Análisis sentimientos

1) Ejercicio 1

Importa el fichero csv en una tabla HIVE. Has de saltarte la primera fila con el nombre de las columnas.

```
CREATE DATABASE IF NOT EXISTS tarea_cuatro;

USE tarea_cuatro;

DROP TABLE IF EXISTS resenas;
CREATE EXTERNAL TABLE resenas (
    id INT,
    producto STRING,
    texto_resena STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
LOCATION '/user/maria_dev/tarea_6_4';

DELETE FROM resenas WHERE id = ‘id’;
```

2) Ejercicio 2

Utilizando SENTENCES , extrae en un array las frases que componen cada reseña en otra tabla (por ejemplo, usando CTAS).

```
CREATE TABLE resenas_frases
AS
SELECT
    id,
    producto,
    SENTENCES(texto_resena) AS frases_array
FROM resenas;
```

resenas_frases.id	resenas_frases.producto	resenas_frases.frases_array
106	Tablet	[["El","material","se","siente","barato"],["La","calidad","es","excelente"]]
107	Laptop	[["El","material","se","siente","barato"],["El","diseño","es","elegante","y","moderno"]]
108	Auriculares	[["El","diseño","es","elegante","y","moderno"],["Las","instrucciones","no","son","claras"]]
109	Monitor 4K	[["Las","instrucciones","no","son","claras"],["Superó","mis","expectativas"]]
110	Smartwatch	[["No","cumple","con","lo","prometido"],["Funciona","perfectamente"]]
111	Ratón Inalámbrico	[["Las","instrucciones","no","son","claras"],["Superó","mis","expectativas"]]
112	Laptop	[["La","configuración","inicial","fue","un","poco","complicada"],["Muy","satisfecho","con","la","compra"]]
113	Smartwatch	[["Gran","compra"],["El","envío","fue","lento"]]
114	Teclado Mecánico	[["El","envío","fue","lento"],["Funciona","perfectamente"]]
115	Laptop	[["Superó","mis","expectativas"],["El","servicio","al","cliente","es","lento"]]

3) Ejercicio 3

Utilizando EXPLODE , aplana la estructura, de modo que cada fila contenga una única frase. Crea una nueva tabla a partir de esta consulta.

```
CREATE TABLE resenas_frases_individuales
AS
SELECT
    id,
    producto,
    frase
FROM resenas_frases
LATERAL VIEW EXPLODE(frases_array) exploded_table AS frase;
```

resenas_frases_individuales.id	resenas_frases_individuales.producto	resenas_frases_individuales.frase
106	Tablet	["El","material","se","siente","barato"]
106	Tablet	["La","calidad","es","excelente"]
107	Laptop	["El","material","se","siente","barato"]
107	Laptop	["El","diseño","es","elegante","y","moderno"]
108	Auriculares	["El","diseño","es","elegante","y","moderno"]
108	Auriculares	["Las","instrucciones","no","son","claras"]
109	Monitor 4K	["Las","instrucciones","no","son","claras"]
109	Monitor 4K	["Superó","mis","expectativas"]
110	Smartwatch	["No","cumple","con","lo","prometido"]
110	Smartwatch	["Funciona","perfectamente"]

4) Ejercicio 4

(INVESTIGA) A partir de la tabla anterior, ¿cómo podrías crear otra tabla que elimine de las frases las siguientes palabras?:

‘el’,‘la’,‘los’,‘las’,‘de’,‘del’,‘al’,‘a’,‘un’,‘una’,‘unos’,‘unas’,
‘que’,‘y’,‘o’,‘en’,‘por’,‘para’,‘con’,‘sin’,‘sobre’,‘tras’,‘entre’,
‘hacia’,‘desde’,‘durante’,‘contra’,‘según’,‘como’,‘muy’,‘todo’,‘todos’,
‘este’,‘esta’,‘estos’,‘estas’,‘ese’,‘esa’,‘esos’,‘esas’,‘aquel’,‘aquella’,
‘lo’,‘le’,‘les’,‘me’,‘te’,‘se’,‘nos’,‘os’,‘lo’,‘los’,‘la’,‘las’,‘me’,‘mi’

```
CREATE TABLE resenas_frases_filtradas AS
SELECT
    id,
    producto,
    COLLECT_LIST(palabra) AS frase_filtrada
FROM (
    SELECT
        id,
        producto,
        LOWER(TRIM(palabra)) AS palabra
```

```
    FROM resenas_frases_individuales
    LATERAL VIEW EXPLODE(frase) exploded_table AS palabra
    WHERE LOWER(TRIM(palabra)) NOT IN (
        'el','la','los','las','de','del','al','a','un','una','unos','unas',
        'que','y','o','en','por','para','con','sin','sobre','tras','entre',

'hacia','desde','durante','contra','según','como','muy','todo','todos',

'este','esta','estos','estas','ese','esa','esos','esas','aquel','aquella',
        'lo','le','les','me','te','se','nos','os','mi'
    )
) palabras_filtradas
GROUP BY id, producto;
```

resenas_frases_filtradas.id	resenas_frases_filtradas.producto	resenas_frases_filtradas.frase_filtrada
106	Tablet	["material","siente","barato","calidad","es","excelente"]
107	Laptop	["material","siente","barato","diseño","es","elegante","moderno"]
108	Auriculares	["diseño","es","elegante","moderno","instrucciones","no","son","claras"]
109	Monitor 4K	["instrucciones","no","son","claras","superó","mis","expectativas"]
110	Smartwatch	["no","cumple","prometido","funciona","perfectamente"]
111	Ratón Inalámbrico	["instrucciones","no","son","claras","superó","mis","expectativas"]
112	Laptop	["configuración","inicial","fue","poco","complicada","satisfecho","compra"]
113	Smartwatch	["gran","compra","envío","fue","lento"]
114	Teclado Mecánico	["envío","fue","lento","funciona","perfectamente"]
115	Laptop	["superó","mis","expectativas","servicio","cliente","es","lento"]

