ESTADO DE LA JUSTICIA

Tercer Informe Estado de la Justicia

Investigación de Base

Extracción de Información Textual de las Resoluciones de la Sala Constitucional de Costa Rica

Investigador:

Kenneth Obando Rodríguez

San José | 2020

Investigación finalizada en el año 2019, se reserva su publicación al año 2020 por embargo de la información contenida.

CONSEJO NACIONAL DE RECTORES

2020 Año de las Universidades Públicas por la Salud Mental

PROGRAMA ESTADO DE LA NACIÓN

PODER JUDICIAL República de Costa Rica

353.4
O12e

Obando Rodríguez, Kenneth
Extracción de información textual de las resoluciones de la Sala Constitucional de Costa Rica / Kenneth Obando Rodríguez. -- Datos electrónicos (1 archivo : 1750 kb). -- San José, C.R. : CONARE - PEN, 2020.

ISBN 978-9930-540-37-4
Formato PDF, 25 páginas.
Investigación de Base para el Tercer Informe del Estado de la Justicia.

1. ANÁLISIS DE CONTENIDOS. 2. BASES DE DATOS. 3. SISTEMA DE INFORMACIÓN JURÍDICA. 4. MINERIA DE DATOS. 5. SALA CONSTITUCIONAL. 6. COSTA RICA. I. Título.

EBV

Contenido

Presentación 3
Introducción 4
Trabajos Previos 4
Dominio del Problema 5
Sala Constitucional 5
Repositorio de las Resoluciones 5
Base de Datos Nexus 5
Metainformación contenida en el texto de resoluciones 6
Número de Voto *6*
Número de Expediente *6*
Fecha y Hora *6*
Redactor *6*
Voto Salvado *6*
Sentencia *6*
Entorno de Programación y Herramientas **7**
Proceso de extracción de la metainformación y resultados **7**
Conversión de los documentos 7
Proceso de Extracción de la Metainformación y Resultados 9
Conversión de los documentos 9
Separación en secciones 10
Expedientes 11
Número de Voto 11
Fecha y Hora 12
Redactor 13
Recurrente 14
Recurrido 15
Sentencia 15
Tipo de proceso 15
Tema 17
Clusterización *17*
Reconocimiento de entidades en el texto *20*
Relación entre la cantidad de Resoluciones presentadas por tema y el porcentaje de éxito. *22*

Conclusión **22**
Trabajo Posterior **23**
Agradecimiento **23**
Referencias **23**

Presentación

Las cifras de esta investigación pueden no coincidir con las consignadas en el Tercer Informe Estado de la Justicia, debido a revisiones posteriores. En caso de encontrarse diferencia entre ambas fuentes, prevalecen las publicadas en el Informe.

Palabras claves

Abstract— En la presente investigación se expone los procesos empleados para la extracción de información textual por medio de técnicas de procesamiento de lenguaje natural a un repositorio con aproximadamente 370 000 resoluciones de la Sala Constitucional de Costa Rica. Esta investigación se encuentra dentro del proyecto de Análisis de Contenido de las Resoluciones desde 1989 de la Sala Constitucional hasta el 2018.

Index Terms—Análisis Automático de Contenido, Procesamiento de Lenguaje Natural, Extracción de Información, Procesamiento de Texto.

Introducción

La aparición de nuevas técnicas de análisis automático de contenido ha abierto nuevas posibilidades al procesamiento de documentos legales a una escala inimaginable"[1]. Cada resolución dictaminada por los Magistrados establece una nueva jurisprudencia que es necesario consultar para los nuevos casos que se estudian en la Sala, lo que hace imperativo el uso de herramientas que permitan una indexación eficiente y una relación correcta entre los documentos.

Por otra parte, con el análisis automatizado de contenido, se han desarrollado nuevas técnicas para relacionar los textos de un repositorio de documentos, la clasificación automatizada y la posibilidad de obtener características de cada texto que difícilmente puede ser realizada de forma manual.

El presente trabajo tiene como objetivo, mostrar las técnicas empleadas para extraer metainformación contenida en el texto crudo de las resoluciones de la Sala Constitucional de la República de Costa Rica y, además, se explora algunas técnicas que serán aplicadas al texto procesado.

Esta etapa inicial no hace sino ampliar aún más las posibilidades para las aplicaciones y la investigación, todo con la meta de crear una imagen fiel de los procesos de la Sala Constitucional y con ello aportar en la mejora del sistema de justicia y la aplicación de los derechos humanos en la sociedad costarricense.

Trabajos Previos

La creación, aplicación y la consecuente interpretación de la Constitución, estatutos y resoluciones, se encuentra en el corazón de los sistemas legales de las más avanzadas democracias modernas. Como resultado, se ha desarrollado un lenguaje propio por el que se puede analizar con el uso de procesos automáticos.

En esta última década, el análisis automático de contenido se ha aplicado en las ciencias judiciales con cierto éxito, y ha facilitado los procesos de clasificación, indexación y, en algunos casos, permite estimar la factibilidad de un proceso legal aún antes de ser presentado en la corte.

Evans et al. [2] explora la aplicación de técnicas de clasificación de casos selectos de la Corte Suprema de Justicia en Estados Unidos. Por su parte, McGuire y Vanberg [3] exploran las técnicas de escalado de textos de diferentes cortes.

En [1] se describe una selección de procedimientos útiles para la clasificación supervisada y no supervisada de textos del caso conocido como “Obamacare”, que al ser sujeta de una fuerte politización, es buen ejemplo para aplicar técnicas de escalado y clasificación.

Dominio del Problema

Sala Constitucional

La Sala Constitucional de Costa Rica es el tribunal de máxima jerarquía en el Poder Judicial, y tiene la obligación de resolver los casos que requieran la interpretación de los artículos de la Constitución Política, los que incluyen[4]:

- Los recursos de hábeas corpus y de amparo.
- Las acciones de inconstitucionalidad.
- Las consultas de constitucionalidad.
- Los conflictos de competencia entre los Poderes del Estado, incluido el Tribunal Supremo de Elecciones y los de competencia constitucional entre éstos y la Contraloría General de la República, municipalidades, entes descentralizados y demás personas de Derecho Público.

Repositorio de las Resoluciones

El repositorio actual consiste en la colección de 364 032 del texto íntegro de las resoluciones de la Sala Constitucional, realizadas entre los años 1989 y 2017. Originalmente, los documentos se encuentran en distintos formatos (Microsoft Word, HTML, PDF, TXT) organizados en años y en su número de serie. Algunos de estos documentos pueden venir vacíos por dificultades propias de la digitalización de la fuente original (menos del 0.5%).

La estructura general de los documentos consiste en cuatro secciones propias de un documento legal: Encabezado, Resultando, Considerando y "Por Tanto". Muchos de los textos no siguen esta estructura como los Desistimientos. Los documentos no cuentan con ningún tipo de metainformación que sirva en el procesamiento, todas las propiedades se deberán obtener del texto propio de la resolución.

