Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Primer Semestre de 2025

IIC 2440 – Procesamiento de Datos Masivos
Tarea 1

1. Enunciado

En esta tarea van a trabajar sobre datos relacionados a sesiones en el parlamento de Chile. En breve, disponen de datos de los parlamentarios y de las intervenciones que han realizado en el congreso. La idea es que carguen los datos en BigQuery y mediante consultas de analítica de datos nos puedan explicar que temas se han tratado a lo largo del último tiempo. Así:

- Primero vas a tener que preprocesar los datos. Esto incluye un proceso de etiquetado de datos, en el que tendrás que crear *keywords* para las intervenciones de los parlamentarios.
- Luego te vamos a indicar las consultas que vas a tener que realizar para que diseñes un modelo de datos que te permita trabajar con los datos.
- Después vas a tener que cargar los datos en BigQuery y correr consultas con las que podrás sacar información valiosa desde los datos.
- Finalmente, vas a tener que plasmar todo este conocimiento en un informe con los resultados del proceso.

Para realizar la tarea vas a recibir un `.csv` con la información de los parlamentarios y otro separado en tres partes con las intervenciones de los mismos.

2. Parte 1 [1 pto] - Procesando los datos

Como queremos saber de qué se habla en el Congreso, lo primero es mirar a los `.csv` llamados `participacion_descripcion`. Hay una columna llamada `TEXTO_PRINCIPAL` y otra llamada `DESCIPCION_DEBATE`, en las que se nos señala sobre qué habló un parlamentario en un momento determinado. La idea es que partas etiquetando los datos: para cada intervención nos interesa tener una o más *keyword* que nos señala qué tema se trató. Algunos ejemplos pueden ser "deportes", "delincuencia", "medio ambiente", entre otros. Además, si encuentras que los datos están sucios o necesitas hacer otro procesamiento, puedes hacerlo.

Para la obtención de *keywords* puedes hacer lo que tú quieras, desde ver si el `TEXTO_PRINCIPAL` contiene una palabra definida por ti, hasta correr un modelo de lenguaje que lea automáticamente la intervención y obtenga *keywords*. En el futuro vas a tener que agrupar por estos tópicos, por lo que es importante que cada tópico se identifique únicamente. Por ejemplo, puedes tener problemas si tienes la *keyword* "deporte" y también "deportes".

3. Parte 2 - [1 pts] - Modelamiento

Para la modelación, queremos que entiendas las entidades y tengas presente las consultas que vas a realizar. Primero, queremos partir de un diagrama normalizado que tenga separado las distintas entidades, que serían los parlamentarios, partidos políticos, debates, intervenciones, y las *keywords* que calculaste. Una

vez contar con un modelo normalizado, puedes crear las vistas que estimes convenientes.

Luego, ten presente que harás las siguientes consultas:

- Cuáles son las top 5 temáticas más tratadas, para cada mes en el que hay registros.
- Media móvil de intervenciones por partido político con un intervalo de 3 meses.
- Para cada trimestre, ver el tema principal tratado por cada partido político.
- Poder pararse en un mes y ver cuál es el top 3 de temáticas tratada por cada partido. Esto vas a querer repetirlo para algunos meses.

4. Parte 3 - [1.5 pts] - BigQuery

Luego, tienes que cargar los datos en BigQuery y ejecutar las cuatro consultas mencionadas en el punto anterior. Se espera que para las consultas lo hagas con *Window Functions*. Si es que no las usas, debes justificar tu decisión.

5. Parte 4 - [2 pts] - Informe

Ahora tienes que responder la pregunta "¿De qué se habla en el congreso?". Esperamos que hagas lo que haría un científico de datos: esto es, navegar los datos y sacar conocimiento que pueda ser de interés para alguien. En este caso, piensa que este análisis se publicaría en una web con el objetivo de que la ciudadanía comprenda de mejor forma lo que se discute en el congreso y rol que juegan los partidos políticos. Tu relato debe ser consistente con los datos.

En concreto, esperamos que hagas un reporte que calce con lo solicitado arriba y que explique cómo llegaste a cada una de tus conclusiones a partir de los datos. Esperamos que tu análisis se apoye en visualizaciones y en formas efectivas de presentar la información.

6. Parte 5 - [0.5 pts] - Resumen ejecutivo

Finalmente, queremos que hagas un resumen ejecutivo de una página en la que nos cuentes tus hallazgos y los presentes de la manera que estimes conveniente. La idea es que alguien con poco tiempo pueda enterarse rápidamente sobre qué hiciste, cómo lo hiciste y qué resultados sacaste.

7. Detalles académicos

Esta tarea debe resolverse en grupos de dos personas. El formato de entrega consta de los siguientes archivos:

- Un archivo que contenga los códigos que usaste para procesar los datos, junto con un `.pdf` que contenga una breve explicación de lo que hiciste para procesarlos.
- Un archivo `.pdf` que contenga el esquema de los datos junto a un diagrama de la modelación. También debes justificar brevemente tu modelación.
- Un archivo `.pdf` que contenga y explique brevemente las consultas ejecutadas en la parte 3. También, analiza brevemente los resultados de las consultas.
- Un reporte en formato `.pdf` que responda la pregunta "¿De qué se ha estado hablando en el Congreso?" según lo descrito en la **Parte 4**.

- El resumen ejecutivo de una página, en formato `.pdf`.

Importante. El enunciado es bastante abierto en algunas partes a propósito. Si tienes que tomar alguna decisión hazlo con confianza mientras la justificación técnica sea razonable.

Fechas. La fecha de entrega de la tarea es el lunes 21 de abril, a las 20:00 hrs.