Homework 1

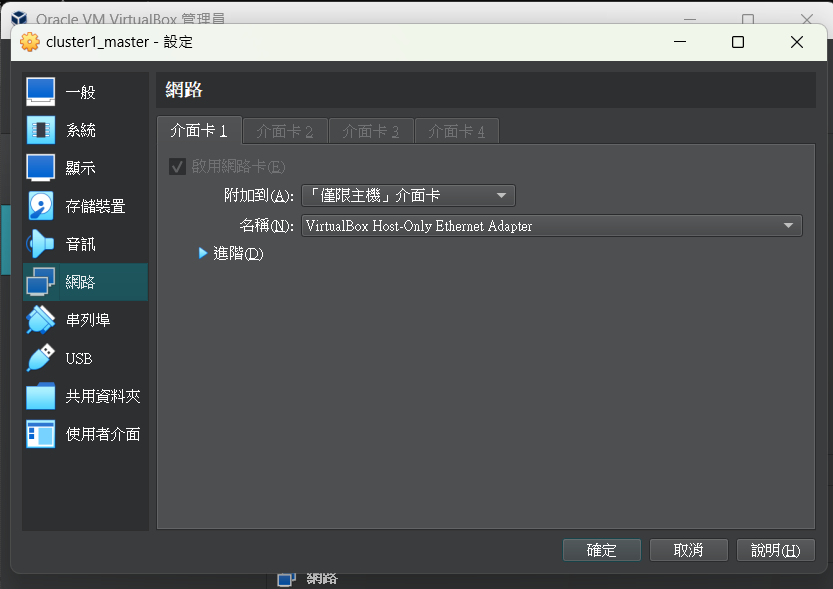
資工碩一 113598065 王丞頤

資工碩一 113598078 廖宥霖

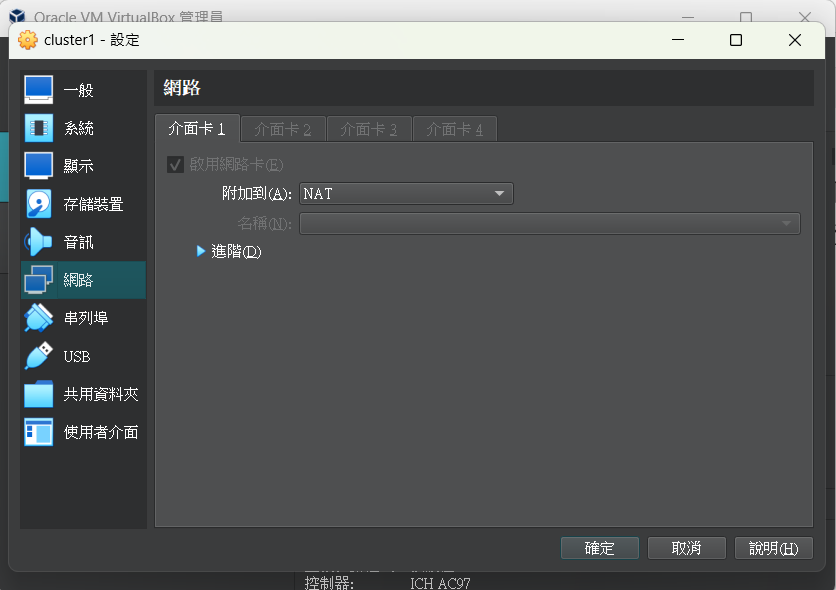
Environment:

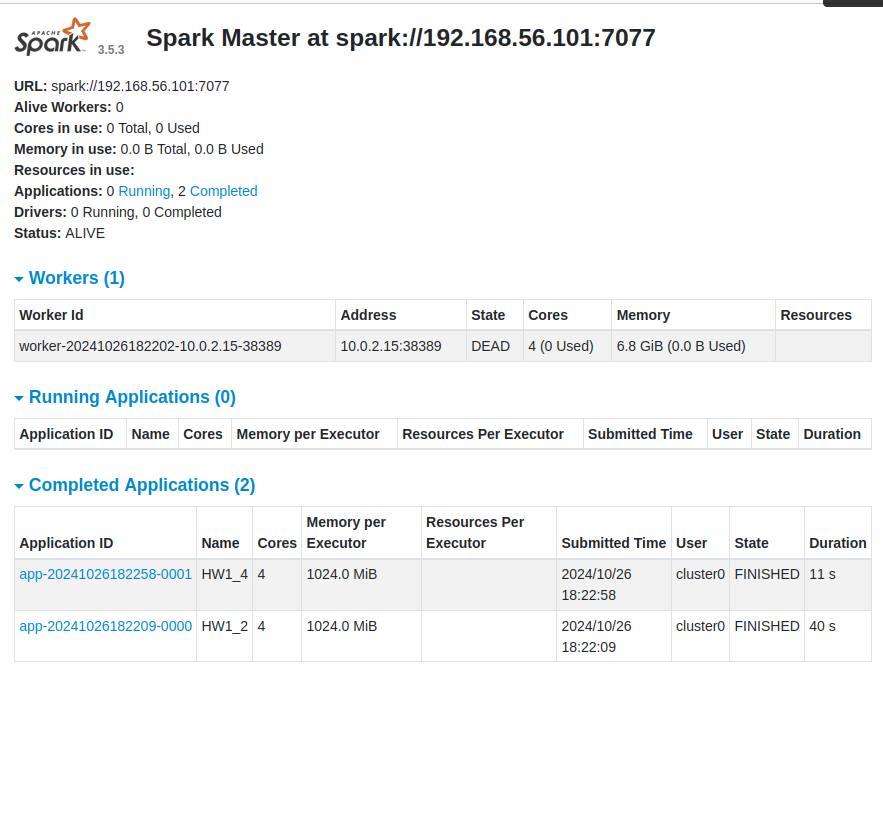
使用兩個VM去做master跟worker，並且由於master本身需要為主機可以用地host所以將其網路介面卡改成host-only使兩個VM可以互相連線形成Multi-Node cluster  
  
如圖:



Cluster-Master:  


Cluster-Worker



  
*how to compile*

* 將zip 解壓縮
* 並從kaggle上去下載spacenews的資料集放進解壓縮後的資料夾
* 此時資料夾內會有py檔案與資料集
* 在資料夾路徑打開cmd
* 在master輸入: spark-class org.apache.spark.deploy.master.Master
* 在worker輸入: spark-class org.apache.spark.deploy.worker.Worker spark://<your IP>:7077
* 下spark-submit指令在master上
* 並且運行各小題的py檔案如:hw1-1.py 並會得到輸出在worker的檔案總管上
* 以下為作業四題之執行方式與結果說明

作業Q1, Q2, Q3, Q4 – 資料預處理:

1. 在date有缺值或者轉換成DateType時資料為NULL(轉換失敗或者無資料)則會被df.filter去過濾掉:  
   
2. 在轉換date時由於雖然透過本地的excel打開來看會是其date的字串長相為: 1-May-24但在spark裡面讀出來格式為MMM dd ,yyyy這是透過打印終端得知因此將其轉換時寫法為: 

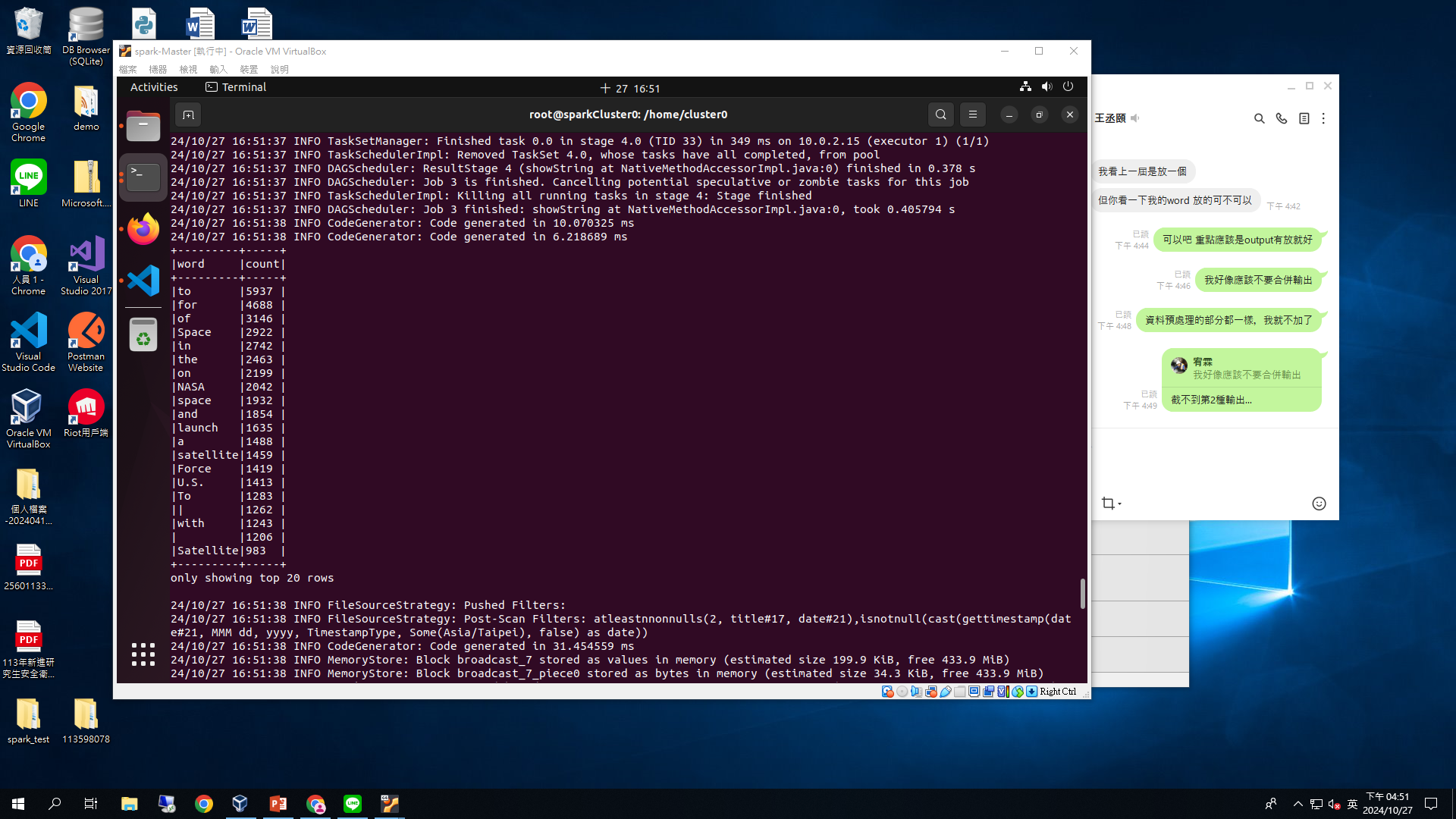
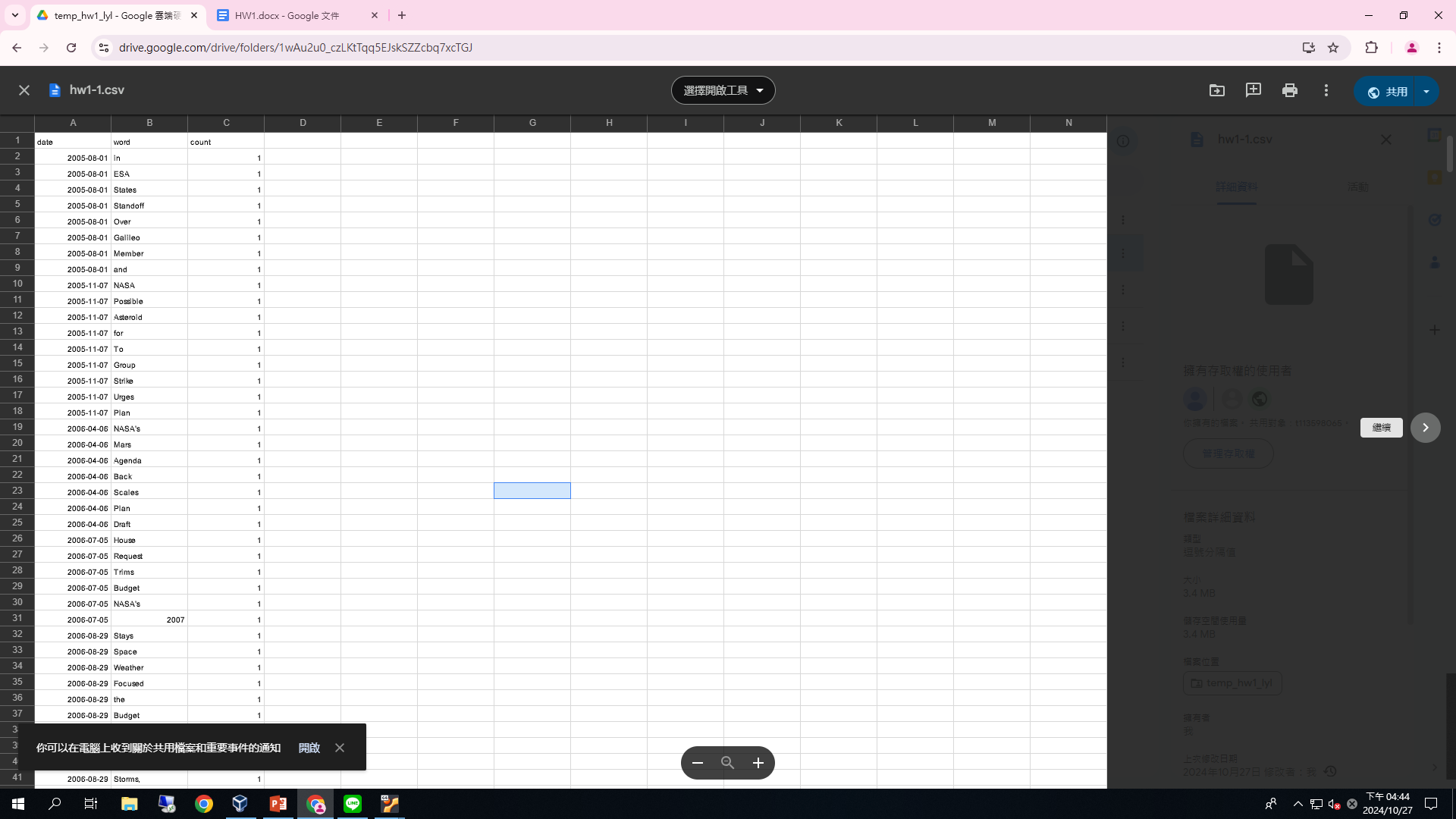
然而由於上面轉換date時這個寫法會遇到spark版本的問題因此終端會出現建議因此我在轉換前加了spark.sql這個語法

Q1:

如果worker因記憶體不足而jvm掛掉，則可以在worker端VM再下一次spark-class生成第二個worker。

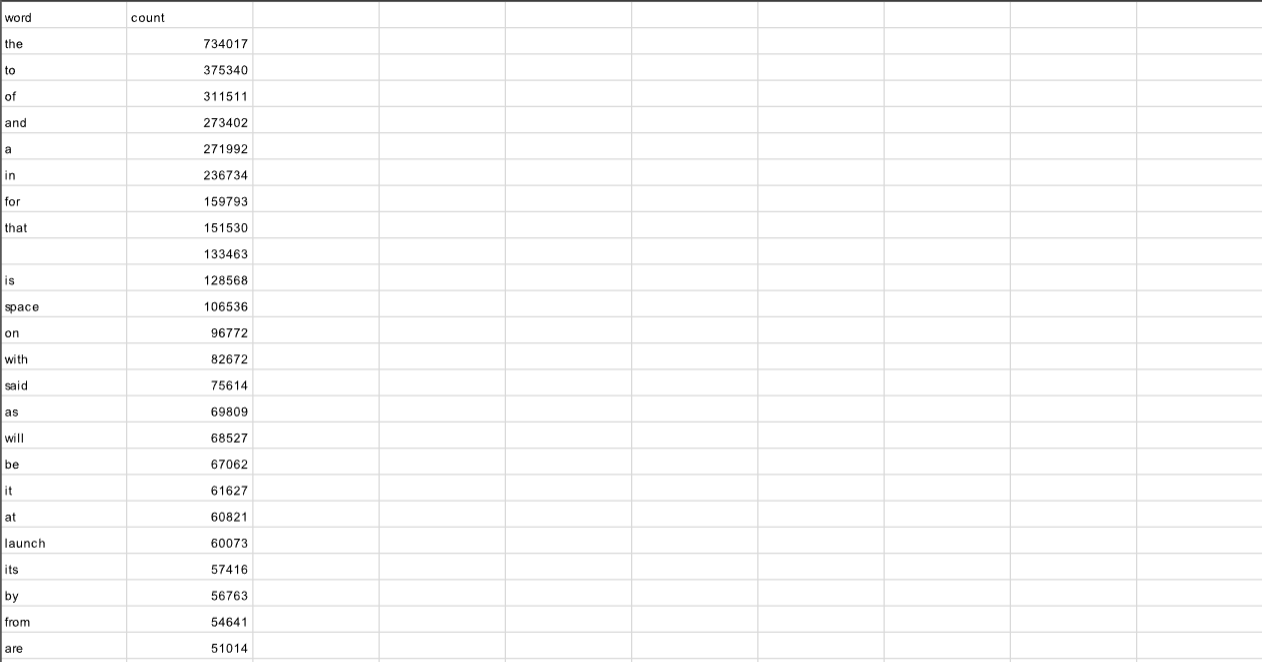
最後有使用unionByName將2種輸出內容(per\_day, in\_total)合併到同一份CSV檔中。

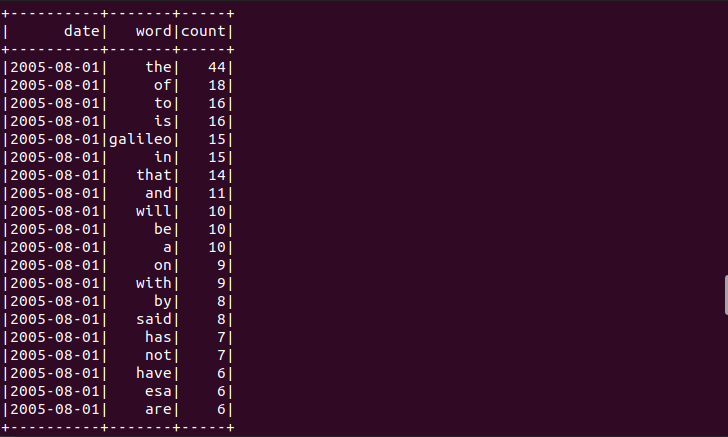
左圖為count\_per\_day，右圖為count\_in\_total



Q2:  
透過將word分組去計算count 並且記錄成檔案後將其透過unpersist去釋放記憶體以防jvm掛掉  
透過explode函數去進行word的紀錄最後透過groupBy去對其用date跟word分組計算count

output: (只能輸出20行因為長度限制)



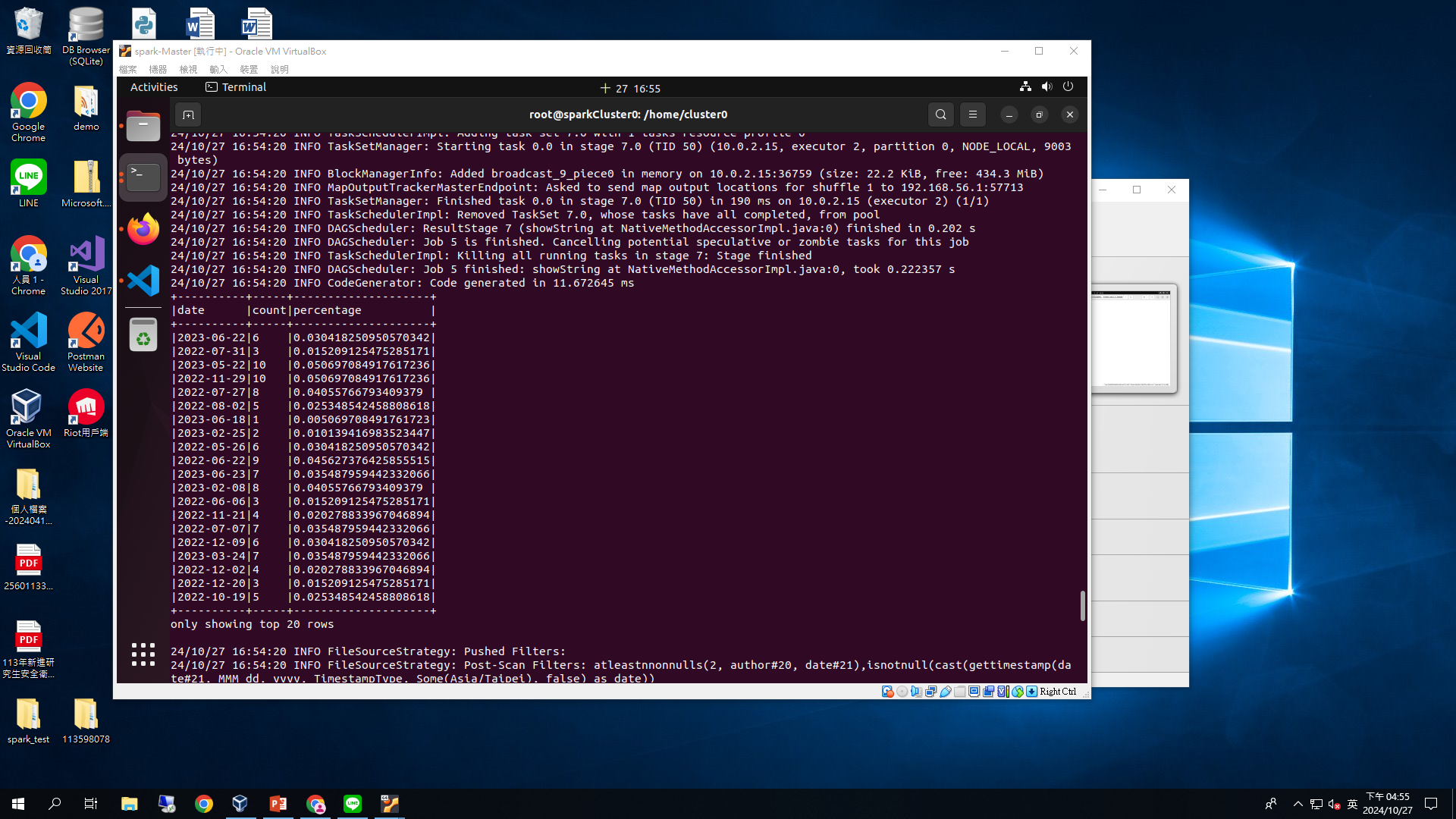
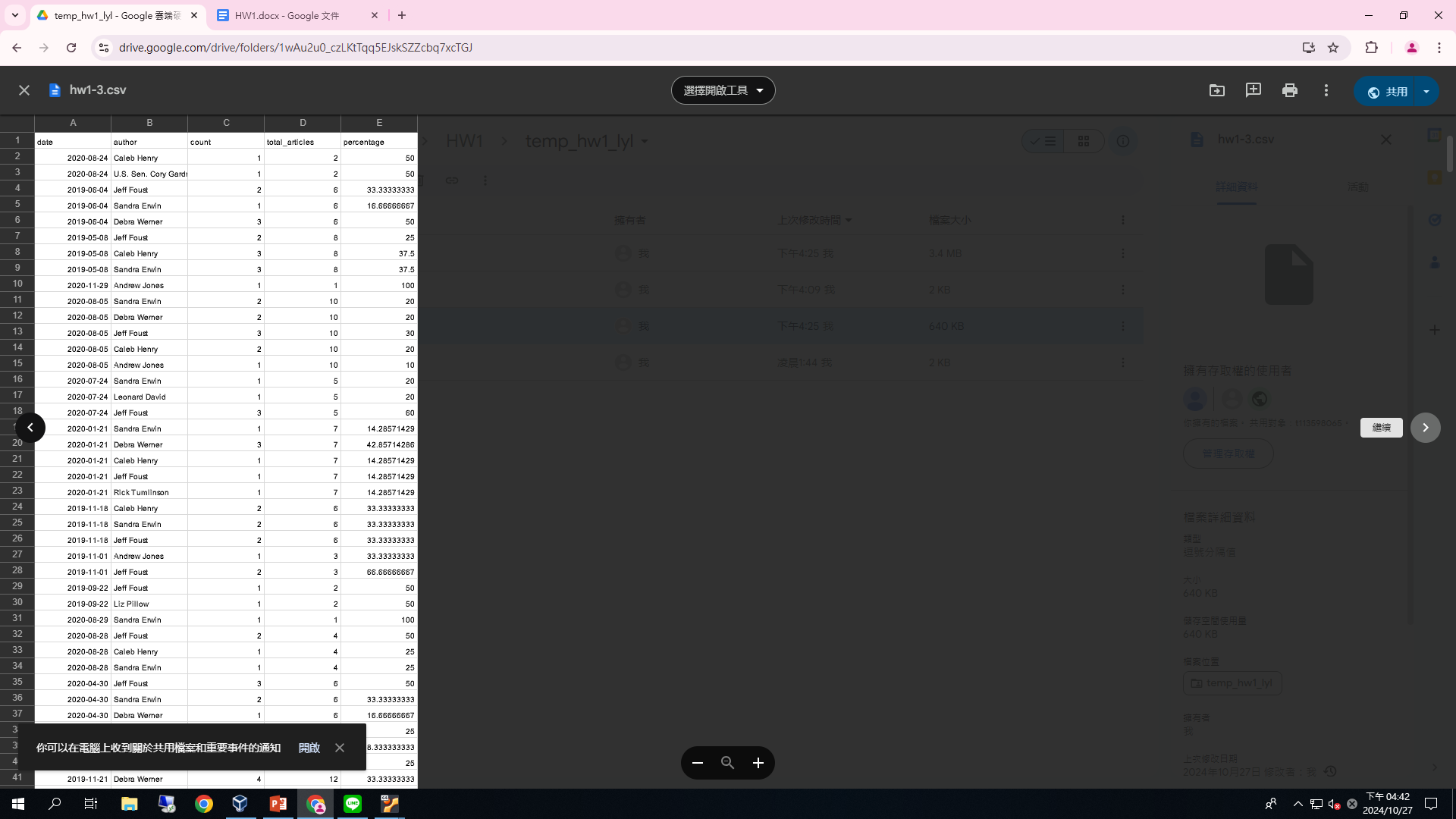


Q3:

如果worker因記憶體不足而jvm掛掉，則可以在worker端VM再下一次spark-class生成第二個worker。

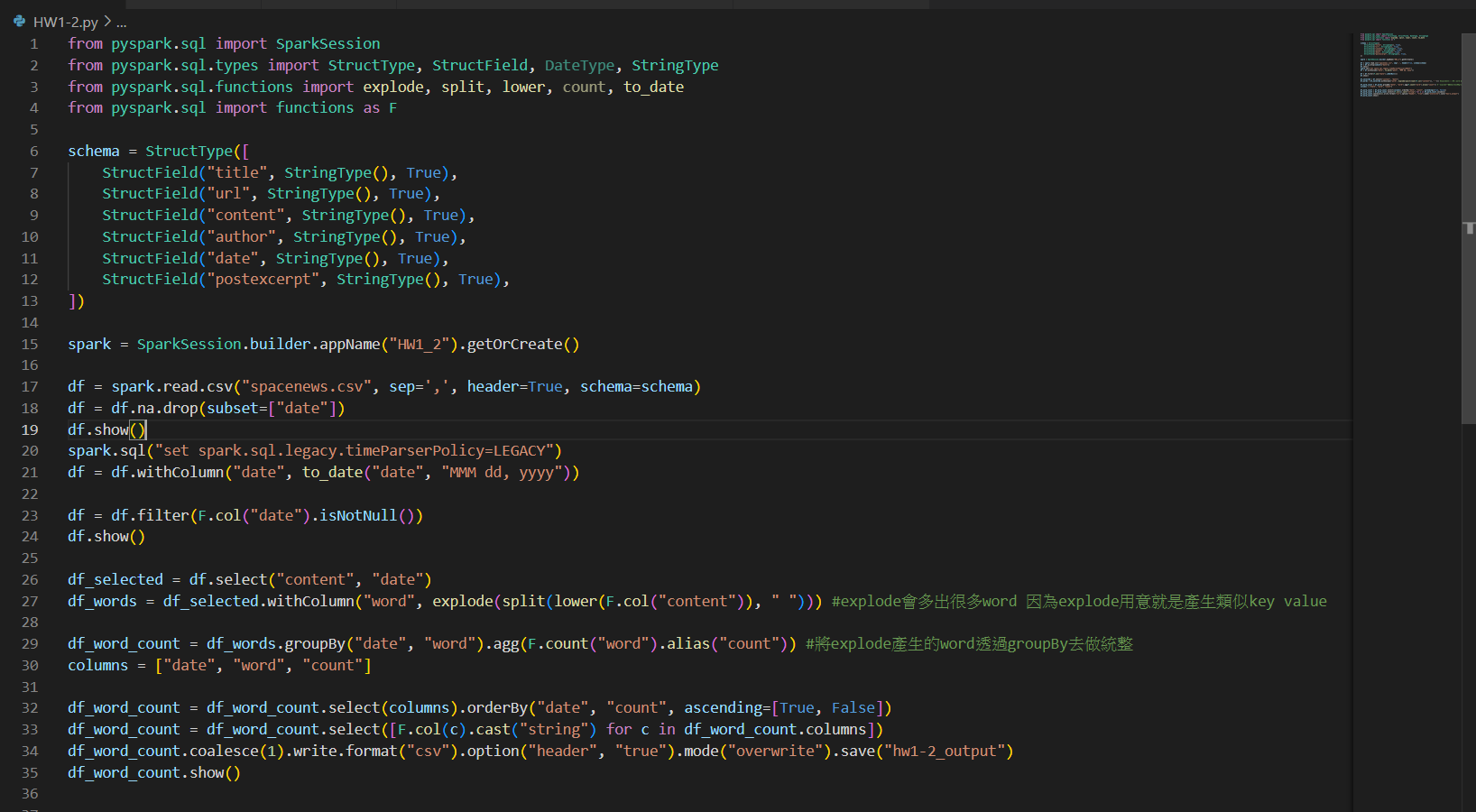
最後有使用unionByName將2種輸出內容(per\_day, in\_total)合併到同一份CSV檔中。

