Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Primer Semestre de 2023

IIC 2440 – Procesamiento de Datos Masivos

Tarea 1

Instrucciones. Esta tarea debe resolverse en grupos de a dos personas, aunque puede ser resuelta en forma individual.

El formato de entrega consta de los siguientes archivos

- Un link a un repositorio público en Github donde se encuentre todo el código/notebooks (salvo los datos) necesario para realizar su tarea.
- Un video en youtube que debe ser publicado en el canal de discord del curso.

Fechas. La fecha de entrega de la tarea es el 31 de Mayo, a las 13:00 hrs.

Problema a resolver En la página encontraras una serie de archivos que contiene (entre otras cosas) una muestra de los tweets escritos a propósito del proceso constitucional 2021-2022, rechazado en plebiscito público, durante esos años, usado en un proyecto llamado *Plataforma Telar*, del Instituto Milenio Fundamento de los Datos. El esquema de los datos te resultará evidente una vez que los veas, y en particular, en que atributo está el texto de los tweets.

Tu deber será encontrar pares de personas que escriban tweets de forma similar. Para eso.

- Para un entero k y un $s \in [0, 1]$, considera que dos tweets son similares cuando la similitud de jaccard de sus k-shingles es al menos s. Debes definir k y s de modo que puedas encontrar una *pequeña cantidad* de tweets similares escritos por personas distintas. Naturalmente, puede que tengas que incurrir en un proceso de búsqueda, o de ensayo y error, para afinar estos números. No daremos una definición de *pequeña cantidad*, pero apelamos al sentido común: debe ser una cantidad que pueda, por ejemplo, ser exportada a excel y revisada por un ser humano sin mayor conocimientos de computación.
- Muestra (uno o mas) grupos de dos personas que, dado tu definición de arriba, hayan escrito una cantidad de tweets similares que te permita aseverar que su estilo de escritura, o su contenido, es similar.

Cuentanos como lo hicieron! Deberán filmar un video corto, de no más de 5 minutos, donde expliquen las definiciones que tomaron, qué fue lo que hicieron, y como lo lograron.