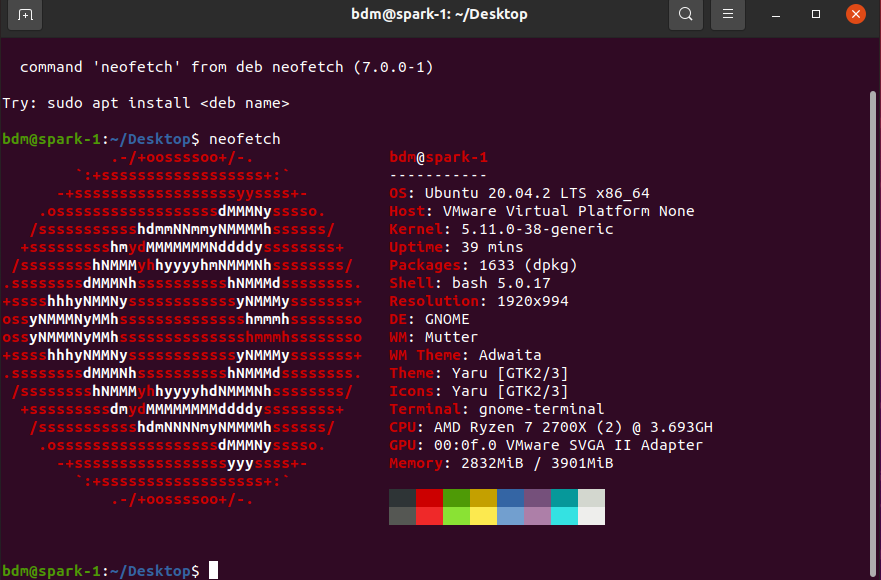
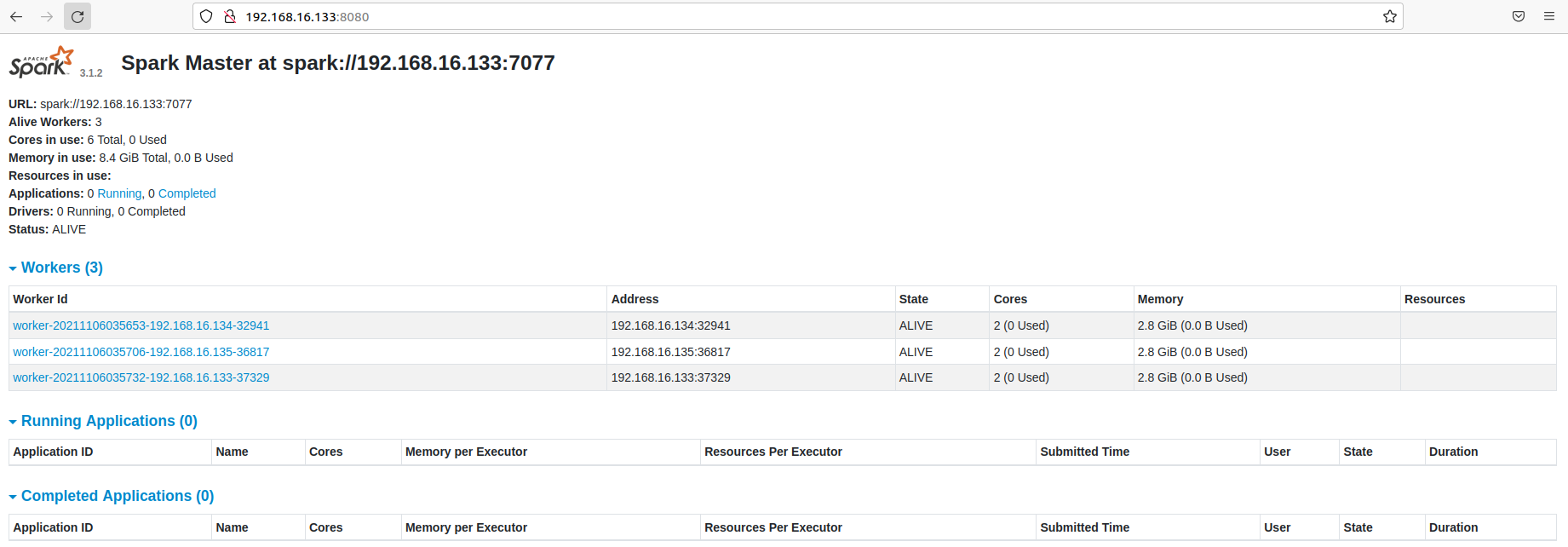
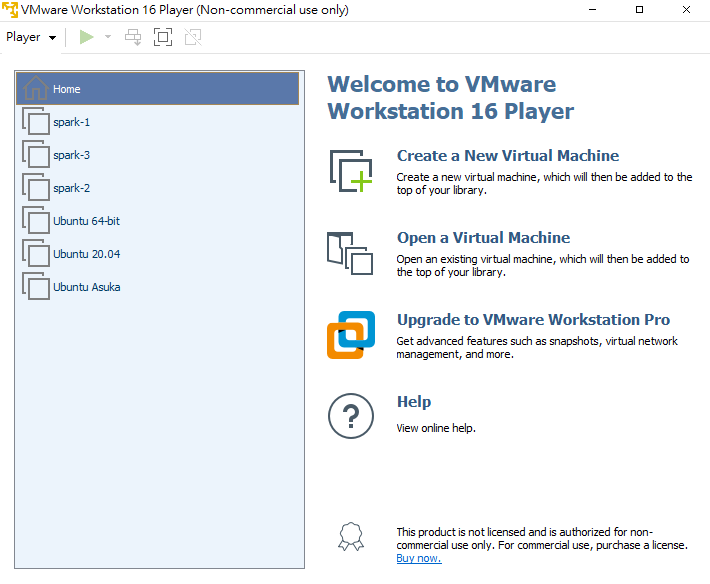
HW3\_助教看看我

