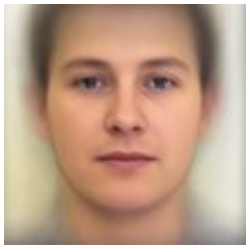
學號：R06921084 系級： 電機碩一 姓名：陳治言

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。



A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

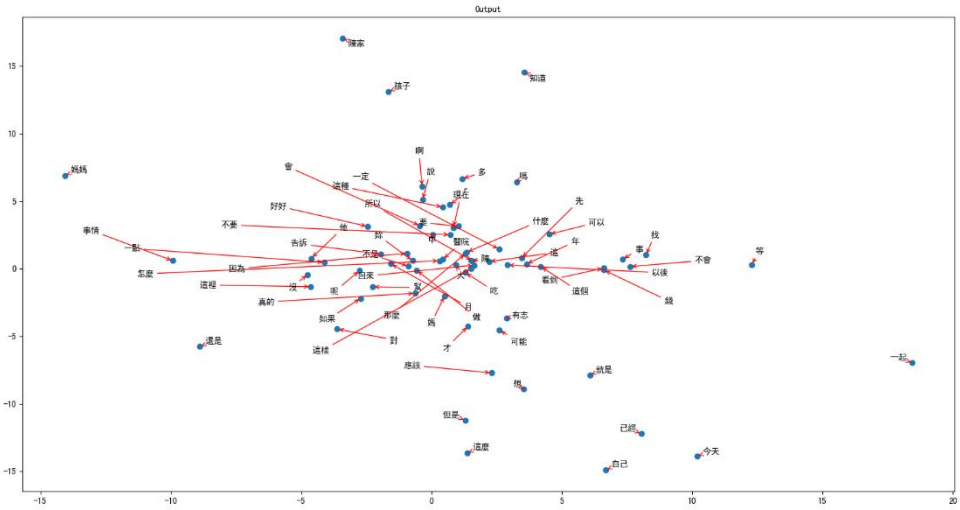
4.2%、2.9%、2.4%、2.2%

B. Visualization of Chinese word embedding

B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

我使用 gensim，size 是 vector 的維度，window 則代表句子中前後看的長度，alpha 則是學習率。size 為 100，window 為 50，Alpha 為 0.0001。

B.2. (.5%) 請在 Report 上放上你 visualization 的結果。



B.3. (.5%) 請討論你從 visualization 的結果觀察到什麼。

代表在我們訓練的文章之中，這些詞彙會意義會非常相近。

C. Image clustering

C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

使用 32 維的 PCA

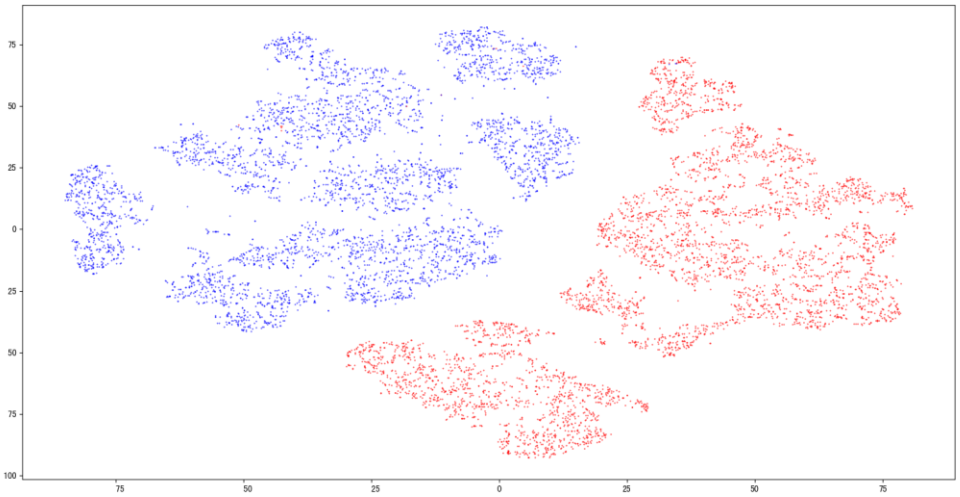
public 為 0.36564，private 為 0.34391

使用 32 維的 auto encoder

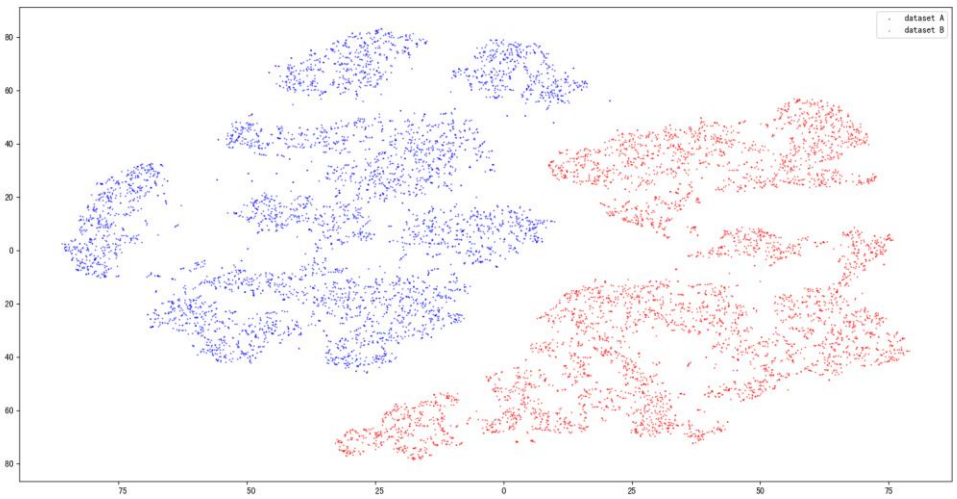
public 為 0.83993，private 為 0.84124

並都使用 k means 來做分類，可是發現到 k means 的效果不是很好，同一個模型訓練出來差異極大，但是以整體效果來說，auto encoder 效果最好。

C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



C.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



其實整體感覺差異不大，因為中間的分隔線算是非常清楚。效果算是非常不錯的。