Homework 1 : Matrix Multiplication

108062608 黃柏翰

Map Reduce Algorithm and Code

已知 P = M*N，則 P 中 element Pij 可寫成

$$P_{ik} = \sum_{j} m_{ij} * n_{jk}$$

a. Map

mik -> (k, (i, value)) , nkj -> (k, (j, value))，k 為共同鍵值，為了下一步的 join

程式碼

```
M = M.map(lambda entry: (entry[1], (entry[0], entry[2])))
N = N.map(lambda entry: (entry[0], (entry[1], entry[2])))
```

結果顯示

```
M :  [(0, (0, 51)), (1, (0, 17)), (2, (0, 83))]
N :  [(0, (0, 73)), (0, (1, 58)), (0, (2, 23))]
```

b. Shuffle

將所有 pair 做 join，得到 MN

程式碼

```
MN = M.join(N)
```

結果顯示

```
MN :  [(0, ((0, 51), (0, 73))), (0, ((0, 51), (1, 58))), (0, ((0, 51), (2, 23)))]
```

再 map 成 ((i,j), mvalue*nvalue) ，用 group by 以(i,j)為基礎來分群

程式碼

```
MN_shuffle =(MN.map(lambda entry: ((entry[1][0][0], entry[1][1][0]) ,(entry[1][0][1]*entry[1][1][1]))).groupByKey())
```

結果顯示

```
MN_shuffle :  [((0, 14), <pyspark.resultiterable.ResultIterable object at 0x0000016352F66588>),
((0, 30), <pyspark.resultiterable.ResultIterable object at 0x0000016352F66A90>), ((0, 46),
<pyspark.resultiterable.ResultIterable object at 0x0000016352F66860>)]
```

其中 groupby 的結果為一 list 物件，所以顯示結果為 address

c. Reduce

合併最後所有屬於同一個 key (i,j)的值,即完成了 Pij

程式碼

```
result = MN_shuffle.map(lambda x: (x[0][0], x[0][1], sum(x[1])))
```

結果顯示

```
result :  [(0, 14, 1169192), (0, 30, 1207314), (0, 46, 1154121)]
```

其結果 return 為非照作業要求順序的結果，在寫入時會先排序再寫檔案，以下是寫檔案的程式碼

寫檔程式碼

```
def filewriter():
    data = reducer()
    with open ('Outputfile.txt', 'w') as f:
        data = sorted(data, key = itemgetter(0,1))
        for i,towrite in enumerate(data):
            f.write(str(towrite)+'\n')
```