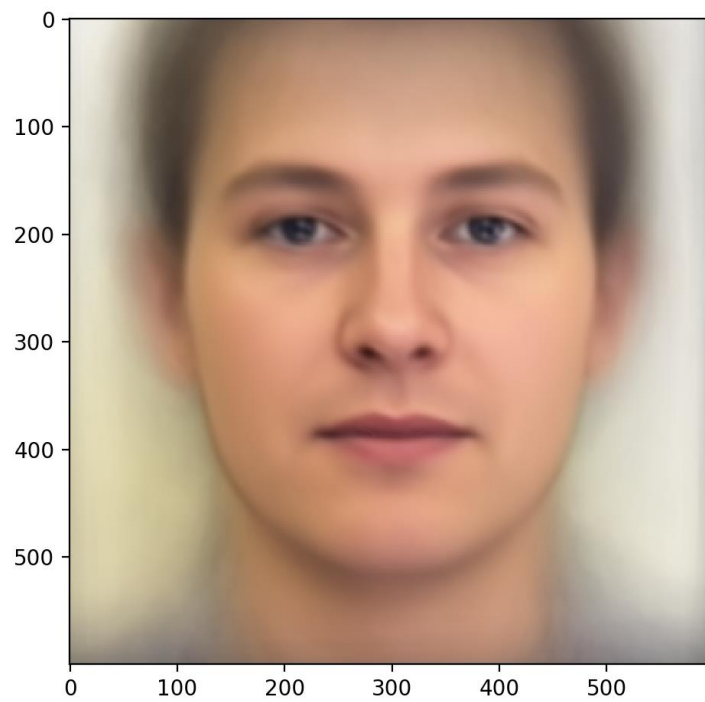


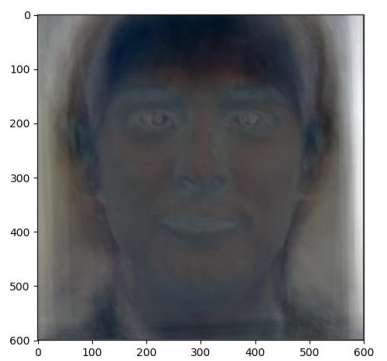
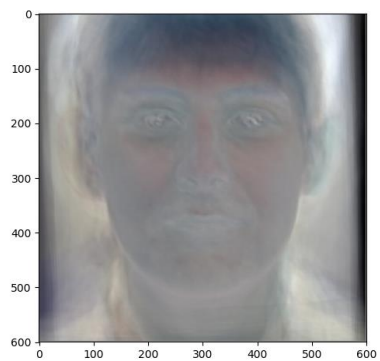
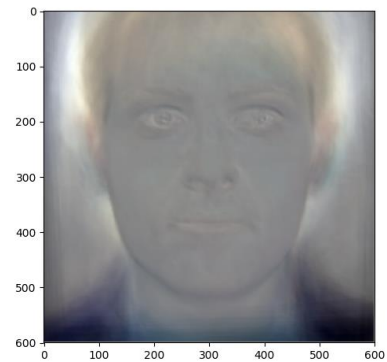
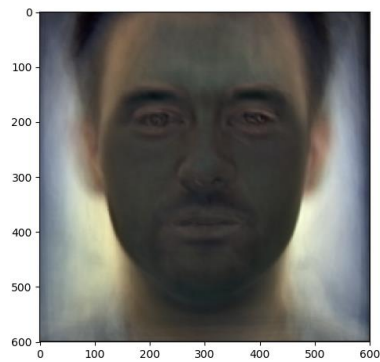
學號：R06922097 系級：資工碩一 姓名：鄭雅文

A. PCA of colored faces

A.1.(.5%) 請畫出所有臉的平均。



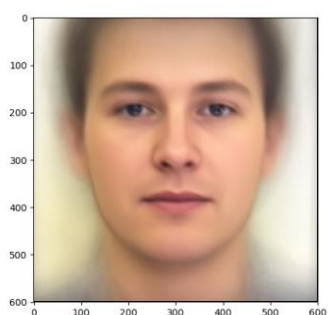
A.2.(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



