08170120 巨資二 A 鄭雅綿

Q1.

程式碼：

```
from mrjob.job import MRJob

import re

class WordFrequencyCbn(MRJob):

    def mapper(self, key, line):

        words = re.split("\s+",line)

        for word in words[1::4]:

            yield(word.lower(),1)

    def combiner(self, word, occrences):

        yield(word, sum(occrences))

    def reducer(self, word, occurences):

        yield(word, sum(occurences))

if __name__ == '__main__':

    WordFrequencyCbn.run()
```

```
from mrjob.job import MRJob
import re

#WORD_REGEXP = re.compile(r"[\w']+")

class WordFrequencyCbn(MRJob):
    def mapper(self, key, line):
        #words = line.findall(WORD_REGEXP)
        #words = WORD_REGEXP.findall(line)
        words = re.split("\s+",line)
        #for w in words:
        #    yield(w.lower(), 1)
        for word in words[1::4]:
            yield(word.lower(),1)



    def combiner(self, word, occrences):
        yield(word, sum(occrences))

    def reducer(self, word, occurences):
        #print(occurences)
        yield(word, sum(occurences))

if __name__ == '__main__':
    WordFrequencyCbn.run()
```

圖片：

```
"959"    50
"96"     240
"960"    18
"961"    24
"962"    17
"963"    35
"964"    6
"965"    14
"966"    18
"967"    8
"968"    15
"969"    62
"97"     208
"970"    7
"971"    28
"972"    22
"973"    4
"974"    25
"975"    33
"976"    7
"977"    33
"978"    20
"979"    26
"98"     300
"980"    18
"981"    8
"982"    16
"983"    12
"984"    33
"985"    19
"986"    17
"987"    4
"988"    77
"989"    29
"99"     136
"990"    27
"991"    23
"992"    3
"993"    57
"994"    6
"995"    22
"996"    12
"997"    11
"998"    10
"999"    6
Removing temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.044314.903320...
(base) [ec2-user@ip-172-31-88-103 mrjob]$
```

cat u1.base | python3 WordFrequencyCbn.py

```
(base) [ec2-user@ip-172-31-88-103 mrjob]$ head -n 15 cat u1.base | python3 WordFrequencyCbn.py
head: cannot open 'cat' for reading: No such file or directory
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.044539.014352
Running step 1 of 1...
reading from STDIN
job output is in /tmp/WordFrequencyCbn.ec2-user.20210502.044539.014352/output
Streaming final output from /tmp/WordFrequencyCbn.ec2-user.20210502.044539.014352/output...
"1"     1
"11"    1
"13"    1
"15"    1
"16"    1
"18"    1
"19"    1
"2"     1
"21"    1
"3"     1
"4"     1
"5"     1
"7"     1
"8"     1
"9"     1
"u1.base"       1
Removing temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.044539.014352...
```

head -n 15 cat u1.base | python3 WordFrequencyCbn.py

Q2.

程式碼(Method1)：

```
from mrjob.job import MRJob

import re

class WordFrequencyCbn(MRJob):

    def mapper(self, key, line):

        words = re.split("\s+",line)

        for word in words[1::4]:

            yield(word.lower(),1)

    def combiner(self, word, occrences):

        yield(word, sum(occrences))

    def reducer(self, word, occurences):

        yield(word, sum(occurences))

if __name__ == '__main__':

    WordFrequencyCbn.run()
```

```python
from mrjob.job import MRJob
import re

#WORD_REGEXP = re.compile(r"[\w']+")

class WordFrequencyCbn(MRJob):
    def mapper(self, key, line):
        #words = line.findall(WORD_REGEXP)
        #words = WORD_REGEXP.findall(line)
        words = re.split("\s+",line)
        #for w in words:
        #    yield(w.lower(), 1)
        for word in words[1::4]:
            yield(word.lower(),1)



    def combiner(self, word, occrences):
        yield(word, sum(occrences))

    def reducer(self, word, occurences):
        #print(occurences)
        yield(word, sum(occurences))

if __name__ == '__main__':
    WordFrequencyCbn.run()
```

