# Laboratorio 5.
# Análisis de sentimientos

## INSTRUCCIONES:

Utilice el data set Grammar and online products review de Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es determinar qué productos tienen mejores reviews. El sistema debe clasificar la calidad de un producto para los clientes basado en las opiniones recibidas. Debe clasificar los reviews en positivos o negativos para el producto y luego determinar qué productos están mejor y peor posicionados dada la opinión de los clientes. Puede utilizar algunos de los recursos que están a disposición en la web para detectar palabras positivas y negativas. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual. Incluya una nube de palabras que le ayude a detectar las que más se repiten. Este laboratorio debe realizarse en grupos de 3. Inscríbase en uno de los grupos que hay en canvas para la actividad.

## DESCRIPCIÓN DEL DATASET

El conjunto de datos contiene 71 045 reviews de 1000 productos diferentes, se incluye el título y el texto de los reviews, el nombre del productor, los metadatos del cliente, y más.
Puede encontrar recursos para el análisis de sentimientos en el siguiente repositorio (https://github.com/laugustyniak/awesome-sentiment-analysis )

## EJERCICIOS

1. Descargue los archivos de datos
2. Cargue los archivos de datos a R o a Python, dependiendo de con qué trabaje.
3. Limpie y preprocese los datos. Describa de forma detallada las actividades de preprocesamiento que llevó a cabo.
   3.1. Se pueden hacer tareas como:
   - Convertir el texto a mayúsculas o a minúsculas
   - Quitar los caracteres especiales que aparecen como “#”,”@” o los apóstrofes.
   - Quitar las url
   - Revisar si hay emoticones y quitarlos (a menos que le den información)
   - Quitar los signos de puntuación
   - Quitar los artículos, preposiciones y conjunciones (stopwords)
   - Quitar números si considera que interferirán en las predicciones.
4. Haga un análisis exploratorio de los datos para entenderlos mejor, documente todos los análisis
   4.1. Puede, para cada archivo:
   - Investigar qué palabra se repite más en cada archivo
   - Hacer una nube de palabras para visualizar las que aparecen con más frecuencia
   - Hacer un histograma con las palabras que más se repiten

- Discutir sobre las palabras que tienen presencia en todos los archivos.
- Determinar las palabras positivas y negativas

5. Teniendo en cuenta la cantidad de palabras positivas y negativas del review determine qué tan positivo, negativo o neutral es el mismo para el producto.
6. Luego de analizar los datos determine:
   6.1. Cuáles son los 10 productos de mejor calidad dado su review.
   6.2. Cuáles son los 10 productos de menor calidad dado su review.
   6.3. Cuáles son los usuarios que dan la mayor cantidad de reviews a distintos productos.
   6.4. Cuáles son los usuarios que más reviews negativos y positivos dan en promedio.
   6.5. Cuáles son los productores que tienen productos de mejor calidad.
   6.6. Cuáles son los productores que tienen productos de peor calidad.
7. Imagine que usted es analista de negocios y que está realizando este análisis para el productor que tiene más productos con malos reviews ¿Qué le propondría a esta empresa para mejorar sus productos? Puede basar su análisis en la frecuencia de las palabras de las opiniones.

## EVALUACIÓN

**(25 puntos)** Análisis exploratorio:
- Se elaboró un análisis exploratorio en el que se explican los cruces de variables, hay gráficos explicativos y análisis que permiten comprender el conjunto de datos.

**(20 puntos)** Limpieza y preprocesamiento de los datos:
- Se documentan las tareas de limpieza, incluyendo los paquetes/módulos que se usaron.

**(10 puntos)** Clasificación de las palabras:
- Clasificación de las palabras en positivas, negativas y neutrales. Explicación de las fuentes de datos o diccionarios utilizados.

**(10 puntos)** Algoritmo de clasificación:
- Se describe el algoritmo que se usó para clasificar el review en positivo, negativo o neutro.

**(25 puntos)** Resultados.
- Se elaboró una función que permite predecir las n posibles palabras que escribirá el usuario tras la frase ingresada.

**(10 puntos)** Estrategia:
- Se elaboró una propuesta de estrategia para el productor que tiene más productos con reviews negativos. El análisis está basado en datos por lo que la estrategia tiene un basamento sólido.

## MATERIAL A ENTREGAR

- Archivo .pdf con el informe que contenga, los resultados de los análisis y las explicaciones.
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó.
- Link del repositorio usado para versionar el código.

## FECHAS DE ENTREGA

- **AVANCE:** Descripción de los datos (tamaño, origen y tipo de corpus), preprocesamiento y sus explicaciones, unigramas, bigramas, modelo preliminar de predicción, conjunto de datos de clasificación de polaridad de términos: jueves 27 de agosto de 2020 a las 19:00.
- **DOCUMENTO FINAL COMPLETO:** lunes 31 de agosto de 2020 a las 23:59
  **NOTA:** Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.