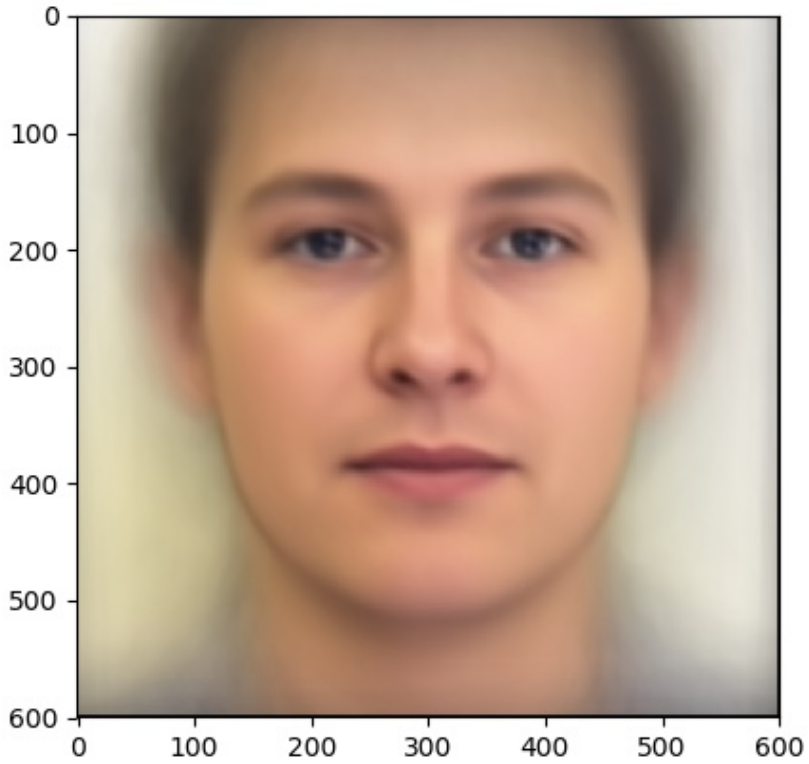
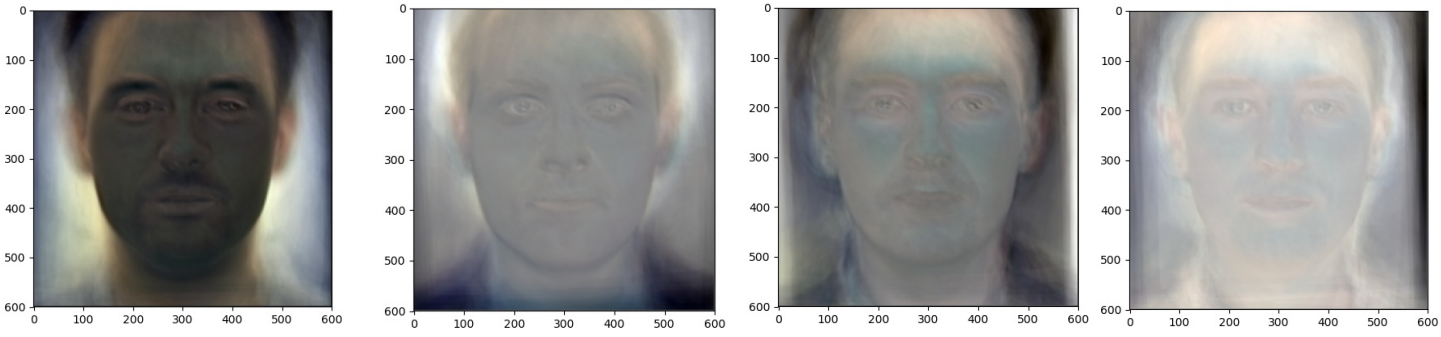
學號：B04901061 系級： 電機三 姓名：蔡忠紘