5) Ejercicio 5: Identificación de Frases Clave (N-gramas)

Utilizando la tabla del ejercicio anterior, desarrolla una consulta que identifique los trigramas (n=3) más frecuentes en todas las reseñas de clientes. La consulta debe usar la función NGRAMS para generar combinaciones de tres palabras que aparezcan consecutivamente, y mostrar el ngram (combinación de palabras) junto con su frecuencia estimada (estfrequency).

Tristemente, la función NGRAMS no existe en nuestra versión de hive. Por lo que la hemos sustituido por una tabla.

Tabla trigramas

```
CREATE TABLE trigramas AS
SELECT
    id,
    producto,
    CONCAT_WS(' ', frase_filtrada[i], frase_filtrada[i+1],
frase_filtrada[i+2]) AS trigram
FROM resenas_frases_filtradas
LATERAL VIEW posexplode(frase_filtrada) exploded AS i, w
WHERE i <= SIZE(frase_filtrada) - 3;
```

Consulta

```
SELECT
    trigram,
    COUNT(*) AS frecuencia
FROM trigramas
GROUP BY trigram
ORDER BY frecuencia DESC
LIMIT 20;
```

trigram	frecuencia
envío fue lento	294
es útil rápido	156
producto es útil	156
configuración inicial fue	155
fue poco complicada	155
inicial fue poco	155
calidad es excelente	148
material siente barato	138
silla oficina ergonómica	125
es compañero viaje	115
aguanta vuelos largos	115
ligero autonomía aguanta	115

6) Ejercicio 6: Análisis de Sentimiento Contextualizado

Vistas las palabras que aparecen con más frecuencia en los trigramas, como:

envío , rendimiento , diseño , producto , etc...

Escoge una de esas palabras, (puedes probar con varias)

Queremos medir el sentimiento en torno a ella.

Crea una consulta que aplique la función CONTEXT_NGRAMS para filtrar y mostrar el contexto de 3 palabras que aparecen en torno al término que elegiste para

mostrar su contexto relevante. Visto el resultado ¿Podemos deducir algo de la opinión de los clientes respecto al término elegido?

Probaremos con la palabra envío:

```
SELECT
    trigram,
    COUNT(*) AS frecuencia
FROM trigramas
WHERE trigram LIKE '%envío%'
GROUP BY trigram
ORDER BY frecuencia DESC
LIMIT 20;
```

trigram	frecuencia
envío fue lento	294
minimalista embargo envío	114
embargo envío tardó	114
envío tardó más	114
gran compra envío	23
compra envío fue	23
perfectamente envío fue	21
funciona perfectamente envío	21
detalles envío fue	18
algunos detalles envío	18
mejorar envío fue	17
necesitan mejorar envío	17
es excelente envío	15
excelente envío fue	15

Viendo los primeros resultados podemos observar que han ocurrido muchos retrasos con los envíos.

7) Ejercicio 7: Análisis de Sentimiento Contextual por nombre del Producto

Modifica la consulta anterior de CONTEXT_NGRAMS para que, además de calcular la frecuencia del contexto de 3 palabras que sigue al término elegido, agrupe estos

resultados por la columna producto . El resultado final debe mostrar el producto, el ngram de contexto y la suma total de su frecuencia en ese grupo de productos.

Visto el resultado ¿Podemos deducir algo de la opinión de los clientes respecto a cada producto?

```
SELECT
    producto,
    trigram AS contexto,
    COUNT(*) AS frecuencia_total
FROM trigramas
WHERE trigram LIKE '%envío%'
GROUP BY producto, trigram
ORDER BY producto, frecuencia_total DESC
LIMIT 50;
```

producto	contexto	frecuencia_total
Altavoz Bluetooth	envío tardó más	11
Altavoz Bluetooth	minimalista embargo envío	11
Altavoz Bluetooth	embargo envío tardó	11
Auriculares	envío fue lento	20
Auriculares	minimalista embargo envío	6
Auriculares	embargo envío tardó	6
Auriculares	envío tardó más	6
Auriculares	funciona perfectamente envío	3
Auriculares	perfectamente envío fue	3
Auriculares	barato envío fue	2
Auriculares	siente barato envío	2
Auriculares	tengo soporte envío	1
Auriculares	algunos detalles envío	1
Auriculares	es excelente envío	1
Auriculares	excelente envío fue	1
Auriculares	detalles envío fue	1

Con esto podemos comprobar que el mayor foco de retrasos en envíos está en los auriculares, algo pasa con el proveedor fijo.