EAFIT

Special topics in telematics

Laboratory 2

Luisa María Álvarez García – Computer Science

(lmalvarez8@eafit.edu.co)

Teacher: Edwin Nelson Montoya

Medellin, November 11, 2024

MAPREDUCE

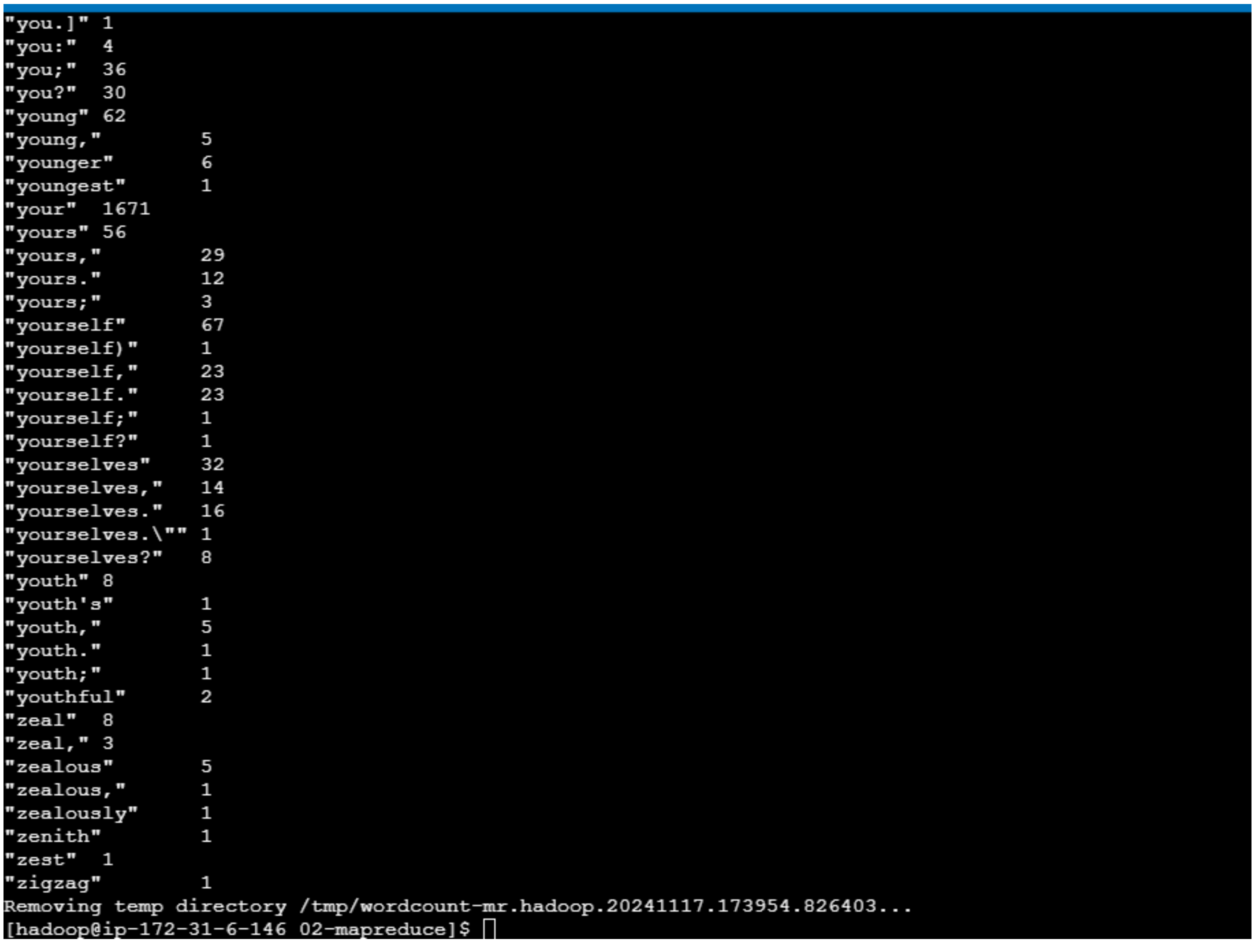
1. First of all, to make the demo of MapReduce using MRJOB create a folder, in this case it is called ***02-mapreduce***, enter into it an run the ***wordcount-local.py*** file that contains:

```
import os
import sys
import glob
import codecs

inputdir = "."

if len(sys.argv) >= 2:
    inputdir = sys.argv[1]

def processdir(dir):
    # Obtiene la lista de archivos
    dirList = glob.glob(dir)
    wordcount = {}

    # Procesa cada archivo
    for f in dirList:
        wordcountfile(f, wordcount)

    # Imprime el conteo de palabras
    for w in sorted(wordcount.keys()):
        print(w, wordcount[w])

def wordcountfile(f, wordcount):
    try:
        # Abre el archivo con codificación UTF-8
        with codecs.open(f, "r", "utf-8") as file:
            for word in file.read().split():
                # Limpia la palabra (elimina puntuación y espacios extra)
                word = word.lower().strip(",.!?;:\"'()[]{}")

                if word:  # Solo cuenta palabras no vacías
                    # Incrementa el contador o inicializa
                    if word not in wordcount:
                        wordcount[word] = 1
                    else:
                        wordcount[word] += 1
    except Exception as e:
        print(f"Error leyendo archivo {f}: {e}", file=sys.stderr)
processdir(inputdir)
```

2. Now, to run the code into the command panel, write:

```
python wordcount-local.py ../datasets/gutenberg-small/*.txt > salida-serial.txt
```

3. Now the terminal should show you:



```
"you.]" 1
"you:"  4
"you;"  36
"you?"  30
"young" 62
"young,"        5
"younger"       6
"youngest"      1
"your"  1671
"yours" 56
"yours,"        29
"yours."        12
"yours;"        3
"yourself"      67
"yourself)"     1
"yourself,"     23
"yourself."     23
"yourself;"     1
"yourself?"     1
"yourselves"    32
"yourselves,"   14
"yourselves."   16
"yourselves.\"" 1
"yourselves?"   8
"youth" 8
"youth's"       1
"youth,"        5
"youth."        1
"youth;"        1
"youthful"      2
"zeal"  8
"zeal," 3
"zealous"       5
"zealous,"      1
"zealously"     1
"zenith"        1
"zest"  1
"zigzag"        1
Removing temp directory /tmp/wordcount-mr.hadoop.20241117.173954.826403...
[hadoop@ip-172-31-6-146 02-mapreduce]$ 
```