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。

<Images picked>: #5 #64 #112 #216

Image #5

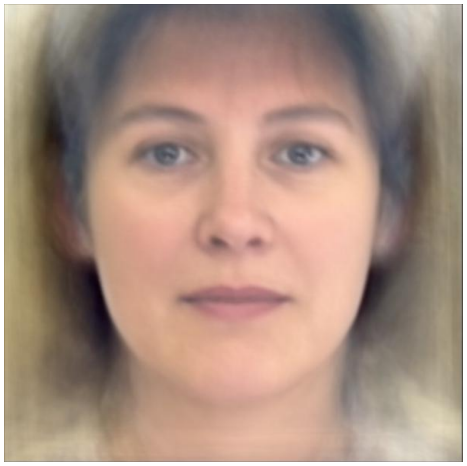


Image #64

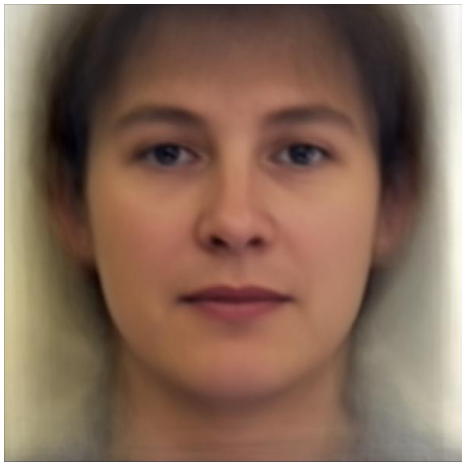


Image #112

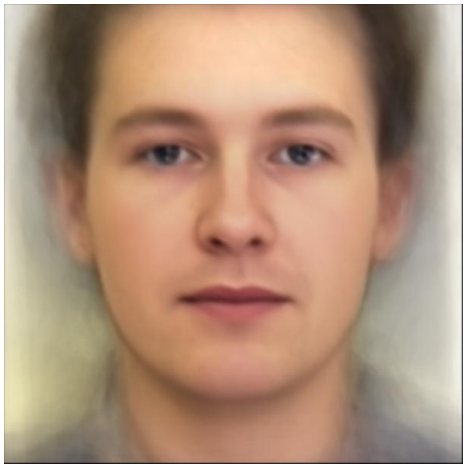
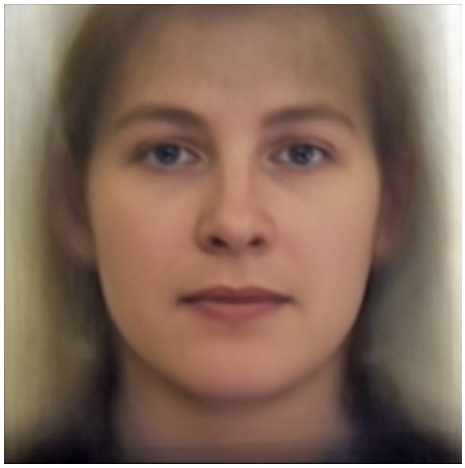


Image #216



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。

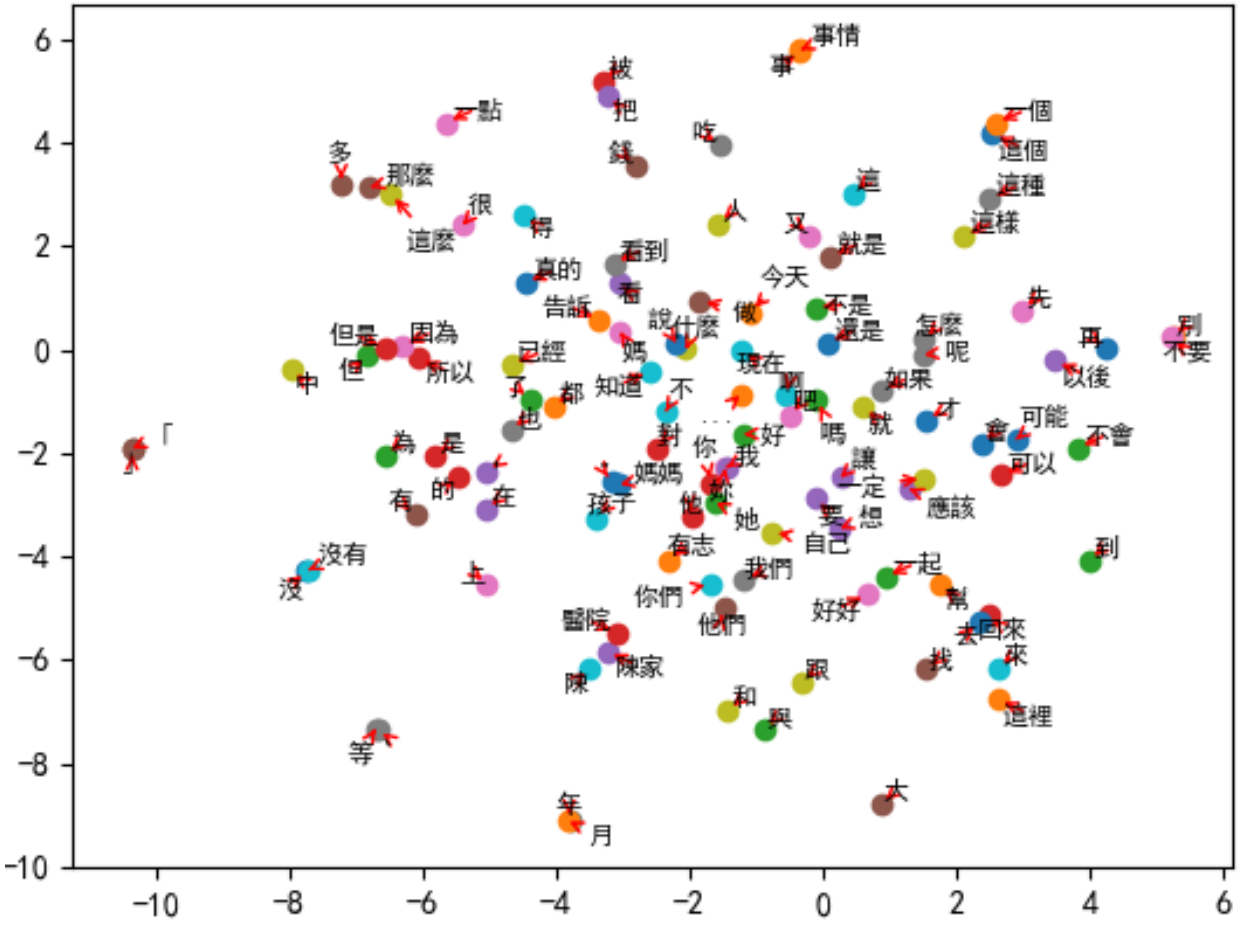
	#5	#64	#112	#216
Eigenface1	0.60	0.38	0.04	0.52
Eigenface2	0.21	0.25	0.11	0.22
Eigenface3	0.14	0.03	0.63	0.10
Eigenface4	0.05	0.33	0.23	0.16

