Alumno: **Bru Montes, Israel**

Edición **Octubre 2024 – Grupo A** a 26/12/2024

# **1. (6 ptos) Dado un dataset que contenga entradas con la forma “persona;método\_pago;dinero\_gastado”, crea un programa llamado personaGastosConTarjetaCredito que para cada persona indique la suma del dinero gastado con tarjeta de crédito, con el formato persona;gastoconTDC. Ejemplo:**

**Entrada Salida**

**Alice;Tarjeta de crédito;100 Alice;250**

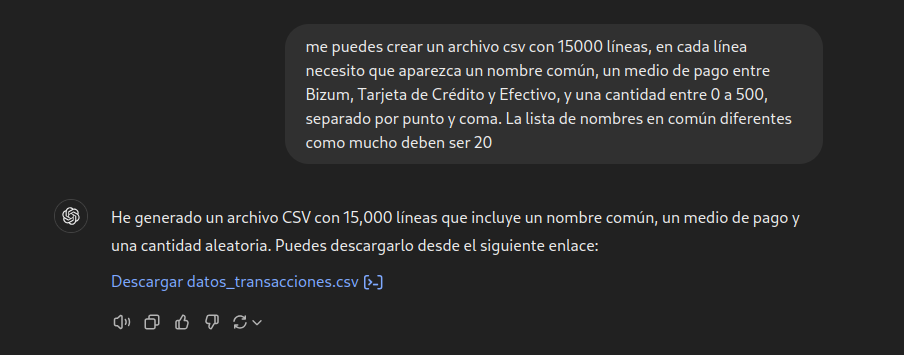
**Alice;Tarjeta de crédito;150 Bob;201**

**Alice;Bizum;200 Luis;0**

**Bob;Tarjeta de crédito;201**

**Luis;Bizum;300**

Generamos con ChatGPT un archivo csv para probar con las condiciones que se muestran en la siguiente captura:



Se genera este archivo para tener más datos.

**2. (7 ptos) Dado un dataset que contiene información sobre los videos de Youtube (https://netsg.cs.sfu.ca/youtubedata/), crear un programa llamado CategoriaDeVideosMenosVista que obtenga cuál es la categoría de videos menos vista de la plataforma Youtube y el número total de visualizaciones que hay en esa categoría. El programa debe recibir dos parámetros de entrada: la carpeta en la que está el dataset y la carpeta en la que se guardará el resultado. En la carpeta donde está el dataset se tienen que descomprimir UNO de los archivos 0222.zip, 0301.zip, etc., que se encuentran en el enlace anterior.**

La estrategia sería leer con el metodo “textFile” la carpeta donde se ubican todos los archivos .txt con los datos. Para procesar posteriormente:

1. Hacemos un split utilizando los “tabs” o tabulaciones por cada línea

2. Utilizamos “filter” para filtrar tanto las líneas del archivo que no contengan al menos en el split 5 elementos para evitar accesos a elementos del split que no existen.

3. Elegimos la columna 3 (categoría) y la 5 (número de visitas) para generar una tupla con los datos necesarios para tratar en la siguiente iteración.

4. Usamos el método ReduceByKey para hacer la suma total de visitas de la categoria e invocamos el método orderBy para ordenar por número total de visitas por categoria

5. Usamos el método first para obtener la primera fila del RDD, y con join convertimos la tupla obtenida en un string para imprimirlo en el fichero de salida.

**3. (7 ptos) Dado un dataset que contenga entradas con la forma “persona;método\_pago;dinero\_gastado”, crea un programa llamado personaYMetodosDePago que:**

**a) Por cada persona indique en cuántas compras pagó más de 1500 euros con un medio de pago diferente a tarjeta de crédito. La solución se tiene que guardar en un archivo llamado comprasSinTDCMayorDe1500.**

**b) Por cada persona indique en cuántas compras pagó menos o igual a 1500 euros con un medio de pago diferente a tarjeta de crédito. La solución se tiene que guardar en un archivo llamado comprasSinTDCMenoroIgualDe1500.**

**Se valorará positivamente la eficiencia del programa, por ejemplo no usar transformaciones innecesarias.**

Generamos como en actividades anteriores un csv con datos para tratar:

