

Práctica de MapReduce en Colab con Python

▼ Paso 1: Instalación de Hadoop en el entorno

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
```

```
--2023-11-27 20:22:55-- https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.95.219, 2a01:4f8:10a:201a::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 730107476 (696M) [application/x-gzip]
Saving to: 'hadoop-3.3.6.tar.gz'

hadoop-3.3.6.tar.gz 100%[=====] 696.28M 19.2MB/s in 38s

2023-11-27 20:23:34 (18.3 MB/s) - 'hadoop-3.3.6.tar.gz' saved [730107476/730107476]
```

Descompresión de la distribución de hadoop

```
[ ] !tar -xzf hadoop-3.3.6.tar.gz
```

Copiar a [/usr/local](#)

```
[ ] #copy hadoop file to user/local
!cp -r hadoop-3.3.6/ /usr/local/
```

▼ Step 2: Configurar el Hadoop JAVA HOME

Hadoop requiere que se establezca la ruta de acceso a Java, ya sea como una variable de entorno o en el archivo de configuración de Hadoop.

1. Buscamos cual es la dirección de Java en la máquina Google Colab

```
!readlink -f /usr/bin/java | sed "s:bin/java::"
```

```
/usr/lib/jvm/java-11-openjdk-amd64/
```

2. Establecemos mediante código Python el valor de esta variable

```
[ ] import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64/"
```

▼ Paso 3: Ejecutando Hadoop

```
#Running Hadoop
!usr/local/hadoop-3.3.6/bin/hadoop
```

```
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
or   hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class

OPTIONS is none or any of:

buildpaths      attempt to add class files from build tree
--config dir    Hadoop config directory
--debug         turn on shell script debug mode
--help         usage information
hostnames list[,of,host,names] hosts to use in worker mode
hosts filename  list of hosts to use in worker mode
loglevel level  set the log4j level for this command
workers        turn on worker mode

SUBCOMMAND is one of:

Admin Commands:

daemonlog      get/set the log level for each daemon

Client Commands:

archive        create a Hadoop archive
checknative    check native Hadoop and compression libraries availability
classpath      prints the class path needed to get the Hadoop jar and the required libraries
conftest       validate configuration XML files
credential      interact with credential providers
distch         distributed metadata changer
distcp         copy file or directories recursively
dtutil         operations related to delegation tokens
envvars        display computed Hadoop environment variables
fs             run a generic filesystem user client
gridmix        submit a mix of synthetic job, modeling a profiled from production load
jar <jar>      run a jar file. NOTE: please use "yarn jar" to launch YARN applications, not this
               command.
jnipath        prints the java.library.path
kdiag         Diagnose Kerberos Problems
```

```
[ ] !mkdir ~/input
!cp /usr/local/hadoop-3.3.6/etc/hadoop/*.* ~/input

[ ] !ls ~/input

capacity-scheduler.xml  hadoop-policy.xml  hdfs-site.xml  kms-acls.xml  mapred-site.xml
core-site.xml           hdfs-rbf-site.xml  httpfs-site.xml  kms-site.xml  yarn-site.xml

[ ] !/usr/local/hadoop-3.3.6/bin/hadoop jar /usr/local/hadoop-3.3.6/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar grep ~/input ~/grep_example 'allowed[.]'
```

```
2023-11-27 20:25:16,752 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-11-27 20:25:17,016 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-11-27 20:25:17,017 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-11-27 20:25:17,388 INFO input.FileInputFormat: Total input files to process : 10
2023-11-27 20:25:17,427 INFO mapreduce.JobSubmitter: number of splits:10
2023-11-27 20:25:17,866 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1156866682_0001
2023-11-27 20:25:17,906 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-27 20:25:18,133 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-11-27 20:25:18,134 INFO mapreduce.Job: Running job: job_local1156866682_0001
2023-11-27 20:25:18,143 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-11-27 20:25:18,153 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2023-11-27 20:25:18,155 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-27 20:25:18,155 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
2023-11-27 20:25:18,156 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-11-27 20:25:18,232 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-11-27 20:25:18,233 INFO mapred.LocalJobRunner: Starting task: attempt_local1156866682_0001_m_000000_0
2023-11-27 20:25:18,283 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2023-11-27 20:25:18,287 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
```

```
!cat ~/grep_example/*
```

```
22 allowed.
1 allowed
```

Download 20newsgroups dataset available at <http://qwone.com/~jason/20Newsgroups>.

```
[ ] !wget http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz
```

```
!tar -xzf 20news-18828.tar.gz
```

```
--2023-11-27 20:25:29-- http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz
Resolving qwone.com (qwone.com)... 173.48.205.131
Connecting to qwone.com (qwone.com)[173.48.205.131]:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 14666916 (14M) [application/x-gzip]
Saving to: '20news-18828.tar.gz'
```

```
20news-18828.tar.gz 100%[=====>] 13.99M 13.0MB/s in 1.1s
```

```
2023-11-27 20:25:30 (13.0 MB/s) - '20news-18828.tar.gz' saved [14666916/14666916]
```

▼ Hadoop Streaming

Hadoop streaming permite crear y ejecutar trabajos Map/Reduce con cualquier ejecutable o script como mapeador y/o reductor.

Mas información sobre Map/Reduce en el siguiente [enlace](#)

```
!find / -name 'hadoop-streaming*.jar'
```

```
/usr/local/hadoop-3.3.6/share/hadoop/tools/sources/hadoop-streaming-3.3.6-sources.jar
/usr/local/hadoop-3.3.6/share/hadoop/tools/sources/hadoop-streaming-3.3.6-test-sources.jar
/usr/local/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar
find: '/proc/59/task/59/net': Invalid argument
find: '/proc/59/net': Invalid argument
/content/hadoop-3.3.6/share/hadoop/tools/sources/hadoop-streaming-3.3.6-sources.jar
/content/hadoop-3.3.6/share/hadoop/tools/sources/hadoop-streaming-3.3.6-test-sources.jar
/content/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar
```

mapper.py

Leerá los datos de *STDIN*, los dividirá en palabras y enviará a *STDOUT* una lista de líneas que asignan las palabras a sus recuentos.