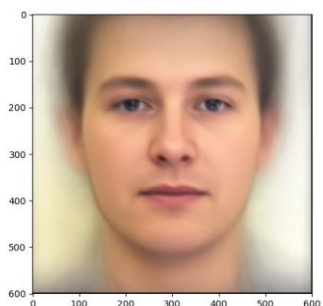
左上: 0 右上: 1 左下: 2 右下: 3

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。

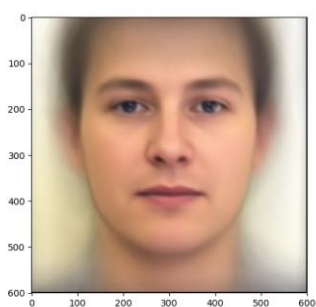
10.jpg



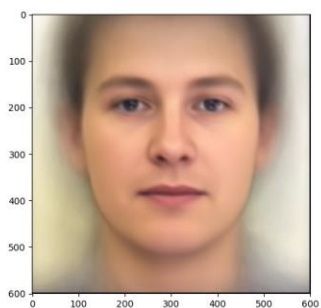
20.jpg



30.jpg



40.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。

$S[0] : 0.041446248382629933 = 4.1\%$

$S[1] : 0.029487322251120808 = 2.9\%$

$s[2] : 0.023877112932084141 = 2.4\%$

$S[3] : 0.022078415569025442 = 2.2\%$

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。

用 gensim 的 word2vec，

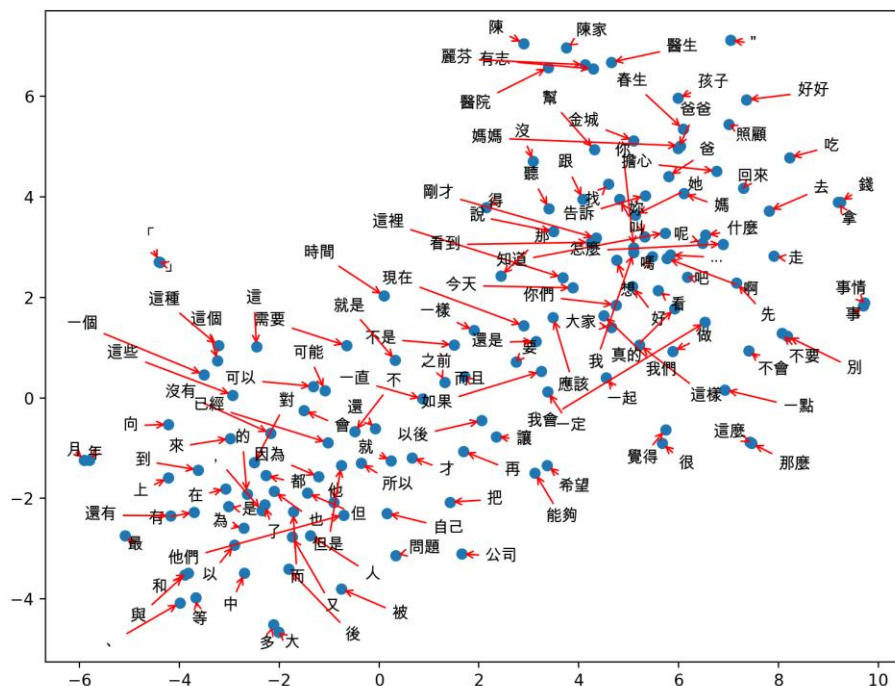
```
model = word2vec.Word2Vec(sentence, size = 64, workers = 4, sg = 1)
```

