

PR_06.4

Dataset de Práctica

Para estos ejercicios utilizaremos el fichero `r3000.csv` que contiene reseñas falsas de una serie de productos generadas por IA .

Ejemplo de sus datos

id	producto	texto_reseña
101	Laptop	"El producto es muy útil y rápido. La batería, sin embargo, dura poco."
102	Tablet	"Gran compra. La atención al cliente fue excelente y el precio es justo."
103	Laptop	"Rápida, pero no tengo soporte. El envío fue lento. El equipo es genial."
104	Tablet	"Funciona bien, aunque el servicio al cliente es lento. Necesitan mejorar."
105	Auriculares	"Sonido increíble. Recomendaría este producto a todos mis amigos."

Ejercicio 1:

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

Ejercicio 2:

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

Ejercicio 3:

Utilizando `EXPLODE` , aplanas la estructura, de modo que **cada fila contenga una única frase**. Crea una nueva tabla apartir de esta consulta.

Ejercicio 4:

(INVESTIGA) A partir de la tabla anterior, ¿cómo podrías crear otra tabla que elimine de las frases las siguiente 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'

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`).

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?

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?