los resultados y evaluación del modelo

Como se hizo mención en el segundo apartado, el método de visualización LDAvis propone una visualización interactiva que permite observar los tópicos mediante círculos en un plano de dos dimensiones, donde el área de cada circulo representa qué tan significativo es cada tópico en el corpus. En la **Imagen 8** se aprecia que el tópico 1 identificado como temáticas

registrales es el que está presente en mayor cantidad de documentos, seguido por el tópico 2 de convenios con municipios, el 3 de cuestiones de género y por último el 4 de programas y subvenciones.

Imagen 8: Visualización en LDAvis para los tópicos obtenidos



Fuente: Elaboración propia mediante Python

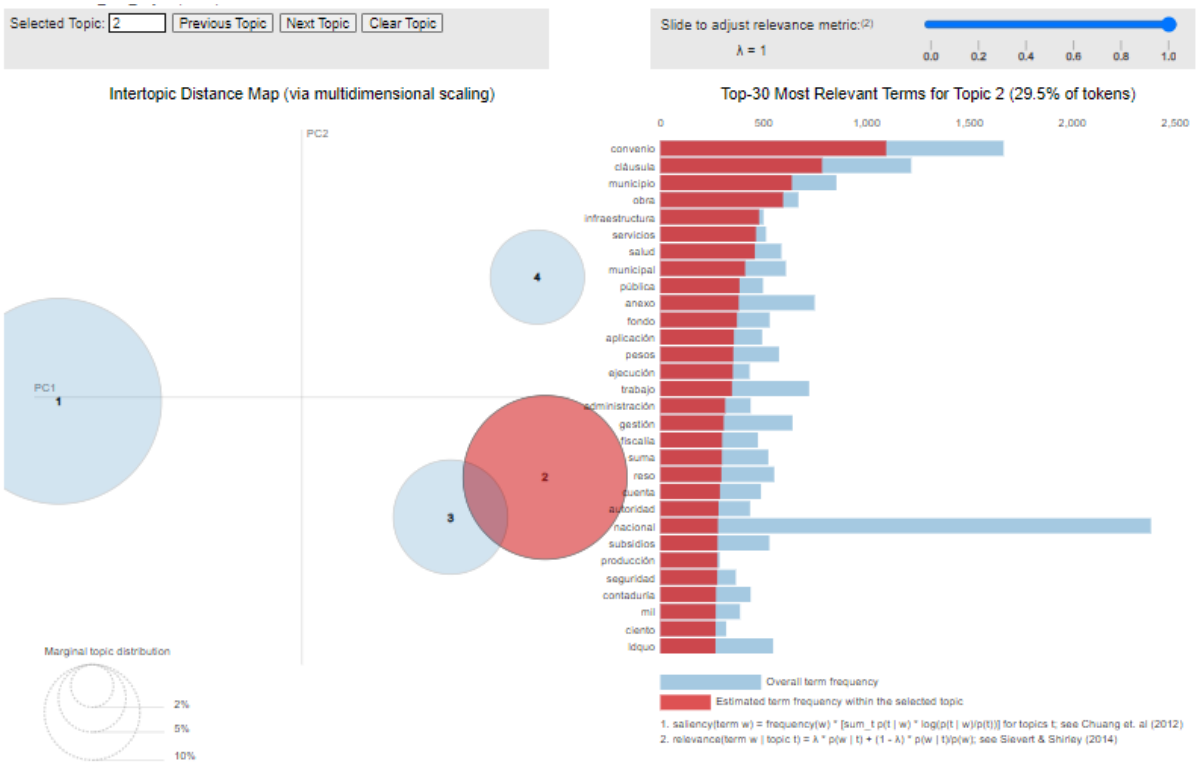
Como la posición de cada tópico se calcula mediante la distancia entre tópicos proyectadas al plano de dos dimensiones, es interesante destacar en este resultado que solo dos tópicos se encuentran superpuestos (el 2 y el 3) y son los únicos en el cuarto cuadrante. Mientras que el tópico 4 es el único ubicado en el primer cuadrante y el tópico 1 es el único en ocupar el segundo y el tercer cuadrante. Esto indica que los círculos que guardan mayor distancia entre sí se refieren a temas diferentes, mientras que los que se encuentran superpuestos tratan temas similares; en este caso los tópicos 2 y 3 comparten la palabra convenio con un alto grado de preponderancia lo que hace a los tópicos cercanos, aunque el contenido que tramita cada tipo de convenio sea diferente.

Del lado derecho del gráfico se presenta información para cada tópico que se quiera observar. El gráfico de barras indica un conjunto de 30 palabras que participan para la identificación de la temática sobre la que versa el tópico detallando la frecuencia de cada una de ellas, e indica

en términos porcentuales la participación relativa en el corpus del conjunto de tokens que representan cada círculo. Por ejemplo, el tópico 2 seleccionado en la **Imagen 9** incluye el 29.5% de los tokens incluidos en el diccionario para conformar un tema mientras que el 4 posee una participación relativa del 9.7%.

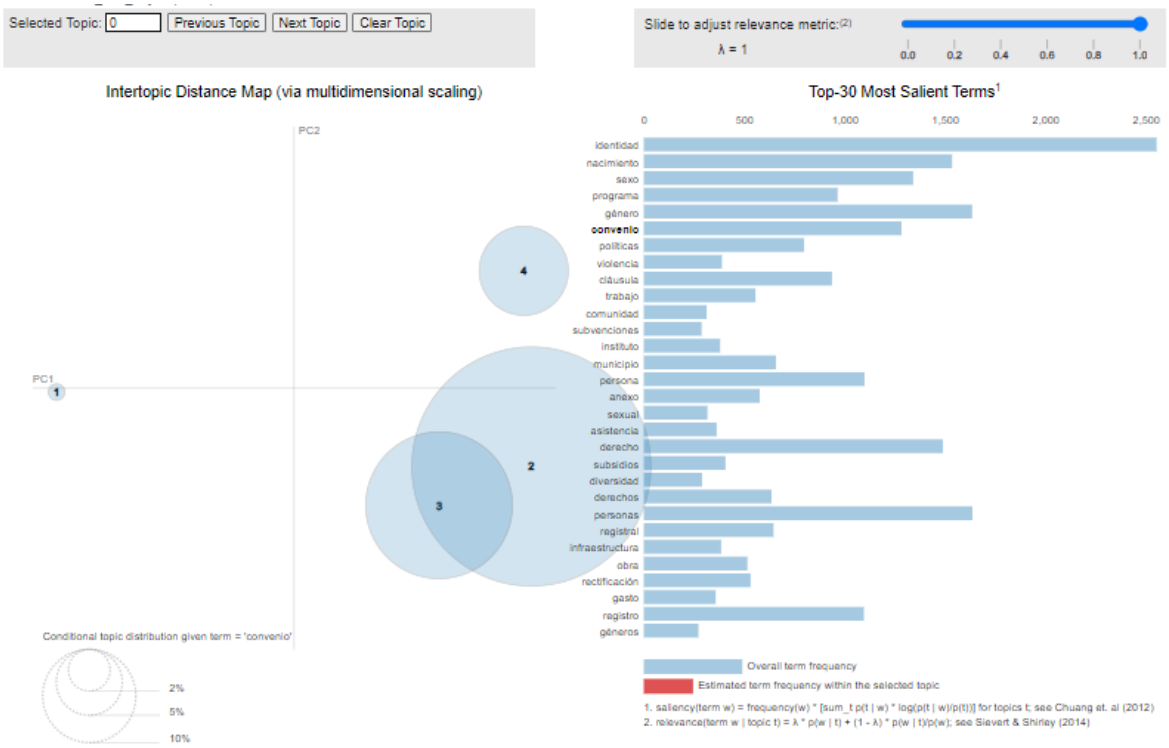
La otra función mencionada que ofrece esta herramienta es la de seleccionar una de las palabras del gráfico de barras para determinar a qué tópicos pertenece. En la **Imagen 10** se seleccionó la palabra 'convenio' mencionada anteriormente, se observa que pertenece principalmente a los tópicos 2 y 3.

