Módulo 12: Arquitecturas y procesos Big Data

Capstone 12. Parte 1: Modelo de *sentiment* sobre Amazon Reviews

Enrique González, Jacinto Arias
Máster en Ciencia de Datos e Ingeniería de Datos en la Nube
Universidad de Castilla-La Mancha

Marta Bellón Castro
Curso 2022-2023

Índice

- 1. Introducción
- 2. Análisis exploratorio
- 3. Modelado

In [1]:
```
# Instalamos algunas librerías útiles para la práctica

import pyspark.sql.functions as sqlf
from pyspark.ml.pipeline import PipelineModel
from pyspark.ml.evaluation import BinaryClassificationEvaluator
```

VBox()

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1691196225143_0001	pyspark	idle	Link (http://ip-172-31-24-52.ec2.internal:20888/proxy/application_1691196225143_0001/)	Link (http://ip-172-31-18-8.ec2.internal:8042/node/containerlogs/container_1691196225143_0001_01_000001/livy)	None	✓

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

SparkSession available as 'spark'.

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

In [2]:
```
sc.install_pypi_package('pandas')
sc.install_pypi_package('seaborn')
sc.install_pypi_package('tabulate')
```

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

```
Collecting pandas
  Downloading pandas-1.3.5-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (11.3 MB)
Collecting python-dateutil>=2.7.3
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
Requirement already satisfied: numpy>=1.17.3; platform_machine != "aarch64" and platform_machine != "arm64" and python_version < "3.10" in /usr/local/lib64/python3.7/site-packages (from pandas) (1.20.0)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/site-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.7.3->pandas) (1.13.0)
Installing collected packages: python-dateutil, pandas
Successfully installed pandas-1.3.5 python-dateutil-2.8.2

Collecting seaborn
  Downloading seaborn-0.12.2-py3-none-any.whl (293 kB)
Requirement already satisfied: pandas>=0.25 in ./tmp/1691196491930-0/lib/python3.7/site-packages (from seaborn) (1.3.5)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in /usr/local/lib64/python3.7/site-packages (from seaborn) (1.20.0)
Collecting typing_extensions; python_version < "3.8"
```

In [3]:
```
# Los siguientes packetes están disponibles en el cluster
sc.list_packages()
```

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

```
Package                    Version
-------------------------- ----------
aws-cfn-bootstrap          2.0
beautifulsoup4             4.9.3
boto                       2.49.0
click                      8.1.3
cycler                     0.11.0
docutils                   0.14
fonttools                  4.38.0
jmespath                   1.0.1
joblib                     1.2.0
kiwisolver                 1.4.4
lockfile                   0.11.0
lxml                       4.9.2
matplotlib                 3.5.3
mysqlclient                1.4.2
nltk                       3.8.1
nose                       1.3.4
numpy                      1.20.0
packaging                  23.1
pandas                     1.3.5
Pillow                     9.5.0
pip                        20.2.2
py-dateutil                2.2
pyparsing                  3.1.1
pystache                   0.5.4
python-daemon              2.2.3
python-dateutil            2.8.2
python37-sagemaker-pyspark 1.4.2
pytz                       2023.3
PyYAML                     5.4.1
regex                      2021.11.10
seaborn                    0.12.2
setuptools                 28.8.0
simplejson                 3.2.0
six                        1.13.0
tabulate                   0.9.0
tqdm                       4.65.0
typing-extensions          4.7.1
wheel                      0.29.0
windmill                   1.6

WARNING: The directory '/home/.cache/pip' or its parent directory is not owned or is not writable by the current user. The cache has been disabled. Check the permi
ssions and owner of that directory. If executing pip with sudo, you may want sudo's -H flag.
```

1. Introducción

En este capstone vamos a aprender un modelo de detección del sentimiento utilizando MLlib y EMR. Una vez aprendido ampliaremos el proyecto serializando este modelo y comparándolo con el modelo de detección de sentimiento disponible en AWS.

Para ello utilizaremos el dataset de **amazon reviews** que está disponible de manera pública

https://s3.amazonaws.com/amazon-reviews-pds/readme.html (https://s3.amazonaws.com/amazon-reviews-pds/readme.html)

Este dataset tiene dos versiones una en tsv y otra en parquet. Nosotros usaremos la que está en parquet que está disponible a través de la ruta de s3: `s3://amazon-reviews-pds/parquet` .

NOTA IMPORTANTE:
El link para descargar los archivos de reviews no funciona, así que se ha realizado el ejercico con los datos descargados de GoogleDrive, pero estos son ligeramente diferentes a los obtenidos utilizados con el dataset de "electronics" que se descargaba desde Amazon, el cual me ha facilitado un compañero.

