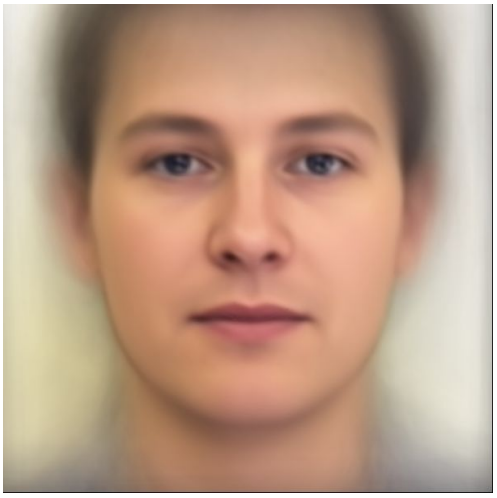
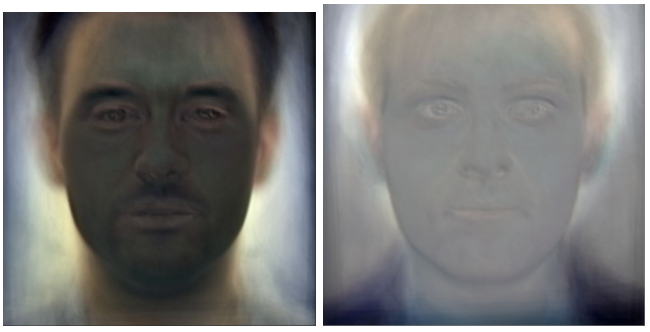
學號：B03902084 系級： 資工四 姓名：王藝霖

A. PCA of colored faces

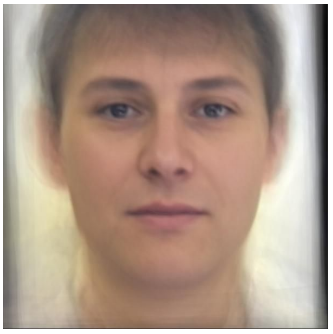
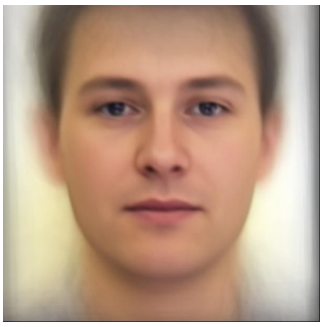
A.1. (.5%) 請畫出所有臉的平均。

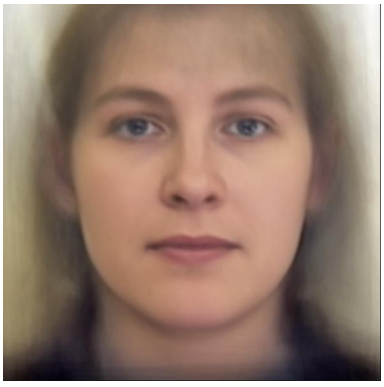
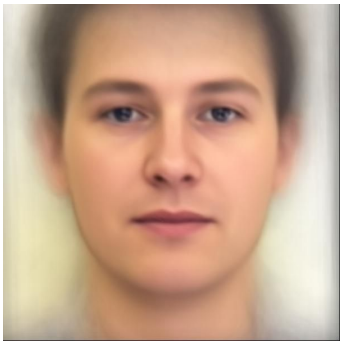


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



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。

52:  104: 

222:  333: 

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。

1: 35.5% 2: 25.2% 3: 20.4% 4: 18.9%

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。

gensim，調整了size也就是vector的大小，我調成2因為想畫2維

min_count=0，也就是在0以上的字才會算數

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

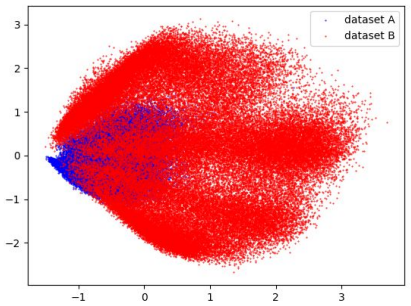
B.3. (.5%) 請討論你從 visualization 的結果觀察到什麼。

C. Image clustering

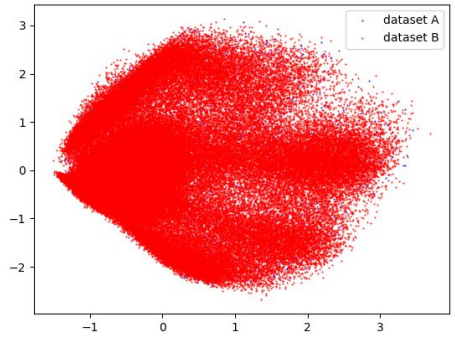
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

1. pca with whiten, kmeans: kaggle 分數 = 0.999
2. pca without whiten, kmeans: kaggle 分數 = 0.14186

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。



因為我用的是降到300維的pca的前兩維的結果，因此可能看不出來太明顯的兩個class，但可以看到我的預測和標準答案是大致相同的，也因此有0.999的高分