ATVSOMATIVA2

July 7, 2024

[1]:
```
import pyarrow as pa
import pyarrow.parquet as pq
import pyarrow.csv as pc
import pyarrow.hdfs as hdfs
import pandas as pd
import pyspark
import pandas as pd
from pyspark.sql.functions import col, split, explode
from pyspark.sql import SparkSession
from pyarrow import fs, hdfs

hdfs.HadoopFileSystem("hdfs://namenode:9000")
conexao = fs.HadoopFileSystem(host="namenode",port=9000)
```

```
/tmp/ipykernel_181/2085559598.py:12: FutureWarning:
pyarrow.hdfs.HadoopFileSystem is deprecated as of 2.0.0, please use
pyarrow.fs.HadoopFileSystem instead.
  hdfs.HadoopFileSystem("hdfs://namenode:9000")
```

[ ]:
```
conexao.get_file_info(fs.FileSelector('/', recursive=True))
```

[2]:
```
# Deletar o diretório se ele já existir
try:
    conexao.delete_dir('/guilherme')
    conexao.delete_file('/guilherme/sentiment.csv')
except Exception as e:
    print(f"O diretório foi excluido!: {e}")

conexao.create_dir('/guilherme')
```

```
O diretório foi excluido!: [Errno 2] Calling GetPathInfo for '/guilherme'
failed. Detail: [errno 2] No such file or directory
```

[3]:
```
conexao.get_file_info(fs.FileSelector('/', recursive=True))
```

[3]:
```
[<FileInfo for '/guilherme': type=FileType.Directory>]
```

[4]:
```
origem = '/home/jovyan/work/sentiment140.csv'
destino = '/guilherme/sentiment140.csv'
```

[5]:
```
conexao.get_file_info(fs.FileSelector('/', recursive=True))
```

[5]:
```
[<FileInfo for '/guilherme': type=FileType.Directory>]
```

[6]:
```
with conexao.open_output_stream(destino) as stream:
    stream.write(open(origem,'rb').read())
```

[5]:
```
conexao.get_file_info(fs.FileSelector('/', recursive=True))
```

[5]:
```
[<FileInfo for '/guilherme': type=FileType.Directory>,
 <FileInfo for '/guilherme/sentiment140.csv': type=FileType.File,
size=44326223>]
```

[7]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
df = spark.read.csv("hdfs://namenode:9000/guilherme/sentiment140.csv",␣
 ↪header=True, inferSchema=True, sep=';')
df.show()
```

```
+------+----------+--------------------+--------+---------------+--------------
-----+
|target|       ids|                date|    flag|           user|
text|
+------+----------+--------------------+--------+---------------+--------------
-----+
|     0|1467810369|Mon Apr 06 22:19:…|NO_QUERY|_TheSpecialOne_|@switchfoot
http:…|
|     0|1467810672|Mon Apr 06 22:19:…|NO_QUERY|  scotthamilton|is upset that
he …|
|     0|1467810917|Mon Apr 06 22:19:…|NO_QUERY|       mattycus|@Kenichan I
dived…|
|     0|1467811184|Mon Apr 06 22:19:…|NO_QUERY|        ElleCTF|my whole body
fee…|
|     0|1467811193|Mon Apr 06 22:19:…|NO_QUERY|
Karoli|@nationwideclass …|
|     0|1467811372|Mon Apr 06 22:20:…|NO_QUERY|       joy_wolf|@Kwesidei not
the…|
|     0|1467811592|Mon Apr 06 22:20:…|NO_QUERY|        mybirch|         Need a
hug |
|     0|1467811594|Mon Apr 06 22:20:…|NO_QUERY|           coZZ|@LOLTrish hey
lo…|
|     0|1467811795|Mon Apr 06 22:20:…|NO_QUERY|2Hood4Hollywood|@Tatiana_K nope
t…|
|     0|1467812025|Mon Apr 06 22:20:…|NO_QUERY|        mimismo|@twittera que
```