Este dataset tiene las siguientes columnas (de su diccionario de datos):

```
marketplace       - 2 letter country code of the marketplace where the review was written.
customer_id       - Random identifier that can be used to aggregate reviews written by a single author.
review_id         - The unique ID of the review.
product_id        - The unique Product ID the review pertains to. In the multilingual dataset the reviews
                    for the same product in different countries can be grouped by the same product_id.
product_parent    - Random identifier that can be used to aggregate reviews for the same product.
product_title     - Title of the product.
product_category  - Broad product category that can be used to group reviews
                    (also used to group the dataset into coherent parts).
star_rating       - The 1-5 star rating of the review.
helpful_votes     - Number of helpful votes.
total_votes       - Number of total votes the review received.
vine              - Review was written as part of the Vine program.
verified_purchase - The review is on a verified purchase.
review_headline   - The title of the review.
review_body       - The review text.
review_date       - The date the review was written.
```

De estas, la columna `product_category` se usa como clave de partición. Podéis encontrar toda la información en el enlace que os proporcionamos más arriba.

2. Análisis exploratorio

Antes de empezar con el modelado exploraremos los datos minimamente para poder estudiar sus propiedades.

Tarea 1: Carga de datos

Carga el dataset completo en formato parquet y cuenta sus registros. De momento, no lo persistas.

In [4]:
```
from pyspark.sql import SparkSession

# Creamos la instancia SparkSession
spark = SparkSession.builder \
    .appName("SentimentAnalysis") \
    .getOrCreate()

# Define la ruta del archivo parquet en S3
parquet_path = "s3://capstone12bucket2/amazon-reviews-pds-parquet"

# Lee el archivo parquet como un DataFrame
data_frame = spark.read.parquet(parquet_path)

# Cuenta el número de registros en el DataFrame
record_count = data_frame.count()

# Print the number of records
print("Total Records:", record_count)
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
Total Records: 16967903
```

Resultado esperado: 160796570 registros

Los resultados son ligeramente diferentes utilizando los datos descargados de GoogleDrive.

Tarea 2: Filtrado

Como el dataset es masivo para entrenar el modelo de sentiment vamos a trabajar únicamente con una partición. Concretamente utilizaremos la partición de `Electronics`. Filtra los datos para quedarte con esta partición y cuenta ahora el total de elementos de este nuevo dataset. No cachees este dataset.

In [6]:
```python
# Solución
# Proceso
from pyspark.sql import SparkSession

# Inicia una instancia de SparkSession
spark = SparkSession.builder \
    .appName("SentimentAnalysis") \
    .getOrCreate()

#Esto es lo que habría que hacer si funcionase el link

# Ruta del archivo parquet en S3
ruta_parquet = "s3://capstone12bucket2/amazon-reviews-pds-parquet"

# Lee el archivo parquet y crea un DataFrame
data_frame = spark.read.parquet(ruta_parquet)

# Filtra los datos para obtener la categoría de Electrónicos
data_frame_electronics = data_frame.filter(data_frame["product_category"] == "Electronics")

# Cuenta el número de registros en el nuevo DataFrame
conteo_electronics = data_frame_electronics.count()

# Imprime el conteo de registros
print("Registros en electroncis:", conteo_electronics)
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

Registros en electroncis: 3105119
```

Resultado esperado: 3120938 registros

Tarea 3: Almacenamiento

Para no seguir trabajando con los datos públicos, vamos a escribir los datos en un bucket de S3 en nuestra cuenta. Para ello, crea un bucket the S3 para este capstone y escribe los datos dentro del bucket en el directorio `electronics`. Utiliza repartition para tener 32 particiones. Tras esto, vuelve a cargar el dataset y cachéalo.

In [7]:
```
# Solución


# Ruta S3 para almacenar los datos
s3_bucket = "s3://capstone12bucket2"

# Escribe los datos en el depósito S3 en la subcarpeta "electronica" con 32 particiones
data_frame_electronics.repartition(32).write.parquet(s3_bucket + "/electronics", mode="overwrite")

# Carga nuevamente el conjunto de datos desde S3
data_frame_electronics = spark.read.parquet(s3_bucket + "/electronics").cache()
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

Tarea 4: Almacenamiento

Obten los siguiente resultados del dataset que acabáis de cargar:

1. Muestra el total de reviews para cada posible número de estrellas recibidas (*star_rating*)
2. Obtén los 10 productos con mayor número de votos (*total_votes*) mostrando su nombre, numero de votos y valoración media (*star_rating*)
3. Obtén la cantidad de reviews (1 registro de dataset -> 1 review) y la valoración media (*star_rating*) por mes y año. Obten los últimos 15 registros ordenador por año y mes.

In [8]:
```
from tabulate import tabulate
import pandas as pd
from pyspark.sql.functions import desc
from pyspark.sql.functions import month, year, avg
from pyspark.sql.window import Window
from pyspark.sql.functions import row_number
from pyspark.sql import functions as F

# Análisis de reseñas por número de estrellas

from pyspark.sql.functions import count

# Calcular el total de reseñas por número de estrellas
reviews_per_star = data_frame_electronics.groupBy("star_rating").agg(count("*").alias("total_reviews"))

review_data = reviews_per_star.toPandas()
print(tabulate(review_data, headers="keys", tablefmt="fancy_grid", floatfmt=".0f"))

