DATASCI W261: Machine Learning at Scale

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

Map

In [9]:

```
%%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():
        if findword.lower() in line.lower():
            count = count + 1
print(count)
```

```
Overwriting mapper.py
```

In [2]:

```
!chmod a+x mapper.py
```

```
'chmod' is not recognized as an internal or external command,
operable program or batch file.
```

Reduce

In [7]:

```
%%writefile reducer.py
#!/usr/bin/python
import sys
sum = 0
for line in sys.stdin:
    sum = sum + int(line)
print(sum)
```

```
Overwriting reducer.py
```

In [60]:

```
!chmod a+x reducer.py
```

Write script to file

In [8]:

```
%%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 &
    python 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 | python reducer.py)
echo "found [$numOfInstances] [$FIND_WORD] in the file [$ORIGINAL_FILE]"
```

```
Overwriting pGrepCount.sh
```

Run the file

In [62]:

```
!chmod a+x pGrepCount.sh
```

Usage: usage: pGrepCount filename word chuncksize

In [5]:

```
!./pGrepCount.sh License.txt COPYRIGHT 4k
```

```
'.' is not recognized as an internal or external command,
operable program or batch file.
```

In [ ]:

```

```