MIDS-W261-2017-EX-Week01-Casper

August 30, 2017

1 DATASCI W261: Machine Learning at Scale

2 This notebook provides a poor man Hadoop through command-line and python. Please insert the python code by yourself.

3 Map

In [2]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re
count = 0
WORD_RE = re.compile(r"[\w']+")
filename = sys.argv[2]
findword = sys.argv[1]
with open (filename, "r") as myfile:
    for line in myfile.readlines():
        # The if and re.search statement below will count the lines containing the find
        # The output will yield the same results when comparing to the grep command. Us
        # wants to count all instances of the findword match.
        if re.search(findword, line, flags = re.I):
            num = 1
        else:
            num = 0
        #print(num)
        if num: count = count + num
print(count)
```

```
Overwriting mapper.py
```

In [3]:
```
!chmod a+x mapper.py
```

4 Reduce

In [4]:
```
%%writefile reducer.py
#!/usr/bin/python
```

```
        import sys
        sum = 0
        for line in sys.stdin:
            # Convert string to an int and sum the running total.
            #print int(line)
            sum = sum + int(line)
        print(sum)
```

```
Overwriting reducer.py
```

```
In [5]: !chmod a+x reducer.py
```

5 Write script to file

```
In [6]: %%writefile pGrepCount.sh
        ORIGINAL_FILE=$1
        FIND_WORD=$2
        BLOCK_SIZE=$3
        CHUNK_FILE_PREFIX=$ORIGINAL_FILE.split
        SORTED_CHUNK_FILES=$CHUNK_FILE_PREFIX*.sorted
        usage()
        {
            echo Parallel grep
            echo usage: pGrepCount filename word chuncksize
            echo greps file file1 in $ORIGINAL_FILE and counts the number of lines
            echo Note: file1 will be split in chunks up to $ BLOCK_SIZE chunks each
            echo $FIND_WORD each chunk will be grepCounted in parallel
        }
        #Splitting $ORIGINAL_FILE INTO CHUNKS
        split -b $BLOCK_SIZE $ORIGINAL_FILE $CHUNK_FILE_PREFIX
        #DISTRIBUTE
        for file in $CHUNK_FILE_PREFIX*
        do
            #grep -i $FIND_WORD $file|wc -l >$file.intermediateCount &
            ./mapper.py $FIND_WORD $file >$file.intermediateCount &
        done
        wait
        #MERGEING INTERMEDIATE COUNT CAN TAKE THE FIRST COLUMN AND TOTOL...
        #numOfInstances=$(cat *.intermediateCount | cut -f 1 | paste -sd+ - |bc)
        numOfInstances=$(cat *.intermediateCount | ./reducer.py)
        echo "found [$numOfInstances] [$FIND_WORD] in the file [$ORIGINAL_FILE]"
```

```
Overwriting pGrepCount.sh
```

6 Run the file

In [7]:
```
!chmod a+x pGrepCount.sh
```

Usage: usage: pGrepCount filename word chuncksize

In [8]:
```
!./pGrepCount.sh License.txt COPYRIGHT 4k
```

```
found [57] [COPYRIGHT] in the file [License.txt]
```

7 Clean up files

In [1]:
```
!rm -rf License.txt.*
```

8 Nbviewer link

http://nbviewer.jupyter.org/github/jenncasper/mids_w261/blob/master/wk1_ex1-DivideAndConquer/MIDS-W261-2017-EX-Week01-Casper.ipynb

In [ ]: