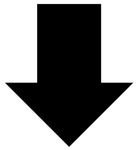
Practice 1

Word Count

Problem

- Separate sentences into words and count how many each words are, using Multi-threading method.

"A long time ago in a galaxy far far away"



A	: 1	a	: 1
long	: 1	galaxy	: 1
time	: 1	far	: 2
ago	: 1	away	: 1
in	: 1		

- Data("testfile1.txt", "testfile2.txt", "LargeTextfile.txt") are provided on I-Campus
- You should submit the results of applying word count results to "LargeTextfile.txt" and report the time between when you use only 1 core and when you use maximum core.

Datatset

1. **The datasets named with “testfile1.txt” and “testfile2.txt” have short simple sentences(you can check whether your code is able to run without problem.**

   Testfile1.txt: “A long time ago in a galaxy far far away”

   Testfile2.txt: “Another episode of Star Wars”

2. **The “LargeTextfile.txt” data is the novel “Animal Farm”. The capacity of this data is 1.15Gb, which is very heavy.**

   (We made this data bigger replicating the novel “Animal Farm” 10 times)

3. **So if you don’t use maximum cores/threads then, you may run the Hadoop File System for a long time.**

Practice 1

1. Make Python code ‘mapper.py’ and ‘reducer.py’ and save in your directory

   mapper.py : Separate sentences into words

   reducer.py : Count the number of words

   NOTE: There must be NO SPACE in your directory

2. Use “sys.stdin” for processing input sentence

   Import sys

3. Input and output file should be processed in HDFS

Practice 1

4. You can use **Multi-threading** method if data is very large.
    - Since, mapper in hdfs uses full cores but reducer uses only one core automatically.
    - So, *if you want to use full cores during reduce process*, use ***-numReduceTasks*** argument in your command line. This argument has value **1 as default**, it means you will use only one core.
    - You can set this number as your maximum number of core.
    - *So test running time when you use **single core** or **maximum number of cores(In our case we have the 8 cores)**.*

Please refer to our practice1_solution pdf file

Submission

- *If you run word count example without problem, you can get result file in hdfs.*
- *Go localhost:50070, then click "utilities" and "Browse the file system"*
- *Click "output" and download "part-00000" .*
- *And you need to submit screenshot of time difference, after you use different number of cores.*
- *For example,*

Application Type	Queue	StartTime	FinishTime	State	FinalStatus
MAPREDUCE	default	Sat Apr 4 02:07:56 +0900 2020	Sat Apr 4 02:11:46 +0900 2020	FINISHED	SUCCEEDED
MAPREDUCE	default	Sat Apr 4 01:59:04 +0900 2020	Sat Apr 4 02:06:10 +0900 2020	FINISHED	SUCCEEDED

Multi-threading

Single-threading

- *Submit YOUR_STUDENT_ID.zip file which includes part-00000 and your screenshot on I-Campus*
- *Submission deadline: April 16 23:59*

Submission

- *You can see your **part-00000** file with the following command*

hdfs dfs -cat part-00000 (Windows)

sudo $HADOOP_HOME/bin/hdfs dfs -cat /output/part-00000 (Linux)

- *Your result file “part-00000” should be as follows.*