# 2-10 productos con mayor número de votos

# 10 productos con mas votos
top_ten_products = data_frame_electronics.groupBy("product_id", "product_title") \
    .agg(F.sum("total_votes").alias("total_votes"), F.avg("star_rating").alias("star_rating")) \
    .orderBy(F.desc("total_votes")) \
    .limit(10)

top_ten_products.show()

# 3-Cantidad de reviews

# Cantidad de reviews y valoración media por mes y año
reviews_avg_rating = data_frame_electronics.select("review_date", "star_rating") \
    .withColumn("year", year("review_date")) \
    .withColumn("month", month("review_date")) \
    .groupBy("year", "month") \
    .agg(count("*").alias("reviews_count"), avg("star_rating").alias("mean_star_rating")) \
    .orderBy(desc("year"), desc("month"))

# Últimos 15 registros por mes y año
window_spec = Window.orderBy(desc("year"), desc("month"))
last_15_records = reviews_avg_rating.withColumn("row_number", row_number().over(window_spec)) \
    .filter("row_number <= 15") \
    .orderBy(desc("year"), desc("month"))

# Obtener los últimos 15 registros ordenados por año y mes
last_15_records_selected = last_15_records.select("year", "month", "reviews_count", "mean_star_rating")

# Mostrar los resultados seleccionados
last_15_records_selected.show()
```

VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

	star_rating	total_reviews
0	3	239459
1	1	359248
2	5	1787754
3	4	538824
4	2	179834

```
+----------+--------------------+-----------+------------------+
|product_id|       product_title|total_votes|       star_rating|
+----------+--------------------+-----------+------------------+
|B000I1X6PM|Denon AKDL1 Dedic...|      46348|3.4917491749174916|
|B000J36XR2|AudioQuest K2 Ter...|      20515|3.9357429718875503|
|B004QK7HI8|Mohu Leaf 30 TV A...|      18198| 4.091006423982869|
|B0001FTVEK|Sennheiser On-Ear...|      18029| 4.034212320982041|
|B000EPLP3C|Zune 30 GB Digita...|      17598|3.7341513292433537|
|B001FA1O18|Apple iPod touch ...|      17103| 4.384232365145229|
|B00D5Q75RC|Bose SoundLink Mi...|      16028| 4.721810699588477|
|B0054JJ0QW|Bose QuietComfort...|      13383| 4.439661515820457|
|B0002L5R78|High Speed HDMI C...|      12859| 4.462750716332378|
|B000WYVBR0|VideoSecu ML531BE...|      11100|4.5796048438495855|
+----------+--------------------+-----------+------------------+

+----+-----+-------------+------------------+
|year|month|reviews_count|  mean_star_rating|
+----+-----+-------------+------------------+
|2015|    8|       102984| 4.093985473471608|
|2015|    7|        99806|  4.08580646454121|
|2015|    6|        91486| 4.093478783639027|
|2015|    5|        89357| 4.100439808856609|
|2015|    4|        93152| 4.102466935760907|
|2015|    3|       108861|  4.11561532596614|
|2015|    2|       107291| 4.118062092813004|
|2015|    1|       120404| 4.152602903558021|
|2014|   12|       107891| 4.120232456831432|
|2014|   11|        77529|  4.10810148460576|
|2014|   10|        78128| 4.114210014335449|
|2014|    9|        77753| 4.116111275449179|
|2014|    8|        82143| 4.115664146671049|
|2014|    7|        79424| 4.118352135374698|
|2014|    6|        48375|4.0157726098191215|
+----+-----+-------------+------------------+
```

Resultados esperados:

1. Muestra el total de reviews para cada posible número de estrellas recibidas (*star_rating*)

star_rating	count
1	360558
3	240859
4	542181
5	1796672
2	180668

2. Obtén los 10 productos con mayor número de votos (*total_votes*) mostrando su nombre, numero de votos y valoración media (*star_rating*)

product_title	total_votes	star_rating
Denon AKDL1 Dedicated Link Cable (Discontinued by Manufacturer)	12944	3
AudioQuest K2 Terminated Speaker Cable - UST 2.44 m Plugs 8' Pair (Discontinued by Manufacturer)	9072	1
Panasonic ErgoFit In-Ear Earbud Headphone	8680	5
Apple iPod touch 8GB (4th Generation)	6353	5
Denon AKDL1 Dedicated Link Cable (Discontinued by Manufacturer)	5546	1
Apple iPod touch 8 GB 2nd Generation	4595	5
Bose QuietComfort 15 Acoustic Noise Cancelling Headphones (Discontinued by Manufacturer)	4556	4
Panasonic ErgoFit In-Ear Earbud Headphone	4341	5
X-Mini II XAM4-B Portable Capsule Speaker, Mono	4260	1
Denon AKDL1 Dedicated Link Cable (Discontinued by Manufacturer)	4242	2

3. Obtén la cantidad de reviews (1 registro de dataset -> 1 review) y la valoración media (*star_rating*) por mes y año. Obten los últimos 15 registros ordenador por año y mes.

year	month	review_count	mean_star_rating
2015	8	103336	4.09441
2015	7	100128	4.08615
2015	6	91815	4.0933
2015	5	89676	4.10048
2015	4	93469	4.10283
2015	3	109175	4.11569
2015	2	107623	4.11792
2015	1	120852	4.15227
2014	12	108294	4.1203
2014	11	77844	4.10784
2014	10	78519	4.11387

year	month	review_count	mean_star_rating
2014	9	78126	4.11593
2014	8	82550	4.11603
2014	7	79816	4.11758
2014	6	48707	4.01573

3. Modelado

Como paso previo al modelado realizaremos dos procesos de limpieza sobre los datos:

Tarea 6: Preparación del texto

Limpiead el texto de las reviews (`review_body`) utilizando expresiones sobre strings o expresiones regulares

- Pasar todo el texto a minusculas.
- Eliminar números y signos de puntuacion.
- Si existen, elimina los registros con valores nulos en el body con las transformaciones anteriores.

Muestra los resultados para las primeras 10 filas del dataframe ordenadas por `review_id`

In [9]:
```python
# Solución
from pyspark.sql import functions as F
from pyspark.sql.functions import lower, regexp_replace, col