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。

我使用的是「gensim」，size用300，min_count =1, K =4500。

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。

從圖中可發現，同一詞性、或性質相似的詞會在附近，例如：「這個、這種、這樣」、「沒、沒有」、「但是、但、因為、所以」、「可以、可能」、「別、不要」、「那麼、這麼」、「事、事情」。

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)

<Method 1>

先使用PCA將資料降到8維，再用TSNE將資料降到2維，KMeans的n_clusters設為2。

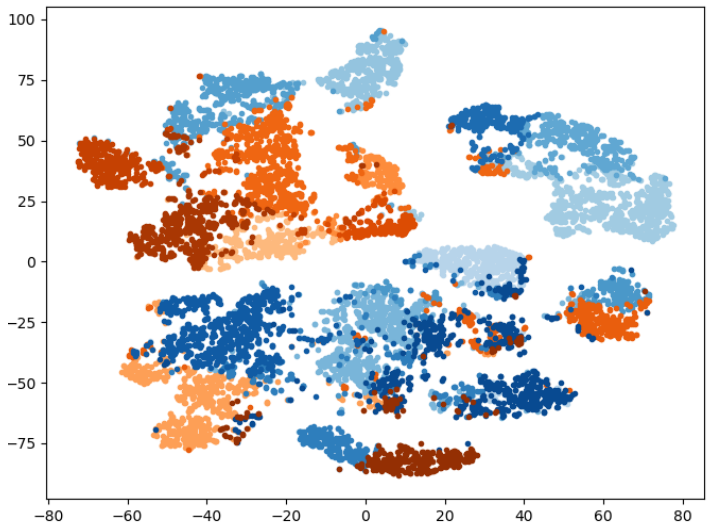
F1 score = 0.40276/0.40264 (private/public)

<Method 2>

用auto encoder將資料從784 -> 128 -> 64 -> 32 -> 64 -> 128 -> 784維，並將sample code最後一層的activation function改成'linear'。將KMeans的n_clusters設為20，逐一看各個cluster的圖片分別是衣服還是數字，如此便能更有效得分為兩大類。

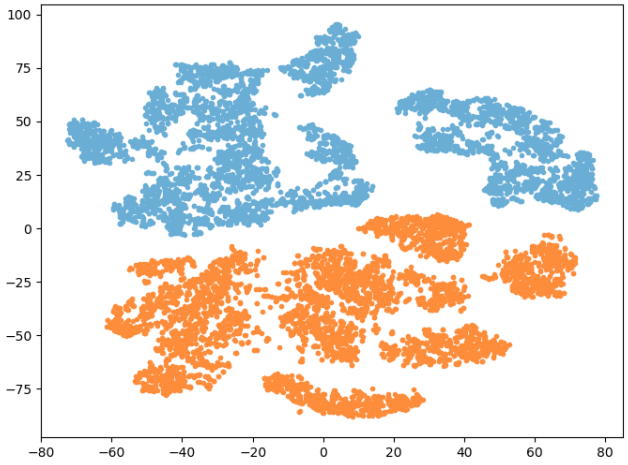
F1 score = 1.0000/1.0000 (private/public)

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



我的方法是將資料分為20個cluster，以不同顏色表示，但仍明顯看出中間有接近水平的分隔線將資料分為兩類。

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。



資料分布大致上與我的預測結果相同。