Es necesario aclarar que esta estructura ha variado a través de los años, determinada tanto por el uso de las personas como de los formatos digitales en los que se almacenan, lo que implica un esfuerzo importante en la normalización del texto.

Base de Datos Nexus

El Poder Judicial ha desarrollado una aplicación Web que permite la consulta de diferentes documentos de las diferentes salas. Esta base de datos contiene una clasificación con base en criterios técnicos realizado por un departamento especializado, aunque no contempla el compendio total de documentos. Además de estos temas, la base de datos cuenta con las siguientes variables:

Año, Tema, Subtema, se menciona se la resolución es clave, es estructural y si es relevante según el criterio técnico, la rama del derecho a la cual pertenece, el redactor, si la resolución “es un cambio de criterio”, si tiene voto salvado y el tipo de contenido de interés.

Metainformación contenida en el texto de resoluciones

A continuación, se enumeran la metainformación relevante en cada una de las resoluciones:

Número de Voto

Este número identifica a cada resolución por separado y se define en el momento de la votación. Se compone de un número consecutivo, el carácter guion ("-") y los dos últimos dígitos del año. Este dato se obtiene fácilmente del nombre de archivo del documento.

Número de Expediente

Este número se define en el momento que se inscribe un proceso judicial en la Sala. Tiene la forma "123456789012AA" aunque aparece en los documentos con diferentes separaciones.

Fecha y Hora

Estos datos se encuentran en forma de texto dentro del encabezado del documento, lo que implica analizar el texto con todas las posibles combinaciones mediante expresiones regulares y construir el dato a partir de esta información.

Redactor

Cada sentencia de la Sala Constitucional viene acompañada de una resolución que es redactada por uno de los magistrados, esta información es importante para realizar un análisis sobre las posturas de cada magistrado con base de su lenguaje y criterio legal.

Voto Salvado

El voto salvado ocurre cuando uno o varios de los magistrados mantienen un criterio distinto a la mayoría, y hace patente este criterio en una nota que se adjunta al documento. Este párrafo puede ser especialmente utilizado para realizar un análisis sobre los criterios de cada magistrado y su postura ante determinados temas y su relación con otras sentencias de la Sala.

Sentencia

La sentencia es la decisión que dictaminó la Sala en un caso particular. No todos los documentos tienen una sentencia definitiva, dado que muchos de ellos pueden ser simplemente recomendaciones de la corte, solicitudes de evidencias u otro tipo de información más relacionada con el quehacer de la Sala.

Entorno de Programación y Herramientas

Para el desarrollo de las herramientas se utiliza el lenguaje de programación Python junto con la librería NLTK (Natural Language ToolKit) y Apache Tika en Java. Los resultados de cada procesamiento se guardan en la base de datos no relacional MongoDB porque permite adaptar cada registro con la información obtenida de forma particular, además de facilitar la extensión del registro de forma dinámica y mantener una eficiencia de procesamiento aceptable.

Python NLTK es una librería originalmente creada en 2001 como parte un curso de lingüística computacional en el Departamento de Ciencias de la Computación y de la información en la Universidad de Pensilvania. Fue diseñada teniendo como metas principales la simplicidad, consistencia, extensibilidad y modularidad, en nuestro caso particular, se utiliza las herramientas de tokenización y segmentación, además de la producción de una representación vectorial del texto de las resoluciones.

Apache Tika es una colección de herramientas utilizadas ampliamente en el análisis de textos. Apache Tika se utiliza para convertir diferentes formatos de texto.

Proceso de extracción de la metainformación y resultados

A continuación, se describe el procedimiento empleado para la extracción de la información de la primera etapa del proyecto,

Conversión de los documentos

El primer paso del procesamiento es la extracción del texto de los diferentes tipos de documentos. Para ello, se utilizan diferentes herramientas según el formato del archivo. En la Tabla 1 se muestra los resultados de esta conversión, cabe resaltar que se tuvo que realizar una rutina usando COM Objects para el procesamiento de los archivos en Word, también se especifican los tipos de archivos que se encuentran en el repositorio pero que no tienen información útil para el procesamiento de las resoluciones. Para otros formatos se utilizó Apache Tika, una librería en Java que reconoce el tipo de archivo y extrae automáticamente el texto.

No todas las resoluciones mantienen el formato legal esperado, lo que implica una gran cantidad de casos de uso que para este primer objetivo no tienen relevancia.

En total existen 365 889 documentos “con tipo de archivo relacionada con texto”, como se mencionó con anterioridad, un 0.5% de estos archivos no tienen contenido. En la Gráfico 1 se puede observar la distribución de estos archivos por año. Como se observa, a partir del 2000 se cuenta con más de 11 000 documentos por año, siendo 2014 el año con mayor cantidad de resoluciones (19 683 documentos) mientras que en la década de 1990 la cantidad es menor, y esto, sumado a problemas de formato y codificación, hace que la muestra de documentos para el análisis sea aún menor que lo esperado. Por otra parte, en 1998 existe un faltante del texto de 727 documentos, lo que representa el 13% de los archivos existentes (representa el 40% del total de documentos faltantes).

Cuadro 1
Resultados y Observaciones de la conversión del formato de los documentos

Ext	Cantidad	Observaciones
.xps	2	Dos archivos con resoluciones en formato de imagen. Se utiliza un programa OCR para procesarlos
.rf	1	Archivo rtf con extensión mal escrita
"	88	Son directorios que no filtra el sistema o copias de archivos word
.rtf	15527	Se utiliza un script con COMObject en Python y Word para su conversión
.pdf	212	Se utiliza la herramienta pdftotext
.doc	172031	Se utiliza un script con COMObject en Python y Word para su conversión
.docx	840	Se utiliza un script con COMObject en Python y Word para su conversión
.txt	240	Simplemente se leen los archivos
.thmx	44	Son con metadatos de algunos archivos html, no contienen información importante para el proceso.
.dotx	1	Se utiliza un script con COMObject en Python y Word para su conversión
.htm	9011	Archivos html, se usa textract para procesarlo
.db	49	No tienen información útil
.bk	1	No tiene información útil
.xml	110	Sólo contienen información de la página web que no es útil para el proceso
.dot	5	Se utiliza un script con COMObject en Python y Word para su conversión
.html	169051	Archivos html, se usa textract para procesarlo
.wbk	85	Se utiliza un script con COMObject en Python y Word para su conversión
.lnk	7	Archivos de enlace sin información útil
.rcv	5	Archivos de tipo desconocido pero que vienen junto a archivos con el mismo nombre de resolución
.gif	92	Las imágenes no tienen información relevante para el proceso
.dpj	5	No se encuentra información sobre este tipo de extensión, se ignoran los archivos.
.rt_	5	Archivos de tipo desconocido pero que vienen junto a archivos con el mismo nombre de resolución
.jpg	8	Las imágenes no tienen información relevante para el proceso
	367420	

Fuente: Elaboración propia.

