Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
IIC2440 – Procesamiento de Datos Masivos

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 11 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 analisis de los datos parlamentarios se fundamento en la utilizacion estrategica de consultas SQL ejecutadas sobre `BigQuery`, permitiendo transformar un conjunto de datos extensos y poco estructurados en informacion clara y relevante. Estas consultas cumplieron un rol central en cada etapa del proceso analitico, desde la exploracion inicial de los datos hasta la validacion empirica 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 permitio consultas mas eficientes. A partir de este modelo, se desarrollaron consultas especificas para responder a las preguntas planteadas en el enunciado, como identificar las tematicas mas discutidas por mes, analizar la evolución de la actividad parlamentaria mediante una media movil 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. Tambien, las consultas se apoyaron en operaciones de agregación y filtros bien definidos, lo que garantizo la precision 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 poloticos especificos o el cambio en las prioridades tematicas a lo largo del tiempo.

4. Resultados

- Se procesaron mas de 60.000 registros de intervenciones parlamentarias.
- La normalización del texto permitió mejorar la precisión en la deteccion de tematicas.
- 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.

De qué se habla en el Congreso?

Tomando la primera query de: Cuáles son las top 5 temáticas más tratadas, para cada mes en el que hay registros. Acumulamos las frecuencias mensuales de las 5 tematicas más frecuentes por mes, con lo que obtenemos la siguiente tabla:

Keyword	Frecuencia
legislacion	842
constitucion	446
derechos humanos	128
economia	217
educacion	57
delincuencia	52
salud	18
justicia	1

Por lo que notamos que con diferencia nuestra Keyword Legislación es mucho más frecuente, probablemente por lo amplio que es el tema. En general vemos que son 4 temas principales, legislación, constitución, economía y derechos humanos. Con lo que entendemos que estos son los temas más hablados en el congreso.

5. 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* tematicas demostro ser efectiva para una clasificacion basica de los discursos.