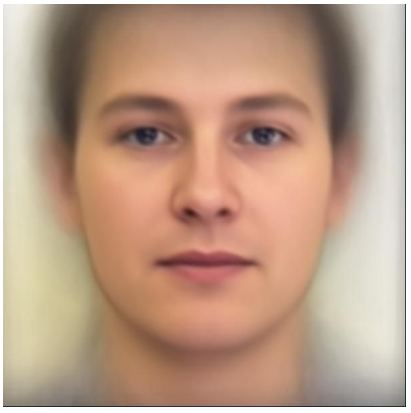
學號：B04501002 系級： 土木三 姓名：謝至宥

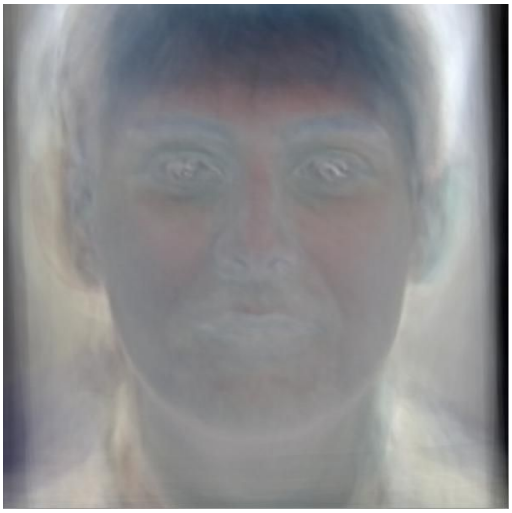
A. PCA of colored faces

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。

這些由左至右，由上而下依序就是前四大eigenvalue對應的圖



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。

原圖依序是403、58、46、325

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。

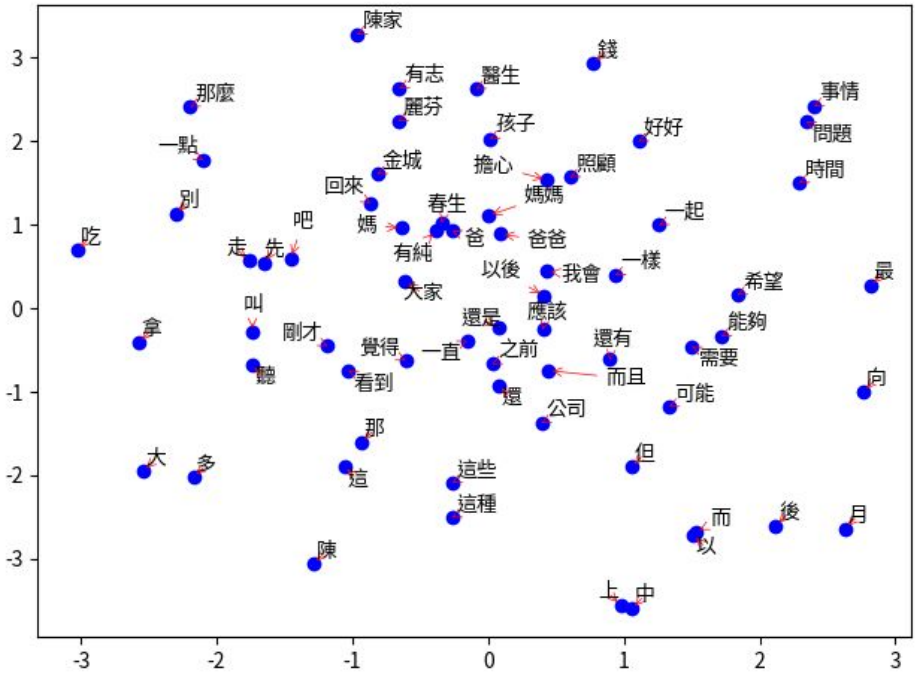
前四大比例依序為：4.1%、2.9%、2.4%、2.2%

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。

我使用jieba分詞，用gensim算word embedding，embedding size設為250，再用TSNE降成兩維。

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。

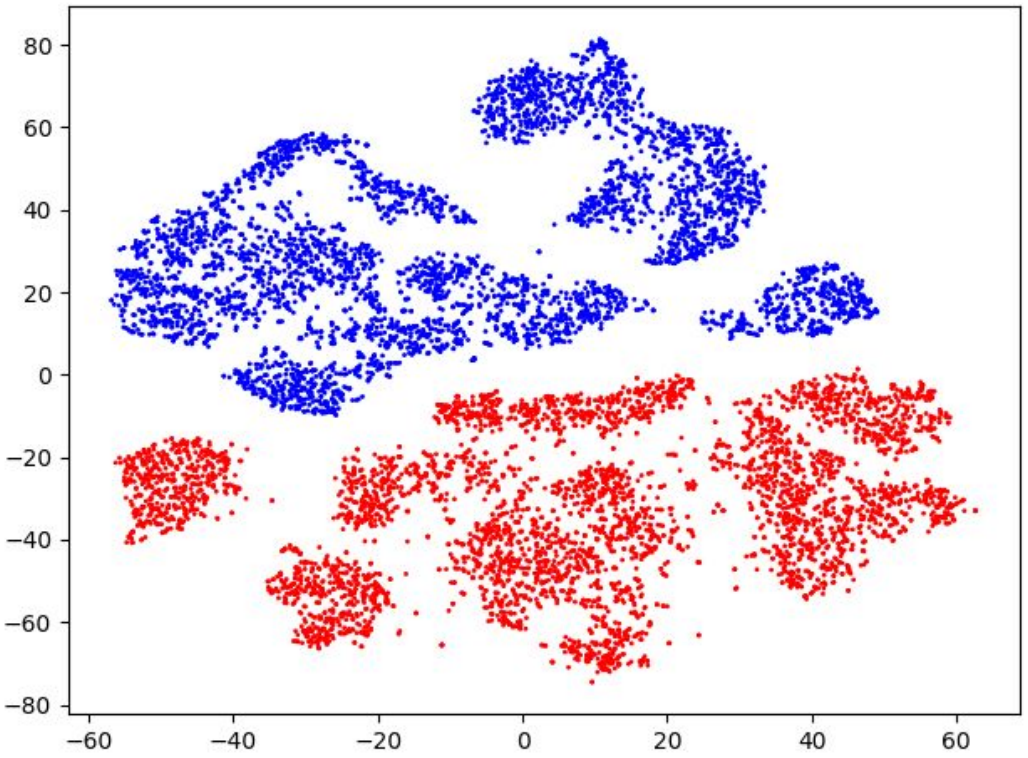
從圖中可以看到人名被聚在一起，人的稱謂也離名字很近，相同詞性的會相鄰，意思類似的也在一起，像是「大」和「多」、「這些」和「這裡」等。

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)

我用了DNN的autoencoder和CNN的autoencoder，經過autoencoder再用kmeans分成兩類。我檢查前10張圖片來看正確率，用CNN的結果比DNN還要糟糕，所以就上傳DNN的預測，在kaggle上拿到0.5的分數。後來分別試了3到7層的DNN autoencoder，發現四層表現的比較好，再實驗依序抽第一層到中間層的feature來做kmeans分類，結果前面幾層的表現都比較好。最後用DNN疊四層，dimension依序是512、256、128、64，再使用512的輸出做kmeans，得到kaggle上1.0的分數。推論這次的資料圖形都較簡單，只需要用基本的feature就可以分好類，中間層的資訊比較反而比較不適合。

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。

預測結果完全正確，所以輸出的圖和前一題一樣