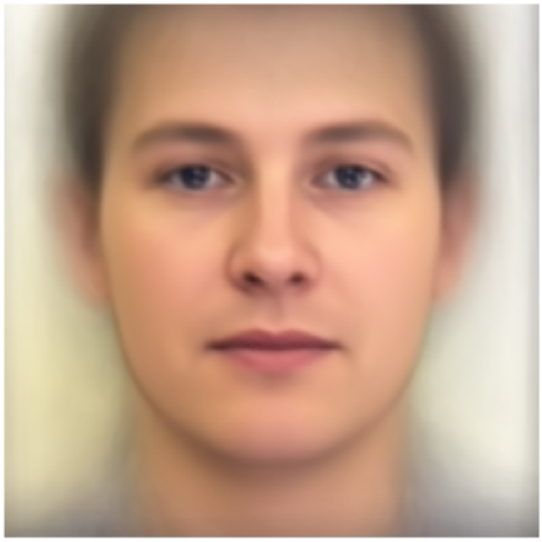
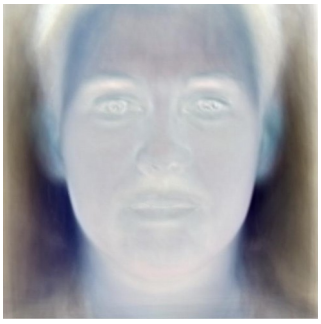
學號：F05942036 系級： 電信博一 姓名：羅翊展

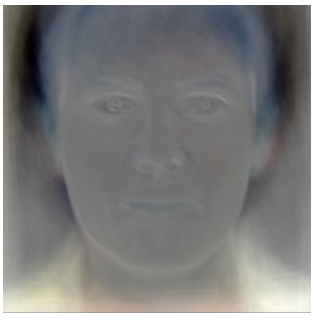
A. PCA of colored faces

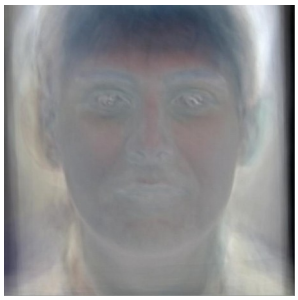
A.1. (.5%) 請畫出所有臉的平均。

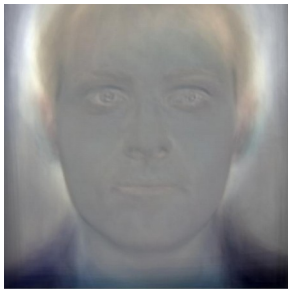


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。（右上、左上、右下、左下）



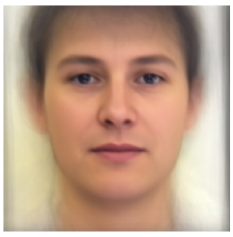
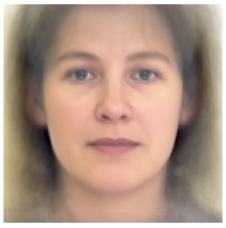
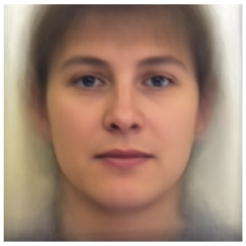
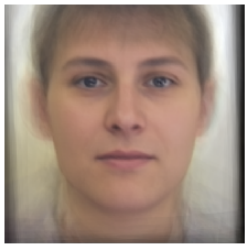






A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

下圖分別是第 100 200 300 400 張還原的結果（右上、左上、右下、左下）

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。

我是取 img100 200 300 400 來算 eigenfaces 的比重，如下：

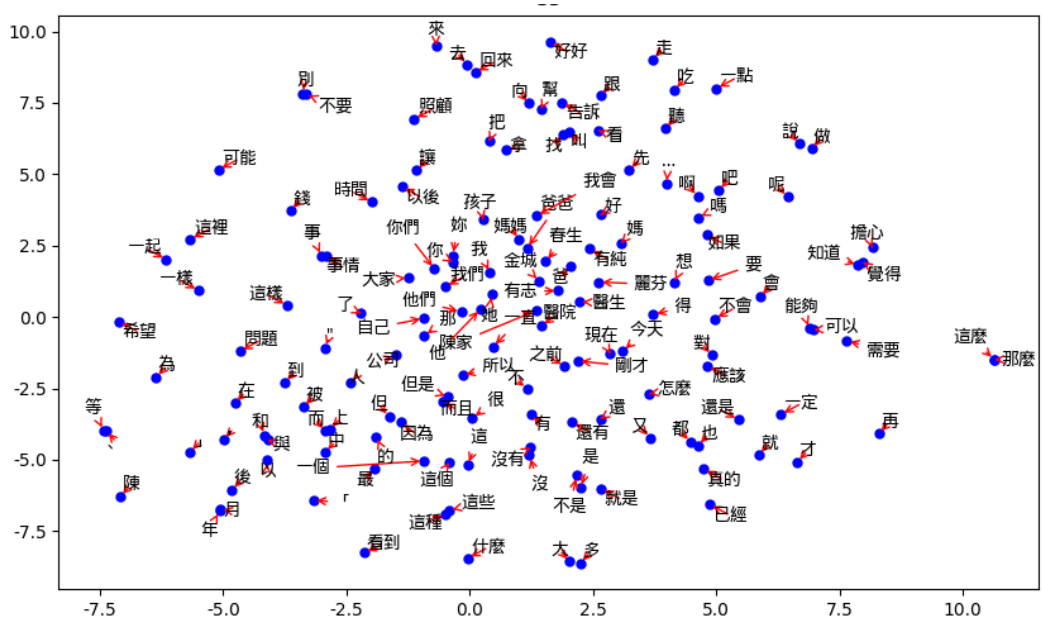
-24.9%，46.9%，85.9%, -7.9%
105.2% ，23.0%, -14.7%, -13.3%
46.8%, -10.9%，23.8%，40.2%
41.1%, -127.6%，265.4%，-78.8%