Imagen 9: Visualización en LDAvis para la selección de un tópico



Fuente: Elaboración propia mediante Python

Imagen 10: Visualización en LDAvis para la selección de una palabra

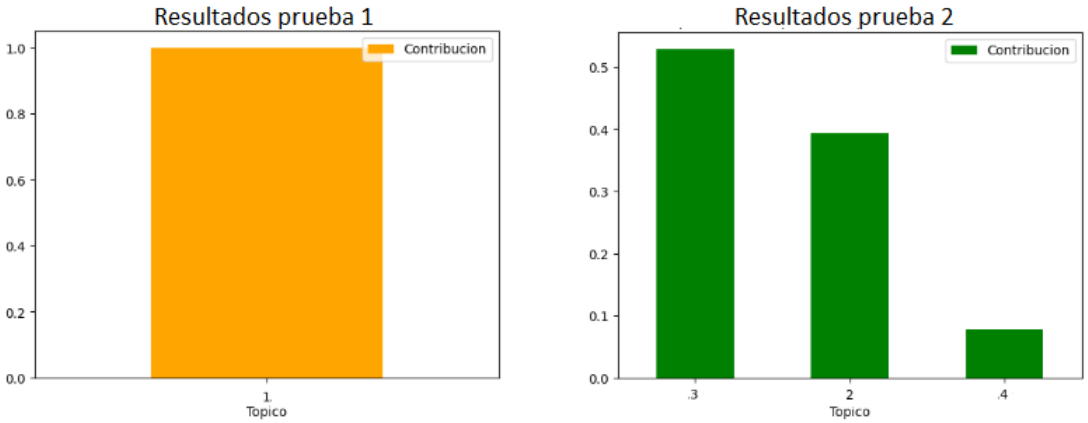


Fuente: Elaboración propia mediante Python

Una vez interpretados los resultados y validados los tópicos obtenidos con los profesionales de la Organización, se realizaron una serie de pruebas para evaluar el modelo. La primera fue tomar uno de los dictámenes ya preprocesado que componen el corpus al azar, y consultarle al modelo los tópicos que lo componen. El dictamen original tomado por el algoritmo tramita una inscripción tardía de un nacimiento, al que le atribuyó una contribución del 100% del tópico 1 como se observa en el gráfico de la izquierda de la **Imagen 11**, lo cual fue considerado correcto por los profesionales de la Organización.

Como segunda prueba se cargó un dictamen nuevo, que no forma parte del corpus original que el modelo utilizó para entrenar. Este dictamen trata de un convenio a realizarse entre una organización y un municipio para crear un centro de atención para casos de violencia de género. Se realizaron sobre el dictamen los pasos de preprocesado que se aplicaron al corpus original, y se consultaron los tópicos resultantes. En este caso el resultado obtenido fue una contribución mayoritaria de los tópicos 3 y 2 como se observa en el gráfico de la derecha de la **Imagen 11**, que habían sido identificados como temáticas de género y convenios con municipios respectivamente. Nuevamente los resultados obtenidos fueron considerados correctos por los agentes de la Organización.

Imagen 11: Contribución de tópicos para los dictámenes de prueba



Fuente: Elaboración propia mediante Python

Con el objetivo de desarrollar la aplicación del algoritmo de LDA sobre el corpus de dictámenes jurídicos estudiados en el trabajo y presentar los resultados obtenidos, en este apartado se comenzó detallando las técnicas de preprocesado y limpieza realizados sobre el conjunto de dictámenes. Como paso siguiente y con el corpus ya preprocesado se desarrolló la aplicación del algoritmo de LDA y se presentaron los tópicos obtenidos y la conformación de estos según la importancia de las palabras. Por último se interpretaron los resultados mediante técnicas de visualización y se realizaron distintas pruebas para comprobar el éxito del modelo.

