TFIDF (Term Frequency-inverse document frequency) is a statistic that measures how import a word is in a document.

In order to compare word occurences in documents of different lengths, we have to use the frequency of the word occurrence to see how the word is prevalent in the document treated now.

The inverse term frequcency part means that the weight of terms which occur a lot diminishes and wthe weight of terms occurring rarely increases.

MAP and reduce implementation:

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

* **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:   
    
  TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).
* **IDF: Inverse Document Frequency**, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:   
    
  IDF(t) = log\_e(Total number of documents / Number of documents with term t in it).

TFIDF: TF\*IDF

Section mapper reducer

1. Calculate the number of occurences of the word in document

In the mapper1, we use regex (get rid of stopwords and non-words) to match words and write <word#DocumentName,1> pairs to the stdOut, which the reducer will take as input.

The reducer1 will aggregate those numbers and output the number of words per documents. The function is designed as follows:

Map()：

Input: <documentLineNumer, contents>

Output: <<word#documentName>, 1>

Reduce()：

Input: <<word#documentName>, 1>

Output: <<word#documentName>, n>

1. Calculate the total number of words of each document

We reorder the input in the mapper, by making the key the documentName and the value word and count: <key, value> output is <documentName,word#n>.

The reducer will count the number of words per documentName, and output th word, filename, count and number of words in document. The function is designed as follows:

Map()：

Input: <<word#documentName>, n>

Output: <documentName, <word#n>>

Reduce()：

Input: <documentName, <word#n>>

Output: <<word#documentName>, <n#N>>

2 possible ways now: C.a (Calculate the TF-IDF by adding an extra Mapper) and C.b (Calculate the TF-IDF by including the calculation the the Reducer.

1. Calculate TF-IDF

Map()：

Input: <<word#documentName>, <n#N>>

Output: <word, <documentNamed#n#N#1>>

Reduce()：

Input: <word, <documentNamed#n#N#1>>

Output: <<word#documentName>, <tfidf >>

Interesting variant to put: calculate the tf-idf in a new mapper, or include it in the last reducer? -> see the runtime difference

To run the cluster:

wget <https://raw.githubusercontent.com/lo1gr/tf-idf_hadoop-Spark/master/1-easy_setup.sh>

sh 1-easy.setup.sh

Download all texts and put them in input:

for i in {0..9}; do

wget <https://raw.githubusercontent.com/lo1gr/tf-idf_hadoop-Spark/master/documents/0/text_$i.txt>

hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

done

Then to run:

wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/2-easy\_run.sh

time sh 2-easy\_run.sh

wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/3-easy\_finish.sh

sh 3-easy\_finish.sh

Then redo:

for i in {0..9}; do

wget <https://raw.githubusercontent.com/lo1gr/tf-idf_hadoop-Spark/master/documents/1/text_$i.txt>

hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

done

Last login: Sun Feb 17 14:24:30 on ttys000

(base) louiss-MacBook-Air:~ louisleguingruand$ ssh -i /Users/louisleguingruand/Desktop/polytechnique/db/project/hadoop\_project.pem hadoop@ec2-54-171-216-21.eu-west-1.compute.amazonaws.com

The authenticity of host 'ec2-54-171-216-21.eu-west-1.compute.amazonaws.com (54.171.216.21)' can't be established.

ECDSA key fingerprint is SHA256:G2MnQDfxauxh7rGN0iMRazjyrt+/4WJCJNdcYPFSl1g.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added 'ec2-54-171-216-21.eu-west-1.compute.amazonaws.com,54.171.216.21' (ECDSA) to the list of known hosts.

Last login: Sun Feb 17 19:04:20 2019

\_\_| \_\_|\_ )

\_| ( / Amazon Linux AMI

\_\_\_|\\_\_\_|\_\_\_|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/

No packages needed for security; 2 packages available

Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRR

E::::::::::::::::::E M:::::::M M:::::::M R::::::::::::::R

EE:::::EEEEEEEEE:::E M::::::::M M::::::::M R:::::RRRRRR:::::R

E::::E EEEEE M:::::::::M M:::::::::M RR::::R R::::R

E::::E M::::::M:::M M:::M::::::M R:::R R::::R

E:::::EEEEEEEEEE M:::::M M:::M M:::M M:::::M R:::RRRRRR:::::R

E::::::::::::::E M:::::M M:::M:::M M:::::M R:::::::::::RR

E:::::EEEEEEEEEE M:::::M M:::::M M:::::M R:::RRRRRR::::R

E::::E M:::::M M:::M M:::::M R:::R R::::R

E::::E EEEEE M:::::M MMM M:::::M R:::R R::::R

EE:::::EEEEEEEE::::E M:::::M M:::::M R:::R R::::R

E::::::::::::::::::E M:::::M M:::::M RR::::R R::::R

EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_$i.txt

> done

--2019-02-17 19:07:14-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_0.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 426875 (417K) [text/plain]

Saving to: ‘text\_0.txt’

text\_0.txt 100%[===================>] 416.87K --.-KB/s in 0.06s

2019-02-17 19:07:14 (7.33 MB/s) - ‘text\_0.txt’ saved [426875/426875]

--2019-02-17 19:07:14-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_1.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 428508 (418K) [text/plain]

Saving to: ‘text\_1.txt’

text\_1.txt 100%[===================>] 418.46K --.-KB/s in 0.05s

2019-02-17 19:07:15 (8.97 MB/s) - ‘text\_1.txt’ saved [428508/428508]

--2019-02-17 19:07:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_2.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 426948 (417K) [text/plain]

Saving to: ‘text\_2.txt’

text\_2.txt 100%[===================>] 416.94K --.-KB/s in 0.05s

2019-02-17 19:07:15 (8.79 MB/s) - ‘text\_2.txt’ saved [426948/426948]

--2019-02-17 19:07:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_3.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 427550 (418K) [text/plain]

Saving to: ‘text\_3.txt’

text\_3.txt 100%[===================>] 417.53K --.-KB/s in 0.05s

2019-02-17 19:07:15 (8.88 MB/s) - ‘text\_3.txt’ saved [427550/427550]

--2019-02-17 19:07:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_4.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 426368 (416K) [text/plain]

Saving to: ‘text\_4.txt’

text\_4.txt 100%[===================>] 416.38K --.-KB/s in 0.05s

2019-02-17 19:07:15 (7.47 MB/s) - ‘text\_4.txt’ saved [426368/426368]

--2019-02-17 19:07:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_5.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 427052 (417K) [text/plain]

Saving to: ‘text\_5.txt’

text\_5.txt 100%[===================>] 417.04K --.-KB/s in 0.05s

2019-02-17 19:07:16 (8.49 MB/s) - ‘text\_5.txt’ saved [427052/427052]

--2019-02-17 19:07:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_6.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 427125 (417K) [text/plain]

Saving to: ‘text\_6.txt’

text\_6.txt 100%[===================>] 417.11K --.-KB/s in 0.05s

2019-02-17 19:07:16 (7.45 MB/s) - ‘text\_6.txt’ saved [427125/427125]

--2019-02-17 19:07:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_7.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 426271 (416K) [text/plain]

Saving to: ‘text\_7.txt’

text\_7.txt 100%[===================>] 416.28K --.-KB/s in 0.05s

2019-02-17 19:07:16 (7.44 MB/s) - ‘text\_7.txt’ saved [426271/426271]

--2019-02-17 19:07:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_8.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 427130 (417K) [text/plain]

Saving to: ‘text\_8.txt’

text\_8.txt 100%[===================>] 417.12K --.-KB/s in 0.05s

2019-02-17 19:07:17 (8.35 MB/s) - ‘text\_8.txt’ saved [427130/427130]

--2019-02-17 19:07:17-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/2/text\_9.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 427613 (418K) [text/plain]

Saving to: ‘text\_9.txt’

text\_9.txt 100%[===================>] 417.59K --.-KB/s in 0.05s

2019-02-17 19:07:17 (7.48 MB/s) - ‘text\_9.txt’ saved [427613/427613]

[hadoop@ip-172-31-7-172 ~]$ ls

text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {1..3}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper$i.py

> done

--2019-02-17 19:07:27-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper1.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1804 (1.8K) [text/plain]

Saving to: ‘mapper1.py’

mapper1.py 100%[===================>] 1.76K --.-KB/s in 0s

2019-02-17 19:07:27 (215 MB/s) - ‘mapper1.py’ saved [1804/1804]

--2019-02-17 19:07:27-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper2.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 626 [text/plain]

Saving to: ‘mapper2.py’

mapper2.py 100%[===================>] 626 --.-KB/s in 0s

2019-02-17 19:07:27 (85.4 MB/s) - ‘mapper2.py’ saved [626/626]

--2019-02-17 19:07:27-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper3.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 353 [text/plain]

Saving to: ‘mapper3.py’

mapper3.py 100%[===================>] 353 --.-KB/s in 0s

2019-02-17 19:07:28 (50.1 MB/s) - ‘mapper3.py’ saved [353/353]

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py text\_0.txt text\_3.txt text\_6.txt text\_9.txt

mapper2.py text\_1.txt text\_4.txt text\_7.txt

mapper3.py text\_2.txt text\_5.txt text\_8.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {1..2}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer$i.py

> done

--2019-02-17 19:07:42-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer1.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1147 (1.1K) [text/plain]

Saving to: ‘reducer1.py’

reducer1.py 100%[===================>] 1.12K --.-KB/s in 0s

2019-02-17 19:07:43 (128 MB/s) - ‘reducer1.py’ saved [1147/1147]

--2019-02-17 19:07:43-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer2.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1000 [text/plain]

Saving to: ‘reducer2.py’

reducer2.py 100%[===================>] 1000 --.-KB/s in 0s

2019-02-17 19:07:43 (137 MB/s) - ‘reducer2.py’ saved [1000/1000]

[hadoop@ip-172-31-7-172 ~]$ wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer3\_A.py

--2019-02-17 19:08:05-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer3\_A.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1155 (1.1K) [text/plain]

Saving to: ‘reducer3\_A.py’

reducer3\_A.py 100%[===================>] 1.13K --.-KB/s in 0s

2019-02-17 19:08:05 (154 MB/s) - ‘reducer3\_A.py’ saved [1155/1155]

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py reducer1.py text\_0.txt text\_3.txt text\_6.txt text\_9.txt

mapper2.py reducer2.py text\_1.txt text\_4.txt text\_7.txt

mapper3.py reducer3\_A.py text\_2.txt text\_5.txt text\_8.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py reducer1.py text\_0.txt text\_3.txt text\_6.txt text\_9.txt

mapper2.py reducer2.py text\_1.txt text\_4.txt text\_7.txt

mapper3.py reducer3\_A.py text\_2.txt text\_5.txt text\_8.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

>

>

>

>

> echo ciao

>

>

>

>

> out

> wget http

> done

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:35-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:38-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:40-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:43-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:45-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:48-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:50-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:53-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:55-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

put: `/user/hadoop/wc/input': No such file or directory: `hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input'

ciao

-bash: out: command not found

--2019-02-17 19:09:58-- http://http/

Resolving http (http)... failed: Name or service not known.

wget: unable to resolve host address ‘http’

[hadoop@ip-172-31-7-172 ~]$ chmod +x \*.py

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc/input

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/input/\*.txt \

> -output /user/hadoop/wc/poutput1 \

> -file /home/hadoop/mapper1.py \

> -mapper /home/hadoop/mapper1.py \

> -file /home/hadoop/reducer1.py \

> -reducer /home/hadoop/reducer1.py

19/02/17 19:11:40 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper1.py, /home/hadoop/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob5279890728215718402.jar tmpDir=null

19/02/17 19:11:41 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:11:41 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:11:42 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:11:42 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:11:42 INFO mapred.FileInputFormat: Total input files to process : 10

19/02/17 19:11:42 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 19:11:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0001

19/02/17 19:11:43 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0001

19/02/17 19:11:43 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0001/

19/02/17 19:11:43 INFO mapreduce.Job: Running job: job\_1550430054052\_0001

19/02/17 19:11:51 INFO mapreduce.Job: Job job\_1550430054052\_0001 running in uber mode : false

19/02/17 19:11:51 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:12:09 INFO mapreduce.Job: map 15% reduce 0%

19/02/17 19:12:10 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 19:12:15 INFO mapreduce.Job: map 35% reduce 0%

19/02/17 19:12:16 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 19:12:18 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 19:12:19 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 19:12:27 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 19:12:30 INFO mapreduce.Job: map 95% reduce 4%