size: 輸出的每個字的 vector 的維度，設為 64 維

workers: 用多少 threads 去 train model，設為 4 個

sg: 用何種演算法，設為 1 是 skip-gram

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。

「事」跟「事情」在圖上位置很相近。

「這麼」、「那麼」、「這樣」在圖上位置很接近。

「陳」、「陳家」、「麗芬」、「有志」、「春生」等人名很接近，向量也指向同個方向。

「這」、「這種」、「這個」、「這些」的位置接近，向量也指向同個方向。

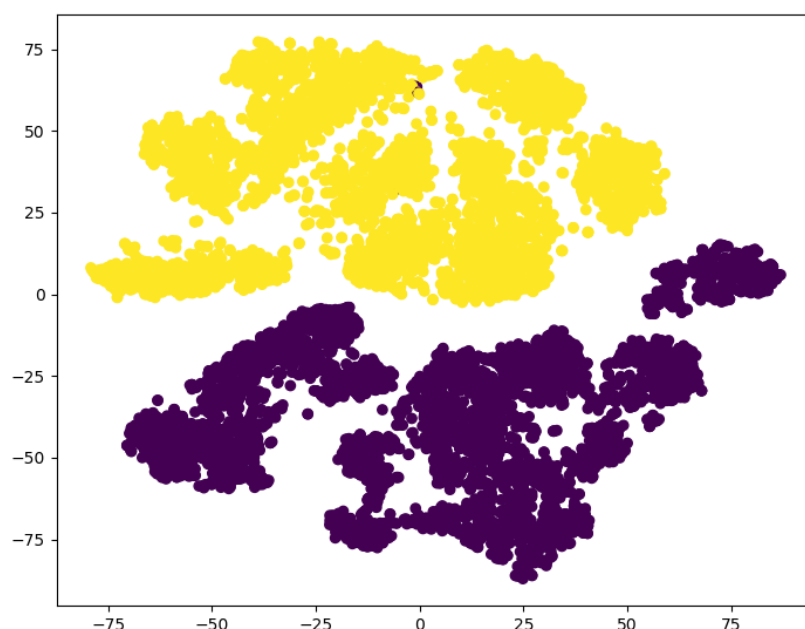
標點符號”「”和”」”相近且互相指向。

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)

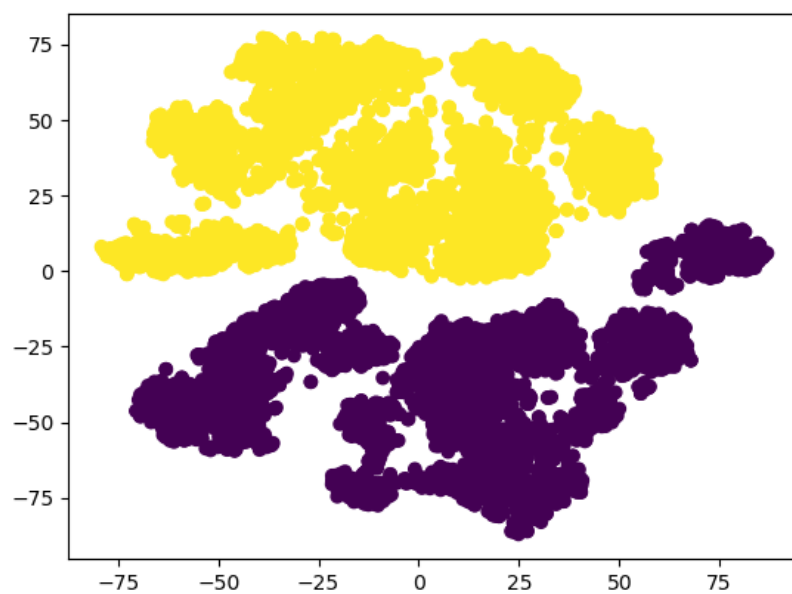
方法描述	Public set score	Private set score
TSNE all vectors to 2 dimension	0.02954	0.02910
Autoencoder+ DNN 128 64 epochs =10, k-means	0.52631	0.52531
Autoencoder+ DNN 128 64 epochs =200, k-means	0.96402	0.96237

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。



自己預測的結果在黃色區塊有一些預測錯誤，但大致上 **autoencoder** 有將兩個 **dataset** 分開，且預測正確。