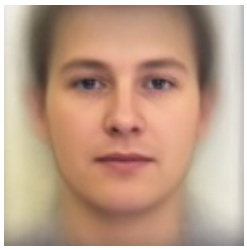
學號：R06921082 系級： 電機所碩一 姓名：陳與賢

A. PCA of colored faces

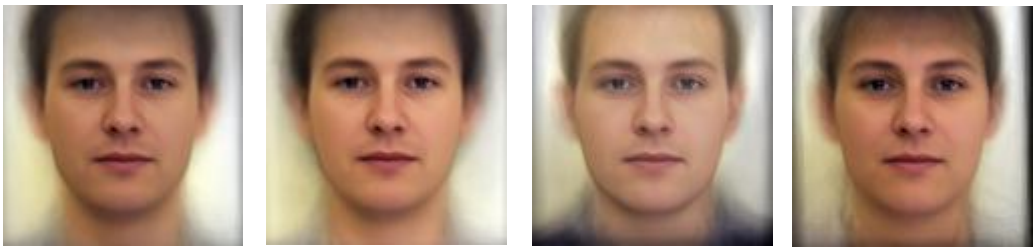
A.1. (.5%) 請畫出所有臉的平均。



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。

55.jpg 123.jpg 289.jpg 355.jpg



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。

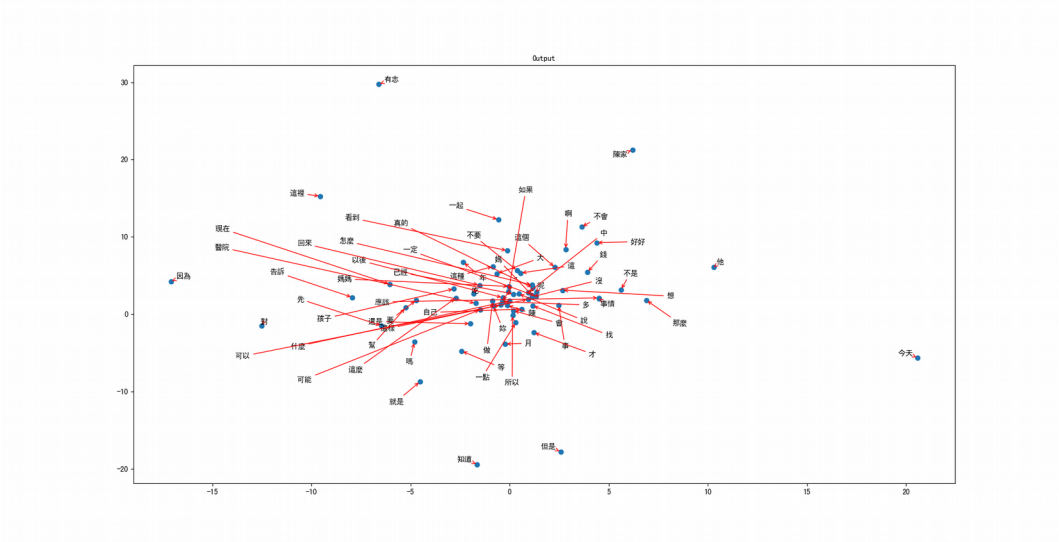
4.2%、2.9%、2.4%、2.2%

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。

使用 gensim 套件，我只有調整 size、window、alpha
size 代表會將每個詞轉成多少維的 vector
window 代表會拿當前詞在句子中的前後幾個詞來一起看
alpha 代表 learning rate