110598040 資工碩一 陳廷豪

**環境:**

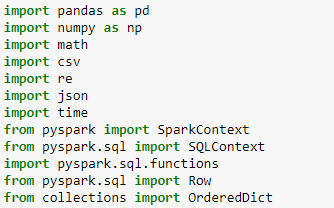




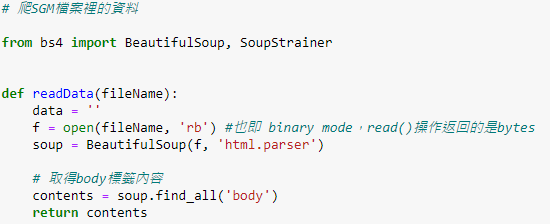


**程式碼:**

先import我要的

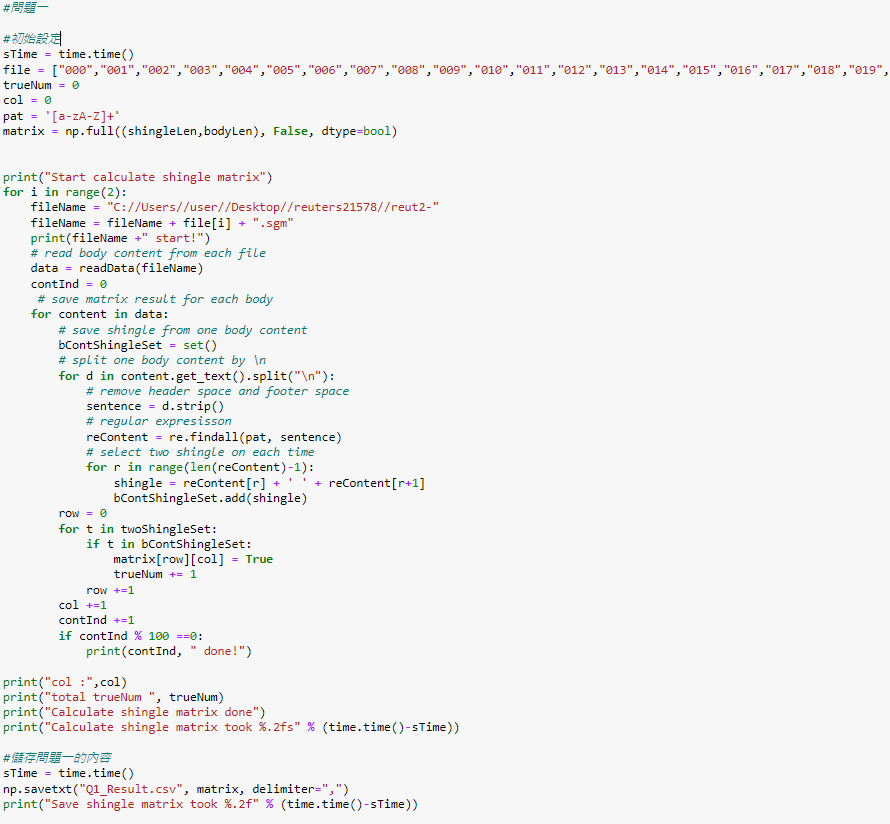


再爬資料:

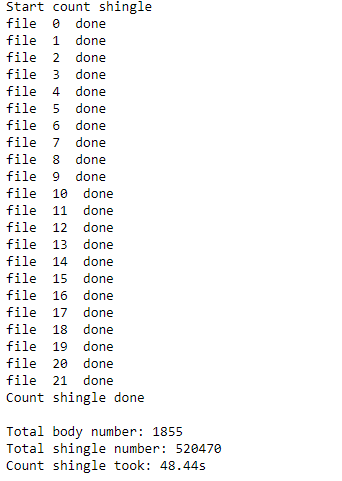


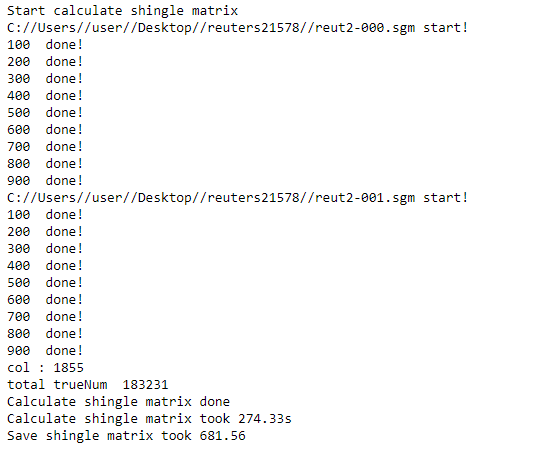
問題一:



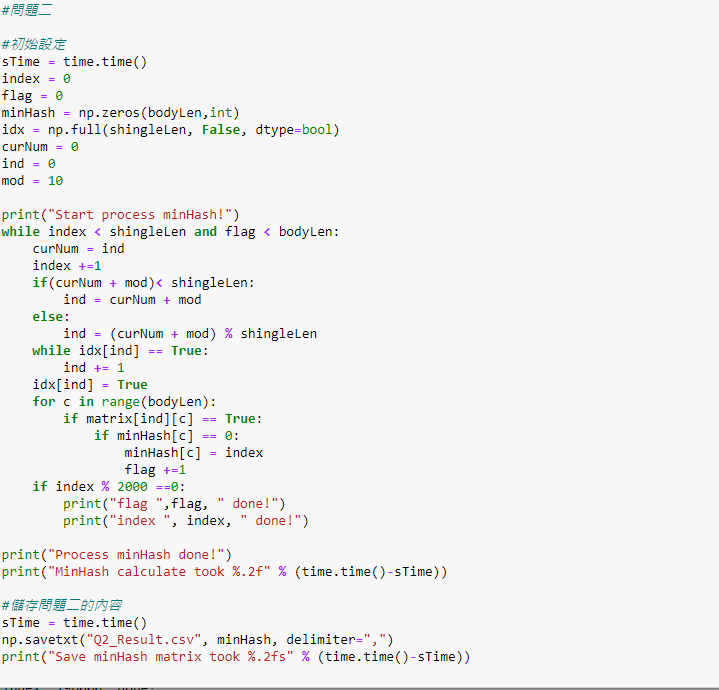


問題一之輸出:

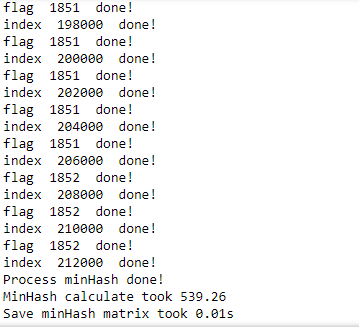




問題二:



問題二之輸出:



資料夾包括此word說明文件以及兩份相同內容的程式碼(.py與.ipynb)

我一樣也是先在jupyter notebook上先執行程式碼來測試自己寫的code能不能成功執行起來

確認能夠執行了之後，再改到虛擬機上面開terminal來執行跑一樣的code來看輸出結果

兩邊確認都是OK的!

虛擬機我是採用VMware，而環境設定就是照著助教PPT那樣子來設置的

弄好了hadoop之後，確認有三個node cluster在run之後，去執行test spark就能在終端機上打開Pyspark來執行我上面的那些程式碼

首先讀取SGM檔案

接著就如同上面code那樣去實作與查看輸出結果這樣子

要注意的地方應該是檔案路徑的不同

jupyter上路徑是: 'C://Users//user//Desktop//reuters21578'

虛擬機終端機上的路徑是: 'file:///home//bdm//Downloads//reuters21578'