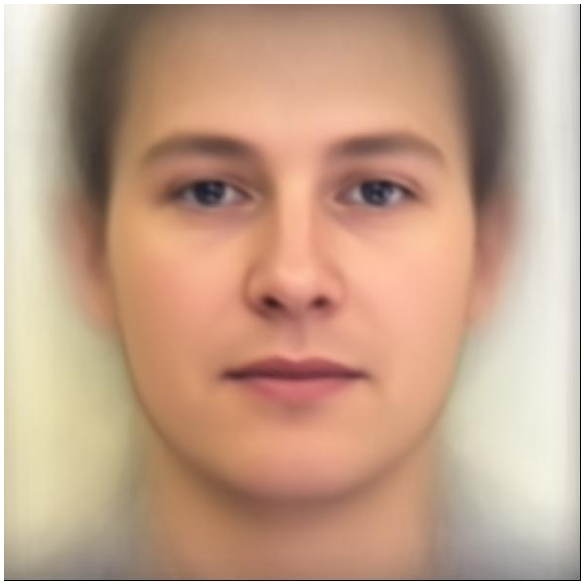
學號：R06723025　系級： 財金碩一　姓名：林耘灣

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。



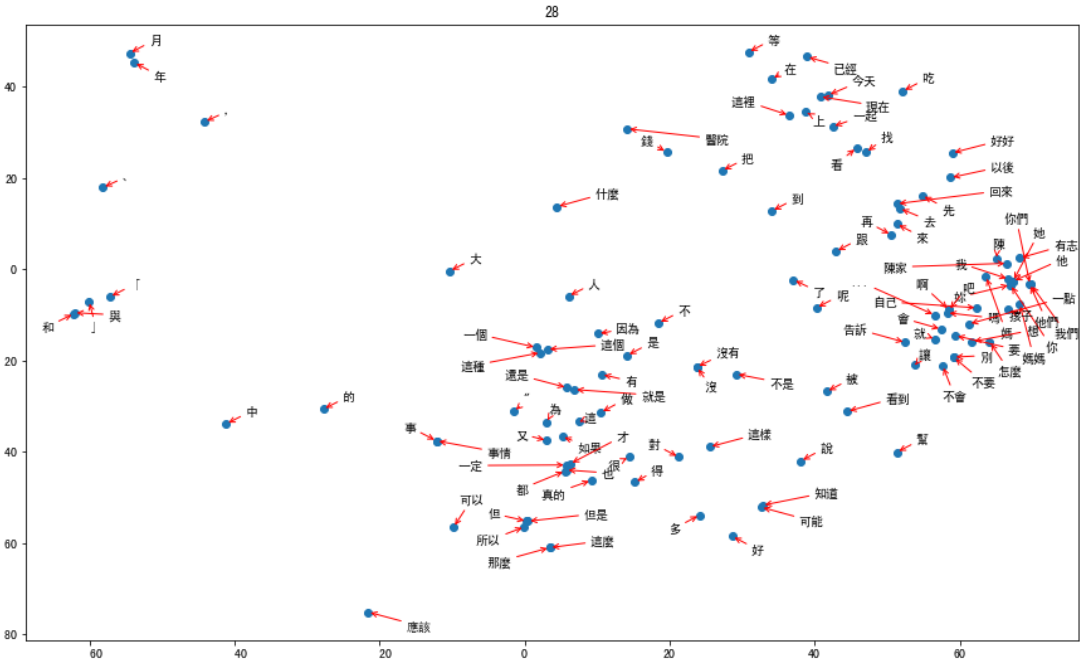

3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，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%

B. Visualization of Chinese word embedding

(collaborators: r06922075 翁瑋)

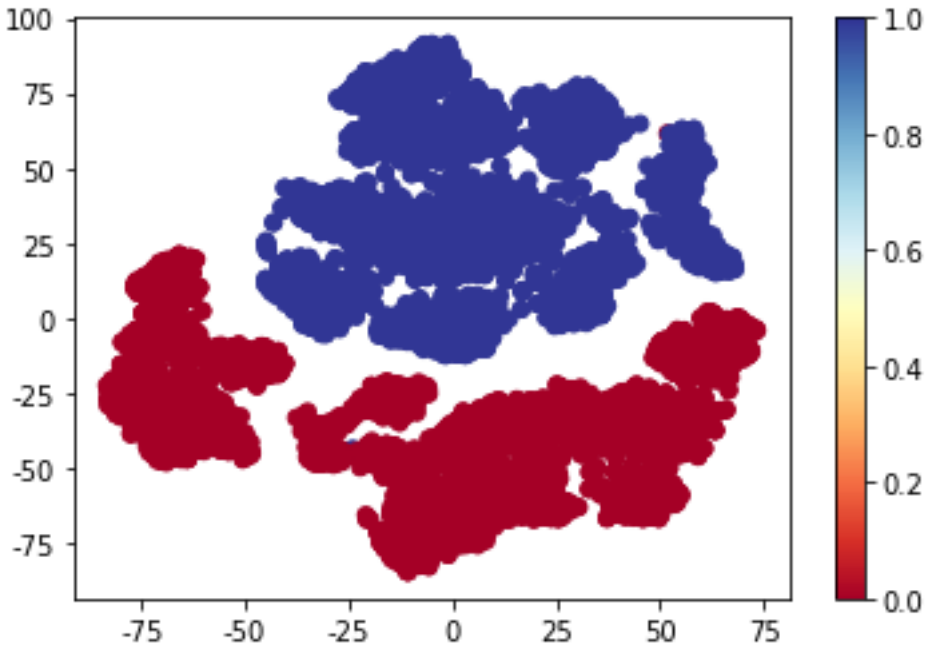
1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。
   使用 gesim 中的 word2vec，調整成為用 100 維代表一個向量。
   model = word2vec.Word2Vec(txt, size=100)
   並且使用 TSNE(n_components = 2)
   將資料降到兩個維度。
2. (.5%) 請在 Report 上放上你 visualization 的結果。



3. (.5%) 請討論你從 visualization 的結果觀察到什麼。
   可以發現相同意思的字還是分得比較類似，比方說月年、和與…。

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)
   利用 autoencoder 加上 kmeans 可得到 1 的結果，但我忘記存這個 model，後來 train 的只能得到 0.98~0.99 的結果。
   利用 pca 加上 kmeans 只能得到 0.05 的結果。
2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



從上圖可以發現模型成功將資料分成兩類。

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。

   利用 kmeans 在做 cluster 後可發現前 5000 筆均預測為 1，後五千筆均預測為 0，與上圖的分類相符。