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。

我用的套件是 gensim 和 jieba 來處理 word2vec 先利用 jieba 來斷句，在用 gensim train 出 word2vec 的 model 我的 size 維持 default 的 100 然後 mincount 設成 4500(剛好是投影片中的中間值)

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。

可以發現相近的字會比較靠近，例如右邊的’一起’’一樣’，他們就靠得很近，意思相近的也可以由箭頭的方向得知，例如左側的’這麼’’那麼’，他們的意思幾乎是一樣的

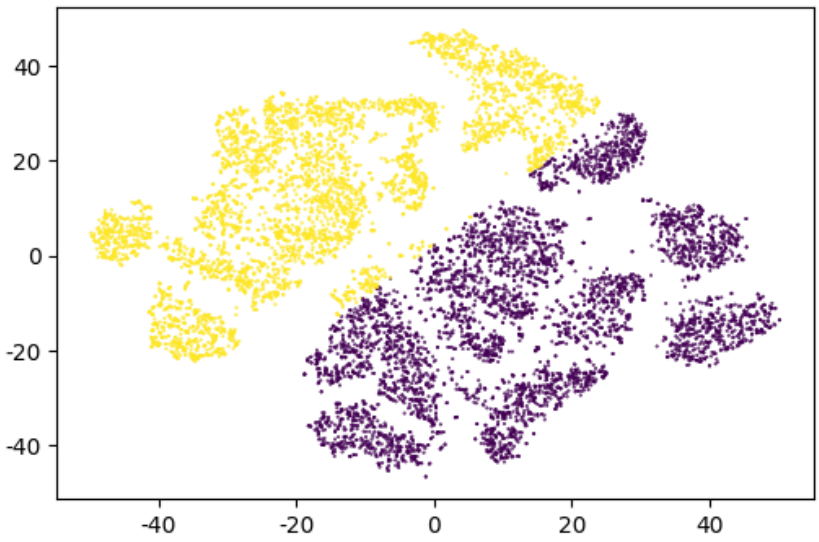
C. Image clustering

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)

我比較了兩種降維的方法：auto encoder 和 pca 在 kaggle 上的結果分別是 0.97 和 0.04(public 和 private 的結果差不多)，而 cluster 的方法則都是用 kmeans，猜測原因是雖然用 pca 和 autoencoder 都是降到 200 維但 autoencoder 可以有效的抽取

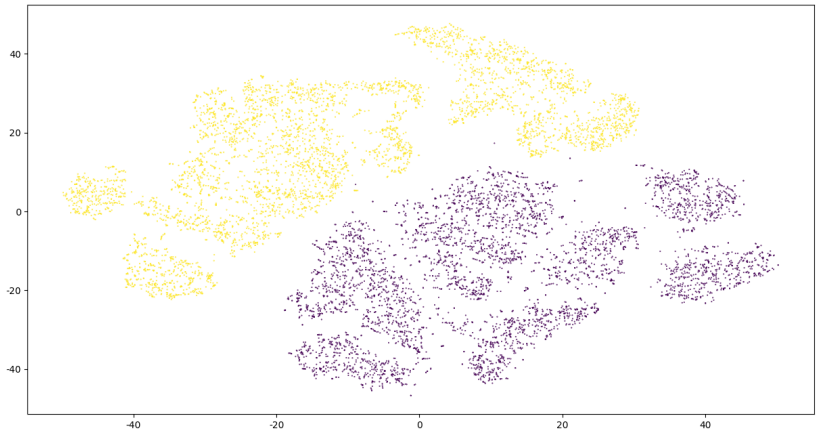
出兩的 data set 最大不同處的 feature，pca 就只能得出變化最大的前 200 個，顯然決定兩個 cluster 怎麼分的重要 feature 無法抽取的很好。

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。

ground truths



我是用 tsne 直接降到 2 維來做，可以發現，跟 ground true 比較有一部分的紫色被預測錯的（變成黃色的）。