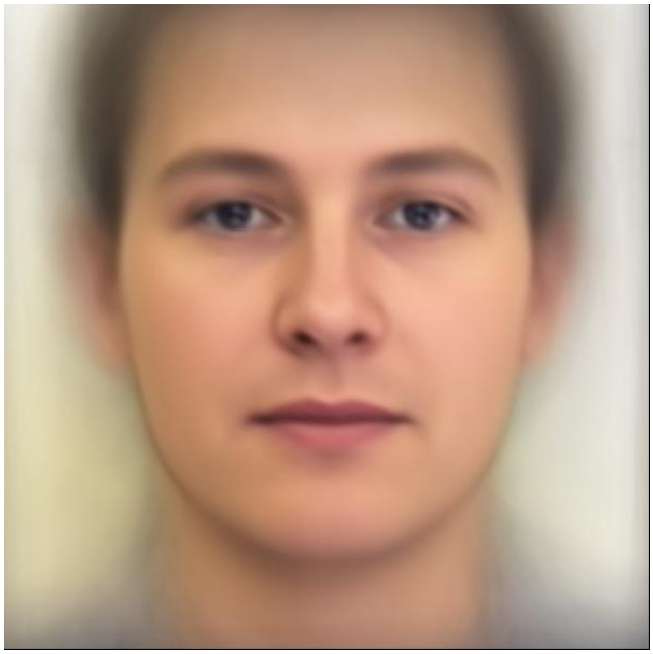
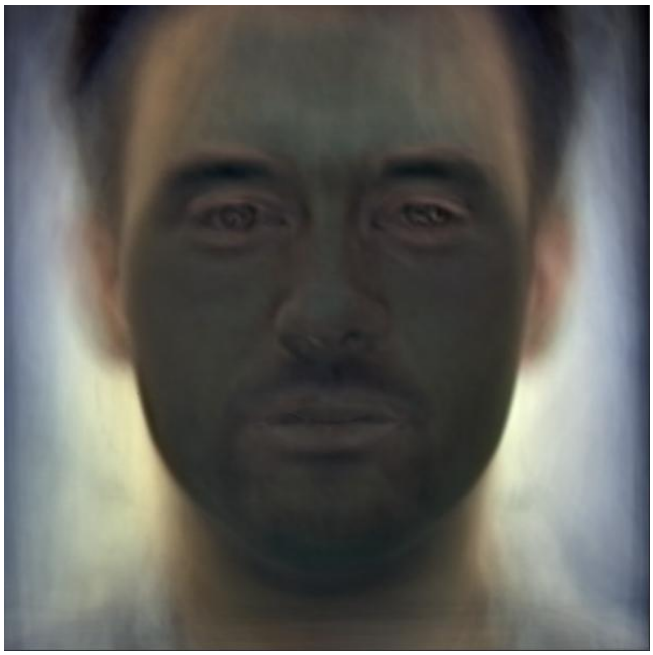
學號：R05942017　系級：　電信碩二　姓名：黃梓鳴