圖片：

```
(base) [ec2-user@ip-172-31-88-103 mrjob]$ cat u1.base | python3 WordFrequencyCbn.py | sort -n -k 2 | ta
il -n 20
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.050849.600833
Running step 1 of 1...
reading from STDIN
job output is in /tmp/WordFrequencyCbn.ec2-user.20210502.050849.600833/output
Streaming final output from /tmp/WordFrequencyCbn.ec2-user.20210502.050849.600833/output...
Removing temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.050849.600833...
"405"   280
"222"   294
"172"   295
"98"    300
"117"   302
"7"     307
"237"   309
"56"    312
"127"   340
"174"   344
"300"   352
"121"   353
"1"     383
"286"   388
"288"   391
"294"   394
"100"   395
"258"   402
"181"   422
"50"    484
```

cat u1.base | python3 WordFrequencyCbn.py | sort -n -k 2 | tail -n 20

程式碼(Method2)：

```
from mrjob.job import MRJob

from mrjob.step import MRStep

import heapq

import mrjob,os

import re

TOPN=20

class WordFrequencyCbn(MRJob):

    def mapper(self, key, line):

        words = re.split("\s+",line)

        for word in words[1::4]:

            yield(word.lower(),1)

    def combiner(self, word, occrences):

        yield(word, sum(occrences))

    def reducer(self, word, occrences):

        heapq.heappush(self.heap,(sum(occrences),word))

        if len(self.heap) > TOPN:

            heapq.heappop(self.heap)

    def reducer_init(self):

        self.heap = []

    def reducer_final(self):

        for (count,word) in self.heap:
```

```
from mrjob.job import MRJob
from mrjob.step import MRStep
import heapq
import mrjob,os
import re

#WORD_REGEXP = re.compile(r"[\w']+")

TOPN=20

class WordFrequencyCbn(MRJob):
    def mapper(self, key, line):
        #words = line.findall(WORD_REGEXP)
        #words = WORD_REGEXP.findall(line)
        words = re.split("\s+",line)
        #for w in words:
        #    yield(w.lower(), 1)
        for word in words[1::4]:
            yield(word.lower(),1)

    def combiner(self, word, occrences):
        yield(word, sum(occrences))

    def reducer(self, word, occrences):
        heapq.heappush(self.heap,(sum(occrences),word))
        if len(self.heap) > TOPN:
            heapq.heappop(self.heap)

    def reducer_init(self):
        self.heap = []

    def reducer_final(self):
        for (count,word) in self.heap:
            yield (word,count)

    def globalTopN_mapper(self,word,count):
        yield "Top"+str(TOPN), (count,word)

    def globalTopN_reducer(self,_,countsAndWords):
        for countAndWord in heapq.nlargest(TOPN,countsAndWords):
            yield _,countAndWord

    def steps(self):
        return [
                MRStep(
                    mapper=self.mapper,
                    combiner=self.combiner,
                    reducer_init=self.reducer_init,
                    reducer=self.reducer,
                    reducer_final=self.reducer_final
                    ),
                MRStep(
                    mapper=self.globalTopN_mapper,
                    reducer=self.globalTopN_reducer
                    )
                ]

if __name__ == '__main__':
    WordFrequencyCbn.run()
```

```
            yield (word,count)
    def globalTopN_mapper(self,word,count):
        yield "Top"+str(TOPN), (count,word)
    def globalTopN_reducer(self,_,countsAndWords):
        for countAndWord in heapq.nlargest(TOPN,countsAndWords):
            yield _,countAndWord
    def steps(self):
        return [
                MRStep(
                    mapper=self.mapper,
                    combiner=self.combiner,
                    reducer_init=self.reducer_init,
                    reducer=self.reducer,
                    reducer_final=self.reducer_final
                    ),
                MRStep(
                    mapper=self.globalTopN_mapper,
                    reducer=self.globalTopN_reducer
                    )
                ]
if __name__ == '__main__':
    WordFrequencyCbn.run()
```

圖片：

```
(base) [ec2-user@ip-172-31-88-103 mrjob]$ cat u1.base | python3 WordFrequencyCbn.py
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.052846.729547
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/WordFrequencyCbn.ec2-user.20210502.052846.729547/output
Streaming final output from /tmp/WordFrequencyCbn.ec2-user.20210502.052846.729547/output...
"Top20" [484, "50"]
"Top20" [422, "181"]
"Top20" [402, "258"]
"Top20" [395, "100"]
"Top20" [394, "294"]
"Top20" [391, "288"]
"Top20" [388, "286"]
"Top20" [383, "1"]
"Top20" [353, "121"]
"Top20" [352, "300"]
"Top20" [344, "174"]
"Top20" [340, "127"]
"Top20" [312, "56"]
"Top20" [309, "237"]
"Top20" [307, "7"]
"Top20" [302, "117"]
"Top20" [300, "98"]
"Top20" [295, "172"]
"Top20" [294, "222"]
"Top20" [280, "405"]
Removing temp directory /tmp/WordFrequencyCbn.ec2-user.20210502.052846.729547...
(base) [ec2-user@ip-172-31-88-103 mrjob]$
```

cat u1.base | python3 WordFrequencyCbn.py