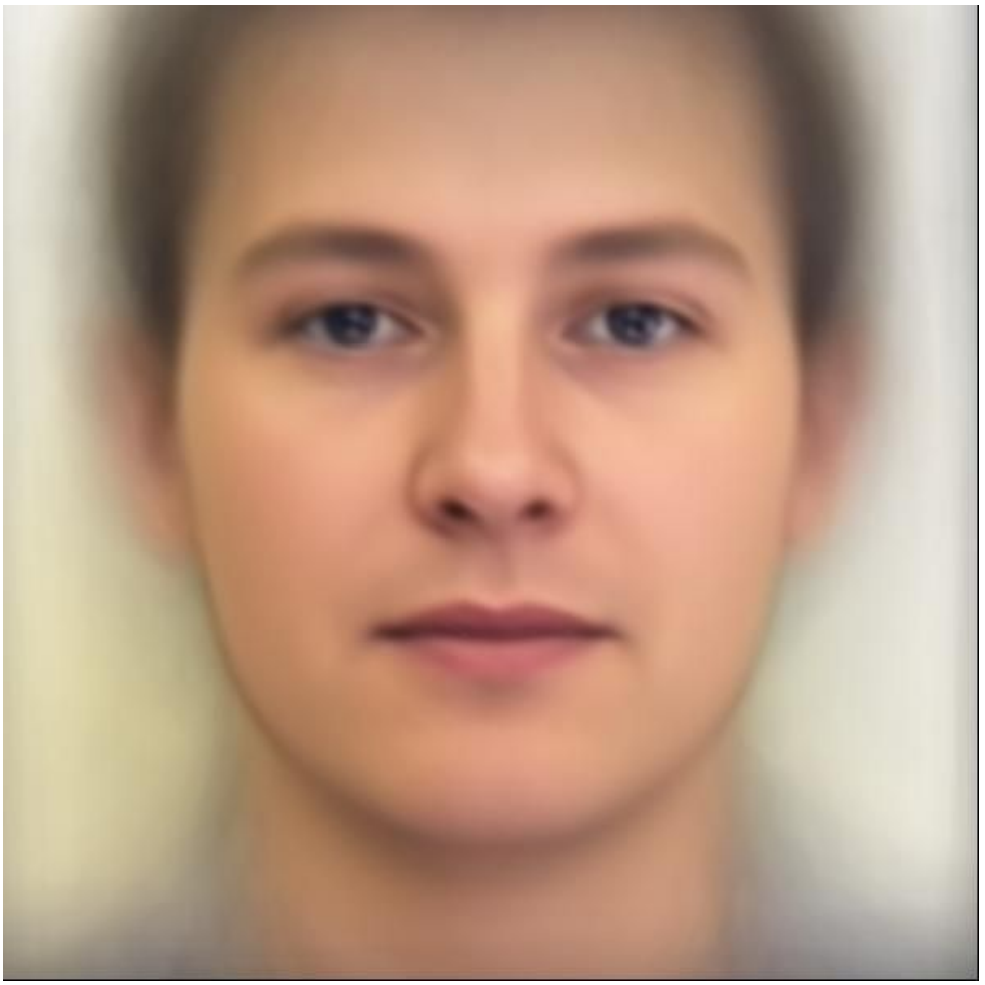
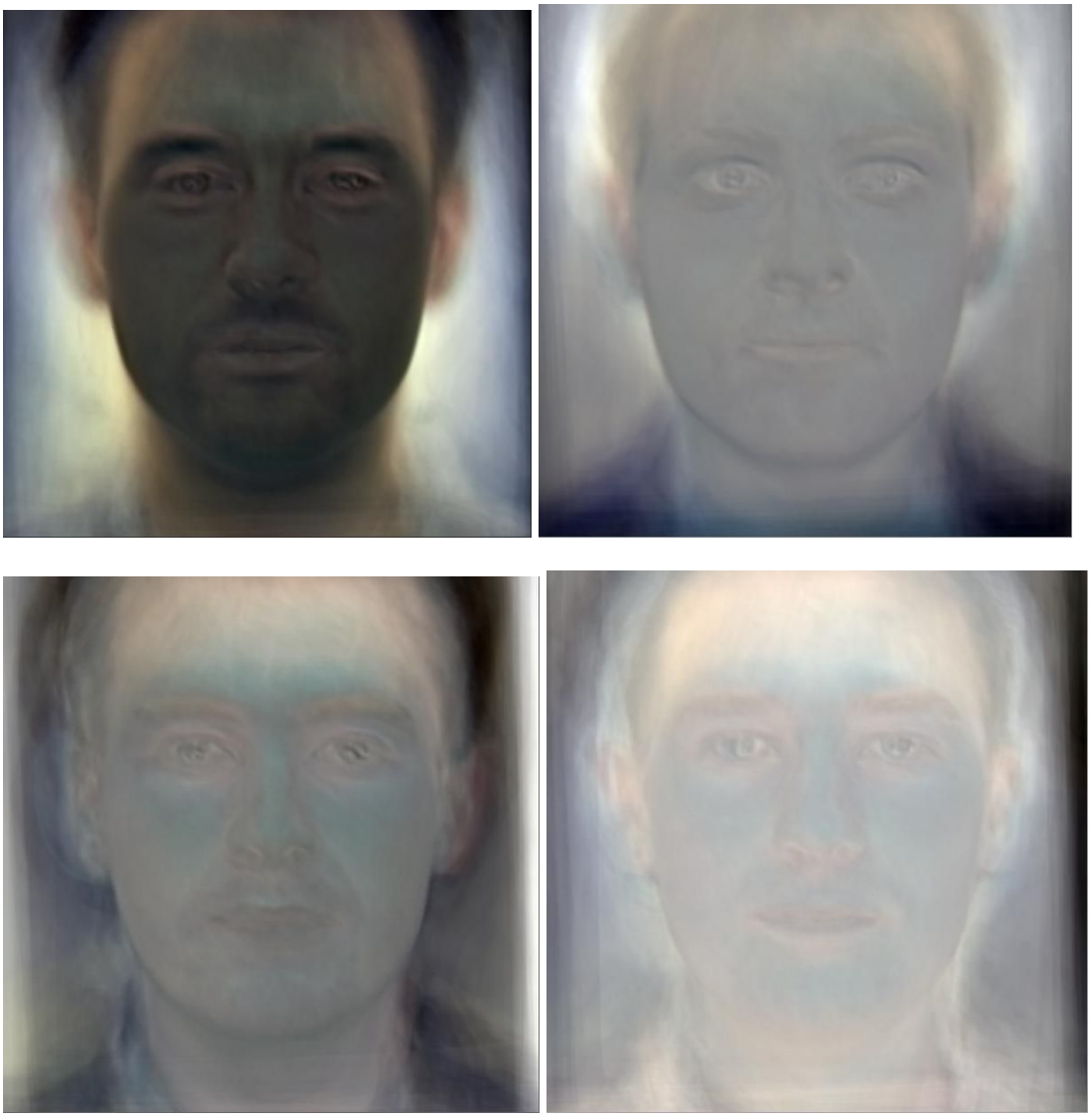
學號：R06922128 系級： 資工碩一 姓名：楊碩碉