Proceso de Extracción de la Metainformación y Resultados

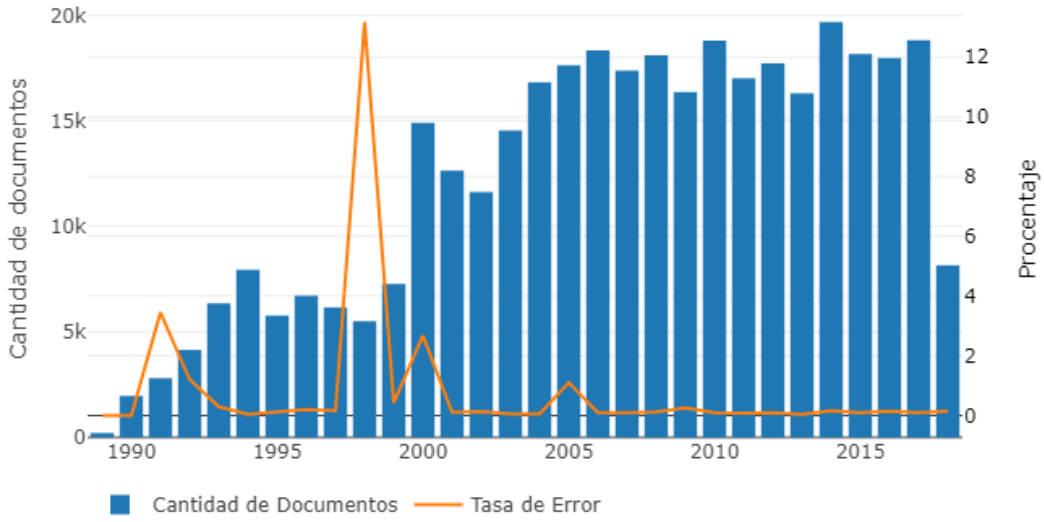
A continuación, se describe el procedimiento empleado para la extracción de la información de la primera etapa del proyecto.

Conversión de los documentos

El primer paso del procesamiento es la extracción del texto de los diferentes tipos de documentos. Para ello, se utilizan diferentes herramientas según el formato del archivo. En la Tabla 1 se muestra los resultados de esta conversión, cabe resaltar que se tuvo que realizar una rutina usando COM Objects para el procesamiento de los archivos en Word, también se especifican los tipos de archivos que se encuentran en el repositorio pero que no tienen información útil para el procesamiento de las resoluciones. Para otros formatos se utilizó Apache Tika, una librería en Java que reconoce el tipo de archivo y extrae automáticamente el texto.

No todas las resoluciones mantienen el formato legal esperado, lo que implica una gran cantidad de casos de uso que para este primer objetivo no tienen relevancia. En el gráfico izquierdo de la Gráfico 1 se presenta el porcentaje de errores encontrados y el número total de resoluciones por año.

Gráfico 1
Distribución de documentos y tasa de error en la conversión a texto



Fuente: Elaboración propio.

En total existen 365 889 documentos “con tipo de archivo relacionada con texto”, como se mencionó con anterioridad, un 0.5% de estos archivos no tienen contenido. En la Gráfico se puede observar la distribución de estos documentos por año,

Claramente se percibe como al inicio de la gestión de la Sala Constitucional no existe una estructura definida para las resoluciones, en medida que los años avanzan, el porcentaje de errores disminuye considerablemente hasta valores inferiores al 1%. En 1998 hay una gran cantidad de resoluciones sin secciones definidas (766 documentos, con un 13% de error) que se debe a un problema en el formato de los archivos elaborados durante ese año.

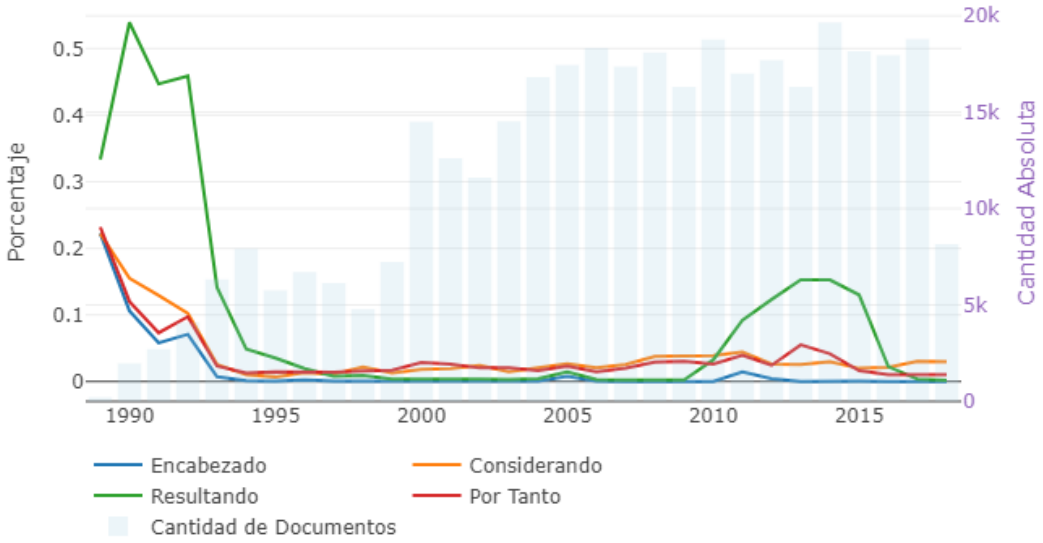
En los resultados se observa como durante los años entre 1989 y 1993 está el mayor porcentaje de error en todos los procesamientos, una posterior revisión de estos casos reveló que estos resultados se deben a que existen muchas resoluciones que no son “sentencias”, sino desestimientos u otro tipo de documento.

Separación en secciones

Para facilitar el procesamiento y el análisis posterior, se separó las diferentes secciones de cada documento (Encabezado, Resultando, Considerando y Por Tanto), lo que a su vez, permite hacer una preselección del tipo de resolución según el formato de la sentencia. Por ejemplo, los desistimientos no cuentan con considerando y resultando.

En la Gráfico 2 se observa como existe una mayor cantidad de documentos irregulares en el inicio de las gestiones de la Sala Constitucional, el caso más sobresaliente es la sección "Resultando" que llega a incluso ser mayor a un 50% en 1991 (1061 documentos), al revisar una muestra de estos errores se observa que se debe a que estos documentos no tienen explícitamente esta sección. Cabe aclarar que en el proceso de extracción de la información no se utiliza esta variable.

Gráfico 2
Tasa de error en la separación de secciones por año



Fuente: Elaboración propia.

Expedientes

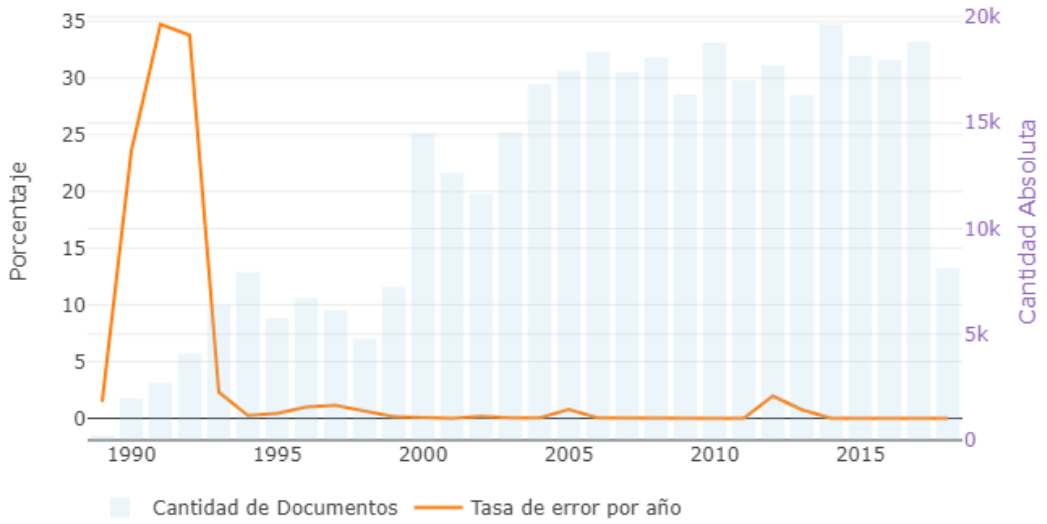
El número expediente es especialmente importante para relacionar los documentos con cada registro de la base de datos Nexus, para su extracción se utilizan varias expresiones regulares diseñadas especialmente para este caso de uso. Algunos documentos no contienen algún número de expediente, estas resoluciones se recolectan en un único archivo para una revisión posterior.

Igual que en el caso anterior, existe una anormalidad en los primeros años de la década de 1990, donde muchos documentos no cuentan con el formato que se estandarizó en años posteriores. En el 2012 el algoritmo no ubicó el 1.3% de los documentos (384) dado que una gran parte no lo tienen escrito de manera explícita.

Número de Voto

Este número identifica a cada resolución por separado y se define en el momento de la votación. Se compone de un número consecutivo, el carácter guion ("-") y los dos últimos dígitos del año. Este dato se obtiene fácilmente del nombre de archivo del documento.

Gráfico 3
Tasa de error en la extracción del número de expediente

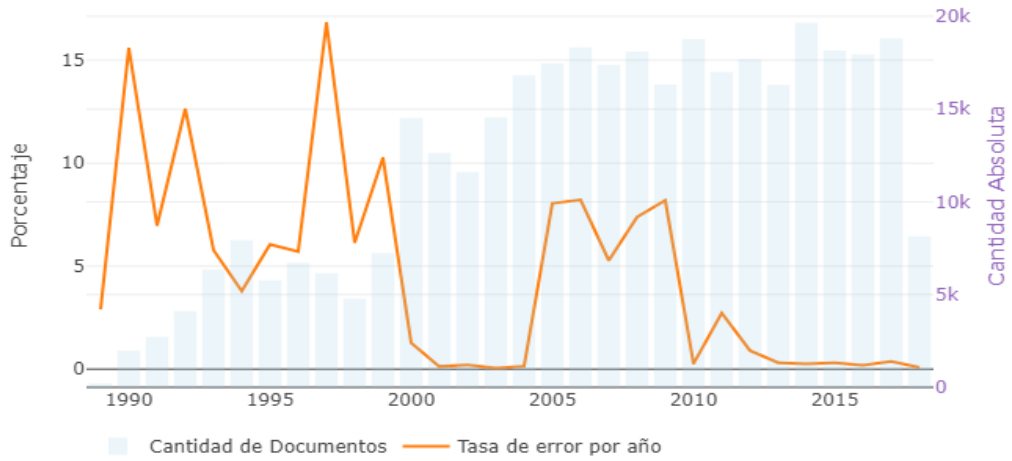


Fuente: Elaboración propia.

Fecha y Hora

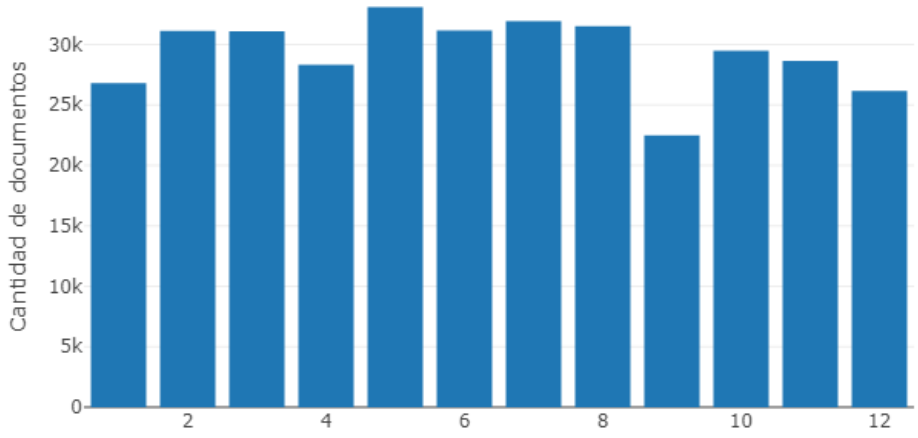
Estos datos se encuentran en forma de texto dentro del encabezado de la resolución (por ejemplo: "a las ocho horas treinta y dos minutos del dos de junio de mil novecientos noventa y cinco."), lo que implica analizar el texto con todas las posibles combinaciones mediante expresiones regulares y construir el dato a partir del texto. Esta variable se utiliza como llave secundaria que relaciona con la base de datos de Nexus en los casos donde no se cuente con el número de expediente.

Gráfico 4
Tasa de error en la extracción de fecha y hora



Fuente: Elaboración propia.

Gráfico 5
Cantidad total de documentos por mes



Fuente: Elaboración propia.

Redactor

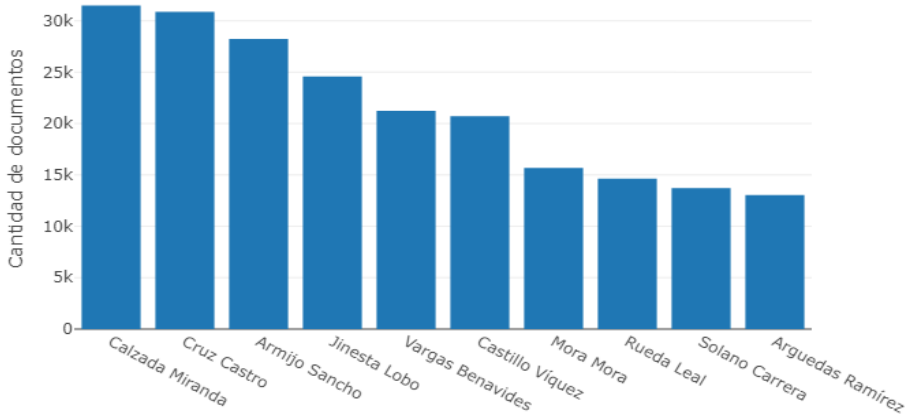
Cada sentencia de la Sala Constitucional viene acompañada de una resolución que es redactada por uno de los magistrados, esta información es importante para hacer un análisis sobre las posturas de cada magistrado con base de su lenguaje y criterio legal. Se puede acceder a esta información desde el texto mismo de la resolución o desde la base de datos Nexus.

Gráfico 6
Tasa de error en la extracción del redactor

Fuente: Elaboración propia.

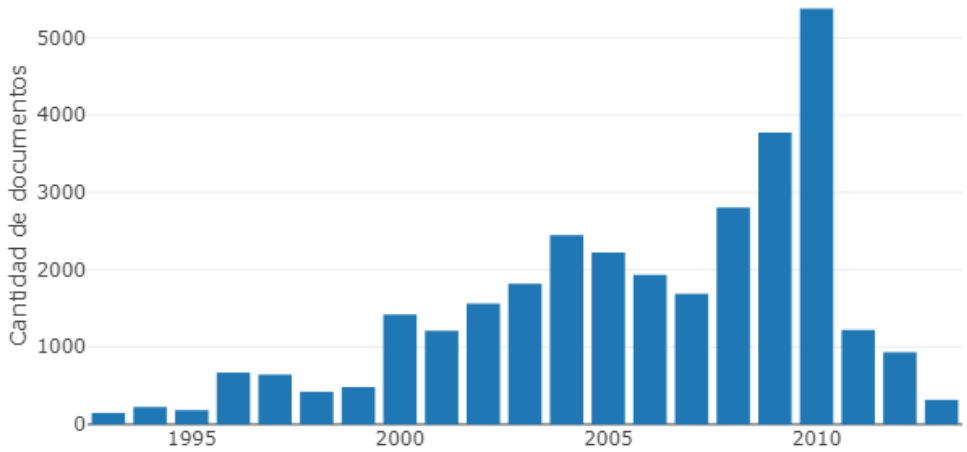
En la Gráfico 8 se presentan los 10 redactores con más cantidad de documentos, la mayor cantidad la presenta "Calzada Miranda" con 31,487 resoluciones a su nombre.

Gráfico 7
Distribución de documentos de los 10 mayores redactores



Fuente: Elaboración propia.

Gráfico 8
Documentos redactados por Calzada Miranda por año

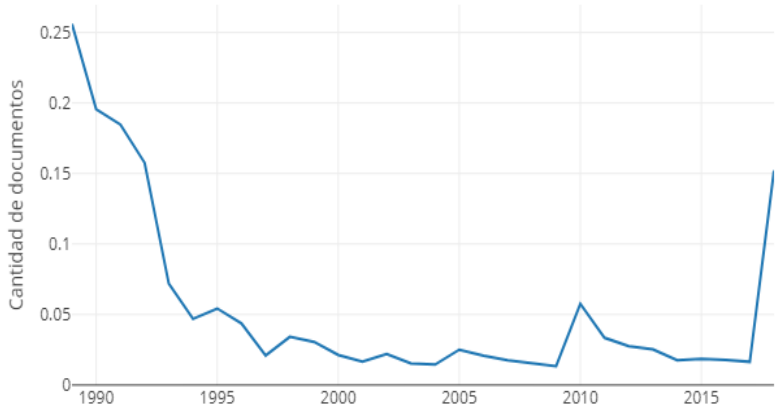


Elaboración propia.

Recurrente

Es la persona que presenta la denuncia ante la Sala Cuarta, se puede obtener información como la cédula y estado civil, aunque muchos de las resoluciones tienen los campos ocultos.

Gráfico 8
Tasa de error en la extracción del recurrente



Fuente: Elaboración propia.

Recurrido

Es la persona o institución contra quien se presenta la denuncia. De la misma manera, se puede obtener información como la cédula y estado civil, aunque de la misma forma puede venir oculta.

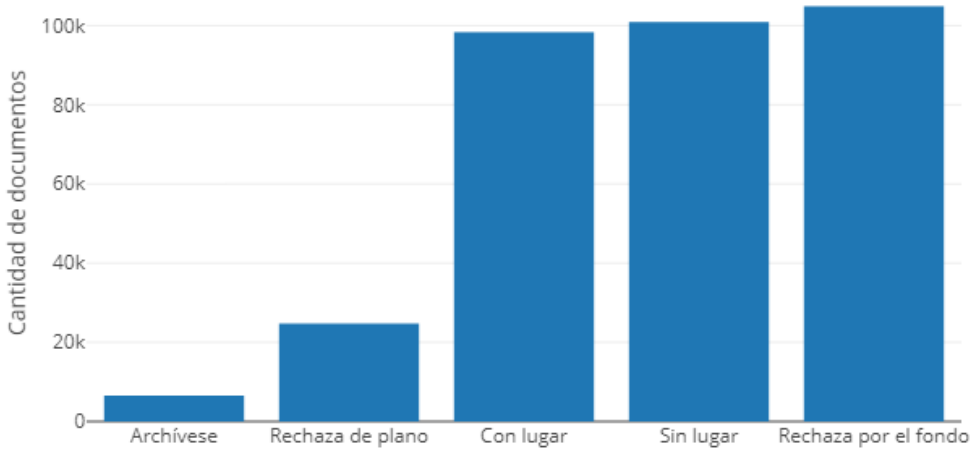
Fuente: Elaboración propia.

Sentencia

La sentencia es la decisión que dictaminó la Sala en un caso particular. No todos los documentos tienen una sentencia, dado que muchos de ellos pueden ser simplemente recomendaciones de la corte, solicitudes de evidencias u otro tipo de información más relacionada con el quehacer de la Sala.

Se presupone que una sentencia se da siempre y cuando exista un "Por tanto" en la sentencia, lo que implica que no se procese una tasa promedio de 10% de los documentos de inicio de las gestiones de la Sala.

Gráfico 9
Documentos por sentencia de la corte



Fuente: elaboración propia.

Tipo de proceso

Son los diferentes tipos de resoluciones que se pueden presentar en la Sala Constitucional, entre ellos se encuentran:

- Hábeas Corpus
- Acción de Inconstitucionalidad

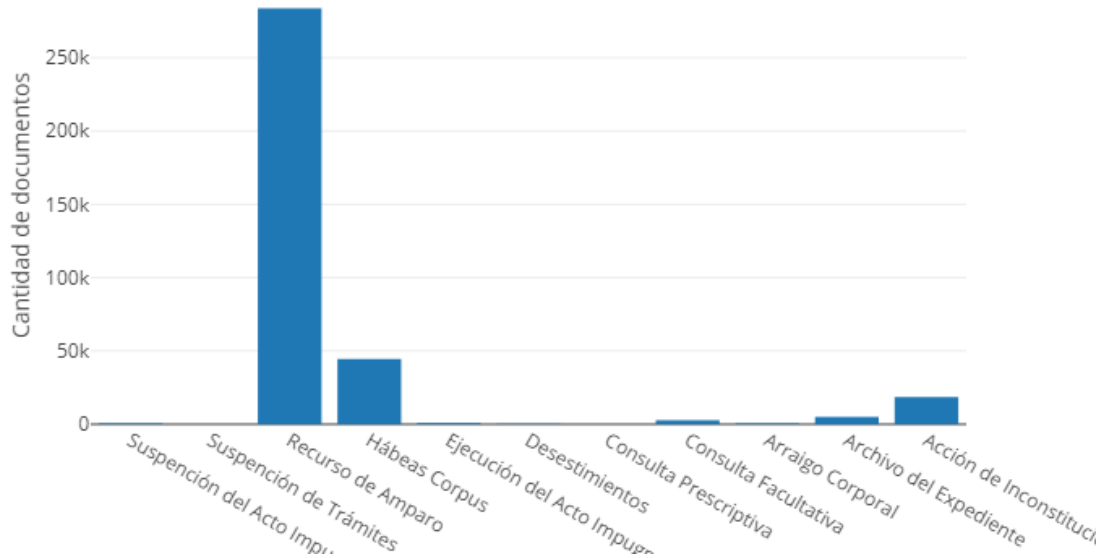
- Consulta Facultativa
- Recurso de Amparo

Por otra parte, dentro de los documentos se han encontrado otro tipo de resoluciones:

- Libertad del Agraviado
- Arraigo Corporal
- Desestimientos
- Suspensión de Trámites
- Suspensión del Acto Impugnado
- Ejecución del Acto Impugnado
- Archivo del Expediente

De toda la distribución de documentos, los recursos de amparo son los que ocupan la mayor cantidad de las resoluciones, con un total de 283,762, los hábeas corpus se compone por 44,607 sentencias.

Gráfico 10
Distribución de los documentos según tipo de proceso



Fuente: elaboración propia.

Tema

A partir de los datos de Nexus, se obtiene el tema y subtema de la resolución que realiza el Departamento de Jurisprudencia del Poder Judicial. Sólo se pudo relacionar 77,232 documentos con las entradas de Nexus. Cada registro tiene un número variables de temas (de 1 a 7).

En total, la base de datos de Nexus tiene 2,181 temas diferentes, que se agregaron en temas afines para un total de 734 temas (ver tabla 2).

Para clasificar el resto de los documentos, se genera un modelo computacional para estimar las categorías para los documentos restantes.

El algoritmo utiliza una red neuronal densa que con una precisión de 99% usando como criterio un set de datos que el modelo no conoce. Con este porcentaje, es confiable clasificar el resto de los documentos.

Clusterización

Como parte de las estrategias y dada la gran importancia que tiene reconocer los temas de las sentencias, se utiliza *Latent Dirichlet Allocation* (LDA)[5], que agrupa los documentos por grandes temas tomando como criterio la frecuencia normalizada de los términos. En la Gráfico 13 se observa como estos documentos giran alrededor de grandes temas como educación, salud, penal y municipal.

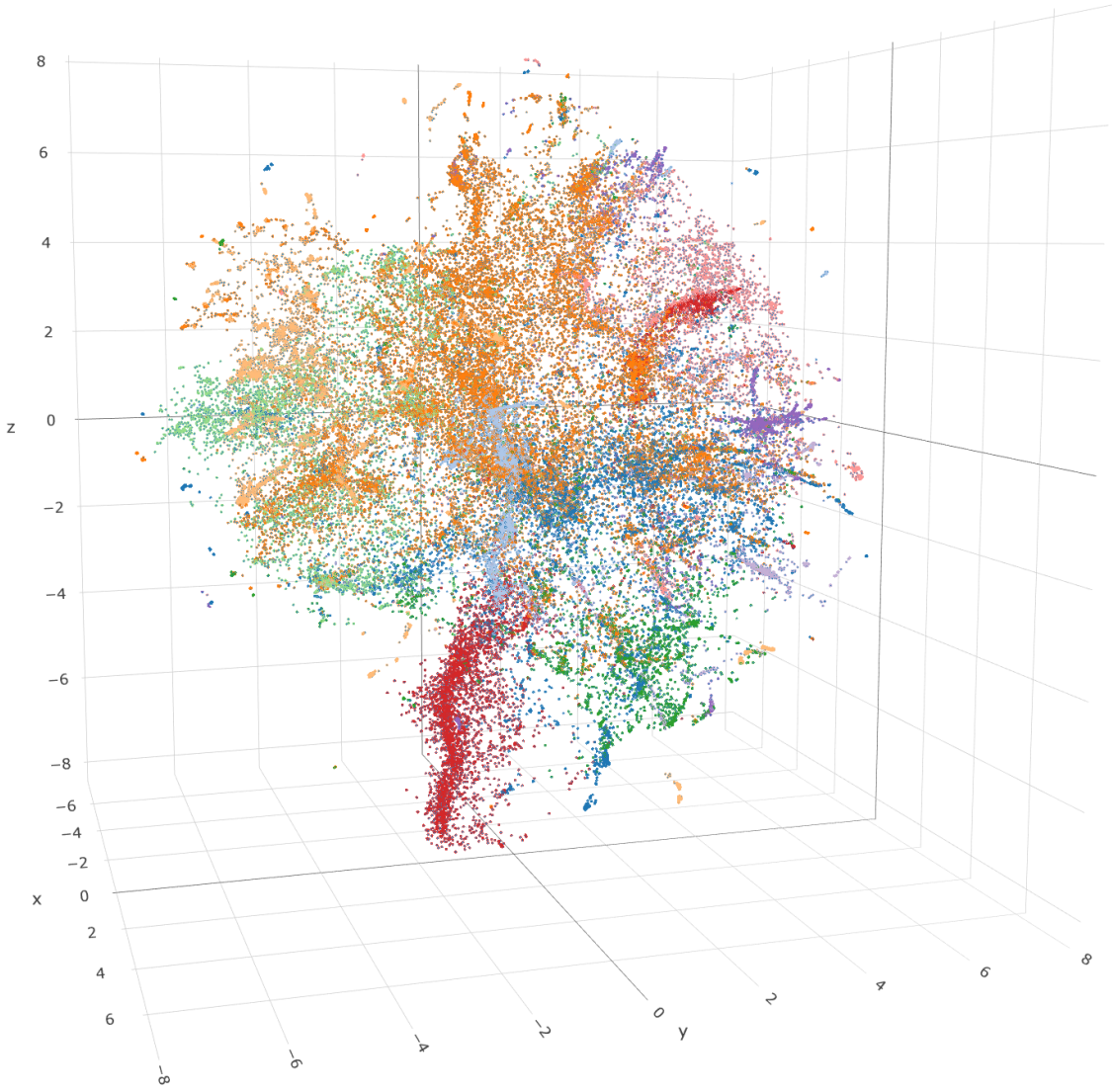
Gráfico 11
Agrupamiento de temas mediante LDA

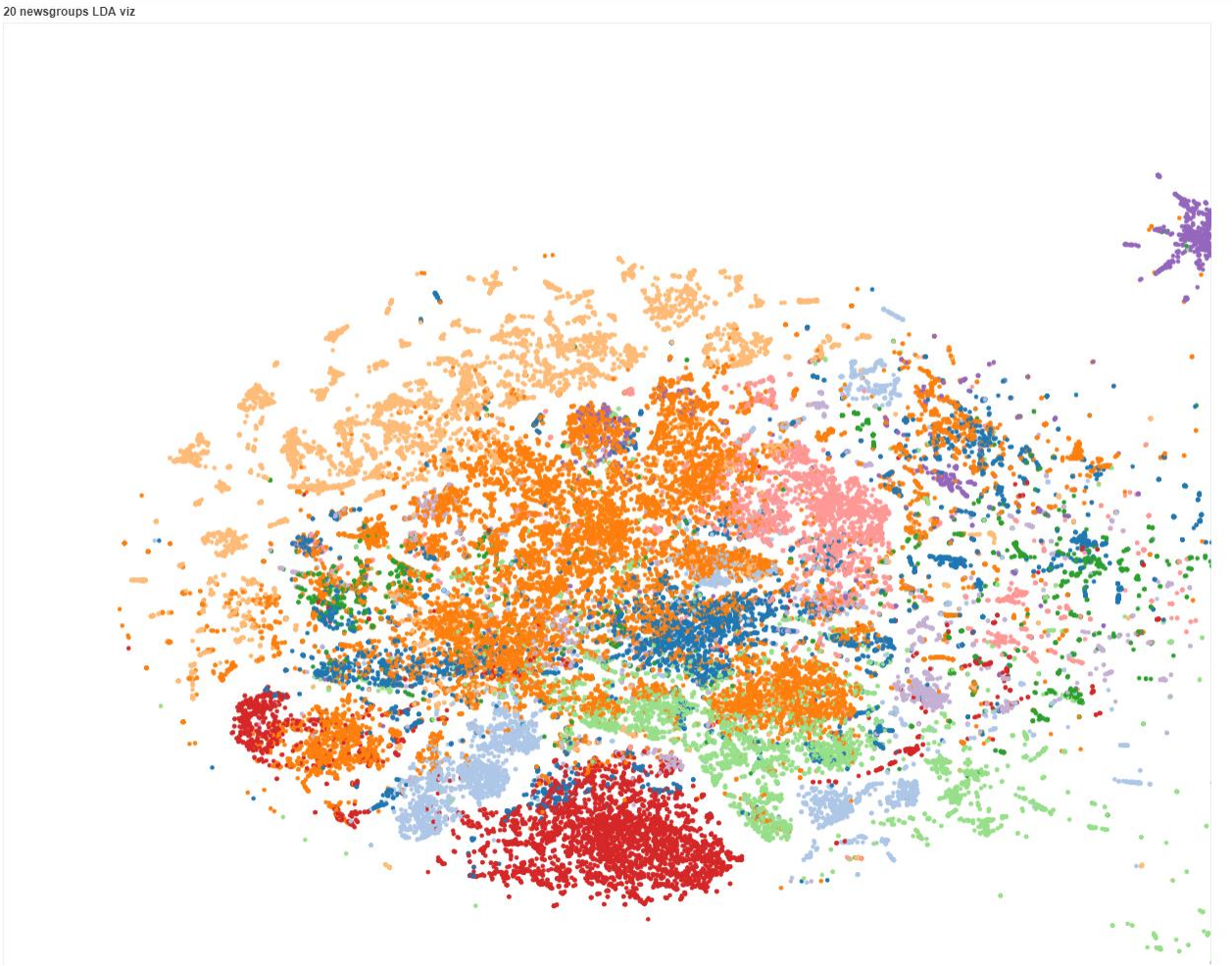
Fuente: elaboración propia.

En la siguiente figura se muestra una representación del espacio vectorial generado para los documentos de la Sala Constitucional, donde se muestran cada uno de los 10 grupos con diferentes colores. Esta representación se realizó utilizando la implementación de Barnes-Hut

del algoritmo tSNE[6], que consiste en obtener una estimación probabilística de la semejanza de los puntos de un espacio de alta dimensionalidad, para con ellos buscar una representación que mantenga las diferencias entre los puntos en un espacio con menos dimensiones (en nuestro caso un espacio de 3 dimensiones).

En la figura se puede observar como los clusters encontrados mantienen una cercanía espacial, esta cercanía se analiza mediante el vocabulario y la misma estructura del texto.





Reconocimiento de entidades en el texto

El reconocimiento de entidades consiste en la identificación de nombres de personas, personas y organizaciones dentro del texto de las resoluciones. Para realizar esto, se utiliza modelos entrenados previamente con datos que son etiquetados por lingüistas o expertos en el campo específico de lo que tratan los documentos. Para realizar este proceso, se utiliza la librería *"SpaCy"*[7].

En la siguiente figura se aprecian las diferentes entidades reconocidas en un documento en particular, las entidades están marcadas con diferentes colores según la categoría en la que son clasificadas (LOC para lugar, ORG para organización y PER para personas, MISC se utiliza cuando el modelo cree encontrar una entidad pero no sabe cómo clasificarla).

Con MISC lugar amparo de Luis Vega PER c/ Dirección Gral. PER Servicio Civil LOC y MAG DESCRIPTORES MISC : Afectación MISC de derechos laborales por la prolongación de procesos de reestructuración o reorganización administrativa. PER Exp. 1539-V-97 Nº MISC 5585-97 SALA MISC CONSTITUCIONAL DE ORG LA CORTE SUPREMA ORG DE JUSTICIA MISC . San José LOC , a las catorce horas con veinticuatro minutos del doce de setiembre de mil novecientos noventa y siete.- Recurso PER de amparo interpuesto el 11 de marzo de 1997 por Luis Román Vega Morales PER , portador de la cédula de identidad nº 9-068-983, contra la Dirección General de Servicio Civil ORG y el departamento de Recursos Humanos del LOC Ministerio de Agricultura LOC y Ganadería LOC . Intervino LOC también en el proceso el señor Ricardo Garrón Figuls PER , en calidad de Ministro de Agricultura y Ganadería LOC .

Resultando: 1.- MISC Señala el recurrente (folio 1) que interpone recurso de amparo contra la Dirección General de Servicio Civil ORG y el Departamento de Recursos Humanos del Ministerio de Agricultura y Ganadería LOC , por estimar ilegítimo y contrario a sus derechos fundamentales, el hecho de que se le hayan rechazado, en forma reiterada, las solicitudes de reasignación del puesto que ocupa –por estimar que tiene mayores responsabilidades que las que establece el manual para ese cargo–, bajo el pretexto de que como el Ministerio de Agricultura y Ganadería LOC se encuentra en proceso de reorganización, no puede darse trámite a su solicitudes, con arreglo a lo que dispone el artículo 110 del Reglamento LOC al Estatuto de Servicio Civil MISC . Reclama LOC además, que el Ministerio LOC tiene ya casi tres años de estar en proceso de reorganización y no se sabe cuando va a concluir, circunstancia que torna imposible el disfrutar de una reorganización en esa Cartera LOC ; estima también que la norma citada es inconstitucional, pues establece una restricción que el Estatuto MISC –la ley– no contempla. Solicita MISC el recurrente que se ordene a los recurridos tramitar su solicitud de reasignación de puestos, junto con la correspondiente condenatoria en costas, daños y perjuicios. PER 2.- Por resolución interlocutoria nº 127-I-97 de las 14:06 hrs del 1 de abril del corriente (folio 41), la Sala MISC dispuso dar curso al amparo en cuanto al cuestionamiento que hace sobre la razonabilidad del plazo hasta ahora observado por la administración para ejecutar el proceso de reestructuración que sufre el Ministerio de Agricultura y Ganadería LOC , y se

En el caso de estudio, se hizo el procesamiento con las tres librerías más utilizados en la actualidad, con resultados diversos. Para mejorar drásticamente la calidad del procesamiento, se debe contar con una muestra de documentos debidamente etiquetados a mano (alrededor de unos 5000.

Relación entre la cantidad de Resoluciones presentadas por tema y el porcentaje de éxito.

El objetivo de este análisis es determinar si existe la tasa de éxito de un caso en algún determinado tema, predice el volumen de nuevos casos presentados a continuación, de tal forma que se pueda determinar si existe un aprendizaje por parte de la población, o si una sentencia desfavorable desincentiva a la población en presentar nuevos casos.

En los gráficos presentados a continuación, se presenta la cantidad de nuevos casos y el porcentaje de éxito de los años 2000 al 2015 para los temas de salud, municipalidad, ambiente, trabajo y penal y penitenciario.

Un ejemplo del comportamiento que se desea observar es el que se observa en el tema Penitenciario entre los años 2005 al 2010, existe una tendencia al alza durante esos años en la cantidad de nuevos casos y la tasa de éxitos, de manera sostenida durante este periodo.

Otro ejemplo es el observado en el tema de Municipal, donde existe una tendencia al alza entre los años 2005 al 2013, pero que cambia de tendencia luego de que durante el 2012 la tasa de éxito bajó drásticamente.

Conclusión

Los resultados presentados marcan un avance en el procesamiento de la información contenida en cada una de las resoluciones y su posterior análisis, lo que permite, a mediano plazo, obtener indicadores confiables, basados en una muestra que se acerca a la totalidad de la población de documentos, todo esto con la facilidad de poder incorporar nuevas resoluciones fácilmente.

Las aplicaciones se extienden desde aplicar los algoritmos implementados en los procesos de "indexación" del Poder Judicial, un "clasificador automático" bajo los criterios ya empleados en el Departamento de Jurisprudencia, hasta la "Estimación del resultado de un caso con base en el comportamiento de la Sala Constitucional en el tiempo".

En el análisis político-judicial, a partir de la información extraída, se puede análisis la evolución del criterio de la Sala para un tema determinado, y, con la asistencia de juristas expertos, se puede analizar cómo la Corte está influenciada por distintas doctrinas de las ciencias jurídicas.

Además, es posible analizar las relaciones que existen entre resoluciones pertenecientes entre distintos temas, incorporando la variable del tiempo.

A nivel computacional, los algoritmos se implementaron utilizando técnicas de paralelización, con herramientas que se encuentran en el "estado del arte". El reto indiscutible, es poder

incorporar todo nuevo avance en el "Procesamiento de Lenguaje Natural" y en el Análisis de Texto.

Entre las dificultades que se deben mitigar se encuentran la gran cantidad de formatos de archivos en lo que se han digitalizado las resoluciones, este problema agrega irregularidades en la platilla utilizada y la consecuente dificultad para identificar los patrones en el texto.

En los años entre 1989 y 1993 hay una gran disparidad en las resoluciones, lo que dificulta la identificación de patrones dentro del texto, y en otros casos, la ausencia de metadatos en el cuerpo del documento.

En el número de expediente hay dos tipos de numeración, y cada uno puede encontrarse con diferentes formatos, aun así, el modelo identifica un gran porcentaje de información de los documentos.

Trabajo Posterior

El proyecto debe continuar en la extracción de los demás campos de información, la identificación de posibles llaves para poder relacionar las sentencias con otras bases de datos y la generación de modelos de datos para un análisis más profundo.

Agradecimiento

Se agradece a Programa Estado de la Nación, y, particularmente al Informe Estado de la Justicia por facilitar y asesorar esta investigación.

Además, se reconoce la asesoría y colaboración de la Dirección de Tecnología del Poder Judicial.

Referencias

[1] J. Todd, "The promises and pitfalls of automated content analysis in judicial politics," pp. 1–43.

[2] M. Evans and C. L. Cates, "Recounting the Courts ? Toward A Text-Centered Computational Approach to Understanding the Dynamics of the Judicial System." 2005.

[3] K. T. McGuire and G. Vanberg, "Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law," *Prep. Deliv. 2005 Annu. Meet. Am. Polit. Sci. Assoc.*, p. Washington, D.C., 2005.

[4] Congreso Constitucional de la República de Costa Rica, *Ley Orgánica del Poder Judicial*. 1937.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, 2003.

[6] L. van der Maaten and G. Hinton, “Visualizing Data using {t-SNE},” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[7] M. Honnibal and I. Montani, “{spaCy 2}: Natural language understanding with {B}loom embeddings, convolutional neural networks and incremental parsing,” 2017.