19/02/17 19:12:31 INFO mapreduce.Job: map 100% reduce 9%

19/02/17 19:12:32 INFO mapreduce.Job: map 100% reduce 33%

19/02/17 19:12:33 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:12:38 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:12:38 INFO mapreduce.Job: Job job\_1550430054052\_0001 completed successfully

19/02/17 19:12:38 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=2116138

FILE: Number of bytes written=8989513

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=4881350

HDFS: Number of bytes written=503482

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=20

Total time spent by all maps in occupied slots (ms)=14972535

Total time spent by all reduces in occupied slots (ms)=8407710

Total time spent by all map tasks (ms)=332723

Total time spent by all reduce tasks (ms)=93419

Total vcore-milliseconds taken by all map tasks=332723

Total vcore-milliseconds taken by all reduce tasks=93419

Total megabyte-milliseconds taken by all map tasks=479121120

Total megabyte-milliseconds taken by all reduce tasks=269046720

Map-Reduce Framework

Map input records=40000

Map output records=400000

Map output bytes=39871440

Map output materialized bytes=2196886

Input split bytes=2760

Combine input records=0

Combine output records=0

Reduce input groups=5000

Reduce shuffle bytes=2196886

Reduce input records=400000

Reduce output records=5000

Spilled Records=800000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5242

CPU time spent (ms)=59410

Physical memory (bytes) snapshot=10733600768

Virtual memory (bytes) snapshot=95563132928

Total committed heap usage (bytes)=9609150464

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=4878590

File Output Format Counters

Bytes Written=503482

19/02/17 19:12:38 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput1 \

> -output /user/hadoop/wc/poutput2 \

> -file /home/hadoop/mapper2.py \

> -mapper /home/hadoop/mapper2.py \

> -file /home/hadoop/reducer2.py \

> -reducer /home/hadoop/reducer2.py

19/02/17 19:12:43 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper2.py, /home/hadoop/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob6573444004348086750.jar tmpDir=null

19/02/17 19:12:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:12:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:12:45 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:12:45 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:12:45 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 19:12:45 INFO mapreduce.JobSubmitter: number of splits:21

19/02/17 19:12:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0002

19/02/17 19:12:46 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0002

19/02/17 19:12:46 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0002/

19/02/17 19:12:46 INFO mapreduce.Job: Running job: job\_1550430054052\_0002

19/02/17 19:12:53 INFO mapreduce.Job: Job job\_1550430054052\_0002 running in uber mode : false

19/02/17 19:12:53 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:13:08 INFO mapreduce.Job: map 5% reduce 0%

19/02/17 19:13:09 INFO mapreduce.Job: map 24% reduce 0%

19/02/17 19:13:10 INFO mapreduce.Job: map 29% reduce 0%

19/02/17 19:13:12 INFO mapreduce.Job: map 43% reduce 0%

19/02/17 19:13:13 INFO mapreduce.Job: map 67% reduce 0%

19/02/17 19:13:19 INFO mapreduce.Job: map 71% reduce 0%

19/02/17 19:13:20 INFO mapreduce.Job: map 81% reduce 0%

19/02/17 19:13:21 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 19:13:22 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:13:23 INFO mapreduce.Job: map 100% reduce 29%

19/02/17 19:13:26 INFO mapreduce.Job: map 100% reduce 43%

19/02/17 19:13:27 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:13:28 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:13:29 INFO mapreduce.Job: Job job\_1550430054052\_0002 completed successfully

19/02/17 19:13:29 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=91922

FILE: Number of bytes written=5045611

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=807870

HDFS: Number of bytes written=533482

HDFS: Number of read operations=84

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=21

Launched reduce tasks=7

Data-local map tasks=20

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12792735

Total time spent by all reduces in occupied slots (ms)=4238370

Total time spent by all map tasks (ms)=284283

Total time spent by all reduce tasks (ms)=47093

Total vcore-milliseconds taken by all map tasks=284283

Total vcore-milliseconds taken by all reduce tasks=47093

Total megabyte-milliseconds taken by all map tasks=409367520

Total megabyte-milliseconds taken by all reduce tasks=135627840

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=503482

Map output materialized bytes=104071

Input split bytes=2961

Combine input records=0

Combine output records=0

Reduce input groups=10

Reduce shuffle bytes=104071

Reduce input records=5000

Reduce output records=5000

Spilled Records=10000

Shuffled Maps =147

Failed Shuffles=0

Merged Map outputs=147

GC time elapsed (ms)=5542

CPU time spent (ms)=31780

Physical memory (bytes) snapshot=10616487936

Virtual memory (bytes) snapshot=98574495744

Total committed heap usage (bytes)=10082058240

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=804909

File Output Format Counters

Bytes Written=533482

19/02/17 19:13:29 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/poutput3 \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_A.py \

> -reducer /home/hadoop/reducer3\_A.py

19/02/17 19:14:30 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_A.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob867011133571525249.jar tmpDir=null

19/02/17 19:14:32 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:14:32 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:14:32 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:14:32 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:14:32 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 19:14:32 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 19:14:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0003

19/02/17 19:14:33 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0003

19/02/17 19:14:33 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0003/

19/02/17 19:14:33 INFO mapreduce.Job: Running job: job\_1550430054052\_0003

19/02/17 19:14:40 INFO mapreduce.Job: Job job\_1550430054052\_0003 running in uber mode : false

19/02/17 19:14:40 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:14:54 INFO mapreduce.Job: map 10% reduce 0%

19/02/17 19:14:55 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 19:14:59 INFO mapreduce.Job: map 40% reduce 0%

19/02/17 19:15:00 INFO mapreduce.Job: map 65% reduce 0%

19/02/17 19:15:01 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 19:15:03 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 19:15:04 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 19:15:11 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:15:12 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 19:15:13 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:15:17 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:15:18 INFO mapreduce.Job: Job job\_1550430054052\_0003 completed successfully

19/02/17 19:15:19 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=72852

FILE: Number of bytes written=4872519

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=829316

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=19

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12049020

Total time spent by all reduces in occupied slots (ms)=5687190

Total time spent by all map tasks (ms)=267756

Total time spent by all reduce tasks (ms)=63191

Total vcore-milliseconds taken by all map tasks=267756

Total vcore-milliseconds taken by all reduce tasks=63191

Total megabyte-milliseconds taken by all map tasks=385568640

Total megabyte-milliseconds taken by all reduce tasks=181990080

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=543482

Map output materialized bytes=123068

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=123068

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5399

CPU time spent (ms)=30940

Physical memory (bytes) snapshot=10296614912

Virtual memory (bytes) snapshot=95417896960

Total committed heap usage (bytes)=9848750080

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=826496

File Output Format Counters

Bytes Written=50840

19/02/17 19:15:19 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/poutput3

Found 8 items

-rw-r--r-- 1 hadoop hadoop 0 2019-02-17 19:15 /user/hadoop/wc/poutput3/\_SUCCESS

-rw-r--r-- 1 hadoop hadoop 7014 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00000

-rw-r--r-- 1 hadoop hadoop 6302 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00001

-rw-r--r-- 1 hadoop hadoop 7331 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00002

-rw-r--r-- 1 hadoop hadoop 8027 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00003

-rw-r--r-- 1 hadoop hadoop 6984 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00004

-rw-r--r-- 1 hadoop hadoop 8363 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00005

-rw-r--r-- 1 hadoop hadoop 6819 2019-02-17 19:15 /user/hadoop/wc/poutput3/part-00006

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -cat /user/hadoop/wc/poutput3/part-00000

argumentative hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

autophthalmoscope hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

bassists hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

benevolentness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

betokener hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

blepharemphysema hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

brachyure hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

cabombas hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

cavitates hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

charactonym hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

counterargument hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

cudweed hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

cy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

cylindroid hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

demonstratively hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

detractory hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

diastasimetry hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

discomedusae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

disinure hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

erethisia hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

etymologizing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

foamlike hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

gonk hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

greatheartedly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

gurus hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt0.0

helicity hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

hormogoneae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

inworn hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.tx0.0

kirsten hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

lickings hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

luminophor hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

merbromin hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

monroe hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.tx0.0

narcissan hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

nebulously hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

newfish hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

nonbrutally hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

nonconfidential hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

nondeluded hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

nonproducing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

opodidymus hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

pandarctos hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

panorama hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

pelites hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

poikile hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

preclusive hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

prepunishment hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

psize hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt0.0

schoolmistressy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

sissone hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

slutchy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

sneaked hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

squinant hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

sticking hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

subthalamic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

successionally hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

tetrathionic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

thalamium hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

tremblement hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

trichological hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

unagricultural hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

unfixity hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

ungaited hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

unpoached hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

vage hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

viritrate hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

weltings hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

wisshe hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.tx0.0

woodenware hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -cat /user/hadoop/wc/poutput3/part-00006

anelectric hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

antianarchic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

aquatic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

autofluorescence hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

bakerlike hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

bescarf hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

cecopexy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

certainly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

clinocephalic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

cockieleekie hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

columnizes hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

compartmentally hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

coniacian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

enterorrhexis hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

episememe hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

eurytropic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

exordium hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

ferric hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

footslog hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

garrulousness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

harquebusier hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

hillcrest hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

hyperidealistically hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

inviably hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

irritatedly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

keout hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

kg hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

lansing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

leglessness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

limbie hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

lipolysis hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

madril hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

maoists hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

millionist hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

monsignorial hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

musculatures hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

nederlands hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

ngai hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

nonfragility hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

oestriols hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

oestrual hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

overfamiliarity hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

overwillingness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

poignado hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

preeing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

prelect hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

ramuscule hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

rayless hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

recklessly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

rectosigmoid hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

rookeried hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

silicoaluminate hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

sketchingly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

skippered hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

skywriting hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

solfeges hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

strictly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

sulphatization hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

supranatural hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

undisproved hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

unforcibleness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

ungalloping hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

ungoggled hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

unrank hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

unsunken hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

varas hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

wickerworked hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop

Found 1 items

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:14 /user/hadoop/wc

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc

Found 4 items

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:11 /user/hadoop/wc/input

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:12 /user/hadoop/wc/poutput1

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:13 /user/hadoop/wc/poutput2

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:15 /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc/poutput1

Deleted /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc/poutput2

Deleted /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc/poutput3

Deleted /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_$i.txt

> done

--2019-02-17 19:20:37-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_0.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853617 (834K) [text/plain]

Saving to: ‘text\_0.txt.1’

text\_0.txt.1 100%[=====================================================>] 833.61K --.-KB/s in 0.06s

2019-02-17 19:20:37 (14.0 MB/s) - ‘text\_0.txt.1’ saved [853617/853617]

--2019-02-17 19:20:37-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_1.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853609 (834K) [text/plain]

Saving to: ‘text\_1.txt.1’

text\_1.txt.1 100%[=====================================================>] 833.60K --.-KB/s in 0.06s

2019-02-17 19:20:38 (14.3 MB/s) - ‘text\_1.txt.1’ saved [853609/853609]

--2019-02-17 19:20:38-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_2.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854680 (835K) [text/plain]

Saving to: ‘text\_2.txt.1’

text\_2.txt.1 100%[=====================================================>] 834.65K --.-KB/s in 0.05s

2019-02-17 19:20:38 (15.8 MB/s) - ‘text\_2.txt.1’ saved [854680/854680]

--2019-02-17 19:20:38-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_3.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854079 (834K) [text/plain]

Saving to: ‘text\_3.txt.1’

text\_3.txt.1 100%[=====================================================>] 834.06K --.-KB/s in 0.06s

2019-02-17 19:20:38 (14.3 MB/s) - ‘text\_3.txt.1’ saved [854079/854079]

--2019-02-17 19:20:38-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_4.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854117 (834K) [text/plain]

Saving to: ‘text\_4.txt.1’

text\_4.txt.1 100%[=====================================================>] 834.10K --.-KB/s in 0.05s

2019-02-17 19:20:38 (15.6 MB/s) - ‘text\_4.txt.1’ saved [854117/854117]

--2019-02-17 19:20:38-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_5.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853866 (834K) [text/plain]

Saving to: ‘text\_5.txt.1’

text\_5.txt.1 100%[=====================================================>] 833.85K --.-KB/s in 0.05s

2019-02-17 19:20:39 (15.1 MB/s) - ‘text\_5.txt.1’ saved [853866/853866]

--2019-02-17 19:20:39-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_6.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 855044 (835K) [text/plain]

Saving to: ‘text\_6.txt.1’

text\_6.txt.1 100%[=====================================================>] 835.00K --.-KB/s in 0.05s

2019-02-17 19:20:39 (15.5 MB/s) - ‘text\_6.txt.1’ saved [855044/855044]

--2019-02-17 19:20:39-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_7.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854250 (834K) [text/plain]

