

PR_06.4 Dani Gayol Rodríguez

PR_06.4 Dani Gayol Rodríguez.....	1
Dataset de Práctica	1
Ejercicio 1: Importa el fichero csv en una tabla HIVE. Has de saltarte la primera fila con el nombre de las columnas.....	2
Ejercicio 2: Utilizando SENTENCES, extrae en un array las frases que componen cada reseña en otra tabla (por ejemplo, usando CTAS).....	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.	3
Ejercicio 4: (INVESTIGA) A partir de la tabla anterior, ¿cómo podrías crear otra tabla que elimine de las frases las siguientes palabras?.....	4
Ejercicio 5: Identificación de Frases Clave (Ngramas) Utilizando la tabla del ejercicio anterior, desarrolla una consulta que identifique los trigramas (n=3) más frecuentes en todas las reseñas de clientes.....	5
Ejercicio 6: Análisis de Sentimiento Contextualizado Vistas las palabras que aparecen con más frecuencia en los trigramas, como: envío, rendimiento, diseño, producto, etc.....	5
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.	6

Dataset de Práctica

Primero de todo, tenemos que subir el archivo “.csv” a la máquina virtual, para ello, voy a utilizar “WinSCP”

C:\Users\Mañana\Documents\			/home/maria_dev/					
Nombre	Tamaño	Tipo	Modificado	Nombre	Tamaño	Modificado	Permisos	Propiet.
..		DIRECTORIO superior	15/12/2025 8:50:53	..		18/06/2018 17:30:38	rvwx-r-x	root
BigData2526		Carpeta de archivos	02/12/2025 10:18:33	archivo_1gb.dat	1.048.57...	11/11/2025 10:13:16	rw-rw-r--	maria..
directorio_prueba		Carpeta de archivos	23/10/2025 11:49:09	consultas_usuarios.pig	2 KB	13/11/2025 10:06:49	rw-rw-r--	maria..
MIA_y_SAA		Carpeta de archivos	14/11/2025 12:08:32	DEPT.java	13 KB	27/11/2025 9:19:15	rw-rw-r--	maria..
ProgramacionIA		Carpeta de archivos	28/10/2025 11:17:03	e_l_quijote.txt	2.174 KB	17/01/2021 11:16:51	rw-rw-r--	maria..
winSCP		Carpeta de archivos	22/10/2025 11:41:54	EMP.java	23 KB	27/11/2025 9:24:30	rw-rw-r--	maria..
admin.pem	2 KB	Archivo PEM	18/11/2025 9:20:05	peliculas.java	13 KB	02/12/2025 9:25:02	rw-rw-r--	maria..
consultas_usuarios.pig	2 KB	Archivo PIG	13/11/2025 10:06:49	pig_1763549292308.log	4 KB	19/11/2025 11:48:28	rw-rw-r--	maria..
ejemplo.txt	1 KB	Documento de tex...	22/10/2025 11:30:15	pig_1763549628980.log	4 KB	19/11/2025 11:54:04	rw-rw-r--	maria..
Horarios.png	86 KB	Archivo PNG	15/10/2025 11:00:15	pig_1763625681821.log	11 KB	20/11/2025 9:09:37	rw-rw-r--	maria..
r3000.csv	162 KB	Archivo de origen ...	15/12/2025 8:34:21	pig_1764575313130.log	4 KB	01/12/2025 8:56:32	rw-rw-r--	maria..
retail_sales_dataset.csv	51 KB	Archivo de origen ...	17/11/2025 9:13:30	pig_1764576066043.log	4 KB	01/12/2025 9:03:22	rw-rw-r--	maria..
script_completo.pig	3 KB	Archivo PIG	17/11/2025 11:52:00	practicas.txt	1 KB	10/11/2025 9:58:57	rw-rw-r--	maria..
script_peliculas.pig	2 KB	Archivo PIG	18/11/2025 11:27:47	QueryResult.java	29 KB	27/11/2025 9:42:30	rw-rw-r--	maria..
script_peliculas2.pig	2 KB	Archivo PIG	19/11/2025 11:33:47	r3000.csv	453 KB	15/12/2025 8:34:21	rw-rw-r--	maria..
script_peliculas3.pig	2 KB	Archivo PIG	19/11/2025 12:11:50	retail_sales_dataset.csv	51 KB	17/11/2025 9:13:30	rw-rw-r--	maria..
script_peliculas4.pig	3 KB	Archivo PIG	20/11/2025 9:22:12	SALGRADE.java	14 KB	27/11/2025 9:26:12	rw-rw-r--	maria..
script_peliculas5.pig	2 KB	Archivo PIG	20/11/2025 10:40:32	script_completo.pig	3 KB	17/11/2025 11:52:00	rw-rw-r--	maria..