```
!cat /content/mapper.py
```

```
# -*- coding: utf-8 -*-
"""mapper.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1yCwGyMX7I2qt3_58a100i7X086IaPcJd
"""

import sys
import io
import re
import nltk
nltk.download('stopwords',quiet=True)
from nltk.corpus import stopwords
punctuations = '''!()-[]{};:'"\,.<>./?@#%*&_~'''

stop_words = set(stopwords.words('english'))
input_stream = io.TextIOWrapper(sys.stdin.buffer, encoding='latin1')
for line in input_stream:
    line = line.strip()
    line = re.sub(r'[\w\s]', ' ',line)
    line = line.lower()
    for x in line:
        if x in punctuations:
            line=line.replace(x, " ")

    words=line.split()
    for word in words:
        if word not in stop_words:
            print('%s\t%s' % (word, 1))
```

reducer.py

Leerá los resultados de mapper.py desde *STDIN* y sumará las ocurrencias de cada palabra hasta llegar a un recuento final, y luego enviará sus resultados a *STDOUT*.

```
[ ] !cat /content/reducer.py
```

```
# -*- coding: utf-8 -*-
"""reducer.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1Yz7-VUs05VVCyMrfPMow3s2IdxXkyQ0i
"""

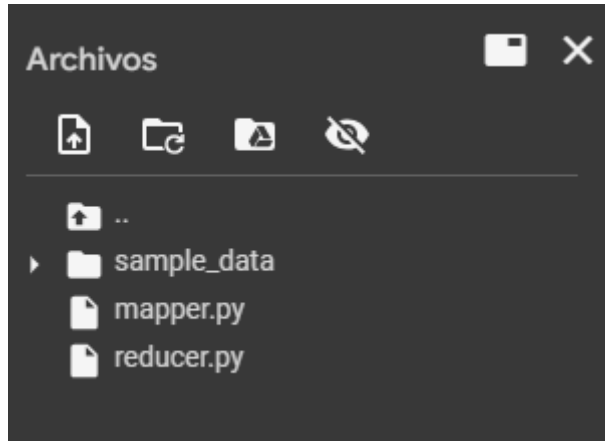
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    line=line.lower()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        #count was not a number, so silently
        #ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
```



```
[ ] !chmod u+rwx /content/mapper.py
[ ] chmod u+rwx /content/reducer.py

!usr/local/hadoop-3.3.6/bin/hadoop jar /usr/local/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /content/20news-18828/alt.atheism/49960 -output /content/output

2023-11-27 20:27:33,288 WARN streaming.StreamJob: file option is deprecated, please use generic option -files instead.
2023-11-27 20:27:33,300 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-11-27 20:27:34,319 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-11-27 20:27:34,319 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-11-27 20:27:34,347 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-11-27 20:27:34,597 INFO mapred.FileInputFormat: Total input files to process : 1
2023-11-27 20:27:34,630 INFO mapreduce.JobSubmitter: number of splits:1
2023-11-27 20:27:34,986 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local755300048_0001
2023-11-27 20:27:34,987 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-27 20:27:35,400 INFO mapred.LocalDistributedCacheManager: Localized file:/content/mapper.py as file:/tmp/hadoop-root/mapred/local/job_local755300048_0001_f829a677-19be-4617-933a-0
2023-11-27 20:27:35,496 INFO mapred.LocalDistributedCacheManager: Localized file:/content/reducer.py as file:/tmp/hadoop-root/mapred/local/job_local755300048_0001_76497ad0-41be-4575-95ff-
2023-11-27 20:27:35,616 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-11-27 20:27:35,618 INFO mapreduce.Job: Running job: job_local755300048_0001
2023-11-27 20:27:35,624 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-11-27 20:27:35,627 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2023-11-27 20:27:35,634 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-27 20:27:35,634 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-27 20:27:35,707 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-11-27 20:27:35,712 INFO mapred.LocalJobRunner: Starting task: attempt_local755300048_0001_m_000000_0
2023-11-27 20:27:35,754 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-11-27 20:27:35,754 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-11-27 20:27:35,797 INFO mapred.Task: Using ResourceCalculatorProcessFree: [ ]
2023-11-27 20:27:35,811 INFO mapred.MapTask: Processing split: file:/content/20news-18828/alt.atheism/49960:0+11599
2023-11-27 20:27:35,831 INFO mapred.MapTask: numReduceTasks: 1
2023-11-27 20:27:35,914 INFO mapred.MapTask: (EQUATION) 0 kv1 26214396(104857584)
```

```
[ ] !ls /content/output

part-000000 _SUCCESS

[ ] !cat /content/output/part-000000

034529887x      1
0511211216     1
071             5
080182494x     1
0801834074     1
0877226423     1
0877227675     1
0908            1
0910309264     1
1              1
10             1
11             1
1266           1
1271           1
14             1
140195         1
14215          1
14226          1
142282197      1
17701900       1
1881           1
1977           1
1981           1
1986           1
```