A. PCA of colored faces

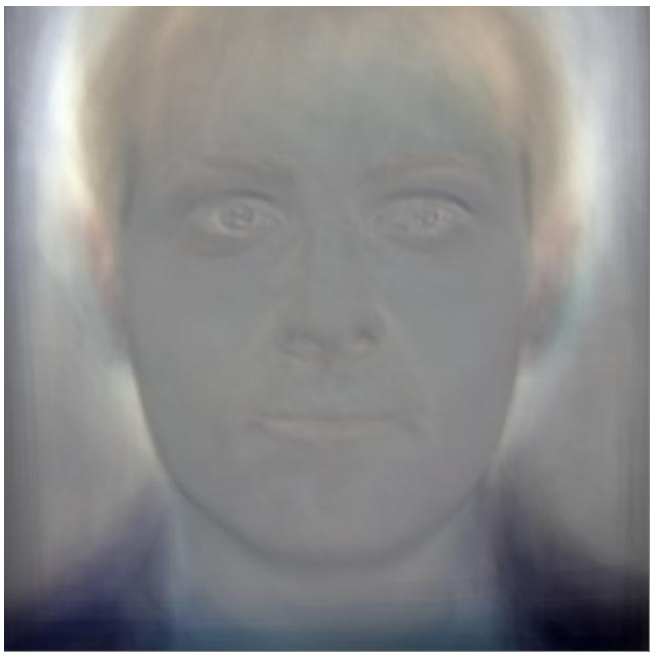
1. (.5%) 請畫出所有臉的平均。



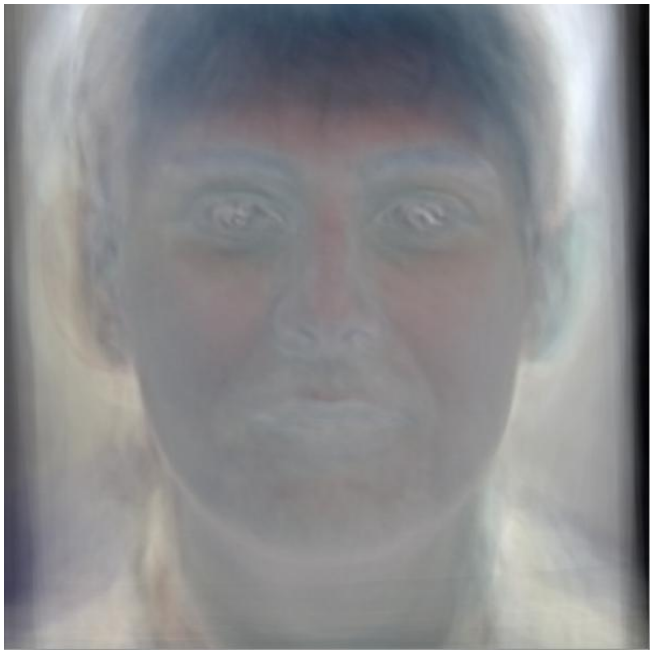
2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。
   第一個:



   第二個:

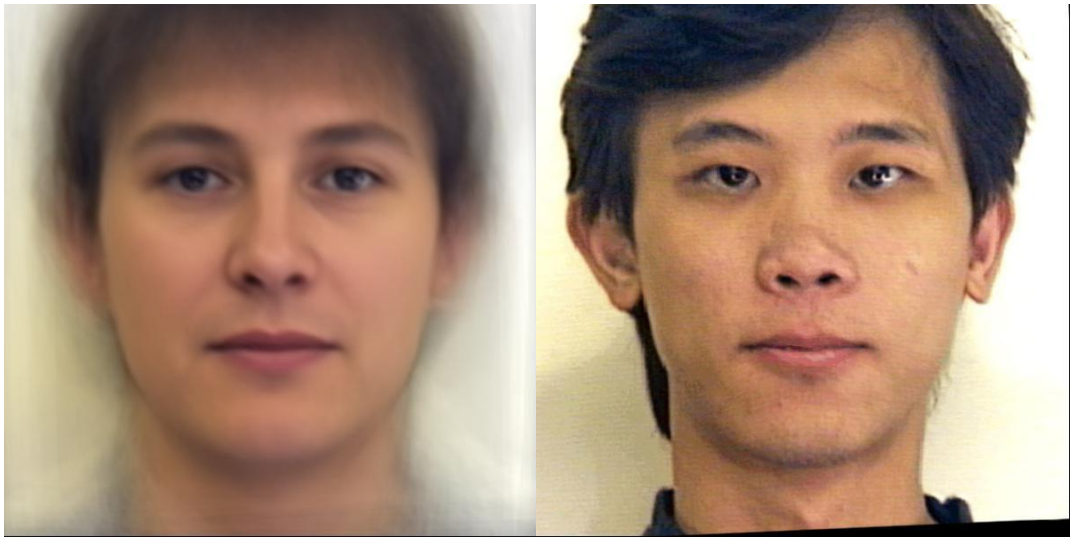


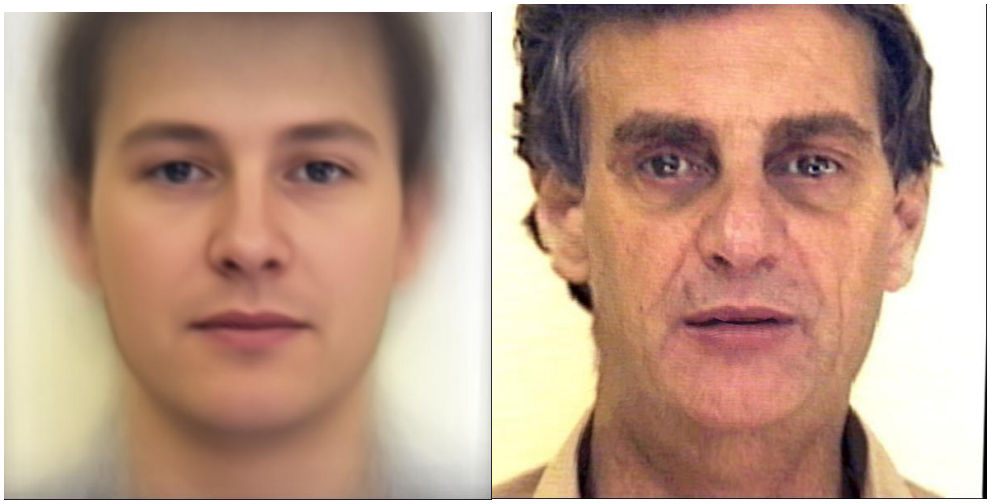
第三個:

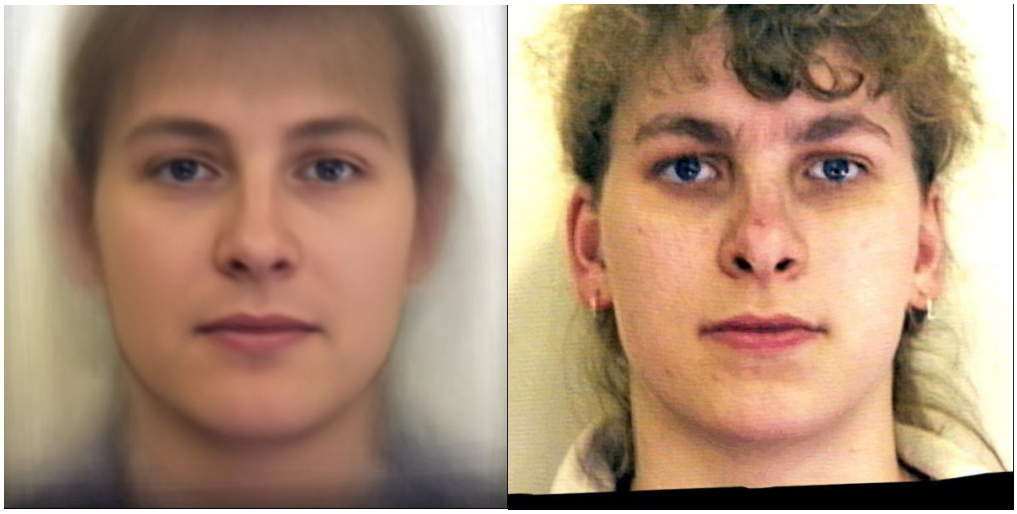


第四個:

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。







4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

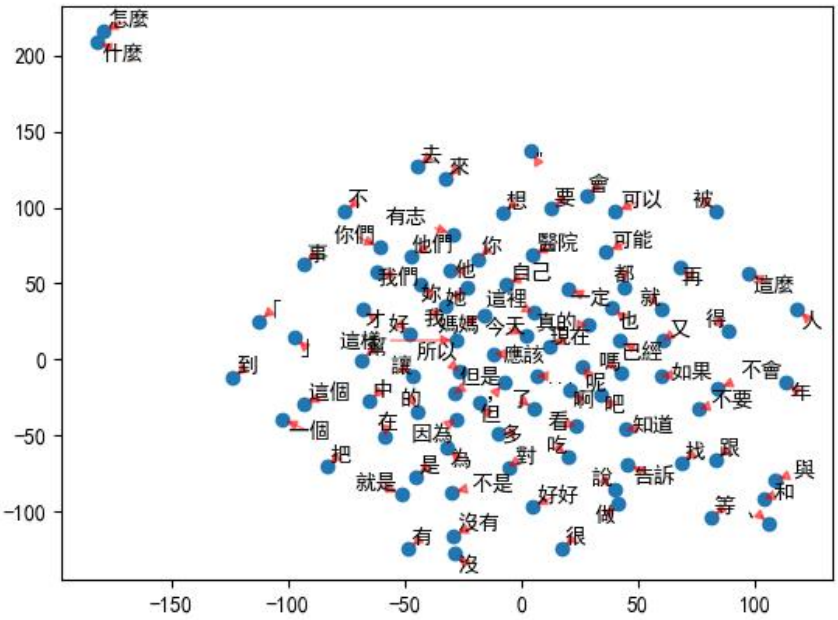
第一大的再前四大裡佔的比重為:4.1%,第二個 2.9%,第三個 2.4%,第四個 2.2%

B. Visualization of Chinese word embedding

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。

我是用 gensim，全部都用 default 值(min_count=5 即最少出現次數,size=100 即維度為 100)

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



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

發現相關的字都會聚在一起，如(怎麼、什麼)都在右上角，(你們、我們、他們)都在中間偏左上的位置，可見相關的字其向量值會很相近

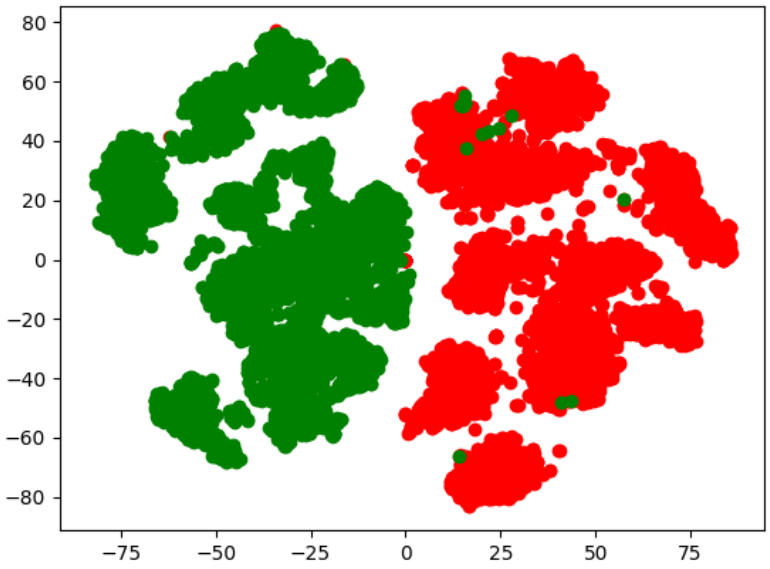
C. Image clustering

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)

我利用 autoencoder 疊 DENSE 層降到 32 維再經過 KMEANS，可以在 Kaggle 上得到 0.81 的分數

而再利用 autocencoder 疊 CNN 層，降成 7*7*32，壓平後進 KMEANS，在 Kaggle 上只有得到 0.03357 的分數

2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。
   我是先經過自己 train 的 autoencoder 降到 32 維，再利用 TSNE 降到 2 維，紅綠代表兩種不同的分類



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。

   降維方法跟上題同，可發現我自己預測的 label，有些被判成綠色的其實是在紅色的區域，而紅綠兩者之間有個很明顯的分界