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

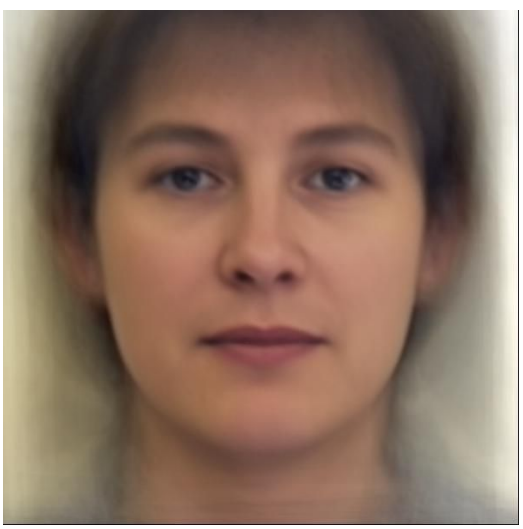
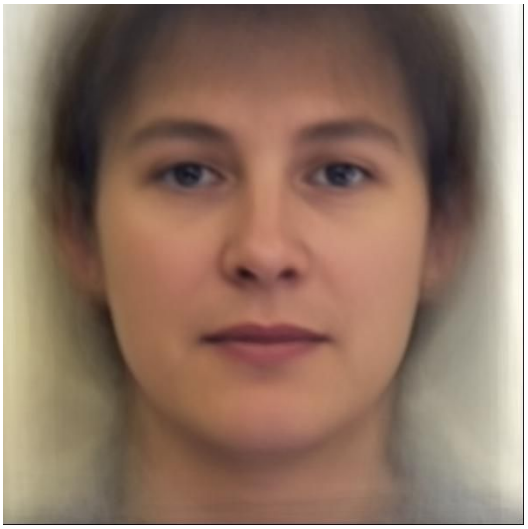


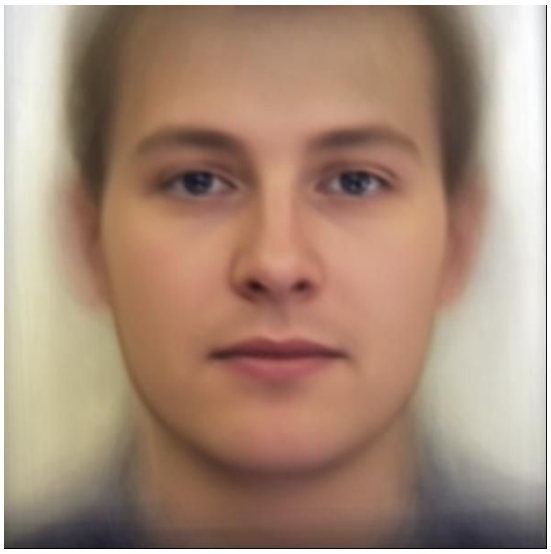
A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



