ANO
2023

UNINTER

CADERNO DE RESPOSTAS DA ATIVIDADE PRÁTICA DE:

PROJETOS BIG DATA

ALUNO:
MOACIR DOMINGOS DA SILVA JUNIOR
RU 3539252

Caderno de Resposta Elaborado por:
Prof. MSc. Guilherme Ditzel Patriota

Prática 01 – Somatório de IDs

Questão – Qual o valor da soma de todos os campos “id” dos filmes classificados como negativos para o banco de dados “imdb-reviews-pt-br.csv”

ENUNCIADO: Nessa prática você deverá descobrir, utilizando sua máquina virtual com o Hadoop ou Spark ou PySpark, qual o valor da soma de todos os campos “id” dos filmes classificados como negativos para o banco de dados “imdb-reviews-pt-br.csv”.

I. Apresentação do Código (não esquecer do identificador pessoal):

Atividade Pratica Big Data.py

```
--Atividade Pratica Big Data

from google.colab import drive
drive.mount('/content/drive')
!cp /content/drive/MyDrive/BIGDATA\
/imdb-reviews-pt-br.csv /content

!pip install pyspark

from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
cam_arq ='/content/drive/MyDrive\
/BIGDATA/imdb-reviews-pt-br.csv'

imdbDf = spark.read.csv\
(cam_arq, header=True, inferSchema=True)

from pyspark.sql.functions import sum
RU = "3539252"
neg_imdbDf = imdbDf.filter\
        (imdbDf.sentiment == "neg")
soma = neg_imdbDf.select\
          (sum("id")).collect()[0][0]
print(soma)
```

Figura 1: Apresentação do Código

II. Apresentação das Imagens/Print do resultado (não esquecer do identificador):

```
[13]  1 imdbDf.show(2)

+---+--------------------+--------------------+---------+
| id|             text_en|             text_pt|sentiment|
+---+--------------------+--------------------+---------+
|  1|Once again Mr. Co...|Mais uma vez, o S...|      neg|
|  2|This is an exampl...|Este é um exemplo...|      neg|
+---+--------------------+--------------------+---------+
only showing top 2 rows
```

```
1 from pyspark.sql.functions import sum
2 RU = "3539252"
3 neg_imdbDf = imdbDf.filter\
4         (imdbDf.sentiment == "neg")
5 soma = neg_imdbDf.select\
6         (sum("id")).collect()[0][0]
7 print(soma)
```

```
247015948
```

```
[ ]  1 RU = "3539252"
```

1s conclusão: 17:22

Figura 2: Print do Resultado do código executado no Google Colab

III. Responda à pergunta: Qual o valor da soma de todos os campos “id” dos filmes classificados como negativos?

Resposta: A soma de todos os campos ‘Id” dos filmes classificados como negativo é 247015948.

Prática 02 – Diferença do número de palavras totais de português para inglês dos textos negativos

Questão – Contar palavras dos textos negativos e achar diferença de quantidade.

ENUNCIADO: Nessa prática você deverá contar todas as palavras existentes nos textos negativos (Português e Inglês) e então deverá encontrar quantas palavras a mais, no total, os textos em português possuem.
Para tal, crie um script em Python ou Scala e rode-o com sua máquina virtual Hadoop ou Spark ou PySpark, como feito na prática 1.
É necessário se preocupar em filtrar corretamente as avaliações de filmes para que apenas os textos marcados como negativos sejam contabilizados.

I. Apresentação do Código (não esquecer do identificador pessoal):

Atividade Pratica Big Data.py

```
30  from pyspark.sql.functions\
31  import explode, split, col
32  neg_imdbDf = imdbDf.filter\
33  (imdbDf.sentiment == "neg")
34  palsPtDf = neg_imdbDf.select\
35  (explode(split\
36  (neg_imdbDf.text_pt, "\\W+"))\
37  .alias("pals_pt"))
38  palsEnDf = neg_imdbDf.select\
39  (explode(split\
40  (neg_imdbDf.text_en, "\\W+"))\
41  .alias("pals_en"))
42  contPalsPtDf = palsPtDf.groupBy\
43  ("pals_pt").count()
44  contPalsEnDf = palsEnDf.groupBy\
45  ("pals_en").count()
46
47  totalPt = contPalsPtDf.agg(\
48  {"count": "sum"}).collect()[0][0]
49  RU = "3539252"
50  totalEn = contPalsEnDf.agg(\
51  {"count": "sum"}).collect()[0][0]
52
53  diferencaPalavras = totalPt - totalEn
54  print(diferencaPalavras)
55
```

Figura 3: Apresentação do Código

II. Apresentação das Imagens/Print do resultado (não esquecer do identificador):

```
[24]  1 from pyspark.sql.functions\
      2 import explode, split, col
      3 neg_imdbDf = imdbDf.filter\
      4 (imdbDf.sentiment == "neg")
      5 palsPtDf = neg_imdbDf.select(explode(split\
      6 (neg_imdbDf.text_pt, "\\W+")).alias("pals_pt"))
      7 palsEnDf = neg_imdbDf.select(explode(split\
      8 (neg_imdbDf.text_en, "\\W+")).alias("pals_en"))
      9 contPalsPtDf = palsPtDf.groupBy\
     10 ("pals_pt").count()
     11 contPalsEnDf = palsEnDf.groupBy\
     12 ("pals_en").count()
```

```
1 totalPt = contPalsPtDf.agg(\
2 {"count": "sum"}).collect()[0][0]
3 RU = "3539252"
4 totalEn = contPalsEnDf.agg(\
5 {"count": "sum"}).collect()[0][0]
6
7 diferencaPalavras = totalPt - totalEn
8 print(diferencaPalavras)
```

```
160085
```

```
1 RU = "3539252"
```

4s conclusão: 19:05

Figura 4: Print do Resultado do código executado no Google Colab

III. Responda à pergunta: Qual o número total de palavras a mais que os textos negativos em português possuem em comparação com a soma total das palavras dos textos negativos em inglês, independentemente de serem repetidas?

Resposta: Os textos negativos em português possuem 160085 palavras a mais do que os textos negativos em inglês.