```
me …|
|     0|1467812416|Mon Apr 06 22:20:…|NO_QUERY| erinx3leannexo|spring break in
p…|
|     0|1467812579|Mon Apr 06 22:20:…|NO_QUERY|   pardonlauren|I just re-
pierced…|
|     0|1467812723|Mon Apr 06 22:20:…|NO_QUERY|           TLeC|@caregiving I
cou…|
|     0|1467812771|Mon Apr 06 22:20:…|NO_QUERY|robrobbierobert|@octolinz16 It
it…|
|     0|1467812784|Mon Apr 06 22:20:…|NO_QUERY|    bayofwolves|@smarrison i
woul…|
|     0|1467812799|Mon Apr 06 22:20:…|NO_QUERY|     HairByJess|@iamjazzyfizzle
I…|
|     0|1467812964|Mon Apr 06 22:20:…|NO_QUERY| lovesongwriter|Hollis' death
sce…|
|     0|1467813137|Mon Apr 06 22:20:…|NO_QUERY|       armotley|about to file
taxes |
|     0|1467813579|Mon Apr 06 22:20:…|NO_QUERY|     starkissed|@LettyA ahh ive
a…|
|     0|1467813782|Mon Apr 06 22:20:…|NO_QUERY|
gi_gi_bee|@FakerPattyPattz …|
+------+----------+--------------------+--------+---------------+--------------
-----+
only showing top 20 rows
```

[9]:
```python
from pyspark.sql.functions import col, lower, regexp_replace, split, explode

# Seleciona a coluna de texto e guarda elas no dataframe tweets
tweets_df = df.select("text")

# Limpa o texto para facilitar entendimento
tweets_df_cleaned = tweets_df.withColumn("text",␣
 ↪lower(regexp_replace(col("text"), "[^a-zA-Z0-9@#\\s]", "")))
# Mostra os primeiros registros para verificar a limpeza
tweets_df_cleaned.show(5, truncate=False)

# Tokenizar os tweets
words_df = tweets_df_cleaned.withColumn("word", explode(split(col("text"),␣
 ↪"\\s+")))
# Conta a frequência de cada palavra
word_freq_df = words_df.groupBy("word").count().orderBy(col("count").desc())
# Mostra as palavras mais frequentes
word_freq_df.show(10)

# Separando usuários mencionados (que começam com @) e os coloca em um␣
 ↪dataframe
```

```
mentions_df = words_df.filter(words_df.word.startswith("@"))
# Conta a frequência de usuário mencionado
mentions_freq_df = mentions_df.groupBy("word").count().orderBy(col("count").
 ↪desc())
# Mostra as menções mais frequentes
mentions_freq_df.show(10)

# Separa as hashtags e colocas elas em um dataframe
hashtags_df = words_df.filter(words_df.word.startswith("#"))
# Conta as hashtags mais frequentes
hashtags_freq_df = hashtags_df.groupBy("word").count().orderBy(col("count").
 ↪desc())
# Mostra as hashtags mais frequentes
hashtags_freq_df.show(10, truncate=False)
```

```
+--------------------------------------------------------------------------------
--------------------------+
|text
|
+--------------------------------------------------------------------------------
--------------------------+
|@switchfoot httptwitpiccom2y1zl  awww thats a bummer  you shoulda got david
carr of third day to do it d |
|is upset that he cant update his facebook by texting it and might cry as a
result  school today also blah|
|@kenichan i dived many times for the ball managed to save 50  the rest go out
of bounds                    |
|my whole body feels itchy and like its on fire
|
|@nationwideclass no its not behaving at all im mad why am i here because i cant
see you all over there   |
+--------------------------------------------------------------------------------
--------------------------+
only showing top 5 rows

+----+------+
|word| count|
+----+------+
|    |236865|
|   i|181426|
|  to|127227|
| the|104178|
|  my| 74140|
|   a| 72945|
| and| 60990|
|  is| 51644|
|  it| 47605|
```

```
|  in| 46196|
+----+------+
only showing top 10 rows

+---------------+-----+
|           word|count|
+---------------+-----+
|              @| 2426|
|    @mileycyrus|  699|
|      @tommcfly|  666|
|      @ddlovato|  418|
|@jonathanrknight|  258|
|  @taylorswift13|  200|
|   @mitchelmusso|  195|
|    @davidarchie|  192|
|  @jonasbrothers|  174|
| @donniewahlberg|  161|
+---------------+-----+
only showing top 10 rows

+-------------+-----+
|word         |count|
+-------------+-----+
|#fb          |376  |
|#fail        |143  |
|#asot400     |135  |
|#bgt         |112  |
|#            |111  |
|#e3          |76   |
|#followfriday|72   |
|#1           |61   |
|#ontd        |59   |
|#2           |58   |
+-------------+-----+
only showing top 10 rows
```

[ ]: