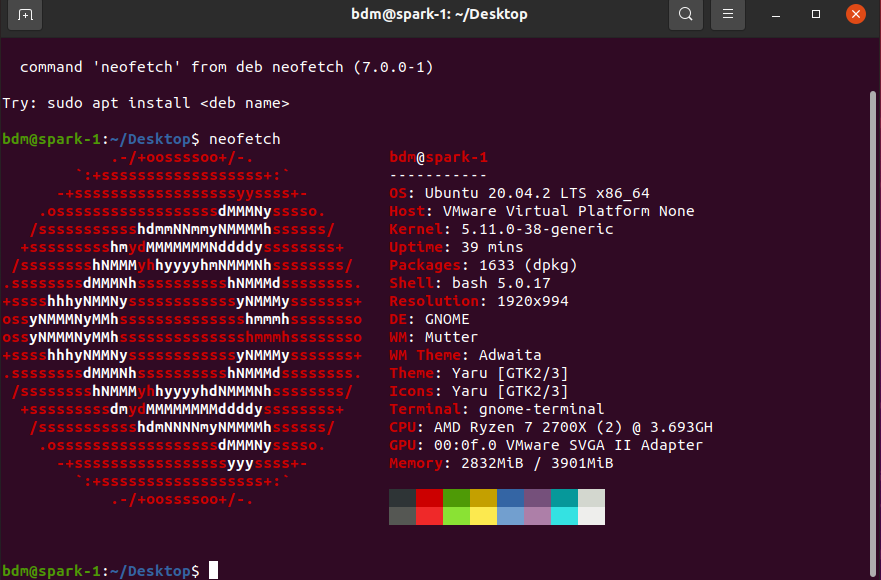
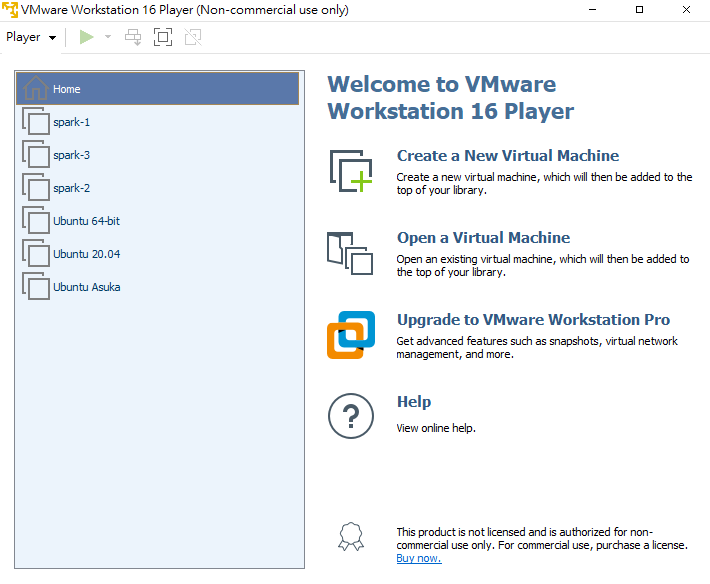
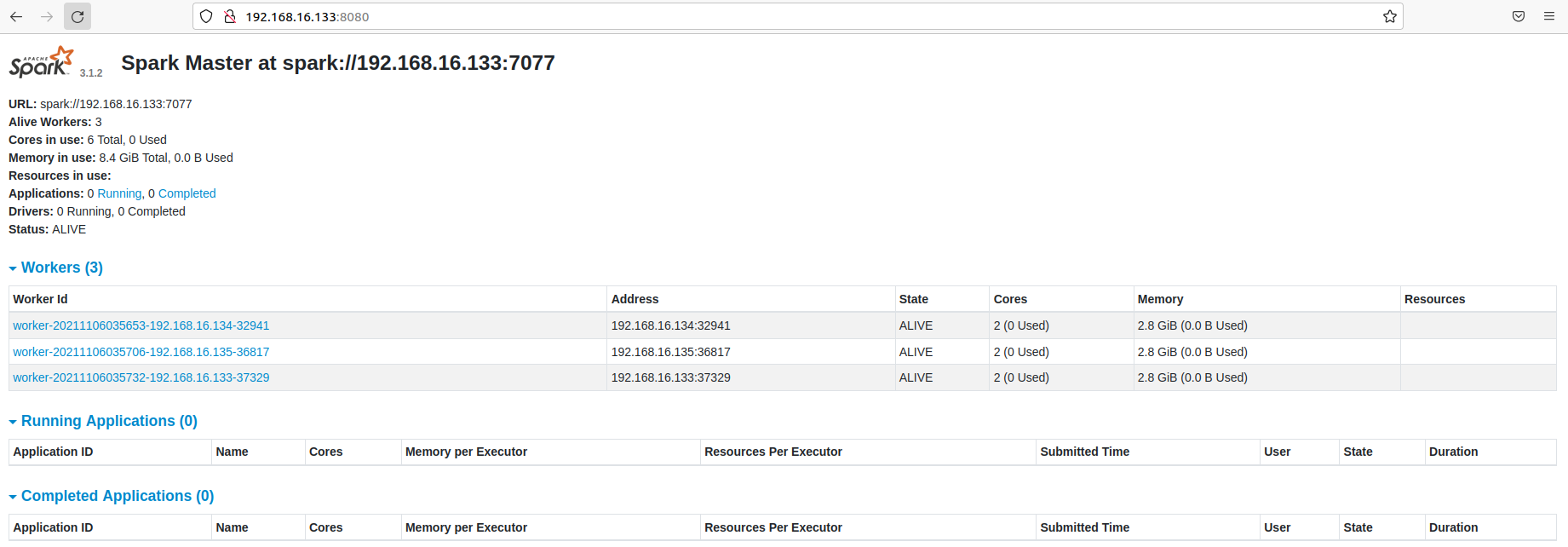
110598040 資工碩一 陳廷豪

#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)

環境設置





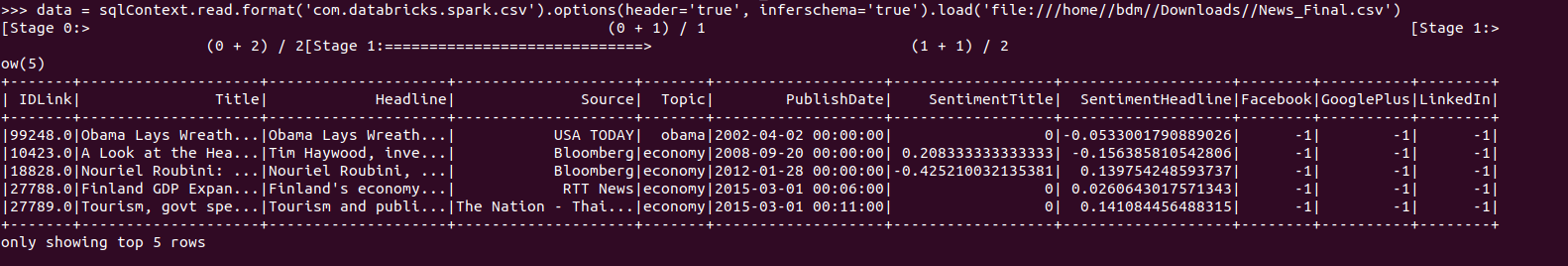


程式碼截圖:

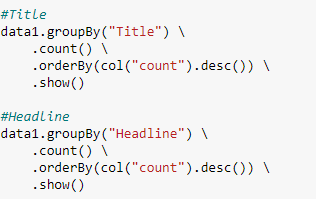
先把我要的東西抓進來，再去對資料做修改並取出Title跟Headline



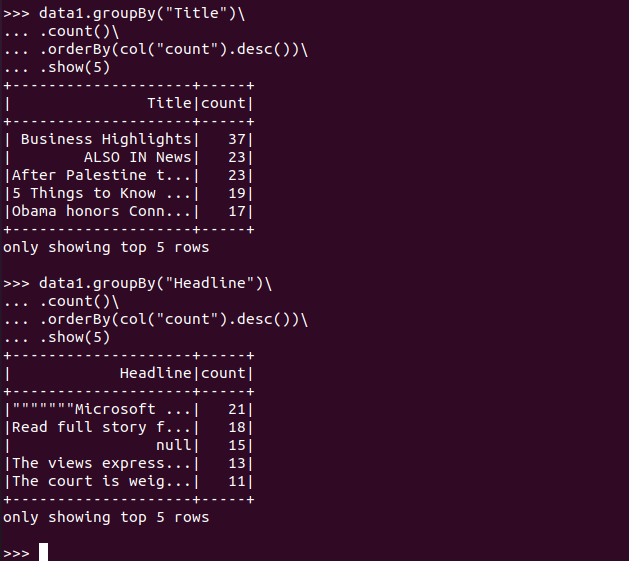
讀取CSV出來的dataframe長這樣子



第一題我一開始的作法是這樣



相對應的輸出是這樣

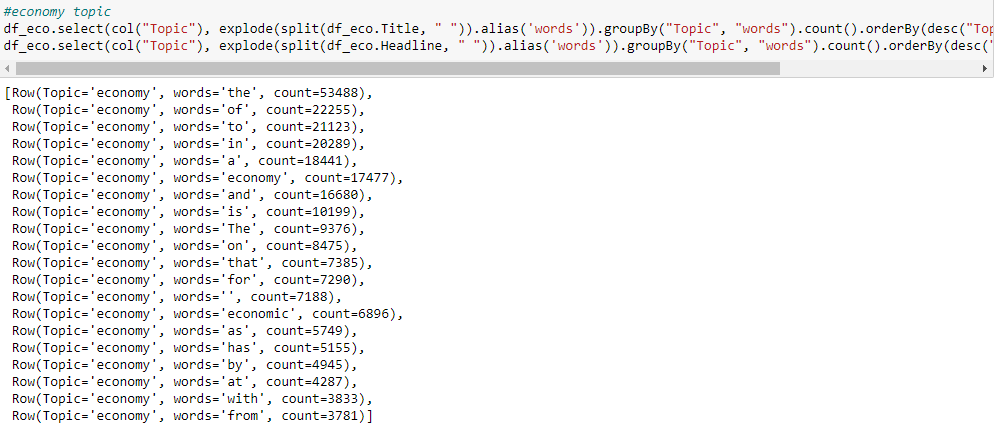


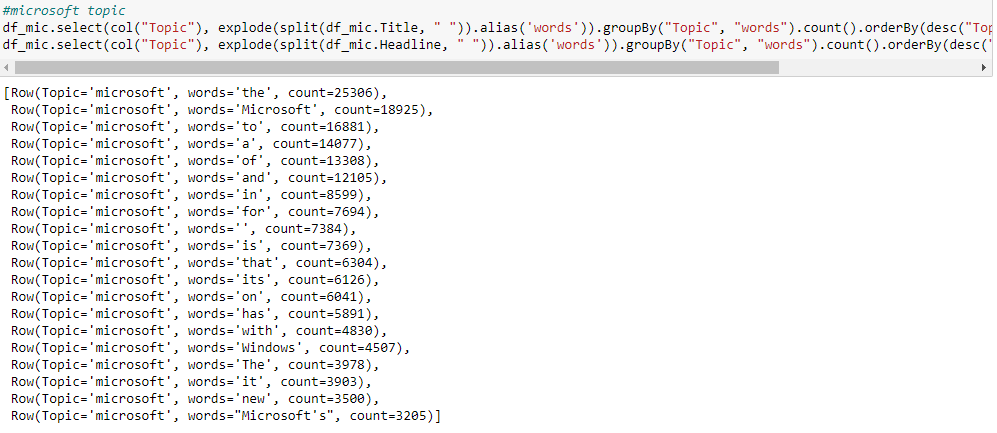
但我認為這樣不太對，所以修改成以下的寫法

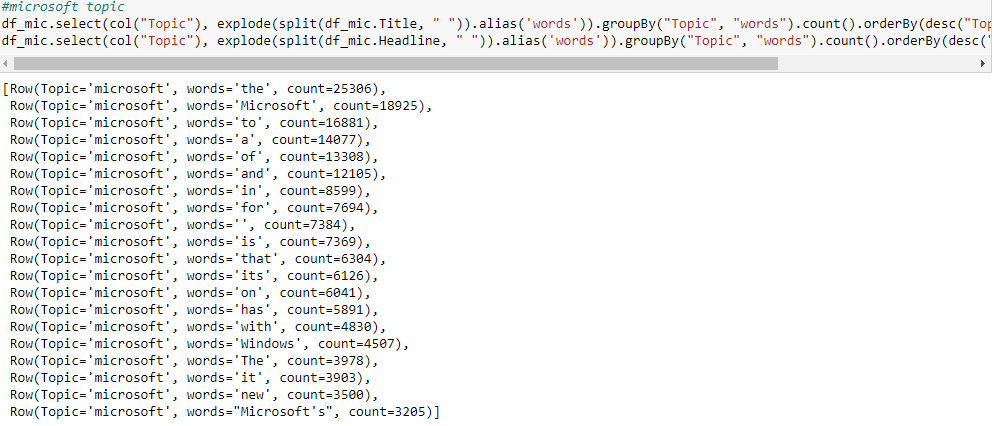
用了filter去分別把Topic中的obama、economy、microsoft、palestine這四個去個別的做words的頻率計算

以下為第一題的Code以及他們的輸出output









下面為第二題的Code:



下面為第三題的Code



以下為第三題的輸出



執行步驟:

我是先在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

首先把pyspark import進來，再讀取CSV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//BigDataHW2//News\_Final.csv'

虛擬機終端機上的路徑是: 'file:///home//bdm//Downloads//News\_Final.csv'