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。

採用前4張圖片 (0~3.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。

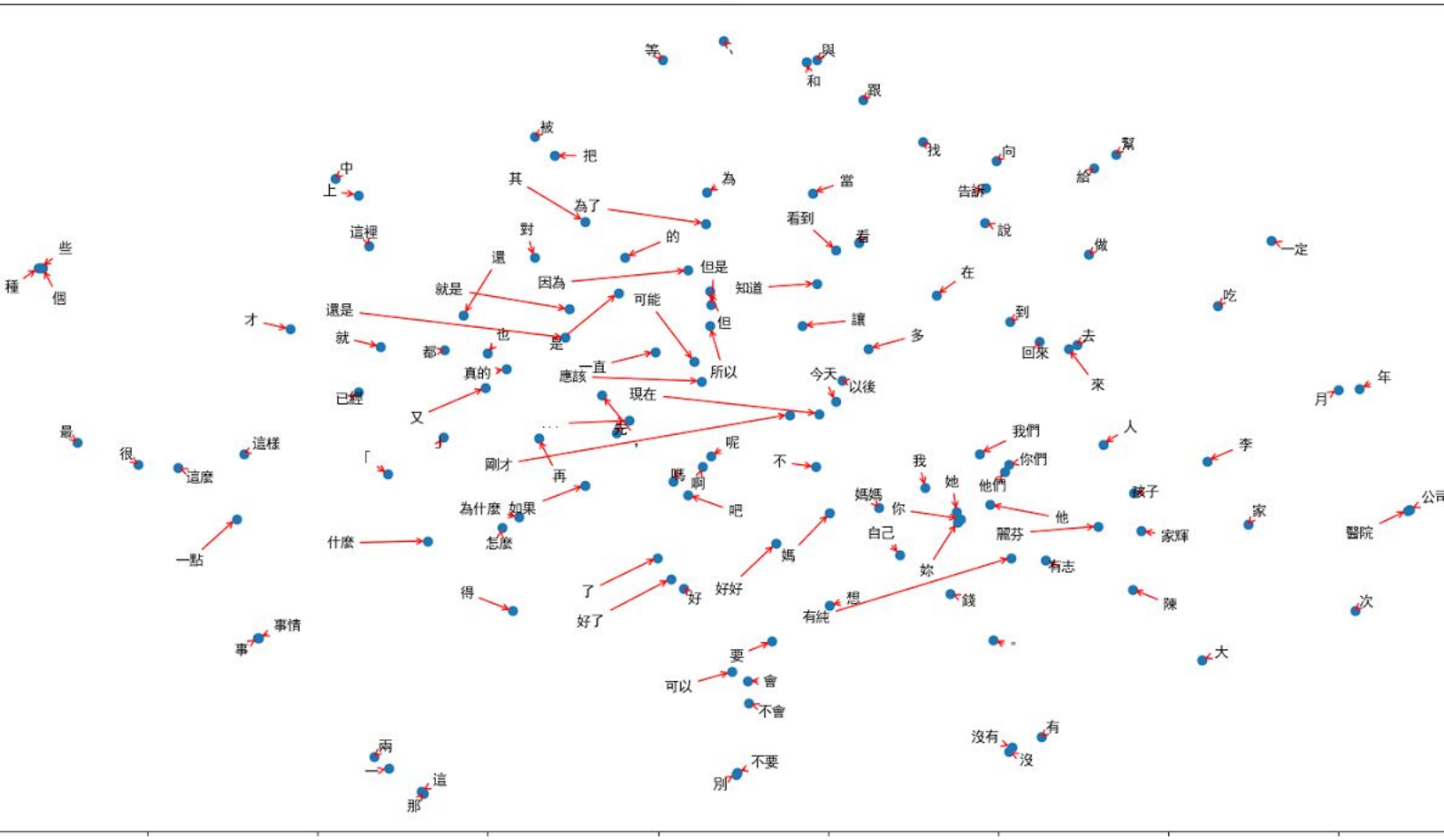
4.1%, 2.9%, 2.4%, 2.2%

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。

使用gensim的Word2Vec來訓練word wmbedding，維度設為128，方法為CBOW，mincount設成10。

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。

「沒有」介在「沒」跟「有」之間，「事」跟「事情」幾乎相疊，「我們」、「你們」、「他們」的位置相近。可以觀察出同樣類型的詞會聚在一起。

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)

第一種：dnn autoencoder，維度降到128，kaggle上可達到0.992

第二種：cnn autocoder，維度一樣128，但是kaggle上只有0.027

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

0為第1類，1為第2類 (Tensorboard作圖)

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。

同上圖，可以將2群完全分開。