# Pasar todo el texto a minusculas.
df_lowercase = data_frame_electronics.withColumn("cleaned_review_body", lower(col("review_body")))

# Eliminar números y signos de puntuacion
df_cleaned = df_lowercase.withColumn("cleaned_review_body", regexp_replace(col("cleaned_review_body"), "[^a-zA-Z\s]", ""))

# Eliminar los registros con valores nulos
df_filtered = df_cleaned.filter(df_cleaned["cleaned_review_body"] != "")

# Mostrar los resultados de las primeras 10 filas ordenadas por review_id
df_filtered.select("review_id", "review_body", "cleaned_review_body").orderBy("review_id").show(10, truncate=False)
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…

+--------------+-----------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------------------------------------------+
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------+
```

Resultado esperado:

review_id	review_body	clean_review_body
R10000WMGXS51T	Great little emergency radio. Very good reception. The weather band is a feature. Can't beat the quality for this price/	great little emergency radio very good reception thebr weather band is a feature cant beat the quality for this price
R10001L4QTCA84	Lives up to its claim, and really does fit bulky phone cases. Braided cable is sturdy but flexible. I think it stays a little more flexible in the cold weather, which is nice. Definitely getting a few more in the future!	lives up to its claim and really does fit bulky phone cases braided cable is sturdy but flexible i think it stays a little more flexible in the cold weather which is nice definitely getting a few more in the future

review_id	review_body	clean_review_body
R10003OLR2P5UE	I've gone through three pairs of these in the last two years. I am in love with the sound quality, and even though I know it's not the best I particularly love how the bass sounds. They're comfortable to wear and very isolating. With these headphones, you don't even need noise canceling. There is very little sound leak, unless you like to listen to music ridiculously loud. All in all, I was very impressed with these. They're without a doubt the best sounding headphones I've ever owned. Now, the problem: The wires are thin and stringy, and do NOT last. On my first pair, the part of the wire that connected to the left cup came apart. I'm not an abuser of headphones, either. On the other two pairs, they wire at the base next to the adapter came apart. I went at them with a soldering iron, desperately trying to make them last as long as I could, but they'd always crap out on me again. The sound quality distorts over time, and the foam around the cups is cheap and wears out quickly. They aren't worth the price for such bad quality. I'd suggest looking around for other pairs, Sony, Denon, and Sennheiser all have superior headphones for a similar price. I myself just ordered a pair of Denon AHD1001's, and here's hoping they last longer!	ive gone through three pairs of these in the last two years i am in love with the sound quality and even though i know its not the best i particularly love how the bass sounds theyre comfortable to wear and very isolating with these headphones you dont even need noise canceling there is very little sound leak unless you like to listen to music ridiculously loud all in all i was very impressed with these theyre without a doubt the best sounding headphones ive ever ownedbr br now the problem the wires are thin and stringy and do not last on my first pair the part of the wire that connected to the left cup came apart im not an abuser of headphones either on the other two pairs they wire at the base next to the adapter came apart i went at them with a soldering iron desperately trying to make them last as long as i could but theyd always crap out on me again the sound quality distorts over time and the foam around the cups is cheap and wears out quicklybr br they arent worth the price for such bad quality id suggest looking around for other pairs sony denon and sennheiser all have superior headphones for a similar price i myself just ordered a pair of denon ahds and heres hoping they last longer
R10005O193PJ6W	stopped working after a while, changed batteries, it worked for a few days, then it quit	stopped working after a while changed batteries it worked for a few days then it quit
R10008LR7CU84N	I ordered this cable and it doesn't work when I contacted them they told me I was doing something wrong. I then had my dad who it a certified computer tech look at it and there is something wrong with the cable. When I told them they never responded to me again.	i ordered this cable and it doesnt work when i contacted them they told me i was doing something wrong i then had my dad who it a certified computer tech look at it and there is something wrong with the cable when i told them they never responded to me again
R10009JN2UWOJC	Have not owned it that long however it has the features , feel and works like a quality unit that would be at a much higher price point	have not owned it that long however it has the features feel and works like a quality unit that would be at a much higher price point
R1000AMVKPW32O	Bought for a gift and it is just what was needed to mount the new 32" TV outdoors. The fact that it has full motion swing makes it even better because we can move it around to see it from different angles and still have a sturdy mount.	bought for a gift and it is just what was needed to mount the new tv outdoors the fact that it has full motion swing makes it even better because we can move it around to see it from different angles and still have a sturdy mount
R1000CJMO2L8X4	Perfect for the gym	perfect for the gym
R1000EDGJUU3CU	Love these !!! The sound quality is amazing ! The price was amazing especially for the quality.	love these the sound quality is amazing the price was amazing especially for the quality

Tarea 7: Obtención del sentiment

Cread la variable `sentiment` en función del número de estrellas asumiendo que una review de menos (<) de 3 estrellas es negativa, usando 1 para el sentiment positivo y 0 para el negativo. Para poder generar la variable que determine el sentiment a partir del número de estrellas podéis utilizar la función de spark `when` . Muestra el resultado para las primeras 10 reviews ordenadas por `review_id` .

In [10]:
```python
# Solución
from pyspark.sql.functions import when

# Agregar la columna "sentiment" basada en la valoración
df_sentiment = data_frame_electronics.withColumn("sentiment", when(data_frame_electronics["star_rating"] < 3, 0).otherwise(1))

# Mostrar los primeros 10 resultados con las columnas relevantes
df_result = df_sentiment.select("review_id", "review_body", "star_rating", "sentiment").orderBy("review_id").limit(10)
df_result.show(truncate=False)
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
+--------------+--------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------------------------+-----------+----
-----+
|review_id     |review_body
|star_rating|sentiment|
+--------------+--------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------------------------+-----------+----
-----+
|R10000WMGXS51T|Great little emergency radio.  Very good reception.  The<br />weather band is a feature.  Can't beat the quality for this price/
|5          |1        |
|R10001L4QTCA84|Lives up to its claim, and really does fit bulky phone cases. Braided cable is sturdy but flexible. I think it stays a little more flexible in the
cold weather, which is nice. Definitely getting a few more in the future!
|5          |1        |
|R10003OLR2P5UE|I've gone through three pairs of these in the last two years. I am in love with the sound quality, and even though I know it's not the best I parti
cularly love how the bass sounds. They're comfortable to wear and very isolating. With these headphones, you don't even need noise canceling. There is very little
sound leak, unless you like to listen to music ridiculously loud. All in all, I was very impressed with these. They're without a doubt the best sounding headphones
I've ever owned.<br /><br />Now, the problem: The wires are thin and stringy, and do NOT last. On my first pair, the part of the wire that connected to the left cu
p came apart. I'm not an abuser of headphones, either. On the other two pairs, they wire at the base next to the adapter came apart. I went at them with a solderin
g iron, desperately trying to make them last as long as I could, but they'd always crap out on me again. The sound quality distorts over time, and the foam around
the cups is cheap and wears out quickly.<br /><br />They aren't worth the price for such bad quality. I'd suggest looking around for other pairs, Sony, Denon, and
Sennheiser all have superior headphones for a similar price. I myself just ordered a pair of Denon AHD1001's, and here's hoping they last longer!|3          |1
|
|R10005O193PJ6W|stopped working after a while, changed batteries, it worked for a few days, then it quit
|3          |1        |
|R10008LR7CU84N|I ordered this cable and it doesn't work when I contacted them they told me I was doing something wrong. I then had my dad who it a certified compu
ter tech look at it and there is something wrong with the cable. When I told them they never responded to me again.
|1          |0        |
|R10009JN2UWOJC|Have not owned it that long however it has the features , feel and works like a quality unit that would be at a much higher price point
|5          |1        |
|R1000AMVKPW32O|Bought for a gift and it is just what was needed to mount the new 32&#34; TV outdoors. The fact that it has full motion swing makes it even better
because we can move it around to see it from different angles and still have a sturdy mount.
|5          |1        |
|R1000CJMO2L8X4|Perfect for the gym
|5          |1        |
|R1000EDGJUU3CU|Love these !!! The sound quality is amazing ! The price was amazing especially for the quality.
|5          |1        |
|R1000EG9XXBLXT|I have had good success with these disks, and have used hundreds of them successfully on both computers and a dedicated Panosonic DVD recorder. The
y seem very reliable, and the lines on the disk label help to keep labeling neat and straight.
|5          |1        |
+--------------+--------------------------------------------------------------------------------------------------------------------------------------------------
```