Conclusión

El principal resultado logrado a través del desarrollo del presente trabajo es la identificación automática de los tópicos que componen un corpus de dictámenes jurídicos, y la representación individual de cada dictamen como una composición de los tópicos hallados. La posibilidad de categorizar el total de documentos emitidos por una organización con tópicos que representen fehacientemente las temáticas tratadas, representa una oportunidad única para obtener información relevante que colabore en la toma de decisiones.

En línea con el objetivo general del presente se logró determinar los 4 tópicos principales que surgen de los dictámenes emitidos, identificándolos con los temas o voces tratados, para luego obtener para cada documento una composición de esos tópicos que lo identifique por completo.

En el trabajo inicialmente se identificaron diferentes campos de aplicación del procesamiento de lenguaje natural en el ámbito jurídico, para luego profundizar en el área de clasificación de textos legales y extracción de información, buscando incorporar nuevas técnicas que permitan extraer tópicos de manera automática de un corpus de dictámenes jurídicos. Se desarrollaron los diferentes enfoques desde donde abordar la clasificación automática de textos en PLN, detectando en la técnica no-supervisada de Modelado de Tópicos la indicada para elaborar el trabajo, debido a que permite inferir temas o tópicos latentes en una colección de documentos considerando la distribución probable de palabras; y se seleccionó el algoritmo de Latent Dirichlet Allocation (LDA) propuesto por David Blei para llevar a cabo el trabajo.

Como paso siguiente se explicó el funcionamiento del algoritmo de Latent Dirichlet Allocation para comprender la forma en que detecta los tópicos presentes en un corpus de documentos. Se comenzó diferenciando dentro del modelado de tópicos al proceso generativo del proceso de inferencia estadística, destacando que este último es el que aplica en el trabajo. Luego se describió conceptualmente el algoritmo de Latent Dirichlet Allocation, utilizando la notación de placas como apoyo gráfico para el entendimiento del mismo, se presentaron técnicas de visualización que permiten interpretar los resultados del modelo como LDAvis y WordCloud, y se determinó que el modelo iba a ser evaluado cualitativamente por los profesionales de la organización encargados de la elaboración del dictamen.

Luego se aplicaron técnicas de preprocesamiento y limpieza de datos fundamentales para el éxito del modelo, de los cuales se obtuvo un diccionario compuesto por todas las palabras obtenidas tras la limpieza y un corpus que representa cada documento con las palabras que lo componen y su frecuencia de aparición. Se llevó a cabo la creación del modelo y su aplicación sobre el corpus obtenido tras el preprocesamiento, de donde se obtuvieron los tópicos que

representan el conjunto de documentos. Por último se analizaron los resultados mediante las técnicas de visualización mencionadas, y se realizaron diferentes pruebas para corroborar el éxito de los tópicos obtenidos.

Tras el desarrollo de cada uno de los capítulos, el trabajo final de especialización realiza un aporte de especial interés para la organización estudiada. Se pudo mostrar que la organización puede recuperar los temas de los más de un millón de dictámenes no categorizados que tiene en su base de datos. Esta situación manifiesta la necesidad de realizar una aplicación integral de este método de clasificación, en la que se involucren todas las áreas de la organización para asegurar que los tópicos obtenidos sean consecuentes con la composición de los dictámenes.

El trabajo presenta diversas posibilidades para continuar ampliando su alcance. Para complementar el análisis realizado en este trabajo sería de utilidad incorporar algoritmos complementarios como el reconocimiento de entidades nombradas. Al tratarse de una organización que se desarrolla en el ámbito jurídico, sería importante poder reconocer en cada dictamen la normativa aplicada que no fue tenida en cuenta en este trabajo ya que en el preprocesamiento se eliminaron los números.

Para ello se propone como línea de trabajo indagar sobre el reconocimiento de entidades nombradas y proponer un modelo que incluya a esta técnica para recuperar la normativa citada en los dictámenes emitidos en años anteriores. Será necesario un desarrollo en conjunto con los profesionales especializados en temática jurídica y posibilitará la posterior implementación de un buscador inteligente con la posibilidad de abrirlo a la comunidad científica para su consulta.

Referencias bibliográficas

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003*). Latent dirichlet allocation*. Journal of machine Learning research

Chaney, A., & D.M. Blei. (2012). *Visualazing Topic Models* Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.

Chuang,J, Manning,C.D y Heer,J. (2012). Termite: Visualization techniques for assessing textual topic models. Stanford University

De Battista, A., Retamar, S., Cristaldo, P., Schab, E., Casanova, C., Ramos, L., ... & Herrera, N. E. (2021). *Aplicaciones de procesamiento de lenguaje natural y Ciencia de Datos*. XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja).

Estevez, E. C., Linares, S., & Fillottrani, P. (2020). *PROMETEA: Transformando la administración de justicia con herramientas de inteligencia artificial*. BID

Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones. Komputer Sapiens, 1, 6-11.

Grimmer, J., & Stewart, B. M. (2013). *Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis*, DOI: https://doi.org/10.1093/pan/mps028

Hernández, M. B., & Gómez, J. M. (2013). *Aplicaciones de procesamiento de lenguaje natural*. Revista Politécnica, 32.

Martínez, G. (2012) *La inteligencia artificial y su aplicación al campo del Derecho*. Alegatos, septiembre-diciembre de 2012

Rincón, E. y Martínez, V. (2021) *Un estudio sobre la posibilidad de aplicar la inteligencia artificial en las decisiones judiciales* Revista Direito GV, vol. 17, núm. 1.

Rodríguez Carrillo, J. M., & Barrón González, M. Á. (2022). *Inteligencia Artificial y su aplicación en los sistemas de justicia en América Latina.* Instituto Belisario Domínguez

Salton, G.M., Wong, A y Yang, C.S. (1975) *A Vector Space Model for Automatic Indexing* Communicatons of the ACM

Sievert & Shirley (2014). *LDAvis: a method for visualizing and interpreting topics*. Procedings of the Workshop on Interactive Lenaguaje Learning, Visualization, and Interfaces

Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (2009). *Procesamiento de lenguaje natural.* Revista de investigación de Sistemas e Informática

Reporte Trabajo Final Integrador de Especialización

"IDENTIFICACIÓN DE TÓPICOS EN DICTÁMENES JURÍDICOS. APLICACIÓN DEL MODELO LDA DE CLASIFICACIÓN AUTOMÁTICA DE TEXTOS"

Autor: Santiago Prandini

Como mentora del presente Trabajo Final Integrador de Especialización del alumno Santiago Prandini, transmito mi opinión sobre la investigación realizada.

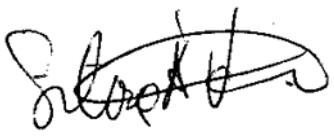
Con respecto al tema de análisis, considero que aborda una problemática relacionada con la gestión de datos en organizaciones; en este caso organización dedicada a emitir opinión jurídica sobre temáticas diversas.

El alumno analiza un corpus de dictámenes jurídicos con una muestra de 400 dictámenes tomados de una de las Secretarías de la Organización en la que se desarrolla el trabajo.

En la investigación, se enmarca la gestión de datos en contextos de Big data señalando la posibilidad de explotar los textos de la base de datos con técnicas de Procesamiento de Lenguaje Natural (NPL o PLN) y analizarlos aplicando metodologías de aprendizaje automático que permitan identificar los tópicos o temas que componen cada documento sin depender del criterio personal de los y las agentes y posibilitando además en un solo proceso categorizar todos los documentos de la base de datos.

A partir del análisis efectuado se deja en evidencia que la implementación de estos procesos en un buscador de documentos inteligente significaría un progreso sustancial en la búsqueda de jurisprudencia que los y las agentes realizan diariamente, mejorando el servicio para la organización y generando la posibilidad de apertura de este sistema de búsqueda para fines académicos.

La coherencia del enfoque planteado, el uso de algoritmos de Minería de Datos, la pertinencia de las referencias bibliográficas y el correcto análisis de los resultados obtenidos, permiten señalar a este trabajo como un aporte relevante en el ámbito jurídico, evidenciando la posibilidad de emprender futuras líneas de investigación.



Dra. Silvia Vietri