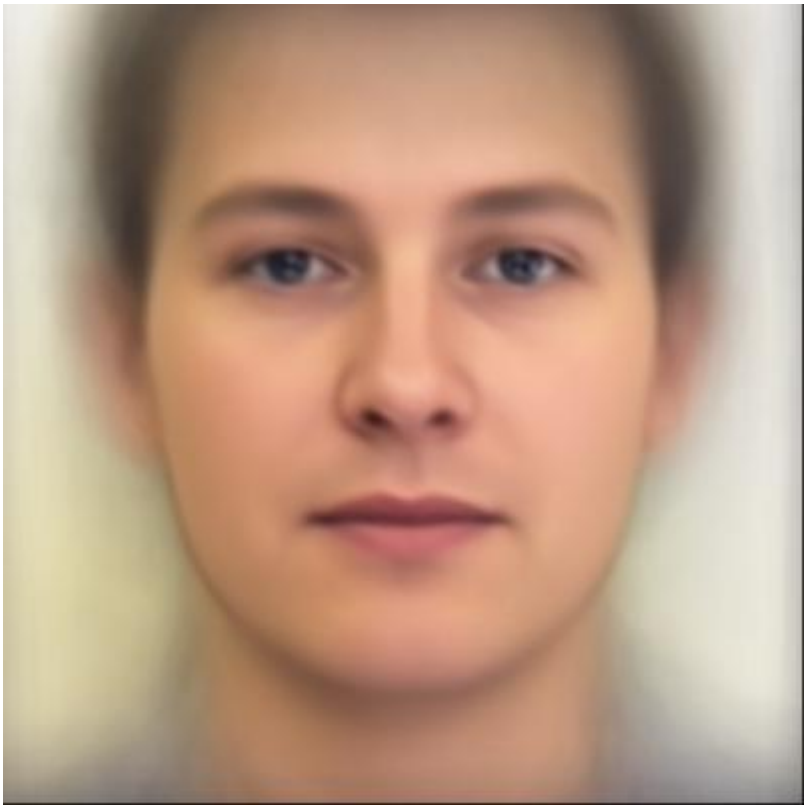
學號：R04921040 系級： 電機碩一 姓名：李中原

A. PCA of colored faces

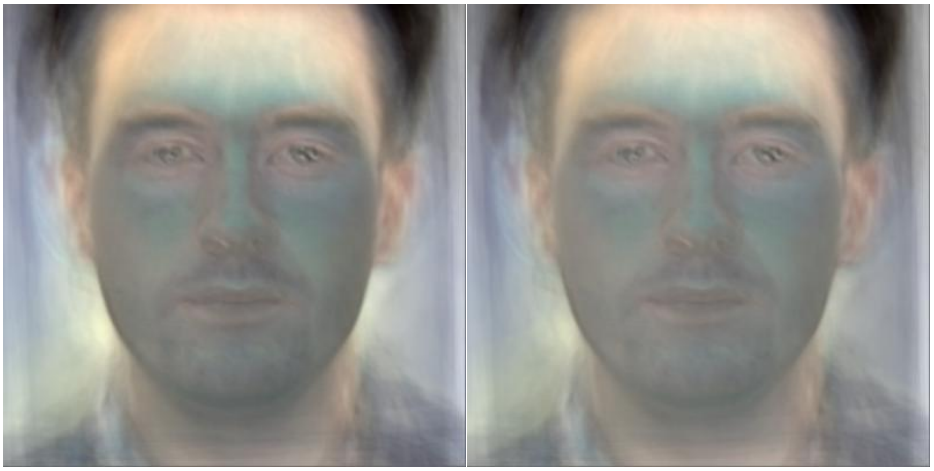
A.1. (.5%) 請畫出所有臉的平均。



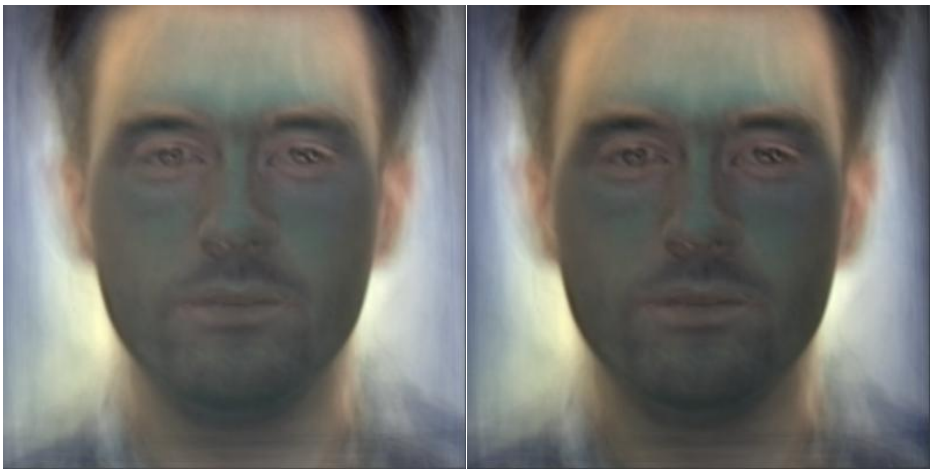
A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

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。

111.jpg: 247.jpg:



303.jpg: 355.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。

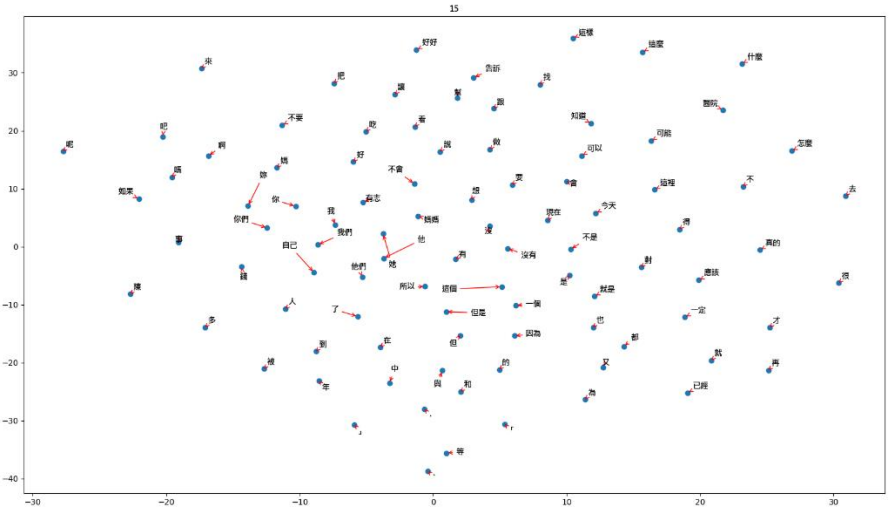
4.2%、3.0%、2.4%、2.2% (resize 成(400 ,400 , 3))

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，min_count = 4700，vec_size = 128，只找出出現超過 4700 次以上的 word 並 mapping 成 128 維的 vector

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。

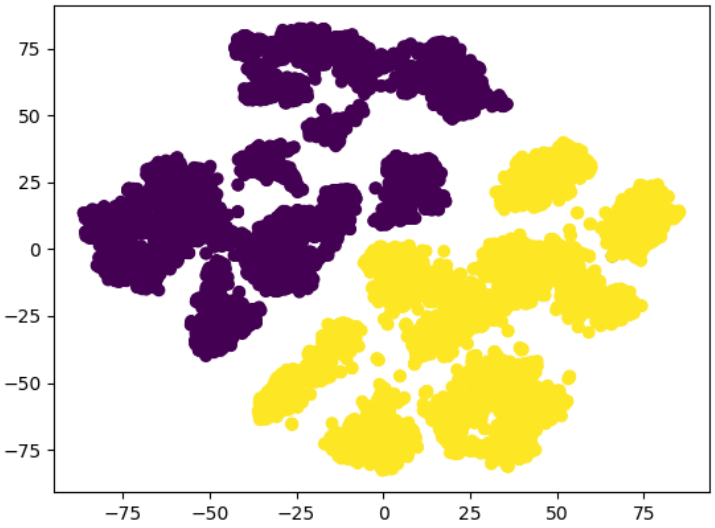
整體來說沒有特別密集的分布，不過像代詞或語助詞等相似詞性的會比較聚集在一個區域

C. Image clustering

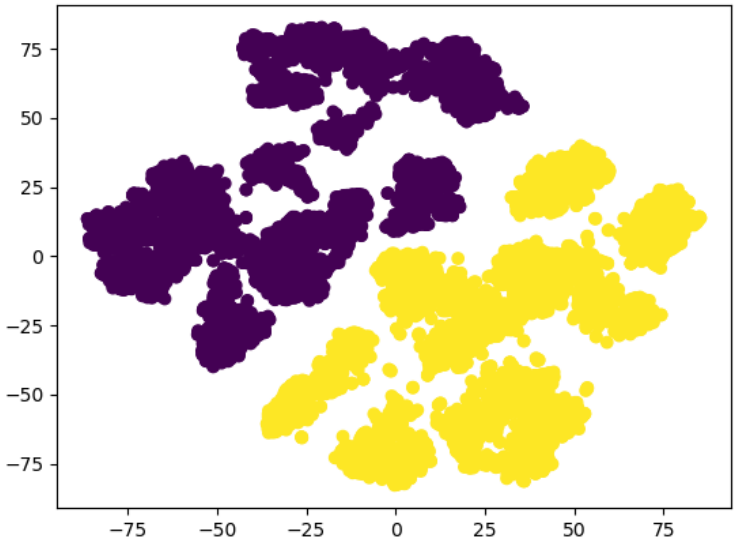
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

有試過幾個不同維度結果的，大致在 30 多維會有較好的結果 (0.978)，16 維的結果大約在 0.815 左右也算不錯，不過要表現特徵的話維度太少了一點。另外有嘗試用先升維到 1024 維再降維的結果，基本上都完全 train 不起來，結果只有 0.1 不到，cluster 都是使用 Kmeans

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。



跟預測的結果看起來幾乎一樣，統計後兩個 cluster 分別有 5001、4999 個，只有一個被分錯，算是非常準確的分類。