```
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------------------------------------------+----------+----
-----+
```

Resultado esperado:

review_id	review_body	star_rating	sentiment
R10000WMGXS51T	Great little emergency radio. Very good reception. The weather band is a feature. Can't beat the quality for this price/	5	1
R10001L4QTCA84	Lives up to its claim, and really does fit bulky phone cases. Braided cable is sturdy but flexible. I think it stays a little more flexible in the cold weather, which is nice. Definitely getting a few more in the future!	5	1
R10003OLR2P5UE	I've gone through three pairs of these in the last two years. I am in love with the sound quality, and even though I know it's not the best I particularly love how the bass sounds. They're comfortable to wear and very isolating. With these headphones, you don't even need noise canceling. There is very little sound leak, unless you like to listen to music ridiculously loud. All in all, I was very impressed with these. They're without a doubt the best sounding headphones I've ever owned. Now, the problem: The wires are thin and stringy, and do NOT last. On my first pair, the part of the wire that connected to the left cup came apart. I'm not an abuser of headphones, either. On the other two pairs, they wire at the base next to the adapter came apart. I went at them with a soldering iron, desperately trying to make them last as long as I could, but they'd always crap out on me again. The sound quality distorts over time, and the foam around the cups is cheap and wears out quickly. They aren't worth the price for such bad quality. I'd suggest looking around for other pairs, Sony, Denon, and Sennheiser all have superior headphones for a similar price. I myself just ordered a pair of Denon AHD1001's, and here's hoping they last longer!	3	1
R10005O193PJ6W	stopped working after a while, changed batteries, it worked for a few days, then it quit	3	1
R10008LR7CU84N	I ordered this cable and it doesn't work when I contacted them they told me I was doing something wrong. I then had my dad who it a certified computer tech look at it and there is something wrong with the cable. When I told them they never responded to me again.	1	0
R10009JN2UWOJC	Have not owned it that long however it has the features , feel and works like a quality unit that would be at a much higher price point	5	1
R1000AMVKPW32O	Bought for a gift and it is just what was needed to mount the new 32" TV outdoors. The fact that it has full motion swing makes it even better because we can move it around to see it from different angles and still have a sturdy mount.	5	1
R1000CJMO2L8X4	Perfect for the gym	5	1
R1000EDGJUU3CU	Love these !!! The sound quality is amazing ! The price was amazing especially for the quality.	5	1
R1000EG9XXBLXT	I have had good success with these disks, and have used hundreds of them successfully on both computers and a dedicated Panosonic DVD recorder. They seem very reliable, and the lines on the disk label help to keep labeling neat and straight.	5	1

Tarea 8: División del conjunto de datos

Divide el conjunto de datos en entrenamiento (70% de los datos) y test (30% de los datos). Una vez hecho esto, guarda los datos de test en el bucket the s3 previamente creado (usa el nombre `electronics_test`)

In [11]:
```
# Solución

#División del conjunto de datos en entrenamiento y prueba
train_data, test_data = df_sentiment.randomSplit([0.7, 0.3], seed=5)
train_data = train_data.na.drop(subset=["review_body"])
test_data = test_data.na.drop(subset=["review_body"])

# Guardar los datos de prueba en el bucket S3
test_data.write.parquet("s3://capstone12bucket2/electronics_test", mode="overwrite")
train_data.show()
```

