Text Mining & Image Recognition
Hoja de Trabajo # 4

Instrucciones: A continuación verá una lista de ejercicios que debe completar para poder entregar el hoja de trabajo #3. Al finalizar, todos sus archivos deben estar contenidos en un archivo lab3-sucarnet.zip. Este archivo lo debe entregar en el link del GES. Por favor cree una carpeta para cada ejercicio que usted realice.

Problema Único:

1. Descargue el **Dataset (click para descargar)** el cual contiene aproximadamente 800,000 tweets de diversos temas.

2. Usando CoLab y expresiones regulares. Determine los 3 usuarios más populares dentro del dataset. Luego arme un corpus el cual contenga los siguientes elementos por cada uno de los 3 usuario seleccionados:

   - Content: texto del tweet.
   - ID de tweeter, si el tweet no contiene ID deberá colocar el string *Undefined.*,
   - Timestamp,
   - Largo del tweet.

   Recuerde que un corpus es una tabla (dataframe) el cual contiene un conjunto de documentos y metadata sobre dichos documentos en formato tabular. Note que deberá terminar con tres dataframes diferentes.

3. Posterior a tener sus 3 corpus creados, responda: ¿Cuál es la razón por la que citan a ese usuario? para esto es necesario que extraiga el contexto de cada tweet y verifique cuales son las palabras que más rodean al nombre de usuario. Recomendamos utilizar un modelo bag of words el cual selecciones de 10 a 20 palabras representativas de cada corpus. Para extraer un contexto valido y debido a la naturaleza del tipo de datos que están disponibles en nuestro dataset, es necesario que le realice el flujo de preprocesamiento visto en clase, esto es:

   - Normalización de texto,
   - Tokenización,
   - Remover Stopwords,
   - Stemming y Lemmatization.

4. Finalmente cree un Word Cloud por cada ID de tweeter selecionaro a partir de su análisis anterior. Note que deberá terminar con 3 Word Clouds.