



# Tarea 1

23 de abril de 2025  
Franco Chiappe - Tomás Pérez

---

## Parte 4 - [2 pts] - Informe

### 1. Introducción

Este informe tiene como objetivo analizar las sesiones parlamentarias registradas entre los años 2023 y 2024, a partir de la información disponible en archivos CSV. Se busca identificar los temas mas repetidos en las intervenciones parlamentarias mediante tecnicas de procesamiento y limpieza de texto.

### 2. Metodología

#### 2.1 Carga de datos

Se utilizaron cuatro archivos CSV:

- `parlamentarios_info_general.csv`
- `participacion_descripcion_2023_1.csv`
- `participacion_descripcion_2023_2.csv`
- `participacion_descripcion_2024.csv`

Los ultimos tres fueron juntafos para formar un unico `DataFrame` de intervenciones parlamentarias.

#### 2.2 Preprocesamiento

Durante el analisis inicial del `DataFrame`, se detectaron columnas con una alta proporción de valores nulos:

- `TEXTO_ANTECEDENTE` (35.768 nulos): fue eliminada por no aportar valor.
- `BOLETIN_ID` (47.568 nulos): se mantuvo porque se va a usar para crer tablas

#### 2.3 Limpieza de texto

Se diseñó una función `limpiar_texto` para normalizar los textos parlamentarios:

- Conversión a minúsculas.
- Eliminar puntos y acentos.
- Reducción de espacios innecesarios.

Esta función la usamos para las columnas `DESCRIPCION_DEBATE` y `TEXTO_PRINCIPAL`. Luego, ambas columnas fueron concatenadas en una nueva llamada `TEXTO_COMBINADO`.

#### 2.4 Detección de temáticas

Se definieron 10 tematicas clave, cada una asociada a un conjunto de palabras clave. Como: **salud**, **educacion**, **delincuencia**, **medio ambiente**, **economia**, entre otras.

La función `asignar_keywords`, se evaluo cada intervencion en base a la presencia de estas palabras clave, generando una columna `KEYWORDS` que asocia cada discurso a uno o mas temas.

### 3. Rol de las Consultas SQL

El análisis de los datos parlamentarios se fundamentó en la utilización estratégica de consultas SQL ejecutadas sobre **BigQuery**, permitiendo transformar un conjunto de datos extensos y poco estructurados en información clara y relevante. Estas consultas cumplieron un rol central en cada etapa del proceso analítico, desde la exploración inicial de los datos hasta la validación empírica de nuestras conclusiones. Para comenzar, se diseñó un modelo de datos relacional que separa las entidades clave —parlamentarios, partidos políticos, intervenciones y tópicos (keywords)—, lo que permitió consultas más eficientes. A partir de este modelo, se desarrollaron consultas específicas para responder a las preguntas planteadas en el enunciado, como identificar las temáticas más discutidas por mes, analizar la evolución de la actividad parlamentaria mediante una media móvil de intervenciones por partido, o determinar los temas dominantes por partido en cada trimestre. En todos los casos se hizo un uso extensivo de funciones analíticas como `ROW_NUMBER()`, `RANK()`, `OVER(PARTITION BY...)` y `ROLLING AVERAGE`, que permitieron trabajar con datos secuenciales y detectar patrones temporales. También, las consultas se apoyaron en operaciones de agregación y filtros bien definidos, lo que garantizó la precisión y relevancia de los resultados. Estos outputs se transformaron luego en visualizaciones que permiten comunicar de forma accesible, como la predominancia de ciertos temas en contextos políticos específicos o el cambio en las prioridades temáticas a lo largo del tiempo.

### 4. Resultados

- Se procesaron más de 60.000 registros de intervenciones parlamentarias.
- La normalización del texto permitió mejorar la precisión en la detección de temáticas.
- Al agrupar las intervenciones por contenido y aplicar las palabras clave, se logró asignar de manera efectiva uno o más temas a cada intervención.

Este enfoque permitió transformar un corpus textual crudo en un conjunto de datos estructurados con metainformación relevante.

### 4. Conclusiones

- El preprocesamiento y limpieza de datos es un paso crucial para cualquier análisis textual, especialmente en corpus grandes como las sesiones parlamentarias.
- La estrategia de usar *keywords* temáticas demostró ser efectiva para una clasificación básica de los discursos.