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。

可以看出有一些點距離十分相近，比如說”所以”及”會”，代表在我訓練的辭彙中這個組合很常出現，而像是”陳家”或”今天”這種辭彙則是較少出現。

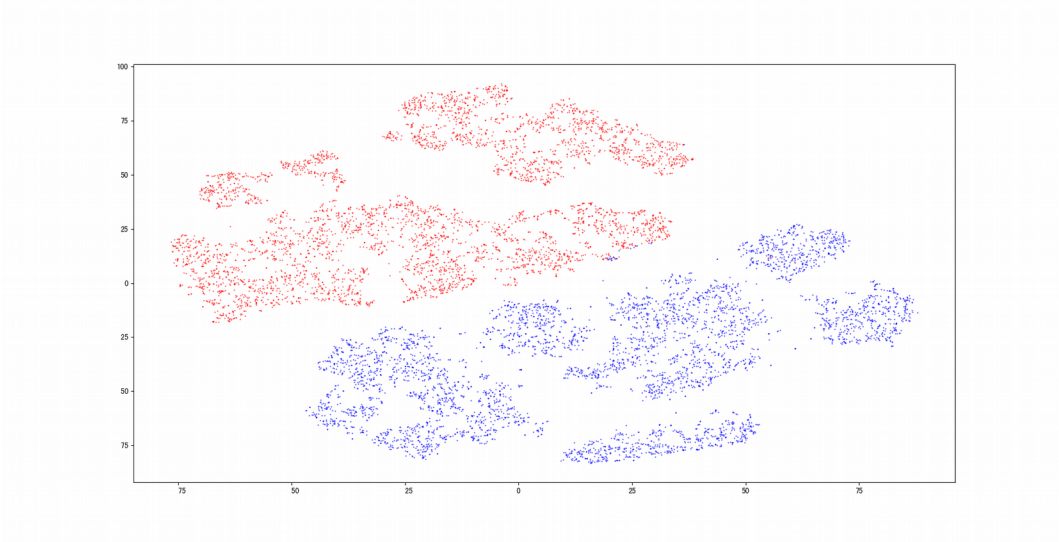
C. Image clustering

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)

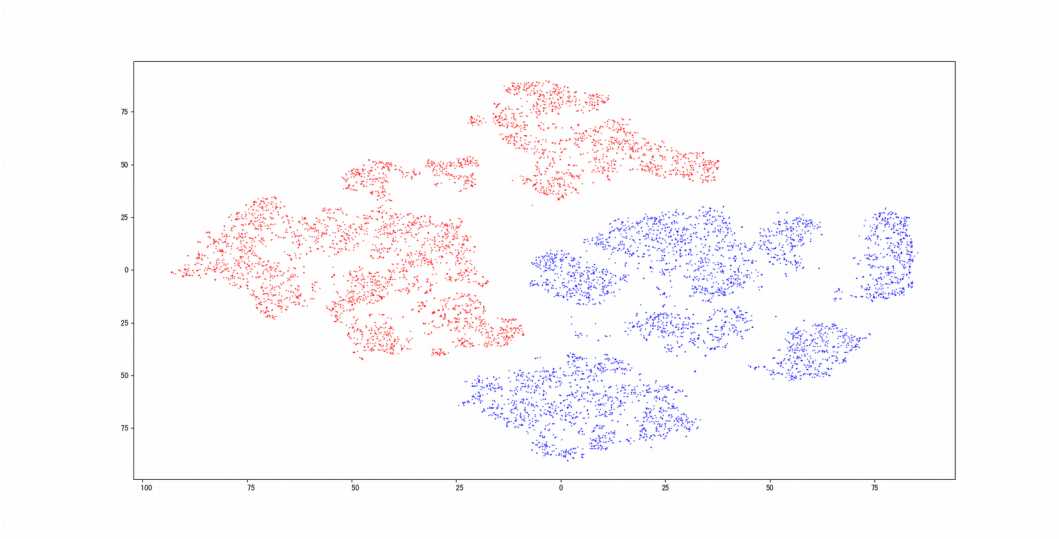
使用維度為 128 的 PCA，public 0.18554、private 0.18842
使用維度為 32 的 auto-encoder，public 0.83993、private 0.84124

兩者都用 k-means 來分群，發現 auto-encoder 的效果較佳，但可能也只是因為我沒有花太多時間研究 PCA（聽說用 PCA 可以達到分數為 1.0000）。

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。



我的預測結果沒有將兩個 dataset 完全區隔開來，所以可以看到上一題的圖在中間會有幾個藍點及紅點混在一起了，而解答的藍點紅點就可以完全區隔開來，視覺化的結果還算合理。