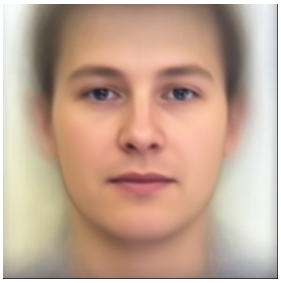
學號：R06921066　系級： 電機碩一　姓名：劉宇閎

## PCA of colored faces

(.5%) 請畫出所有臉的平均。



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

(.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。

在此題進行實驗時發現使用少量 eigenface 做重建時出現的圖會激似平均臉，但在使用更多 eigenface 做重建可以得到幾乎原圖的效果。

(.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。

W1 = 21.7%, W2 = 2.7%, W3 = 2.4%, W4 = 1.8%

## Visualization of Chinese word embedding

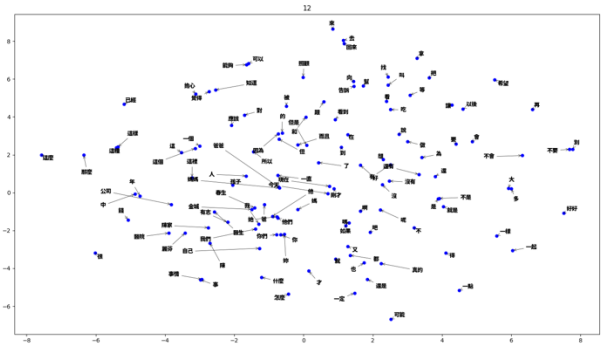
(.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

使用 gensim 套件並調整以下參數：

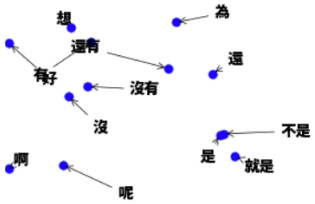
'size': 向量的維度

'min_count':限制最少的出現次數

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



(.5%) 請討論你從 visualization 的結果觀察到什麼。



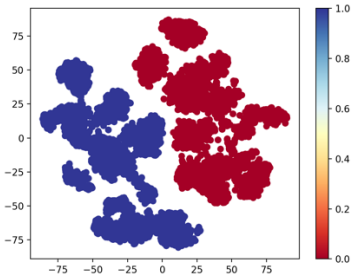

截出上圖的一部分可以看出有相近意義的詞分佈會相當接近，代表他們在高維空間中應該會被分別得更加清楚。

## Image clustering

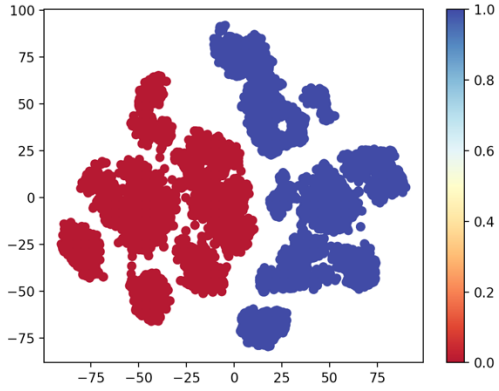
(.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

曾使用 C N N來 train autoencoder 但效果並不彰顯，得到的 F1-score 為 38.3%推測一種原因為雜訊過多，因圖檔的重點元素所佔比例較小故推測在使用 CNN train 的時候可能會需要使用更加特別的技巧，但後來是使用 DNN 並且用高達 2500 個 epochs 去做 training 竟然得到了幾乎 100%的準確率，所以也有可能是因為訓練次數仍然不足所導致，因硬體設備不足所以沒有將 CNN 也使用大量 epochs 去做嘗試。

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




(.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。




使用自己預測的值加上正解作圖後可以發現這次的 training 非常的成功，將兩筆不同 label 的資料區分開來且沒有模糊不清的區塊