Saving to: ‘text\_7.txt.1’

text\_7.txt.1 100%[=====================================================>] 834.23K --.-KB/s in 0.05s

2019-02-17 19:20:39 (15.9 MB/s) - ‘text\_7.txt.1’ saved [854250/854250]

--2019-02-17 19:20:39-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_8.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854017 (834K) [text/plain]

Saving to: ‘text\_8.txt.1’

text\_8.txt.1 100%[=====================================================>] 834.00K --.-KB/s in 0.06s

2019-02-17 19:20:40 (14.0 MB/s) - ‘text\_8.txt.1’ saved [854017/854017]

--2019-02-17 19:20:40-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_9.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853794 (834K) [text/plain]

Saving to: ‘text\_9.txt.1’

text\_9.txt.1 100%[=====================================================>] 833.78K --.-KB/s in 0.06s

2019-02-17 19:20:40 (13.9 MB/s) - ‘text\_9.txt.1’ saved [853794/853794]

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/input

Found 10 items

-rw-r--r-- 1 hadoop hadoop 426875 2019-02-17 19:11 /user/hadoop/wc/input/text\_0.txt

-rw-r--r-- 1 hadoop hadoop 428508 2019-02-17 19:11 /user/hadoop/wc/input/text\_1.txt

-rw-r--r-- 1 hadoop hadoop 426948 2019-02-17 19:11 /user/hadoop/wc/input/text\_2.txt

-rw-r--r-- 1 hadoop hadoop 427550 2019-02-17 19:11 /user/hadoop/wc/input/text\_3.txt

-rw-r--r-- 1 hadoop hadoop 426368 2019-02-17 19:11 /user/hadoop/wc/input/text\_4.txt

-rw-r--r-- 1 hadoop hadoop 427052 2019-02-17 19:11 /user/hadoop/wc/input/text\_5.txt

-rw-r--r-- 1 hadoop hadoop 427125 2019-02-17 19:11 /user/hadoop/wc/input/text\_6.txt

-rw-r--r-- 1 hadoop hadoop 426271 2019-02-17 19:11 /user/hadoop/wc/input/text\_7.txt

-rw-r--r-- 1 hadoop hadoop 427130 2019-02-17 19:11 /user/hadoop/wc/input/text\_8.txt

-rw-r--r-- 1 hadoop hadoop 427613 2019-02-17 19:11 /user/hadoop/wc/input/text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc/input

Deleted /user/hadoop/wc/input

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc/input

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/input

Found 10 items

-rw-r--r-- 1 hadoop hadoop 426875 2019-02-17 19:21 /user/hadoop/wc/input/text\_0.txt

-rw-r--r-- 1 hadoop hadoop 428508 2019-02-17 19:21 /user/hadoop/wc/input/text\_1.txt

-rw-r--r-- 1 hadoop hadoop 426948 2019-02-17 19:21 /user/hadoop/wc/input/text\_2.txt

-rw-r--r-- 1 hadoop hadoop 427550 2019-02-17 19:21 /user/hadoop/wc/input/text\_3.txt

-rw-r--r-- 1 hadoop hadoop 426368 2019-02-17 19:21 /user/hadoop/wc/input/text\_4.txt

-rw-r--r-- 1 hadoop hadoop 427052 2019-02-17 19:21 /user/hadoop/wc/input/text\_5.txt

-rw-r--r-- 1 hadoop hadoop 427125 2019-02-17 19:21 /user/hadoop/wc/input/text\_6.txt

-rw-r--r-- 1 hadoop hadoop 426271 2019-02-17 19:21 /user/hadoop/wc/input/text\_7.txt

-rw-r--r-- 1 hadoop hadoop 427130 2019-02-17 19:21 /user/hadoop/wc/input/text\_8.txt

-rw-r--r-- 1 hadoop hadoop 427613 2019-02-17 19:21 /user/hadoop/wc/input/text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/input/\*.txt \

> -output /user/hadoop/wc/poutput1 \

> -file /home/hadoop/mapper1.py \

> -mapper /home/hadoop/mapper1.py \

> -file /home/hadoop/reducer1.py \

> -reducer /home/hadoop/reducer1.py

19/02/17 19:22:22 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper1.py, /home/hadoop/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob5600586350247193828.jar tmpDir=null

19/02/17 19:22:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:22:23 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:22:24 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:22:24 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:22:24 INFO mapred.FileInputFormat: Total input files to process : 10

19/02/17 19:22:24 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 19:22:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0004

19/02/17 19:22:24 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0004

19/02/17 19:22:24 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0004/

19/02/17 19:22:24 INFO mapreduce.Job: Running job: job\_1550430054052\_0004

19/02/17 19:22:32 INFO mapreduce.Job: Job job\_1550430054052\_0004 running in uber mode : false

19/02/17 19:22:32 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:22:42 INFO mapreduce.Job: map 5% reduce 0%

19/02/17 19:22:43 INFO mapreduce.Job: map 10% reduce 0%

19/02/17 19:22:46 INFO mapreduce.Job: map 15% reduce 0%

19/02/17 19:22:50 INFO mapreduce.Job: map 20% reduce 0%

19/02/17 19:22:52 INFO mapreduce.Job: map 25% reduce 0%

19/02/17 19:22:54 INFO mapreduce.Job: map 40% reduce 0%

19/02/17 19:22:55 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 19:22:56 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 19:23:01 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 19:23:05 INFO mapreduce.Job: map 90% reduce 0%

19/02/17 19:23:06 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 19:23:07 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:23:08 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 19:23:09 INFO mapreduce.Job: map 100% reduce 43%

19/02/17 19:23:10 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 19:23:12 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:23:13 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 19:23:14 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:23:14 INFO mapreduce.Job: Job job\_1550430054052\_0004 completed successfully

19/02/17 19:23:14 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=2116138

FILE: Number of bytes written=8989512

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=4881350

HDFS: Number of bytes written=503482

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=13

Rack-local map tasks=7

Total time spent by all maps in occupied slots (ms)=12586815

Total time spent by all reduces in occupied slots (ms)=6464880

Total time spent by all map tasks (ms)=279707

Total time spent by all reduce tasks (ms)=71832

Total vcore-milliseconds taken by all map tasks=279707

Total vcore-milliseconds taken by all reduce tasks=71832

Total megabyte-milliseconds taken by all map tasks=402778080

Total megabyte-milliseconds taken by all reduce tasks=206876160

Map-Reduce Framework

Map input records=40000

Map output records=400000

Map output bytes=39871440

Map output materialized bytes=2196886

Input split bytes=2760

Combine input records=0

Combine output records=0

Reduce input groups=5000

Reduce shuffle bytes=2196886

Reduce input records=400000

Reduce output records=5000

Spilled Records=800000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=4649

CPU time spent (ms)=56460

Physical memory (bytes) snapshot=10471698432

Virtual memory (bytes) snapshot=95659917312

Total committed heap usage (bytes)=9571401728

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=4878590

File Output Format Counters

Bytes Written=503482

19/02/17 19:23:14 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput1 \

> -output /user/hadoop/wc/poutput2 \

> -file /home/hadoop/mapper2.py \

> -mapper /home/hadoop/mapper2.py \

> -file /home/hadoop/reducer2.py \

> -reducer /home/hadoop/reducer2.py

19/02/17 19:28:17 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper2.py, /home/hadoop/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob4022118501782493265.jar tmpDir=null

19/02/17 19:28:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:28:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:28:19 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:28:19 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:28:19 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 19:28:19 INFO mapreduce.JobSubmitter: number of splits:21

19/02/17 19:28:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0005

19/02/17 19:28:19 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0005

19/02/17 19:28:19 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0005/

19/02/17 19:28:19 INFO mapreduce.Job: Running job: job\_1550430054052\_0005

19/02/17 19:28:26 INFO mapreduce.Job: Job job\_1550430054052\_0005 running in uber mode : false

19/02/17 19:28:26 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:28:41 INFO mapreduce.Job: map 19% reduce 0%

19/02/17 19:28:42 INFO mapreduce.Job: map 29% reduce 0%

19/02/17 19:28:46 INFO mapreduce.Job: map 43% reduce 0%

19/02/17 19:28:47 INFO mapreduce.Job: map 67% reduce 0%

19/02/17 19:28:51 INFO mapreduce.Job: map 71% reduce 0%

19/02/17 19:28:52 INFO mapreduce.Job: map 81% reduce 0%

19/02/17 19:28:54 INFO mapreduce.Job: map 86% reduce 0%

19/02/17 19:28:58 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:28:59 INFO mapreduce.Job: map 100% reduce 29%

19/02/17 19:29:00 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 19:29:01 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:29:04 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:29:04 INFO mapreduce.Job: Job job\_1550430054052\_0005 completed successfully

19/02/17 19:29:04 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=91932

FILE: Number of bytes written=5045631

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=807870

HDFS: Number of bytes written=533482

HDFS: Number of read operations=84

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed reduce tasks=1

Launched map tasks=21

Launched reduce tasks=7

Data-local map tasks=20

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12886020

Total time spent by all reduces in occupied slots (ms)=4973940

Total time spent by all map tasks (ms)=286356

Total time spent by all reduce tasks (ms)=55266

Total vcore-milliseconds taken by all map tasks=286356

Total vcore-milliseconds taken by all reduce tasks=55266

Total megabyte-milliseconds taken by all map tasks=412352640

Total megabyte-milliseconds taken by all reduce tasks=159166080

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=503482

Map output materialized bytes=104071

Input split bytes=2961

Combine input records=0

Combine output records=0

Reduce input groups=10

Reduce shuffle bytes=104071

Reduce input records=5000

Reduce output records=5000

Spilled Records=10000

Shuffled Maps =147

Failed Shuffles=0

Merged Map outputs=147

GC time elapsed (ms)=5324

CPU time spent (ms)=30190

Physical memory (bytes) snapshot=10460344320

Virtual memory (bytes) snapshot=98564591616

Total committed heap usage (bytes)=9965142016

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=804909

File Output Format Counters

Bytes Written=533482

19/02/17 19:29:04 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/poutput3 \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_A.py \

> -reducer /home/hadoop/reducer3\_A.py

19/02/17 19:29:13 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_A.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob832102837109220351.jar tmpDir=null

19/02/17 19:29:14 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:29:15 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:29:15 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:29:15 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:29:15 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 19:29:15 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 19:29:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0006

19/02/17 19:29:16 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0006

19/02/17 19:29:16 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0006/

19/02/17 19:29:16 INFO mapreduce.Job: Running job: job\_1550430054052\_0006

19/02/17 19:29:23 INFO mapreduce.Job: Job job\_1550430054052\_0006 running in uber mode : false

19/02/17 19:29:23 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:29:38 INFO mapreduce.Job: map 25% reduce 0%

19/02/17 19:29:39 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 19:29:43 INFO mapreduce.Job: map 50% reduce 0%

19/02/17 19:29:44 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 19:29:46 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 19:29:47 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 19:29:53 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 19:29:55 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:29:56 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 19:29:57 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 19:29:59 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 19:30:00 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:30:01 INFO mapreduce.Job: Job job\_1550430054052\_0006 completed successfully

19/02/17 19:30:01 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=72739

FILE: Number of bytes written=4872453

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=829316

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=19

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=11979405

Total time spent by all reduces in occupied slots (ms)=5801400

Total time spent by all map tasks (ms)=266209

Total time spent by all reduce tasks (ms)=64460

Total vcore-milliseconds taken by all map tasks=266209

Total vcore-milliseconds taken by all reduce tasks=64460

Total megabyte-milliseconds taken by all map tasks=383340960

Total megabyte-milliseconds taken by all reduce tasks=185644800

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=543482

Map output materialized bytes=123105

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=123105

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=4836

CPU time spent (ms)=30930

Physical memory (bytes) snapshot=10173304832

Virtual memory (bytes) snapshot=95424876544

Total committed heap usage (bytes)=9607577600

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=826496

File Output Format Counters

Bytes Written=50840

19/02/17 19:30:01 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/poutput3

Found 8 items

-rw-r--r-- 1 hadoop hadoop 0 2019-02-17 19:29 /user/hadoop/wc/poutput3/\_SUCCESS

-rw-r--r-- 1 hadoop hadoop 7014 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00000

-rw-r--r-- 1 hadoop hadoop 6302 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00001

-rw-r--r-- 1 hadoop hadoop 7331 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00002

-rw-r--r-- 1 hadoop hadoop 8027 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00003

-rw-r--r-- 1 hadoop hadoop 6984 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00004

-rw-r--r-- 1 hadoop hadoop 8363 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00005

-rw-r--r-- 1 hadoop hadoop 6819 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00006