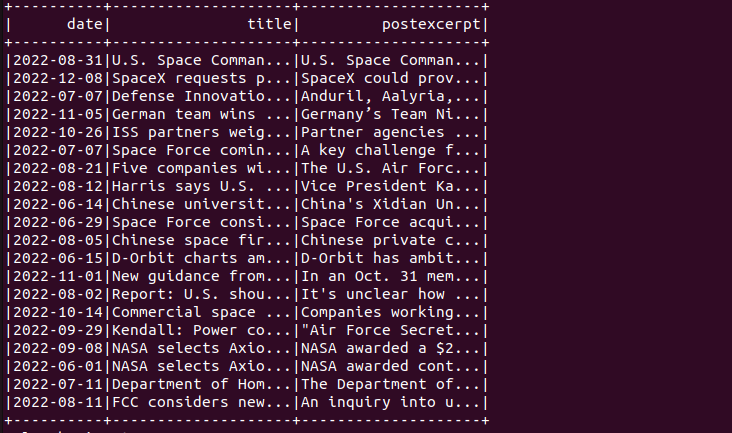
上圖為percentage\_by\_author\_in\_a\_day，下圖為percentage\_in\_a\_day

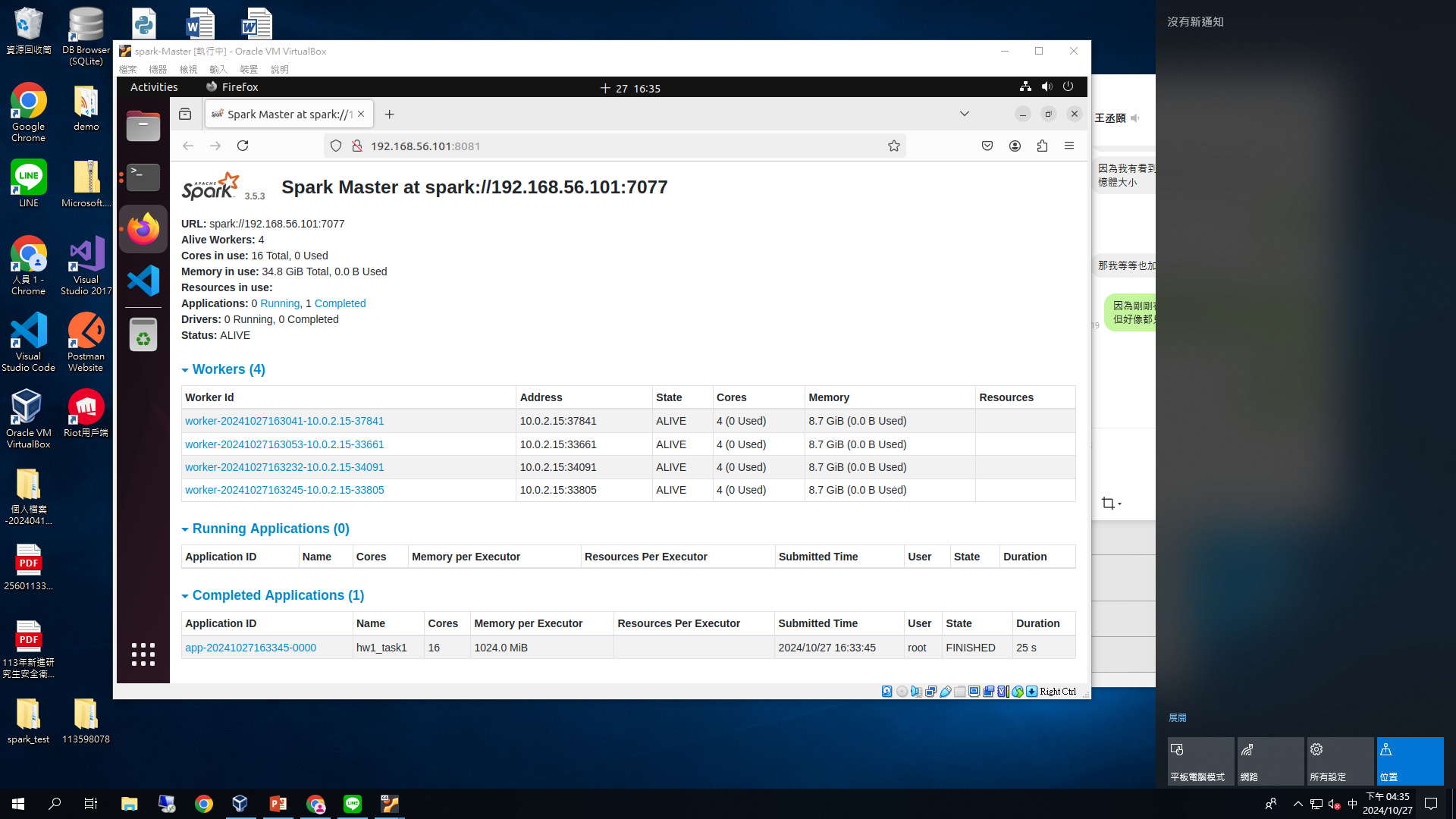
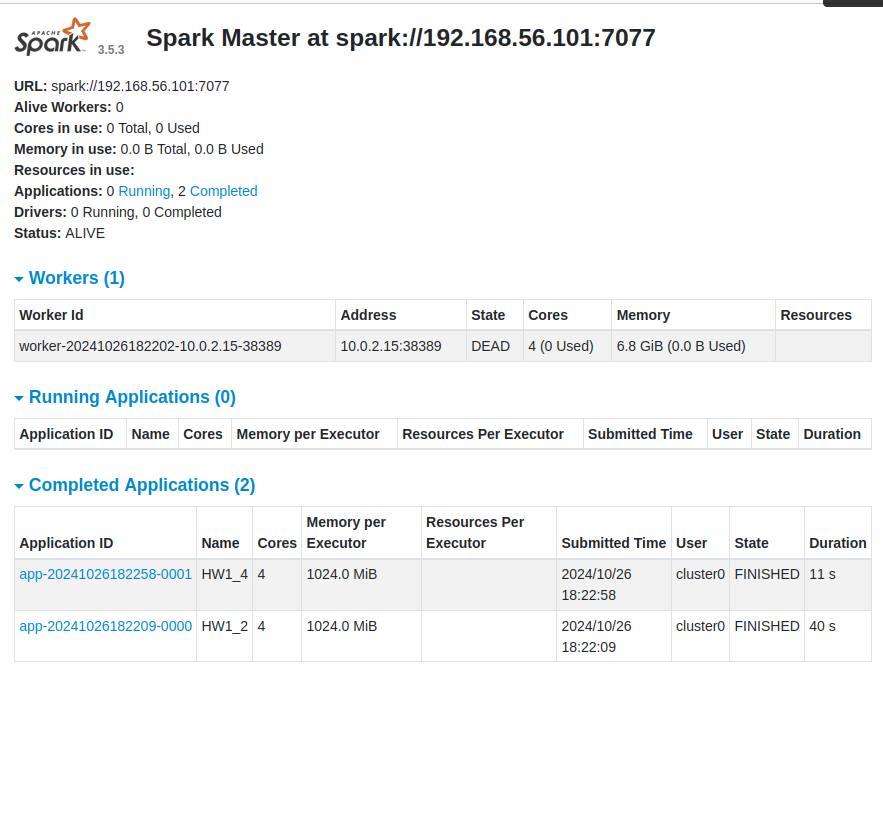


Q4:

透過將其所要判斷的columns裡面的內容都轉成小寫並透過contains判斷有沒有包含space去輸出均有符合的rows





Q2 , Q4, Q1, Q3完成jobs的圖片:  


團隊成員:

資工碩一 113598065 王丞頤: Q2, Q4

資工碩一 113598078 廖宥霖: Q1, Q3