wordcount-spark-colab

November 14, 2025

[ ]:
```
#configuración de Spark en Google Colab
from google.colab import drive
drive.mount('/content/gdrive')
```

[ ]:
```
#configuración de Spark en Google Colab
#instalar java y spark
!apt-get install openjdk-17-jdk-headless -qq > /dev/null
!wget -q https://downloads.apache.org/spark/spark-4.0.1/spark-4.0.1-bin-hadoop3.
 ↪tgz
!tar xf spark-4.0.1-bin-hadoop3.tgz
!pip install -q findspark
```

[ ]:
```
#configuración de Spark en Google Colab
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-17-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-4.0.1-bin-hadoop3"
```

[ ]:
```
#configuración de Spark en Google Colab
import findspark
findspark.init()
```

[ ]:
```
#configuración de Spark en Google Colab
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
sc = spark.sparkContext
```

[ ]:
```
# WORDCOUNT COMPACTO
# en AWS S3
#files_rdd = sc.textFile("s3a://username_datalake/datasets/gutenberg-small/*.
 ↪txt")

# en gdrive:
files_rdd = sc.textFile("gdrive/MyDrive/st0263-252/bigdata/datasets/
 ↪gutenberg-small/*.txt")

# local:
#files_rdd = sc.textFile("../datasets/gutenberg-small/*.txt")
```

```
wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word:␣
 ↪(word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])
for tupla in wc.take(10):
        print(tupla)
```

[ ]:
```
# WORDCOUNT PASO A PASO
```

[ ]:
```
# en AWS S3
#files = sc.textFile("s3a://username_datalake/datasets/gutenberg-small/*.txt")
files = sc.textFile("gdrive/MyDrive/st0263-252/bigdata/datasets/gutenberg-small/
 ↪*.txt")
for f in files.take(10):
    print(f)
```

[ ]:
```
tokens = files.flatMap(lambda line: line.split())
for t in tokens.take(10):
    print(t)
```

[ ]:
```
wc1 = tokens.map(lambda word: (word, 1))
for c in wc1.take(10):
    print(c)
```

[ ]:
```
wc = wc1.reduceByKey(lambda a, b: a + b)
for c in wc.take(10):
    print(c)
```

[ ]:
```
wcsort = wc.sortBy(lambda a: -a[1])
for c in wcsort.take(10):
    print(c)
```