Ahora voy a pasar el archivo desde el directorio “/home” hacia el HDFS

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/r3000.csv /user/maria_dev/  
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/r3*  
-rw-r--r-- 1 maria_dev hdfs 463816 2025-12-15 07:54 /user/maria_dev/r3000.csv
```

Ejercicio 1: Importa el fichero csv en una tabla HIVE. Has de saltarte la primera fila con el nombre de las columnas

Antes de importar el fichero, hay que crear la base de datos junto con las tablas

```
1 CREATE DATABASE r3000
```

Ahora creamos la tabla para importarle los datos del archivo csv

```
1 CREATE TABLE reviews (  
2     id STRING,  
3     producto STRING,  
4     resena STRING  
5 )  
6 ROW FORMAT DELIMITED  
7 FIELDS TERMINATED BY '|'  
8 STORED AS TEXTFILE  
9 TBLPROPERTIES ("skip.header.line.count"="1");
```

Una vez creada la tabla, importamos los datos

```
1 LOAD DATA INPATH '/user/maria_dev/r3000.csv' INTO TABLE reviews;
```

Finalmente, comprobamos que se cargaron correctamente los datos

reviews.id	reviews.producto	reviews.resena
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."

Ejercicio 2: Utilizando SENTENCES, extrae en un array las frases que componen cada reseña en otra tabla (por ejemplo, usando CTAS).

```
1 CREATE TABLE reviews_sentences
2 AS
3 SELECT
4     id,
5     producto,
6     sentences(resena) AS frases
7 FROM reviews;
```

Comprobamos que se hizo correctamente

reviews_sentences.id	reviews_sentences.producto	reviews_sentences.frases
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"]]

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.

```
1 CREATE TABLE reviews_aplanar_sentences
2 AS
3 SELECT
4     id,
5     producto,
6     concat_ws(' ', frase) AS frase
7 FROM reviews_sentences
8 LATERAL VIEW explode(frases) t AS frase;
```

Comprobamos que se hizo correctamente

	reviews_aplanar_sentences.id	reviews_aplanar_sentences.producto	reviews_aplanar_sentences.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

Ejercicio 4: (INVESTIGA) A partir de la tabla anterior, ¿cómo podrías crear otra tabla que elimine de las frases las siguientes palabras?

```
1 CREATE TABLE reviews_borrar_sentences
2 AS
3 SELECT
4     id,
5     producto,
6     trim(
7         regexp_replace(
8             lower(frase),
9             '\b(el|la|los|las|de|del|al|a|un|una|unos|unas|que|y|o|en|por|para|' ..
10            ..
11        )
12    ) AS frase_limpia
13 FROM reviews_aplanar_sentences;
```

Comprobamos que se hizo correctamente

reviews_borrar_sentences.id	reviews_borrar_sentences.producto	reviews_borrar_sentences.frase_limpia
106	Tablet	material siente barato
106	Tablet	calidad es excelente
107	Laptop	material siente barato
107	Laptop	diseño es elegante moderno
108	Auriculares	diseño es elegante moderno

Ejercicio 5: Identificación de Frases Clave (Ngramas)

Utilizando la tabla del ejercicio anterior, desarrolla una consulta que identifique los trigramas (n=3) más frecuentes en todas las reseñas de clientes.

```

1 SELECT
2     ng.ngram      AS ngram,
3     ng.estfrequency AS estfrequency
4 FROM (
5     SELECT
6         ngrams(split(frase_limpia, ' '), 3, 1000) AS ng
7     FROM reviews_borrar_sentences
8 ) t
9 ORDER BY estfrequency DESC;

```

Comprobamos que se hizo correctamente

trigram	frecuencia
[["envío","fue","lento"],[],[],["hace","",""], [,"útil","",["es","","útil"],["producto","es","", ["útil","","rápido"],[],["poco","complicada"], ["configuración","inicial","fue"], ["fue","","poco"],["inicial","fue","", [,"","","compra"],["calidad","es","excelente"], [,"siente","barato"],["material","","siente"], [,"oficina","ergonómica"],	[294.0,223.0,196.0,156.0,156.0,156.0,155.0,155.0,155.0,155.0,152.0,

Ejercicio 6: Análisis de Sentimiento Contextualizado Vistas las palabras que aparecen con más frecuencia en los trigramas, como: envío, rendimiento, diseño, producto, etc.

```
1 SELECT
2     ctx.ngram           AS contexto,
3     ROUND(ctx.estfrequency, 0) AS frecuencia
4 FROM (
5     SELECT
6         context_ngrams(
7             split(frase_limpia, ' '),
8             array('envío'),
9             3,
10            1000
11        ) AS ctx
12     FROM reviews_borrar_sentences
13 ) t
14 ORDER BY frecuencia DESC
15 LIMIT 20;
```

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.

```
1 SELECT
2     producto,
3     ctx.ngram AS contexto,
4     ROUND(SUM(ctx.estfrequency), 0) AS frecuencia_total
5 FROM (
6     SELECT
7         producto,
8         context_ngrams(
9             split(frase_limpia, ' '),
10            array('envío'),
11            3,
12            1000
13        ) AS ctx
14     FROM reviews_borrar_sentences
15 ) t
16 GROUP BY producto, ctx.ngram
17 ORDER BY producto, frecuencia_total DESC;
```