**Tarea 2 Procesamiento de Lenguaje Natural**

**Magister en Data Science, UDD**

**Instrucciones**

* La tarea es individual. Debe ser entregada a través de Canvas hasta las 23:59 del día 3 de agosto de 2025.
* Si utiliza otras librerías o códigos encontrados en internet u otra fuente, ponga la referencia y explique brevemente qué hace el código o librería.
* Deberá entregar un código comentado (.py o .ipynb)

**Comparando topic modeling y clustering**

1. (10 puntos) Los modelos LDA que ejecutamos en clases usaron una matriz DTM de tokens, esto es, palabras solas. Es posible que la interpretación de los tópicos mejore al incorporar bigramas. Lo primero que harás entonces es ejecutar un modelo LDA incorporando bigramas. Puedes ayudarte de este link <https://radimrehurek.com/gensim_3.8.3/auto_examples/tutorials/run_lda.html>. (aunque no tienes necesariamente que hacerlo así). Entonces:

1. Usaremos un subset de la base de datos “El Chile que Queremos” El subset está disponible en Canvas.
2. Añadir bigramas al corpus. (4 puntos)
3. Ejecutar diagnóstico de tópicos y escoger un número de tópicos para el modelo. Justificar. (3 puntos)
4. Ejecutar LDA con el número de tópicos escogidos y etiquetar los tópicos, justificando en no más de tres líneas, la etiqueta propuesta. (3 puntos)

2. (5 puntos) Los modelos LDA a menudo se usan para hacer clustering. A diferencia de otros métodos de machine learning, LDA es un método pensado para trabajar con textos. Pero uno perfectamente podría hacer clustering con otras técnicas (como Kmeans) a partir de la misma matriz DTM que alimenta un modelo LDA. Se le pide, entonces, que pruebe con Kmeans sobre la matriz DTM y compare los resultados con LDA. ¿Es el número óptimo de cluster el mismo que el número óptimo de tópicos? ¿Qué partición hace más sentido?