```
VBox()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-----------+-------------+-----------+----+-----------------+--------------
-----+--------------------+-----------+---------------+---------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|star_rating|helpful_votes|total_votes|vine|verified_purchase|     review_hea
dline|         review_body|review_date|product_category|sentiment|
+-----------+-----------+--------------+----------+--------------+--------------------+-----------+-------------+-----------+----+-----------------+--------------
-----+--------------------+-----------+---------------+---------+
|         US|      10048|R2C754BDSEUSES|B00A3RVNXI|     890589353|TETC Blue G4000 P...|          5|            0|          0|   N|                Y|          Five
Stars|Great Sound !!!!<...| 2014-11-23|     Electronics|        1|
|         US|      10276| RFRWB5TK2MIUT|B003EM6AOG|     921315865|Panasonic ErgoFit...|          5|            0|          0|   N|                Y|           Grea
t fit|These are super d...| 2013-08-16|     Electronics|        1|
|         US|      10562|R38W9DP4IRB6MM|B003M2YP6S|     508377575|Case Logic Nylon ...|          5|            1|          1|   N|                Y|I love this! I
bo...|I love this! I bo...| 2015-07-05|     Electronics|        1|
|         US|      10610|R16DGTL6PT6BVX|B00H2QUD9S|     374159046|Jarv NMotion Spor...|          4|            0|          0|   N|                Y|          Four
Stars|It's good but it ...| 2014-09-10|     Electronics|        1|
|         US|      11953|R2WGX3ONH2O4XO|B00LBUUH1K|     323236313|Universal Replace...|          5|            1|          1|   N|                Y|          Five
Stars|         Works great| 2014-09-28|     Electronics|        1|
|         US|      11997|R2XEBXDP1GGI92|B00A16BT4E|     874883760|SHURE Sound Isola...|          5|            7|          7|   N|                N|BEST PURCHASE E
VE...|I bought these on...| 2015-03-21|     Electronics|        1|
|         US|      12150| RZQSIB37PHU1I|B0096LZ5QI|     515180177|AmazonBasics High...|          5|            0|          0|   N|                Y|          Five
Stars|nice worked for m...| 2015-04-05|     Electronics|        1|
|         US|      12784|R37XN60RVORZZ8|B005E2Y0KK|     819401893|Circle cable clip...|          1|            0|          5|   N|                Y|      pay atten
tion!|Not really a revi...| 2015-04-29|     Electronics|        0|
|         US|      12784|R3RJWG5KKMJ902|B000MDJU06|      82954439|C2G / Cables to G...|          5|            0|          0|   N|                Y|   Best connect
ivity|The connectivity ...| 2012-11-20|     Electronics|        1|
|         US|      12917|R1FP67JTRAV0R4|B001DL9IBC|     295363842|UPG UB1250 Sealed...|          5|            0|          0|   N|                Y|          Five
Stars|        Great price.| 2015-03-09|     Electronics|        1|
|         US|      13092|R2AVQYP42YCLII|B00NVO7IDY|     755664102|MOYAKA Replacemen...|          1|            1|          5|   N|                Y|volume control
an...|These are probabl...| 2015-03-24|     Electronics|        0|
|         US|      13176|R2MSREFS80SRET|B005LXS7OM|     963657797|Generic Replaceme...|          1|            0|          0|   N|                Y|Original part r
ec...|The product didn'...| 2013-11-28|     Electronics|        0|
|         US|      13184|R1XBMWGLDRVEH9|B00CMB8JDA|     930220373|Bose Acoustic Wav...|          5|            0|          0|   N|                Y|          Five
Stars|      Great product.| 2015-04-09|     Electronics|        1|
|         US|      13315|R2BBV1ZNPNNMAD|B001E75I6E|     978464869|Ematic 2GB Color ...|          1|            0|          0|   N|                Y|only buy if you
d...|its a piece of s*...| 2013-11-14|     Electronics|        0|
|         US|      13397|R2SRJ3WJ8P719D|B004COIBFG|     224733578|SanDisk SDMX18R-0...|          4|            0|          0|   N|                Y|          great
deal|good deal save me...| 2014-05-14|     Electronics|        1|
|         US|      13461|R1943WN4S7A5BM|B00F5NE2KG|     175809283|Bluetooth Speaker...|          5|            0|          0|   N|                Y|          Five
Stars|great product and...| 2014-12-14|     Electronics|        1|
|         US|      13760|R3KU2E7ZJVEPR8|B00IF70TSI|     677825593|1byone Amplified ...|          5|            0|          0|   N|                Y|and I would rec
om...|This product work...| 2015-03-11|     Electronics|        1|
|         US|      13775| RFFUYAASNZ4H8|B0052SCU8U|     572574607|AmazonBasics High...|          5|            0|          0|   N|                Y|
good|Even when is good...| 2015-05-15|     Electronics|        1|
|         US|      14298|R2WU2XJ91S7TVR|B00BN0N08K|      21856130|Sony MDRAS200 Act...|          4|            0|          0|   N|                Y|Good for the Ac
ti...|This was bought a...| 2015-03-11|     Electronics|        1|
|         US|      14393|R2XJLHIRY1SRJI|B001TICH08|     975283485|VideoSecu Full Mo...|          5|            0|          0|   N|                Y|       Great pr
oduct|Got here quickly ...| 2015-04-24|     Electronics|        1|
+-----------+-----------+--------------+----------+--------------+--------------------+-----------+-------------+-----------+----+-----------------+--------------
-----+--------------------+-----------+---------------+---------+
only showing top 20 rows
```

A continuación vamos a entrenar el modelo, para ello utilizaremos diferentes opciones de preprocesamiento. Para poder entrenar un clasificador de sentimiento necesitamos contruir una representación del texto que nos permita entrenar el modelo. Para ello utilizaremos podemos utilizar algoritmos de extracción de características como TF-IDF o Word2Vec que vienen implementados en Spark MLlib y que nos permitirá transformar una cadena de texto a un vector para utilizarlo como datos de entrenamiento de un clasificador. Una vez implementado el modelo, lo serializaremos y guardaremos en S3 para la parte 2

Tarea 8: Modelo de sentiment

Construye un **pipeline** de entrenamiento de un modelo de sentiment a partir de los datos preparados anteriormente. Deberas utilizar una secuencia de diferentes **transformadores** y **estimadores**:

- `Tokenizer` nos permitirá construir un vector de palabras a partir de nuestras sentencias
- `StopWordsRemover` nos permitirá limpiar de nuestros vectores de palabras las de menor significado
- Construcción de características dos alternativas:
  - Modelo TF-IDF usando `HashingTF e IDF`
  - `Word2Vec` nos permitirá crear un vector a partir de la lista de palabras
- Clasificación binaria, basada en la variable sentiment que hemos utilizado, aplica un clasificador (LogisticRegregession, DecisionTree) evita ensembles por su alto tiempo de aprendizaje.

Buscando en la documentación, encuentra los distintos elementos y conectalos en un pipeline junto a un algorimo de clasificación

- Recomendamos utilizar una muestra (método `sample`) pues el tiempo puede ser excesivo
- Es posible ajustar hiperparámetros, pero igualmente puede ser bastante lento

Valida el modelo con el conjunto de test anterior usando el area bajo la curva ROC

In [12]:

```
# Solución

from pyspark.ml import Pipeline
from pyspark.ml.feature import Tokenizer, StopWordsRemover, HashingTF, IDF
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.classification import LogisticRegression


# Creamos objeto Tokenizer
tokenizer = Tokenizer(inputCol="review_body", outputCol="words")

# Creamos objeto StopWordsRemover
remover = StopWordsRemover(inputCol="words", outputCol="filtered_words")
# Datos
df_tokenized = tokenizer.transform(train_data)
df_clean = remover.transform(df_tokenized)
df_clean.select("review_body", "words", "filtered_words").show()

# Vectores de frecuencia
hashingTF = HashingTF(inputCol="filtered_words", outputCol="rawFeatures", numFeatures=8000)
idf = IDF(inputCol="rawFeatures", outputCol="features")

df_features = hashingTF.transform(df_clean)
df_rescaled = idf.fit(df_features).transform(df_features)

# Clasificación binaria y creación del pipeline
classifier = LogisticRegression(featuresCol="features", labelCol="sentiment")
pipeline = Pipeline(stages=[tokenizer, remover, hashingTF, idf, classifier])

# Entrenamiento del modelo
model = pipeline.fit(train_data)

# Realizamos las precciones
predictions = model.transform(test_data)
predictions.select("sentiment", "prediction").show()

# Crear un evaluador de clasificación
evaluator = MulticlassClassificationEvaluator(labelCol="sentiment", predictionCol="prediction", metricName="accuracy")

# Calcular la precisión del modelo
accuracy = evaluator.evaluate(predictions)
print("Accuracy:", accuracy)

# Calcular la recuperación (recall)
evaluator.setMetricName("weightedRecall")
recall = evaluator.evaluate(predictions)
print("Recall:", recall)

# Calcular la puntuación F1
evaluator.setMetricName("f1")
f1 = evaluator.evaluate(predictions)
```

```
print("F1 Score:", f1)
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
+--------------------+--------------------+--------------------+
|         review_body|               words|      filtered_words|
+--------------------+--------------------+--------------------+
|Great Sound !!!!<...|[great, sound, !!...|[great, sound, !!...|
|These are super d...|[these, are, supe...|[super, durable, ...|
|I love this! I bo...|[i, love, this!, ...|[love, this!, bou...|
|It's good but it ...|[it's, good, but,...|   [good, fix, ears]|
|         Works great|      [works, great]|      [works, great]|
|I bought these on...|[i, bought, these...|[bought, ebay, lo...|
|nice worked for m...|[nice, worked, fo...|[nice, worked, am...|
|Not really a revi...|[not, really, a, ...|[really, review.,...|
|The connectivity ...|[the, connectivit...|[connectivity, pe...|
|        Great price.|     [great, price.]|     [great, price.]|
|These are probabl...|[these, are, prob...|[probably, defect...|
|The product didn'...|[the, product, di...|[product, work, 1...|
|      Great product.|   [great, product.]|   [great, product.]|
|its a piece of s*...|[its, a, piece, o...|[piece, s***, tur...|
|good deal save me...|[good, deal, save...|[good, deal, save...|
|great product and...|[great, product, ...|[great, product, ...|
|This product work...|[this, product, w...|[product, works, ...|
|Even when is good...|[even, when, is, ...|[even, good, got,...|
|This was bought a...|[this, was, bough...|[bought, gift, so...|
|Got here quickly ...|[got, here, quick...|[got, quickly, pr...|
+--------------------+--------------------+--------------------+
only showing top 20 rows

+---------+----------+
|sentiment|prediction|
+---------+----------+
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        0|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        0|       0.0|
|        1|       1.0|
|        0|       0.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        0|       1.0|
|        1|       1.0|
+---------+----------+
only showing top 20 rows
```

```
Accuracy: 0.896269414822578
Recall: 0.8962694148225782
F1 Score: 0.8885258974153066
```

Tarea 9: Serialización

Guarda el modelo entrenado en S3 (al bucket que creaste anteriormente) utilizando la opción nativa de Spark.

In [15]:
```
# Solución
# Guardo el modelo en mi S3
model.save("s3://capstone12bucket2/trained_model")
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [ ]: