Equipo de Estado Abierto

Ciudad de México, 5 de febrero de 2025

Nota Metodológica 1
Clasificación y análisis temático de las Solicitudes de Información Pública (SIP) 2024

Este documento describe el proceso metodológico implementado para la categorización y análisis de las solicitudes de información pública recibidas en 2024. El objetivo principal fue desarrollar un sistema de clasificación que permita identificar patrones temáticos relevantes en las solicitudes, permitiendo que se homologue su organización y sistematización para análisis estadístico (y comparación trimestral y anual) y para la generación de reportes. La metodología combina técnicas de procesamiento de texto con herramientas de análisis de datos para construir categorías generales y específicas, utilizando un enfoque replicable y adaptado a las características del conjunto de datos.

La gestión de solicitudes de información pública implica el manejo de grandes volúmenes de datos textuales, que requieren metodologías sólidas para su organización y análisis. En este estudio, se diseñaron y aplicaron dos variables principales: (1) temas_nuevos, que clasifica las solicitudes en categorías generales basadas en patrones temáticos amplios, y (2) temas_adicionales, que refina la clasificación de solicitudes que inicialmente no pudieron ser categorizadas con precisión. Este enfoque busca responder a la necesidad de estructurar datos textuales no uniformes en temas relevantes para la gestión pública.

Metodología

El proceso metodológico se desarrolló en tres fases principales: estructuración y carga de datos, diseño de variables de clasificación y generación del archivo procesado. Cada fase se detalla a continuación.

Fase 1: Estructuración y carga de datos

El análisis partió de un archivo en formato Excel con información relativa a las solicitudes de información pública obtenidas del Sistema de Solicitudes de Acceso a la Información (SISAI 2.0) de la Plataforma Nacional de Transparencia, el cual contenía una columna denominada SOLICITUD, que almacena las descripciones textuales de las solicitudes de información. Se utilizó el software R para la lectura y procesamiento de los datos, empleando la librería readxl. Durante esta etapa, se verificó la calidad del archivo y se corrigieron posibles inconsistencias estructurales, asegurando que todos los registros fueran procesables.

Fase 2: Diseño de las Variables de Clasificación

Variable temas_nuevos

La variable temas_nuevos se diseñó como una categorización inicial de las solicitudes en siete categorías generales:

- Relación con la sociedad
- Actos de gobierno

- Organización interna
- Programático, presupuestal y financiero
- Informes y programas
- Regulatorio
- Otros (categoría residual para solicitudes no clasificables en las anteriores).

Para asignar estas categorías, se utilizaron expresiones regulares (regex), una herramienta para la identificación de patrones textuales. Por ejemplo, términos como "relación con sociedad" o "sociedad y relación" fueron indicativos para asignar la categoría "Relación con la sociedad".

Variable temas_adicionales

La variable temas_adicionales refina la clasificación de aquellas solicitudes categorizadas como "Otros" en temas_nuevos. Para su diseño, se establecieron 13 categorías específicas:

A continuación, se detalla el diseño metodológico de cada categoría temática definida en la variable temas_adicionales, incluyendo ejemplos de las palabras clave utilizadas para identificar las solicitudes correspondientes. Estas categorías reflejan áreas específicas de interés relacionadas con las solicitudes de información pública.

Agua

Esta categoría agrupa solicitudes relacionadas con temas hídricos, incluyendo la gestión de recursos y problemáticas asociadas al agua. Los términos utilizados para identificarla incluyen:

- *Río*
- *Tratamiento*
- *Pozo*
- *Sequía*
- *Reservorio*
- *Manto freático*
- *Inundación*
- *Tuberías*
- *Canalización*

Construcción

Las solicitudes categorizadas como construcción están vinculadas con proyectos de infraestructura, arquitectura y obras civiles. Palabras clave empleadas:

- *Edificio*
- *Infraestructura*
- *Remodelación*
- *Carretera*
- *Puente*
- *Ampliación*
- *Zona*
- *Instalación*

Contrato

Incluye solicitudes relacionadas con acuerdos legales, contratos y procesos de licitación. Los términos identificados son:

- *Acuerdo*
- *Convenio*
- *Licitación*
- *Pacto*
- *Arrendamiento*
- *Rescisión*
- *Servicio*

Informe

Las solicitudes en esta categoría están asociadas a documentos oficiales y reportes informativos. Ejemplos de palabras clave:

- *Informe*
- *Diagnóstico*
- *Resultados*
- *Memoria*
- *Análisis*
- *Boletín*
- *Presentación*

Obra

Incluye solicitudes relacionadas específicamente con la planificación, ejecución o evaluación de obras públicas. Palabras clave utilizadas:

- *Obra*
- *Zona de obra*
- *Gestión*
- *Plan*
- *Instalación*
- *Construcción*

Presupuesto

Categoría enfocada en solicitudes relacionadas con la asignación y uso de recursos financieros. Términos identificados:

- *Presupuesto*
- *Gasto*
- *Inversión*
- *Asignación*
- *Financiamiento*
- *Recorte*
- *Costo*
- *Saldo*

Programa

Esta categoría agrupa solicitudes que hacen referencia a programas o políticas públicas. Palabras clave:

- *Programa*

- *Campaña*
- *Estrategia*
- *Actividad*
- *Intervención*
- *Política pública*

Proyecto

Incluye solicitudes vinculadas al diseño, propuesta o implementación de proyectos específicos. Ejemplos de términos utilizados:

- *Proyecto*
- *Iniciativa*
- *Propuesta*
- *Innovación*
- *Modelo*
- *Plan piloto*

Recursos

Esta categoría comprende solicitudes relacionadas con la disponibilidad, administración y uso de recursos materiales o financieros. Palabras clave empleadas:

- *Materiales*
- *Insumos*
- *Activos*
- *Capital*
- *Fondos*
- *Suministro*

Seguridad

Incluye solicitudes enfocadas en temas de vigilancia, prevención de riesgos y protección. Los términos considerados fueron:

- *Seguridad*
- *Protección*
- *Vigilancia*
- *Prevención*
- *Riesgo*
- *Alarma*
- *Policía*

Servicios

Categoría que agrupa solicitudes relacionadas con la prestación de servicios públicos y asistencia. Palabras clave utilizadas:

- *Servicio*
- *Mantenimiento*
- *Gestión*
- *Asistencia*
- *Capacitación*
- *Soporte*
- *Atención*

Trabajo

Las solicitudes en esta categoría se centran en temas laborales, empleo y actividades profesionales. Ejemplos de términos clave:

- *Trabajo*
- *Empleo*
- *Labor*
- *Ocupación*
- *Recursos humanos*
- *Actividades*
- *Oficio*

Transparencia

Incluye solicitudes relacionadas con el acceso a la información y la rendición de cuentas. Palabras clave utilizadas:

- *Transparencia*
- *Acceso*
- *Publicidad*
- *Apertura*
- *Responsabilidad*
- *Rendición de cuentas*
- *Control*
- *Claridad*

Estas categorías permiten una clasificación organizada de las solicitudes, de manera que se facilita el análisis temático de los datos. El diseño de esta variable propone que las solicitudes más complejas o menos estructuradas se incluyeran en el análisis, asignándoles una categoría que refleje mejor su contenido.

Es importante describir que se utilizó un enfoque jerárquico: temas_adicionales sólo se aplicó a solicitudes previamente clasificadas como “Otros” en temas_nuevos. Esto evitó redundancias y que la asignación temática siguiera una estructura lógica y secuencial.

Fase 3: Implementación y generación del archivo procesado

El procesamiento técnico se realizó utilizando las librerías de R: dplyr para la manipulación de datos, stringr para la detección de patrones textuales y writexl para guardar los resultados. El flujo de trabajo incluyó:

1. Clasificación temática: Aplicación de expresiones regulares para detectar coincidencias textuales en la columna SOLICITUD.
2. Asignación condicional: Uso de mutate y case_when para asignar valores a las variables temas_nuevos y temas_adicionales.
3. Exportación de resultados: Generación de un nuevo archivo Excel que incluye las columnas originales y las variables creadas.

El archivo resultante proporciona una representación estructurada de los datos.

Es importante mencionar que existen limitaciones inherentes a este método, por ejemplo, que las solicitudes que no incluyan términos del banco de palabras pueden quedar sin clasificar.