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -cat /user/hadoop/wc/poutput3/part-00003

agaricaceae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

alidades hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

anhydridization hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

archduke hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

arthrolith hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

ataghans hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

babishness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

banting hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

blowfishes hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

broiden hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

canvassy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

chronica hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

circulator hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

clof hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

coercing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

counterpoising hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

cyanogenetic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

cyanurin hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

cypraeoid hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

dean hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

disclosures hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

dumpiness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

emblemish hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

encave hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

endoderms hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

enharmonical hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

footcandles hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

forestalls hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

frenghi hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

galoch hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

heliotypic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

impressional hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

interfoliaceous hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

lithopaedium hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

mack hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

mastauxe hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

maturant hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

megatherm hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

mellow hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

milleporous hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

mountainer hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

myotacismus hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

naphthionate hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

navy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

noninfusibleness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

nosography hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

ocydromine hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

orthoepically hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

outranks hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

pastoralness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

pectization hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

pericardiophrenic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

perisphinctes hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

platinode hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

premillenarian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

protohistoric hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

reactivated hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

removedness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

schrank hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

scolopophore hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

semimercerized hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

sextantal hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

siskins hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

spigeliaceae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

stycerin hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

tarpeian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

triglyph hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

tripodial hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

tuberculomas hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

tufter hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

uncubbed hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

underhand hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

unnettled hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

unshook hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

ussingite hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

utilidor hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

virulented hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

vivificated hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

wallman hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/poutput3

Found 8 items

-rw-r--r-- 1 hadoop hadoop 0 2019-02-17 19:29 /user/hadoop/wc/poutput3/\_SUCCESS

-rw-r--r-- 1 hadoop hadoop 7014 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00000

-rw-r--r-- 1 hadoop hadoop 6302 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00001

-rw-r--r-- 1 hadoop hadoop 7331 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00002

-rw-r--r-- 1 hadoop hadoop 8027 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00003

-rw-r--r-- 1 hadoop hadoop 6984 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00004

-rw-r--r-- 1 hadoop hadoop 8363 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00005

-rw-r--r-- 1 hadoop hadoop 6819 2019-02-17 19:29 /user/hadoop/wc/poutput3/part-00006

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc

Found 4 items

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:21 /user/hadoop/wc/input

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:23 /user/hadoop/wc/poutput1

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:29 /user/hadoop/wc/poutput2

drwxr-xr-x - hadoop hadoop 0 2019-02-17 19:29 /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc

Deleted /user/hadoop/wc

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc

^[[A[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc/input

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py reducer1.py text\_0.txt text\_1.txt.1 text\_3.txt text\_4.txt.1 text\_6.txt text\_7.txt.1 text\_9.txt

mapper2.py reducer2.py text\_0.txt.1 text\_2.txt text\_3.txt.1 text\_5.txt text\_6.txt.1 text\_8.txt text\_9.txt.1

mapper3.py reducer3\_A.py text\_1.txt text\_2.txt.1 text\_4.txt text\_5.txt.1 text\_7.txt text\_8.txt.1

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> rm text\_$i.txt

> done

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer2.py text\_0.txt.1 text\_2.txt.1 text\_4.txt.1 text\_6.txt.1 text\_8.txt.1

mapper2.py reducer1.py reducer3\_A.py text\_1.txt.1 text\_3.txt.1 text\_5.txt.1 text\_7.txt.1 text\_9.txt.1

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> rm text\_$i.txt.1

> done

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper2.py mapper3.py reducer1.py reducer2.py reducer3\_A.py

[hadoop@ip-172-31-7-172 ~]$ cat reducer3\_A.py

#!/usr/bin/env python

from operator import itemgetter

import sys

from math import log10, sqrt

prev\_word = None

count = 1

word = None

df = {}

l1 = []

D = 10.0

tfidf=0

# input: <word, <documentNamed#n#N#1>>

# input comes from STDIN

for line in sys.stdin:

line = line.strip()

w, z = line.split('\t', 1)

f, nNc = z.split(' ', 1)

n, Nc = nNc.split(' ', 1)

N, c = Nc.split(' ', 1)

if prev\_word == w:

count = count + int(c) # count occurences of word in all the docs

else:

if prev\_word != None:

q = n + ' ' + N + ' ' + str(count)

df[prev\_word] = q

j = prev\_word + ' ' + f

l1.append(j)

count = 1

prev\_word = w

q = n + ' ' + N + ' ' + str(count)

df[prev\_word] = q

j = prev\_word + ' ' + f

l1.append(j)

# l1: word#f

# df: {word: n#N#count}

for h in l1:

w, f = h.split(' ', 1)

for d in df:

if w == d:

n, N, m = df[d].split(' ', 2)

n = float(n)

N = float(N)

m = float(m)

tfidf = (n / N) \* log10(D / m)

print '%s\t%s' % (h, tfidf) # <<word#documentName>, <tfidf>>

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper2.py mapper3.py reducer1.py reducer2.py reducer3\_A.py

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_$i.txt

> done

--2019-02-17 19:57:28-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_0.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853617 (834K) [text/plain]

Saving to: ‘text\_0.txt’

text\_0.txt 100%[=====================================================>] 833.61K --.-KB/s in 0.05s

2019-02-17 19:57:28 (15.8 MB/s) - ‘text\_0.txt’ saved [853617/853617]

--2019-02-17 19:57:28-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_1.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853609 (834K) [text/plain]

Saving to: ‘text\_1.txt’

text\_1.txt 100%[=====================================================>] 833.60K --.-KB/s in 0.06s

2019-02-17 19:57:28 (14.3 MB/s) - ‘text\_1.txt’ saved [853609/853609]

--2019-02-17 19:57:28-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_2.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854680 (835K) [text/plain]

Saving to: ‘text\_2.txt’

text\_2.txt 100%[=====================================================>] 834.65K --.-KB/s in 0.06s

2019-02-17 19:57:28 (14.1 MB/s) - ‘text\_2.txt’ saved [854680/854680]

--2019-02-17 19:57:28-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_3.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854079 (834K) [text/plain]

Saving to: ‘text\_3.txt’

text\_3.txt 100%[=====================================================>] 834.06K --.-KB/s in 0.06s

2019-02-17 19:57:29 (14.0 MB/s) - ‘text\_3.txt’ saved [854079/854079]

--2019-02-17 19:57:29-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_4.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854117 (834K) [text/plain]

Saving to: ‘text\_4.txt’

text\_4.txt 100%[=====================================================>] 834.10K --.-KB/s in 0.06s

2019-02-17 19:57:29 (14.1 MB/s) - ‘text\_4.txt’ saved [854117/854117]

--2019-02-17 19:57:29-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_5.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853866 (834K) [text/plain]

Saving to: ‘text\_5.txt’

text\_5.txt 100%[=====================================================>] 833.85K --.-KB/s in 0.06s

2019-02-17 19:57:29 (14.3 MB/s) - ‘text\_5.txt’ saved [853866/853866]

--2019-02-17 19:57:29-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_6.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 855044 (835K) [text/plain]

Saving to: ‘text\_6.txt’

text\_6.txt 100%[=====================================================>] 835.00K --.-KB/s in 0.06s

2019-02-17 19:57:29 (14.4 MB/s) - ‘text\_6.txt’ saved [855044/855044]

--2019-02-17 19:57:29-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_7.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854250 (834K) [text/plain]

Saving to: ‘text\_7.txt’

text\_7.txt 100%[=====================================================>] 834.23K --.-KB/s in 0.06s

2019-02-17 19:57:30 (14.2 MB/s) - ‘text\_7.txt’ saved [854250/854250]

--2019-02-17 19:57:30-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_8.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854017 (834K) [text/plain]

Saving to: ‘text\_8.txt’

text\_8.txt 100%[=====================================================>] 834.00K --.-KB/s in 0.06s

2019-02-17 19:57:30 (14.3 MB/s) - ‘text\_8.txt’ saved [854017/854017]

--2019-02-17 19:57:30-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_9.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853794 (834K) [text/plain]

Saving to: ‘text\_9.txt’

text\_9.txt 100%[=====================================================>] 833.78K --.-KB/s in 0.06s

2019-02-17 19:57:30 (14.0 MB/s) - ‘text\_9.txt’ saved [853794/853794]

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc

^[[A/inpumkdir: `/user/hadoop/wc': File exists

t

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -mkdir /user/hadoop/wc/input

mkdir: `/user/hadoop/wc/input': File exists

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/input

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/input/\*.txt \

> -output /user/hadoop/wc/poutput1 \

> -file /home/hadoop/mapper1.py \

> -mapper /home/hadoop/mapper1.py \

> -file /home/hadoop/reducer1.py \

> -reducer /home/hadoop/reducer1.py

19/02/17 19:58:41 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper1.py, /home/hadoop/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob8651708060013969804.jar tmpDir=null

19/02/17 19:58:42 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:58:43 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 19:58:43 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 19:58:43 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 19:58:43 INFO mapred.FileInputFormat: Total input files to process : 10

19/02/17 19:58:43 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 19:58:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0007

19/02/17 19:58:44 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0007

19/02/17 19:58:44 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0007/

19/02/17 19:58:44 INFO mapreduce.Job: Running job: job\_1550430054052\_0007

19/02/17 19:58:51 INFO mapreduce.Job: Job job\_1550430054052\_0007 running in uber mode : false

19/02/17 19:58:51 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 19:59:07 INFO mapreduce.Job: map 5% reduce 0%

19/02/17 19:59:08 INFO mapreduce.Job: map 10% reduce 0%

19/02/17 19:59:09 INFO mapreduce.Job: map 25% reduce 0%

19/02/17 19:59:10 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 19:59:15 INFO mapreduce.Job: map 35% reduce 0%

19/02/17 19:59:16 INFO mapreduce.Job: map 55% reduce 0%

19/02/17 19:59:17 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 19:59:18 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 19:59:19 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 19:59:22 INFO mapreduce.Job: map 90% reduce 0%

19/02/17 19:59:27 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 19:59:28 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 19:59:30 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 19:59:31 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 19:59:34 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 19:59:35 INFO mapreduce.Job: Job job\_1550430054052\_0007 completed successfully

19/02/17 19:59:35 INFO mapreduce.Job: Counters: 52

File System Counters

FILE: Number of bytes read=4157841

FILE: Number of bytes written=13073086

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=9103903

HDFS: Number of bytes written=508400

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=18

Rack-local map tasks=2

Total time spent by all maps in occupied slots (ms)=15285465

Total time spent by all reduces in occupied slots (ms)=7158150

Total time spent by all map tasks (ms)=339677

Total time spent by all reduce tasks (ms)=79535

Total vcore-milliseconds taken by all map tasks=339677

Total vcore-milliseconds taken by all reduce tasks=79535

Total megabyte-milliseconds taken by all map tasks=489134880

Total megabyte-milliseconds taken by all reduce tasks=229060800

Map-Reduce Framework

Map input records=80000

Map output records=800000

Map output bytes=79741073

Map output materialized bytes=4238754

Input split bytes=2760

Combine input records=0

Combine output records=0

Reduce input groups=5000

Reduce shuffle bytes=4238754

Reduce input records=800000

Reduce output records=5000

Spilled Records=1600000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5362

CPU time spent (ms)=71990

Physical memory (bytes) snapshot=10630529024

Virtual memory (bytes) snapshot=95614472192

Total committed heap usage (bytes)=9727115264

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=9101143

File Output Format Counters

Bytes Written=508400

19/02/17 19:59:35 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput1 \

> -output /user/hadoop/wc/poutput2 \

> -file /home/hadoop/mapper2.py \

> -mapper /home/hadoop/mapper2.py \

> -file /home/hadoop/reducer2.py \

> -reducer /home/hadoop/reducer2.py

19/02/17 20:01:47 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper2.py, /home/hadoop/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob3172341555153974436.jar tmpDir=null

19/02/17 20:01:49 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:01:49 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:01:49 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:01:49 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:01:49 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:01:49 INFO mapreduce.JobSubmitter: number of splits:21

19/02/17 20:01:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0008

19/02/17 20:01:50 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0008

19/02/17 20:01:50 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0008/

19/02/17 20:01:50 INFO mapreduce.Job: Running job: job\_1550430054052\_0008

19/02/17 20:01:57 INFO mapreduce.Job: Job job\_1550430054052\_0008 running in uber mode : false

19/02/17 20:01:57 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:02:12 INFO mapreduce.Job: map 19% reduce 0%

19/02/17 20:02:13 INFO mapreduce.Job: map 29% reduce 0%

19/02/17 20:02:18 INFO mapreduce.Job: map 38% reduce 0%

19/02/17 20:02:19 INFO mapreduce.Job: map 57% reduce 0%

19/02/17 20:02:20 INFO mapreduce.Job: map 67% reduce 0%

19/02/17 20:02:22 INFO mapreduce.Job: map 76% reduce 0%

19/02/17 20:02:23 INFO mapreduce.Job: map 81% reduce 0%

19/02/17 20:02:24 INFO mapreduce.Job: map 86% reduce 0%

19/02/17 20:02:31 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 20:02:32 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 20:02:33 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 20:02:37 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:02:38 INFO mapreduce.Job: Job job\_1550430054052\_0008 completed successfully

19/02/17 20:02:38 INFO mapreduce.Job: Counters: 52

File System Counters

FILE: Number of bytes read=93766

FILE: Number of bytes written=5050100

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=811238

HDFS: Number of bytes written=538400

HDFS: Number of read operations=84

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Killed reduce tasks=1

Launched map tasks=21

Launched reduce tasks=7

Data-local map tasks=20

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=13698990

Total time spent by all reduces in occupied slots (ms)=6327180

Total time spent by all map tasks (ms)=304422

Total time spent by all reduce tasks (ms)=70302

Total vcore-milliseconds taken by all map tasks=304422

Total vcore-milliseconds taken by all reduce tasks=70302

Total megabyte-milliseconds taken by all map tasks=438367680

Total megabyte-milliseconds taken by all reduce tasks=202469760

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=508400

Map output materialized bytes=106710

Input split bytes=2961

Combine input records=0

Combine output records=0

Reduce input groups=10

Reduce shuffle bytes=106710

Reduce input records=5000

Reduce output records=5000

Spilled Records=10000

Shuffled Maps =147

Failed Shuffles=0

Merged Map outputs=147

GC time elapsed (ms)=5782

CPU time spent (ms)=32340

Physical memory (bytes) snapshot=10564112384

Virtual memory (bytes) snapshot=98604789760

Total committed heap usage (bytes)=10020716544

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=808277

File Output Format Counters

Bytes Written=538400

19/02/17 20:02:38 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/poutput3 \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_A.py \

> -reducer /home/hadoop/reducer3\_A.py

19/02/17 20:03:26 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_A.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob6889569332813673050.jar tmpDir=null

19/02/17 20:03:27 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:03:27 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:03:28 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:03:28 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:03:28 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:03:28 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:03:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0009

19/02/17 20:03:28 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0009

19/02/17 20:03:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0009/

19/02/17 20:03:28 INFO mapreduce.Job: Running job: job\_1550430054052\_0009

19/02/17 20:03:35 INFO mapreduce.Job: Job job\_1550430054052\_0009 running in uber mode : false

19/02/17 20:03:35 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:03:52 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:03:54 INFO mapreduce.Job: map 35% reduce 0%

19/02/17 20:03:55 INFO mapreduce.Job: map 55% reduce 0%

19/02/17 20:03:56 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:04:05 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 20:04:06 INFO mapreduce.Job: map 90% reduce 0%

19/02/17 20:04:07 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:04:08 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 20:04:13 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 20:04:14 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:04:14 INFO mapreduce.Job: Job job\_1550430054052\_0009 completed successfully

19/02/17 20:04:14 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=71784

FILE: Number of bytes written=4872397

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=832376

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=20

Total time spent by all maps in occupied slots (ms)=13159665

Total time spent by all reduces in occupied slots (ms)=5920380

Total time spent by all map tasks (ms)=292437

Total time spent by all reduce tasks (ms)=65782

Total vcore-milliseconds taken by all map tasks=292437

Total vcore-milliseconds taken by all reduce tasks=65782

Total megabyte-milliseconds taken by all map tasks=421109280

Total megabyte-milliseconds taken by all reduce tasks=189452160

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=548400

Map output materialized bytes=123990

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=123990

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5605

CPU time spent (ms)=30150

Physical memory (bytes) snapshot=10051661824

Virtual memory (bytes) snapshot=95421960192

Total committed heap usage (bytes)=9484894208

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=829556

File Output Format Counters

Bytes Written=50840

19/02/17 20:04:14 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper4\_B.py

--2019-02-17 20:06:49-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper4\_B.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 384 [text/plain]

Saving to: ‘mapper4\_B.py’

mapper4\_B.py 100%[=====================================================>] 384 --.-KB/s in 0s

2019-02-17 20:06:49 (46.2 MB/s) - ‘mapper4\_B.py’ saved [384/384]

[hadoop@ip-172-31-7-172 ~]$ wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer3\_B.py

--2019-02-17 20:07:09-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/reducer3\_B.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1030 (1.0K) [text/plain]

Saving to: ‘reducer3\_B.py’

reducer3\_B.py 100%[=====================================================>] 1.01K --.-KB/s in 0s

2019-02-17 20:07:10 (163 MB/s) - ‘reducer3\_B.py’ saved [1030/1030]

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/output3\_test \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_B.py \

> -reducer /home/hadoop/reducer3\_B.py

19/02/17 20:08:38 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_B.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob1049952976066627478.jar tmpDir=null

19/02/17 20:08:39 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:08:39 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:08:40 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:08:40 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:08:40 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:08:40 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:08:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0010

19/02/17 20:08:40 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0010

19/02/17 20:08:40 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0010/

19/02/17 20:08:40 INFO mapreduce.Job: Running job: job\_1550430054052\_0010

19/02/17 20:08:48 INFO mapreduce.Job: Job job\_1550430054052\_0010 running in uber mode : false

19/02/17 20:08:48 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:09:03 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:09:08 INFO mapreduce.Job: map 50% reduce 0%

19/02/17 20:09:09 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:09:11 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 20:09:12 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 20:09:19 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:09:20 INFO mapreduce.Job: map 100% reduce 29%

19/02/17 20:09:21 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 20:09:25 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:09:25 INFO mapreduce.Job: Job job\_1550430054052\_0010 completed successfully

19/02/17 20:09:25 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=71793

FILE: Number of bytes written=4872525

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=832376

HDFS: Number of bytes written=55340

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=19

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12171420

Total time spent by all reduces in occupied slots (ms)=5409540

Total time spent by all map tasks (ms)=270476

Total time spent by all reduce tasks (ms)=60106

Total vcore-milliseconds taken by all map tasks=270476

Total vcore-milliseconds taken by all reduce tasks=60106

Total megabyte-milliseconds taken by all map tasks=389485440

Total megabyte-milliseconds taken by all reduce tasks=173105280

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=548400

Map output materialized bytes=123990

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=123990

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5174

CPU time spent (ms)=29670

Physical memory (bytes) snapshot=10155225088

Virtual memory (bytes) snapshot=95387787264

Total committed heap usage (bytes)=9715056640

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=829556

File Output Format Counters

Bytes Written=55340

19/02/17 20:09:25 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/output3\_test

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/output3\_test \

> -output /user/hadoop/wc/output4\_test \

> -file /home/hadoop/mapper4\_B.py \

> -mapper /home/hadoop/mapper4\_B.py

19/02/17 20:09:49 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper4\_B.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob2640495407442771758.jar tmpDir=null

19/02/17 20:09:50 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:09:51 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:09:51 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:09:51 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:09:51 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:09:51 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:09:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0011

19/02/17 20:09:52 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0011

19/02/17 20:09:52 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0011/

19/02/17 20:09:52 INFO mapreduce.Job: Running job: job\_1550430054052\_0011

19/02/17 20:09:59 INFO mapreduce.Job: Job job\_1550430054052\_0011 running in uber mode : false

19/02/17 20:09:59 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000011\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000009\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000003\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000004\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000002\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:13 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000010\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:17 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000006\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:17 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000005\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000007\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000001\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000008\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000017\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000015\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:18 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000000\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:21 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000011\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:25 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000009\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:28 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000003\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:30 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000004\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:30 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000017\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:31 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000015\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:33 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000011\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:33 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000000\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:34 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000010\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:35 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000007\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:35 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000006\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:35 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000002\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:35 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000001\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:35 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000005\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:36 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000008\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:38 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000009\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:41 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000003\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:42 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000004\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:42 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000017\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:44 INFO mapreduce.Job: Task Id : attempt\_1550430054052\_0011\_m\_000015\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 2

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:455)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:344)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:175)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1844)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:169)

19/02/17 20:10:47 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:10:48 INFO mapreduce.Job: Job job\_1550430054052\_0011 failed with state FAILED due to: Task failed task\_1550430054052\_0011\_m\_000011

Job failed as tasks failed. failedMaps:1 failedReduces:0

19/02/17 20:10:48 INFO mapreduce.Job: Counters: 17

Job Counters

Failed map tasks=35

Killed map tasks=19

Killed reduce tasks=7

Launched map tasks=48

Other local map tasks=33

Data-local map tasks=15

Total time spent by all maps in occupied slots (ms)=25317855

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=562619

Total time spent by all reduce tasks (ms)=0

Total vcore-milliseconds taken by all map tasks=562619

Total vcore-milliseconds taken by all reduce tasks=0

Total megabyte-milliseconds taken by all map tasks=810171360

Total megabyte-milliseconds taken by all reduce tasks=0

Map-Reduce Framework

CPU time spent (ms)=0

Physical memory (bytes) snapshot=0

Virtual memory (bytes) snapshot=0

19/02/17 20:10:48 ERROR streaming.StreamJob: Job not successful!

Streaming Command Failed!

[hadoop@ip-172-31-7-172 ~]$ cat mapper4\_B.py

import sys

import os

from math import log10,sqrt

D=10.0

# input comes from STDIN (standard input)

for line in sys.stdin:

# remove leading and trailing whitespace

line = line.strip()

# split the line into words

wf,nNm=line.split('\t',1)

n,N,m=nNm.split(' ',2)

n=float(n)

N=float(N)

m=float(m)

tfidf= (n/N)\*log10(D/m)

print '%s\t%s' % (wf,tfidf)[hadoop@ip-172-31-7-172 ~]$

[hadoop@ip-172-31-7-172 ~]$ cat mapper3.py

#!/usr/bin/env python

import sys

import os

#input: <<word#documentName>, <n#N>>

# input comes from STDIN (standard input)

for line in sys.stdin:

# remove leading and trailing whitespace

line = line.strip()

# split the line into words

wf,nN=line.split('\t',1)

w,f=wf.split(' ',1)

z=f+' '+nN+' '+str(1)

print '%s\t%s' % (w,z)[hadoop@ip-172-31-7-172 ~]$

[hadoop@ip-172-31-7-172 ~]$ cat mapper4.py

cat: mapper4.py: No such file or directory

[hadoop@ip-172-31-7-172 ~]$ cat mapper4\_B.py

import sys

import os

from math import log10,sqrt

D=10.0

# input comes from STDIN (standard input)

for line in sys.stdin:

# remove leading and trailing whitespace

line = line.strip()

# split the line into words

wf,nNm=line.split('\t',1)

n,N,m=nNm.split(' ',2)

n=float(n)

N=float(N)

m=float(m)

tfidf= (n/N)\*log10(D/m)

print '%s\t%s' % (wf,tfidf)[hadoop@ip-172-31-7-172 ~]$

[hadoop@ip-172-31-7-172 ~]$

[hadoop@ip-172-31-7-172 ~]$

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer1.py reducer3\_A.py text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

mapper2.py mapper4\_B.py reducer2.py reducer3\_B.py text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ rm mapper4\_B.py

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer2.py reducer3\_B.py text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

mapper2.py reducer1.py reducer3\_A.py text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

[hadoop@ip-172-31-7-172 ~]$ wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper4\_B.py

--2019-02-17 20:13:47-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/mapper4\_B.py

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 407 [text/plain]

Saving to: ‘mapper4\_B.py’

mapper4\_B.py 100%[=====================================================>] 407 --.-KB/s in 0s

2019-02-17 20:13:47 (50.1 MB/s) - ‘mapper4\_B.py’ saved [407/407]

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -input /user/hadoop/wc/output3\_test -output /user/hadoop/wc/output4\_test -file /home/hadoop/mapper4\_B.py -mapper /home/hadoop/mapper4\_B.py

19/02/17 20:13:59 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper4\_B.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob6297526782450884525.jar tmpDir=null

19/02/17 20:14:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:14:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:14:01 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/output4\_test already exists

Streaming Command Failed!

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm R /user/hadoop/wc/output4\_test

rm: `R': No such file or directory

rm: `/user/hadoop/wc/output4\_test': Is a directory

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -rm -R /user/hadoop/wc/output4\_test

Deleted /user/hadoop/wc/output4\_test

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -input /user/hadoop/wc/output3\_test -output /user/hadoop/wc/output4\_test -file /home/hadoop/mapper4\_B.py -mapper /home/hadoop/mapper4\_B.py

19/02/17 20:14:43 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper4\_B.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob7770085791289034098.jar tmpDir=null

19/02/17 20:14:45 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:14:45 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:14:45 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:14:45 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:14:45 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:14:45 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:14:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0012

19/02/17 20:14:46 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0012

19/02/17 20:14:46 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0012/

19/02/17 20:14:46 INFO mapreduce.Job: Running job: job\_1550430054052\_0012

19/02/17 20:14:53 INFO mapreduce.Job: Job job\_1550430054052\_0012 running in uber mode : false

19/02/17 20:14:53 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:15:08 INFO mapreduce.Job: map 20% reduce 0%

19/02/17 20:15:09 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:15:12 INFO mapreduce.Job: map 55% reduce 0%

19/02/17 20:15:13 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:15:18 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 20:15:19 INFO mapreduce.Job: map 85% reduce 0%

19/02/17 20:15:20 INFO mapreduce.Job: map 90% reduce 0%

19/02/17 20:15:23 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:15:24 INFO mapreduce.Job: map 100% reduce 43%

19/02/17 20:15:25 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 20:15:28 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:15:28 INFO mapreduce.Job: Job job\_1550430054052\_0012 completed successfully

19/02/17 20:15:28 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=10570

FILE: Number of bytes written=4685200

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=96358

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=19

Rack-local map tasks=1

Total time spent by all maps in occupied slots (ms)=12374910

Total time spent by all reduces in occupied slots (ms)=4928940

Total time spent by all map tasks (ms)=274998

Total time spent by all reduce tasks (ms)=54766

Total vcore-milliseconds taken by all map tasks=274998

Total vcore-milliseconds taken by all reduce tasks=54766

Total megabyte-milliseconds taken by all map tasks=395997120

Total megabyte-milliseconds taken by all reduce tasks=157726080

Map-Reduce Framework

Map input records=500

Map output records=500

Map output bytes=50840

Map output materialized bytes=24519

Input split bytes=2900

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=24519

Reduce input records=500

Reduce output records=500

Spilled Records=1000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5272

CPU time spent (ms)=27870

Physical memory (bytes) snapshot=10146099200

Virtual memory (bytes) snapshot=95418732544

Total committed heap usage (bytes)=9586081792

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=93458

File Output Format Counters

Bytes Written=50840

19/02/17 20:15:28 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/output4\_test

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/output4\_test

Found 8 items

-rw-r--r-- 1 hadoop hadoop 0 2019-02-17 20:15 /user/hadoop/wc/output4\_test/\_SUCCESS

-rw-r--r-- 1 hadoop hadoop 7744 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00000

-rw-r--r-- 1 hadoop hadoop 7155 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00001

-rw-r--r-- 1 hadoop hadoop 7276 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00002

-rw-r--r-- 1 hadoop hadoop 6924 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00003

-rw-r--r-- 1 hadoop hadoop 6691 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00004

-rw-r--r-- 1 hadoop hadoop 8754 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00005

-rw-r--r-- 1 hadoop hadoop 6296 2019-02-17 20:15 /user/hadoop/wc/output4\_test/part-00006

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -cat /user/hadoop/wc/output4\_test/part-00000

aboveproof hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

agaricaceae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

alerse hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

alidades hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

autofluorescence hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

belite hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

blepharemphysema hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

cangia hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

chamois hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

characteristical hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

chronica hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

completory hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

crassier hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

dauncy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

demersed hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

dissceptre hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

egeran hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

eighteenthly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

embay hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

encave hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

eurytropic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

eustele hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

ferric hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

fingerboards hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

footcandles hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

footslog hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

forwearying hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

fowlings hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

fumigates hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

haplopetalous hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

hedriophthalmous hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

historism hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

hylism hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

hyperidealistically hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

immarginate hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

inextirpable hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

leglessness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

mack hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

mandragvn hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

microdetector hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

misbiasing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

monsignorial hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

myotacismus hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

nebulously hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

nederlands hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

ngai hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

noninterrupted hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

ornithopod hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

overwillingness hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

physalis hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

poristic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

preeing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

recklessly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

rectifiers hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

repurified hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

riparian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

rotogravure hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

scalelet hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

semimercerized hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

seroprognosis hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

sinopite hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

skyscraping hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

stonesmich hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

stormiest hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

supersulfureted hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

timbermonger hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

transaminase hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

tripodial hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

unconscionable hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

ungalloping hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

vage hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

vermilingues hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

vernition hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

wickerworked hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

xenoplastic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

yare hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -cat /user/hadoop/wc/output4\_test/part-00006

adjourns hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

antianarchic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

aponeurotome hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

aquatic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

archduke hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

arthrolith hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

bedighting hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

brayers hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

bundlings hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

callisto hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

cavitates hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

chorial hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

cloyment hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

compassionately hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

coniacian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

cornily hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

counselorship hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

counterpoising hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

cudweed hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

cyanurin hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

cylindroid hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

dean hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

decomponent hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

detractory hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

devotee hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

electing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

episiotomy hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

excursionize hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

firewater hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

forestalls hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

geraniol hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

greatheartedly hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

hormogoneae hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

jumbler hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

lansing hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

leptomeningeal hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_3.txt 0.0

lorgnon hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

luminophor hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

maturant hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

memorializer hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

monactinellid hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

monosulphide hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

monroe hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

naphthionate hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

oestriols hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_0.txt 0.0

orthodomatic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

phlorone hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_1.txt 0.0

premillenarian hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

ramuscule hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

rondelle hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_2.txt 0.0

schrank hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_8.txt 0.0

scolopophore hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_5.txt 0.0

sissone hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

subthalamic hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

successionally hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_6.txt 0.0

thimbles hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

tuberculomas hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

tutory hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_4.txt 0.0

undisproved hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

unfixity hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_9.txt 0.0

verray hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

woodenware hdfs://ip-172-31-7-172.eu-west-1.compute.internal:8020/user/hadoop/wc/input/text\_7.txt 0.0

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer1.py reducer3\_A.py text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

mapper2.py mapper4\_B.py reducer2.py reducer3\_B.py text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> rm text\_$i.txt

> done

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -rm /user/hadoop/wc/input/text\_$i.txt

> done

Deleted /user/hadoop/wc/input/text\_0.txt

Deleted /user/hadoop/wc/input/text\_1.txt

Deleted /user/hadoop/wc/input/text\_2.txt

Deleted /user/hadoop/wc/input/text\_3.txt

Deleted /user/hadoop/wc/input/text\_4.txt

Deleted /user/hadoop/wc/input/text\_5.txt

Deleted /user/hadoop/wc/input/text\_6.txt

for i in {1..3}; do

hdfs dfs -rm -R /user/hadoop/wc/poutput$i

done

Deleted /user/hadoop/wc/input/text\_7.txt

Deleted /user/hadoop/wc/input/text\_8.txt

Deleted /user/hadoop/wc/input/text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {1..3}; do

> hdfs dfs -rm -R /user/hadoop/wc/poutput$i

> done

Deleted /user/hadoop/wc/poutput1

Deleted /user/hadoop/wc/poutput2

for i in {0..9}; do

wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_$i.txt

done

Deleted /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_$i.txt

> done

--2019-02-17 20:27:14-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_0.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853617 (834K) [text/plain]

Saving to: ‘text\_0.txt’

text\_0.txt 100%[=====================================================>] 833.61K --.-KB/s in 0.07s

2019-02-17 20:27:14 (12.2 MB/s) - ‘text\_0.txt’ saved [853617/853617]

--2019-02-17 20:27:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_1.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853609 (834K) [text/plain]

Saving to: ‘text\_1.txt’

text\_1.txt 100%[=====================================================>] 833.60K --.-KB/s in 0.07s

2019-02-17 20:27:15 (12.5 MB/s) - ‘text\_1.txt’ saved [853609/853609]

--2019-02-17 20:27:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_2.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854680 (835K) [text/plain]

Saving to: ‘text\_2.txt’

text\_2.txt 100%[=====================================================>] 834.65K --.-KB/s in 0.07s

2019-02-17 20:27:15 (11.1 MB/s) - ‘text\_2.txt’ saved [854680/854680]

--2019-02-17 20:27:15-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_3.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854079 (834K) [text/plain]

Saving to: ‘text\_3.txt’

text\_3.txt 100%[=====================================================>] 834.06K --.-KB/s in 0.06s

2019-02-17 20:27:16 (13.5 MB/s) - ‘text\_3.txt’ saved [854079/854079]

--2019-02-17 20:27:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_4.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854117 (834K) [text/plain]

Saving to: ‘text\_4.txt’

text\_4.txt 100%[=====================================================>] 834.10K --.-KB/s in 0.07s

2019-02-17 20:27:16 (12.5 MB/s) - ‘text\_4.txt’ saved [854117/854117]

--2019-02-17 20:27:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_5.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853866 (834K) [text/plain]

Saving to: ‘text\_5.txt’

text\_5.txt 100%[=====================================================>] 833.85K --.-KB/s in 0.06s

2019-02-17 20:27:16 (12.8 MB/s) - ‘text\_5.txt’ saved [853866/853866]

--2019-02-17 20:27:16-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_6.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 855044 (835K) [text/plain]

Saving to: ‘text\_6.txt’

text\_6.txt 100%[=====================================================>] 835.00K --.-KB/s in 0.06s

2019-02-17 20:27:17 (14.1 MB/s) - ‘text\_6.txt’ saved [855044/855044]

--2019-02-17 20:27:17-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_7.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854250 (834K) [text/plain]

Saving to: ‘text\_7.txt’

text\_7.txt 100%[=====================================================>] 834.23K --.-KB/s in 0.06s

2019-02-17 20:27:17 (14.1 MB/s) - ‘text\_7.txt’ saved [854250/854250]

--2019-02-17 20:27:17-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_8.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 854017 (834K) [text/plain]

Saving to: ‘text\_8.txt’

text\_8.txt 100%[=====================================================>] 834.00K --.-KB/s in 0.07s

2019-02-17 20:27:17 (12.5 MB/s) - ‘text\_8.txt’ saved [854017/854017]

--2019-02-17 20:27:17-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/3/text\_9.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.16.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.16.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 853794 (834K) [text/plain]

Saving to: ‘text\_9.txt’

text\_9.txt 100%[=====================================================>] 833.78K --.-KB/s in 0.06s

2019-02-17 20:27:18 (13.5 MB/s) - ‘text\_9.txt’ saved [853794/853794]

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer1.py reducer3\_A.py text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

mapper2.py mapper4\_B.py reducer2.py reducer3\_B.py text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

-input /user/hadoop/wc/input/\*.txt \

-output /user/hadoop/wc/poutput1 \

-file /home/hadoop/mapper1.py \

-mapper /home/hadoop/mapper1.py \

-file /home/hadoop/reducer1.py \

-reducer /home/hadoop/reducer1.py

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/input/\*.txt \

> -output /user/hadoop/wc/poutput1 \

> -file /home/hadoop/mapper1.py \

> -mapper /home/hadoop/mapper1.py \

> -file /home/hadoop/reducer1.py \

> -reducer /home/hadoop/reducer1.py

19/02/17 20:27:56 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper1.py, /home/hadoop/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob8516546166157398675.jar tmpDir=null

19/02/17 20:27:57 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:27:58 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:27:58 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:27:58 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:27:58 INFO mapred.FileInputFormat: Total input files to process : 10

19/02/17 20:27:58 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:27:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0013

19/02/17 20:27:59 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0013

19/02/17 20:27:59 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0013/

19/02/17 20:27:59 INFO mapreduce.Job: Running job: job\_1550430054052\_0013

19/02/17 20:28:06 INFO mapreduce.Job: Job job\_1550430054052\_0013 running in uber mode : false

19/02/17 20:28:06 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:28:26 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:28:28 INFO mapreduce.Job: map 40% reduce 0%

19/02/17 20:28:29 INFO mapreduce.Job: map 45% reduce 0%

19/02/17 20:28:30 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:28:42 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 20:28:43 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:28:46 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 20:28:47 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 20:28:52 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:28:53 INFO mapreduce.Job: Job job\_1550430054052\_0013 completed successfully

19/02/17 20:28:53 INFO mapreduce.Job: Counters: 50

File System Counters

FILE: Number of bytes read=4157841

FILE: Number of bytes written=13073100

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=9103903

HDFS: Number of bytes written=508400

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=2

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=20

Total time spent by all maps in occupied slots (ms)=16517025

Total time spent by all reduces in occupied slots (ms)=7998120

Total time spent by all map tasks (ms)=367045

Total time spent by all reduce tasks (ms)=88868

Total vcore-milliseconds taken by all map tasks=367045

Total vcore-milliseconds taken by all reduce tasks=88868

Total megabyte-milliseconds taken by all map tasks=528544800

Total megabyte-milliseconds taken by all reduce tasks=255939840

Map-Reduce Framework

Map input records=80000

Map output records=800000

Map output bytes=79741073

Map output materialized bytes=4238754

Input split bytes=2760

Combine input records=0

Combine output records=0

Reduce input groups=5000

Reduce shuffle bytes=4238754

Reduce input records=800000

Reduce output records=5000

Spilled Records=1600000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5500

CPU time spent (ms)=76300

Physical memory (bytes) snapshot=10729553920

Virtual memory (bytes) snapshot=95611179008

Total committed heap usage (bytes)=9629597696

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=9101143

File Output Format Counters

Bytes Written=508400

19/02/17 20:28:53 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput1 \

> -output /user/hadoop/wc/poutput2 \

> -file /home/hadoop/mapper2.py \

> -mapper /home/hadoop/mapper2.py \

> -file /home/hadoop/reducer2.py \

> -reducer /home/hadoop/reducer2.py

19/02/17 20:29:01 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper2.py, /home/hadoop/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob3552656363670839068.jar tmpDir=null

19/02/17 20:29:03 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:29:03 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:29:03 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:29:03 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:29:03 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:29:03 INFO mapreduce.JobSubmitter: number of splits:21

19/02/17 20:29:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0014

19/02/17 20:29:04 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0014

19/02/17 20:29:04 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0014/

19/02/17 20:29:04 INFO mapreduce.Job: Running job: job\_1550430054052\_0014

19/02/17 20:29:11 INFO mapreduce.Job: Job job\_1550430054052\_0014 running in uber mode : false

19/02/17 20:29:11 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:29:27 INFO mapreduce.Job: map 29% reduce 0%

19/02/17 20:29:30 INFO mapreduce.Job: map 52% reduce 0%

19/02/17 20:29:31 INFO mapreduce.Job: map 67% reduce 0%

19/02/17 20:29:35 INFO mapreduce.Job: map 71% reduce 0%

19/02/17 20:29:36 INFO mapreduce.Job: map 81% reduce 0%

19/02/17 20:29:43 INFO mapreduce.Job: map 86% reduce 0%

19/02/17 20:29:44 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 20:29:45 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 20:29:50 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:29:50 INFO mapreduce.Job: Job job\_1550430054052\_0014 completed successfully

19/02/17 20:29:50 INFO mapreduce.Job: Counters: 50

File System Counters

FILE: Number of bytes read=93770

FILE: Number of bytes written=5050086

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=811238

HDFS: Number of bytes written=538400

HDFS: Number of read operations=84

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=21

Launched reduce tasks=7

Data-local map tasks=21

Total time spent by all maps in occupied slots (ms)=12905865

Total time spent by all reduces in occupied slots (ms)=6066810

Total time spent by all map tasks (ms)=286797

Total time spent by all reduce tasks (ms)=67409

Total vcore-milliseconds taken by all map tasks=286797

Total vcore-milliseconds taken by all reduce tasks=67409

Total megabyte-milliseconds taken by all map tasks=412987680

Total megabyte-milliseconds taken by all reduce tasks=194137920

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=508400

Map output materialized bytes=106710

Input split bytes=2961

Combine input records=0

Combine output records=0

Reduce input groups=10

Reduce shuffle bytes=106710

Reduce input records=5000

Reduce output records=5000

Spilled Records=10000

Shuffled Maps =147

Failed Shuffles=0

Merged Map outputs=147

GC time elapsed (ms)=5698

CPU time spent (ms)=32260

Physical memory (bytes) snapshot=10576797696

Virtual memory (bytes) snapshot=98559676416

Total committed heap usage (bytes)=10181672960

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=808277

File Output Format Counters

Bytes Written=538400

19/02/17 20:29:50 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/poutput3 \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_A.py \

> -reducer /home/hadoop/reducer3\_A.py

19/02/17 20:30:03 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_A.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob1181591934327296568.jar tmpDir=null

19/02/17 20:30:05 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:30:05 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:30:05 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:30:05 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:30:05 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:30:05 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:30:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0015

19/02/17 20:30:06 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0015

19/02/17 20:30:06 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0015/

19/02/17 20:30:06 INFO mapreduce.Job: Running job: job\_1550430054052\_0015

19/02/17 20:30:13 INFO mapreduce.Job: Job job\_1550430054052\_0015 running in uber mode : false

19/02/17 20:30:13 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:30:28 INFO mapreduce.Job: map 25% reduce 0%

19/02/17 20:30:29 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:30:33 INFO mapreduce.Job: map 50% reduce 0%

19/02/17 20:30:34 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:30:41 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 20:30:42 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 20:30:43 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:30:44 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 20:30:49 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 20:30:50 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:30:50 INFO mapreduce.Job: Job job\_1550430054052\_0015 completed successfully

19/02/17 20:30:50 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=71699

FILE: Number of bytes written=4872324

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=832376

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Killed reduce tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=20

Total time spent by all maps in occupied slots (ms)=13090770

Total time spent by all reduces in occupied slots (ms)=5073120

Total time spent by all map tasks (ms)=290906

Total time spent by all reduce tasks (ms)=56368

Total vcore-milliseconds taken by all map tasks=290906

Total vcore-milliseconds taken by all reduce tasks=56368

Total megabyte-milliseconds taken by all map tasks=418904640

Total megabyte-milliseconds taken by all reduce tasks=162339840

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=548400

Map output materialized bytes=123989

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=123989

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5452

CPU time spent (ms)=30980

Physical memory (bytes) snapshot=10097049600

Virtual memory (bytes) snapshot=95379099648

Total committed heap usage (bytes)=9480699904

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=829556

File Output Format Counters

Bytes Written=50840

19/02/17 20:30:50 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> rm text\_$i.txt

> done

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -rm /user/hadoop/wc/input/text\_$i.txt

> done

Deleted /user/hadoop/wc/input/text\_0.txt

Deleted /user/hadoop/wc/input/text\_1.txt

for i in {1..3}; do

hdfs dfs -rm -R /user/hadoop/wc/poutput$i

doneDeleted /user/hadoop/wc/input/text\_2.txt

Deleted /user/hadoop/wc/input/text\_3.txt

Deleted /user/hadoop/wc/input/text\_4.txt

Deleted /user/hadoop/wc/input/text\_5.txt

Deleted /user/hadoop/wc/input/text\_6.txt

Deleted /user/hadoop/wc/input/text\_7.txt

Deleted /user/hadoop/wc/input/text\_8.txt

Deleted /user/hadoop/wc/input/text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {1..3}; do

> hdfs dfs -rm -R /user/hadoop/wc/poutput$i

> done

Deleted /user/hadoop/wc/poutput1

Deleted /user/hadoop/wc/poutput2

for i in {0..9}; do

wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_$i.txt

done

Deleted /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> wget https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_$i.txt

> done

--2019-02-17 20:34:43-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_0.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1710050 (1.6M) [text/plain]

Saving to: ‘text\_0.txt’

text\_0.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:43 (24.7 MB/s) - ‘text\_0.txt’ saved [1710050/1710050]

--2019-02-17 20:34:43-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_1.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1707941 (1.6M) [text/plain]

Saving to: ‘text\_1.txt’

text\_1.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.06s

2019-02-17 20:34:44 (25.4 MB/s) - ‘text\_1.txt’ saved [1707941/1707941]

--2019-02-17 20:34:44-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_2.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1710345 (1.6M) [text/plain]

Saving to: ‘text\_2.txt’

text\_2.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:44 (24.3 MB/s) - ‘text\_2.txt’ saved [1710345/1710345]

--2019-02-17 20:34:44-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_3.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1708676 (1.6M) [text/plain]

Saving to: ‘text\_3.txt’

text\_3.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:45 (22.9 MB/s) - ‘text\_3.txt’ saved [1708676/1708676]

--2019-02-17 20:34:45-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_4.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1708186 (1.6M) [text/plain]

Saving to: ‘text\_4.txt’

text\_4.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:45 (24.9 MB/s) - ‘text\_4.txt’ saved [1708186/1708186]

--2019-02-17 20:34:45-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_5.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1710104 (1.6M) [text/plain]

Saving to: ‘text\_5.txt’

text\_5.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.06s

2019-02-17 20:34:45 (25.1 MB/s) - ‘text\_5.txt’ saved [1710104/1710104]

--2019-02-17 20:34:45-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_6.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1707914 (1.6M) [text/plain]

Saving to: ‘text\_6.txt’

text\_6.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.08s

2019-02-17 20:34:46 (19.8 MB/s) - ‘text\_6.txt’ saved [1707914/1707914]

--2019-02-17 20:34:46-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_7.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1710126 (1.6M) [text/plain]

Saving to: ‘text\_7.txt’

text\_7.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:46 (24.5 MB/s) - ‘text\_7.txt’ saved [1710126/1710126]

--2019-02-17 20:34:46-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_8.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1710575 (1.6M) [text/plain]

Saving to: ‘text\_8.txt’

text\_8.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.07s

2019-02-17 20:34:47 (24.2 MB/s) - ‘text\_8.txt’ saved [1710575/1710575]

--2019-02-17 20:34:47-- https://raw.githubusercontent.com/lo1gr/tf-idf\_hadoop-Spark/master/documents/4/text\_9.txt

Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.60.133

Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.60.133|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 1709629 (1.6M) [text/plain]

Saving to: ‘text\_9.txt’

text\_9.txt 100%[=====================================================>] 1.63M --.-KB/s in 0.08s

2019-02-17 20:34:47 (19.5 MB/s) - ‘text\_9.txt’ saved [1709629/1709629]

[hadoop@ip-172-31-7-172 ~]$ ls

mapper1.py mapper3.py reducer1.py reducer3\_A.py text\_0.txt text\_2.txt text\_4.txt text\_6.txt text\_8.txt

mapper2.py mapper4\_B.py reducer2.py reducer3\_B.py text\_1.txt text\_3.txt text\_5.txt text\_7.txt text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ for i in {0..9}; do

> hdfs dfs -put text\_$i.txt /user/hadoop/wc/input

> done

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/input/\*.txt \

> -output /user/hadoop/wc/poutput1 \

> -file /home/hadoop/mapper1.py \

> -mapper /home/hadoop/mapper1.py \

> -file /home/hadoop/reducer1.py \

> -reducer /home/hadoop/reducer1.py

19/02/17 20:35:25 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper1.py, /home/hadoop/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob7581691775242661927.jar tmpDir=null

19/02/17 20:35:27 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:35:27 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:35:27 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:35:27 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:35:27 INFO mapred.FileInputFormat: Total input files to process : 10

19/02/17 20:35:27 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:35:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0016

19/02/17 20:35:28 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0016

19/02/17 20:35:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0016/

19/02/17 20:35:28 INFO mapreduce.Job: Running job: job\_1550430054052\_0016

19/02/17 20:35:35 INFO mapreduce.Job: Job job\_1550430054052\_0016 running in uber mode : false

19/02/17 20:35:35 INFO mapreduce.Job: map 0% reduce 0%

hedfs dfs -ls /user/hadoop/wc/input

hdfs dfs -ls/us19/02/17 20:35:54 INFO mapreduce.Job: map 5% reduce 0%

er19/02/17 20:35:55 INFO mapreduce.Job: map 15% reduce 0%

/19/02/17 20:35:56 INFO mapreduce.Job: map 30% reduce 0%

hadoop/inp19/02/17 20:36:01 INFO mapreduce.Job: map 35% reduce 0%

19/02/17 20:36:03 INFO mapreduce.Job: map 45% reduce 0%

19/02/17 20:36:04 INFO mapreduce.Job: map 75% reduce 0%

19/02/17 20:36:07 INFO mapreduce.Job: map 80% reduce 0%

19/02/17 20:36:15 INFO mapreduce.Job: map 80% reduce 4%

19/02/17 20:36:16 INFO mapreduce.Job: map 80% reduce 8%

19/02/17 20:36:17 INFO mapreduce.Job: map 85% reduce 8%

19/02/17 20:36:19 INFO mapreduce.Job: map 100% reduce 8%

19/02/17 20:36:21 INFO mapreduce.Job: map 100% reduce 30%

19/02/17 20:36:22 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 20:36:27 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:36:27 INFO mapreduce.Job: Job job\_1550430054052\_0016 completed successfully

19/02/17 20:36:27 INFO mapreduce.Job: Counters: 51

File System Counters

FILE: Number of bytes read=8240004

FILE: Number of bytes written=21237509

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=17553966

HDFS: Number of bytes written=508400

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=18

Rack-local map tasks=2

Total time spent by all maps in occupied slots (ms)=17141985

Total time spent by all reduces in occupied slots (ms)=9751950

Total time spent by all map tasks (ms)=380933

Total time spent by all reduce tasks (ms)=108355

Total vcore-milliseconds taken by all map tasks=380933

Total vcore-milliseconds taken by all reduce tasks=108355

Total megabyte-milliseconds taken by all map tasks=548543520

Total megabyte-milliseconds taken by all reduce tasks=312062400

Map-Reduce Framework

Map input records=160000

Map output records=1600000

Map output bytes=159493546

Map output materialized bytes=8321034

Input split bytes=2760

Combine input records=0

Combine output records=0

Reduce input groups=5000

Reduce shuffle bytes=8321034

Reduce input records=1600000

Reduce output records=5000

Spilled Records=3200000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5132

CPU time spent (ms)=83190

Physical memory (bytes) snapshot=10593902592

Virtual memory (bytes) snapshot=95554584576

Total committed heap usage (bytes)=9609150464

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=17551206

File Output Format Counters

Bytes Written=508400

19/02/17 20:36:27 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput1

[hadoop@ip-172-31-7-172 ~]$ hedfs dfs -ls /user/hadoop/wc/input

-bash: hedfs: command not found

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls/user/hadoop/input

-ls/user/hadoop/input: Unknown command

Usage: hadoop fs [generic options]

[-appendToFile <localsrc> ... <dst>]

[-cat [-ignoreCrc] <src> ...]

[-checksum <src> ...]

[-chgrp [-R] GROUP PATH...]

[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]

[-chown [-R] [OWNER][:[GROUP]] PATH...]

[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]

[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]

[-createSnapshot <snapshotDir> [<snapshotName>]]

[-deleteSnapshot <snapshotDir> <snapshotName>]

[-df [-h] [<path> ...]]

[-du [-s] [-h] [-x] <path> ...]

[-expunge]

[-find <path> ... <expression> ...]

[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-getfacl [-R] <path>]

[-getfattr [-R] {-n name | -d} [-e en] <path>]

[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]

[-help [cmd ...]]

[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]

[-mkdir [-p] <path> ...]

[-moveFromLocal <localsrc> ... <dst>]

[-moveToLocal <src> <localdst>]

[-mv <src> ... <dst>]

[-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-renameSnapshot <snapshotDir> <oldName> <newName>]

[-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]

[-rmdir [--ignore-fail-on-non-empty] <dir> ...]

[-setfacl [-R] [{-b|-k} {-m|-x <acl\_spec>} <path>]|[--set <acl\_spec> <path>]]

[-setfattr {-n name [-v value] | -x name} <path>]

[-setrep [-R] [-w] <rep> <path> ...]

[-stat [format] <path> ...]

[-tail [-f] <file>]

[-test -[defsz] <path>]

[-text [-ignoreCrc] <src> ...]

[-touchz <path> ...]

[-truncate [-w] <length> <path> ...]

[-usage [cmd ...]]

Generic options supported are

-conf <configuration file> specify an application configuration file

-D <property=value> use value for given property

-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.

-jt <local|resourcemanager:port> specify a ResourceManager

-files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster

-libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.

-archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is

command [genericOptions] [commandOptions]

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/input

ls: `/user/hadoop/input': No such file or directory

[hadoop@ip-172-31-7-172 ~]$ hdfs dfs -ls /user/hadoop/wc/input

Found 10 items

-rw-r--r-- 1 hadoop hadoop 1710050 2019-02-17 20:34 /user/hadoop/wc/input/text\_0.txt

-rw-r--r-- 1 hadoop hadoop 1707941 2019-02-17 20:34 /user/hadoop/wc/input/text\_1.txt

-rw-r--r-- 1 hadoop hadoop 1710345 2019-02-17 20:35 /user/hadoop/wc/input/text\_2.txt

-rw-r--r-- 1 hadoop hadoop 1708676 2019-02-17 20:35 /user/hadoop/wc/input/text\_3.txt

-rw-r--r-- 1 hadoop hadoop 1708186 2019-02-17 20:35 /user/hadoop/wc/input/text\_4.txt

-rw-r--r-- 1 hadoop hadoop 1710104 2019-02-17 20:35 /user/hadoop/wc/input/text\_5.txt

-rw-r--r-- 1 hadoop hadoop 1707914 2019-02-17 20:35 /user/hadoop/wc/input/text\_6.txt

-rw-r--r-- 1 hadoop hadoop 1710126 2019-02-17 20:35 /user/hadoop/wc/input/text\_7.txt

-rw-r--r-- 1 hadoop hadoop 1710575 2019-02-17 20:35 /user/hadoop/wc/input/text\_8.txt

-rw-r--r-- 1 hadoop hadoop 1709629 2019-02-17 20:35 /user/hadoop/wc/input/text\_9.txt

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput1 \

> -output /user/hadoop/wc/poutput2 \

> -file /home/hadoop/mapper2.py \

> -mapper /home/hadoop/mapper2.py \

> -file /home/hadoop/reducer2.py \

> -reducer /home/hadoop/reducer2.py

19/02/17 20:37:33 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper2.py, /home/hadoop/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob1362325041808547028.jar tmpDir=null

19/02/17 20:37:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:37:34 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:37:35 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:37:35 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:37:35 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:37:35 INFO mapreduce.JobSubmitter: number of splits:21

19/02/17 20:37:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0017

19/02/17 20:37:35 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0017

19/02/17 20:37:35 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0017/

19/02/17 20:37:35 INFO mapreduce.Job: Running job: job\_1550430054052\_0017

19/02/17 20:37:42 INFO mapreduce.Job: Job job\_1550430054052\_0017 running in uber mode : false

19/02/17 20:37:42 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:37:58 INFO mapreduce.Job: map 10% reduce 0%

19/02/17 20:37:59 INFO mapreduce.Job: map 29% reduce 0%

19/02/17 20:38:01 INFO mapreduce.Job: map 33% reduce 0%

19/02/17 20:38:02 INFO mapreduce.Job: map 43% reduce 0%

19/02/17 20:38:03 INFO mapreduce.Job: map 67% reduce 0%

19/02/17 20:38:11 INFO mapreduce.Job: map 76% reduce 0%

19/02/17 20:38:12 INFO mapreduce.Job: map 86% reduce 0%

19/02/17 20:38:13 INFO mapreduce.Job: map 100% reduce 14%

19/02/17 20:38:14 INFO mapreduce.Job: map 100% reduce 29%

19/02/17 20:38:15 INFO mapreduce.Job: map 100% reduce 43%

19/02/17 20:38:17 INFO mapreduce.Job: map 100% reduce 57%

19/02/17 20:38:18 INFO mapreduce.Job: map 100% reduce 71%

19/02/17 20:38:20 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:38:21 INFO mapreduce.Job: Job job\_1550430054052\_0017 completed successfully

19/02/17 20:38:21 INFO mapreduce.Job: Counters: 50

File System Counters

FILE: Number of bytes read=94807

FILE: Number of bytes written=5051984

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=811238

HDFS: Number of bytes written=543400

HDFS: Number of read operations=84

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=21

Launched reduce tasks=7

Data-local map tasks=21

Total time spent by all maps in occupied slots (ms)=13352850

Total time spent by all reduces in occupied slots (ms)=4728690

Total time spent by all map tasks (ms)=296730

Total time spent by all reduce tasks (ms)=52541

Total vcore-milliseconds taken by all map tasks=296730

Total vcore-milliseconds taken by all reduce tasks=52541

Total megabyte-milliseconds taken by all map tasks=427291200

Total megabyte-milliseconds taken by all reduce tasks=151318080

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=508400

Map output materialized bytes=107580

Input split bytes=2961

Combine input records=0

Combine output records=0

Reduce input groups=10

Reduce shuffle bytes=107580

Reduce input records=5000

Reduce output records=5000

Spilled Records=10000

Shuffled Maps =147

Failed Shuffles=0

Merged Map outputs=147

GC time elapsed (ms)=5287

CPU time spent (ms)=31590

Physical memory (bytes) snapshot=10538676224

Virtual memory (bytes) snapshot=98588942336

Total committed heap usage (bytes)=9906946048

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=808277

File Output Format Counters

Bytes Written=543400

19/02/17 20:38:21 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput2

[hadoop@ip-172-31-7-172 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

> -input /user/hadoop/wc/poutput2 \

> -output /user/hadoop/wc/poutput3 \

> -file /home/hadoop/mapper3.py \

> -mapper /home/hadoop/mapper3.py \

> -file /home/hadoop/reducer3\_A.py \

> -reducer /home/hadoop/reducer3\_A.py

19/02/17 20:39:18 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.

packageJobJar: [/home/hadoop/mapper3.py, /home/hadoop/reducer3\_A.py] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-1.jar] /tmp/streamjob2939794222121036476.jar tmpDir=null

19/02/17 20:39:19 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:39:20 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-172.eu-west-1.compute.internal/172.31.7.172:8032

19/02/17 20:39:20 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library

19/02/17 20:39:20 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 1546b8dc0ca6f1ffd26a812d52bd7b80915e0a25]

19/02/17 20:39:20 INFO mapred.FileInputFormat: Total input files to process : 7

19/02/17 20:39:20 INFO mapreduce.JobSubmitter: number of splits:20

19/02/17 20:39:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1550430054052\_0018

19/02/17 20:39:20 INFO impl.YarnClientImpl: Submitted application application\_1550430054052\_0018

19/02/17 20:39:20 INFO mapreduce.Job: The url to track the job: http://ip-172-31-7-172.eu-west-1.compute.internal:20888/proxy/application\_1550430054052\_0018/

19/02/17 20:39:20 INFO mapreduce.Job: Running job: job\_1550430054052\_0018

19/02/17 20:39:28 INFO mapreduce.Job: Job job\_1550430054052\_0018 running in uber mode : false

19/02/17 20:39:28 INFO mapreduce.Job: map 0% reduce 0%

19/02/17 20:39:42 INFO mapreduce.Job: map 5% reduce 0%

19/02/17 20:39:43 INFO mapreduce.Job: map 30% reduce 0%

19/02/17 20:39:48 INFO mapreduce.Job: map 45% reduce 0%

19/02/17 20:39:49 INFO mapreduce.Job: map 70% reduce 0%

19/02/17 20:39:54 INFO mapreduce.Job: map 90% reduce 0%

19/02/17 20:39:59 INFO mapreduce.Job: map 95% reduce 0%

19/02/17 20:40:00 INFO mapreduce.Job: map 100% reduce 0%

19/02/17 20:40:01 INFO mapreduce.Job: map 100% reduce 86%

19/02/17 20:40:04 INFO mapreduce.Job: map 100% reduce 100%

19/02/17 20:40:05 INFO mapreduce.Job: Job job\_1550430054052\_0018 completed successfully

19/02/17 20:40:05 INFO mapreduce.Job: Counters: 50

File System Counters

FILE: Number of bytes read=73207

FILE: Number of bytes written=4876204

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=837048

HDFS: Number of bytes written=50840

HDFS: Number of read operations=81

HDFS: Number of large read operations=0

HDFS: Number of write operations=14

Job Counters

Killed map tasks=1

Launched map tasks=20

Launched reduce tasks=7

Data-local map tasks=20

Total time spent by all maps in occupied slots (ms)=12488490

Total time spent by all reduces in occupied slots (ms)=5490180

Total time spent by all map tasks (ms)=277522

Total time spent by all reduce tasks (ms)=61002

Total vcore-milliseconds taken by all map tasks=277522

Total vcore-milliseconds taken by all reduce tasks=61002

Total megabyte-milliseconds taken by all map tasks=399631680

Total megabyte-milliseconds taken by all reduce tasks=175685760

Map-Reduce Framework

Map input records=5000

Map output records=5000

Map output bytes=553400

Map output materialized bytes=126373

Input split bytes=2820

Combine input records=0

Combine output records=0

Reduce input groups=500

Reduce shuffle bytes=126373

Reduce input records=5000

Reduce output records=500

Spilled Records=10000

Shuffled Maps =140

Failed Shuffles=0

Merged Map outputs=140

GC time elapsed (ms)=5230

CPU time spent (ms)=31180

Physical memory (bytes) snapshot=10270760960

Virtual memory (bytes) snapshot=95383240704

Total committed heap usage (bytes)=9798942720

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=834228

File Output Format Counters

Bytes Written=50840

19/02/17 20:40:05 INFO streaming.StreamJob: Output directory: /user/hadoop/wc/poutput3

[hadoop@ip-172-31-7-172 ~]$