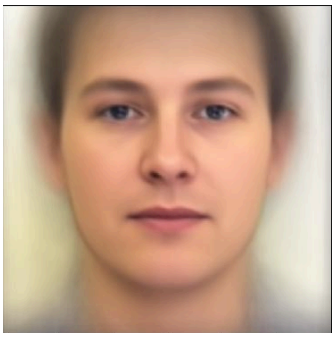
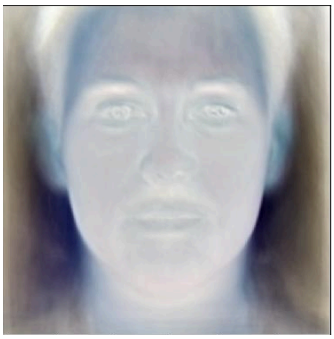
學號：r06946004 系級： 資科學程碩一 姓名：蔡尚錡

1.PCA of colored faces

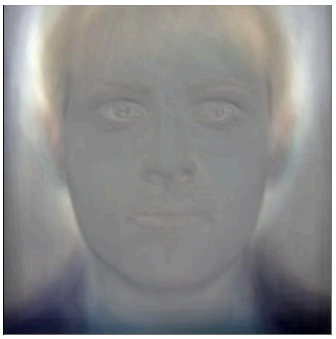
1.(.5%) 請畫出所有臉的平均。



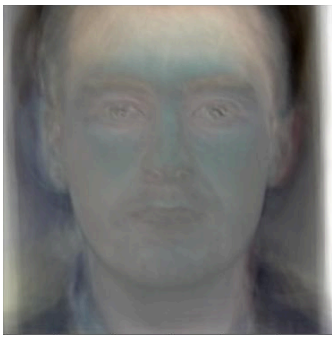
2.(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



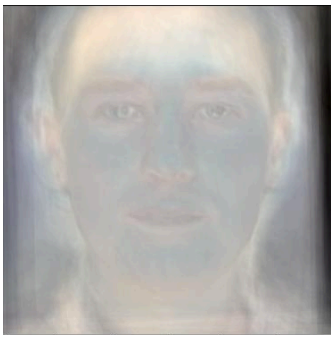
第一個



第二個

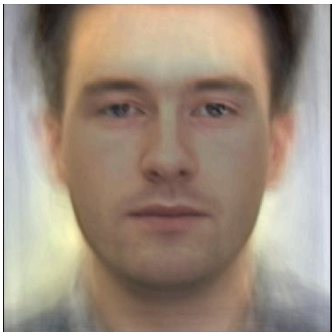


第三個

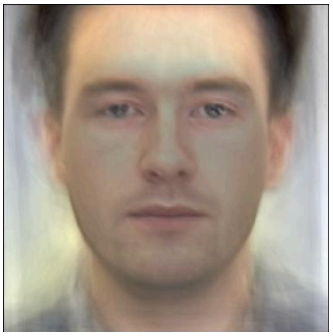


第四個

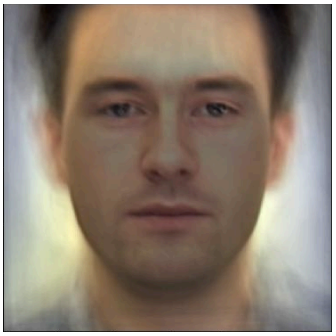
3.(.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。



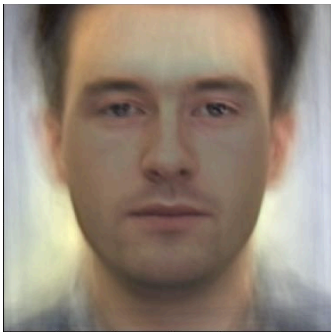
6.jpg



35.jpg



87.jpg



106.jpg

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。

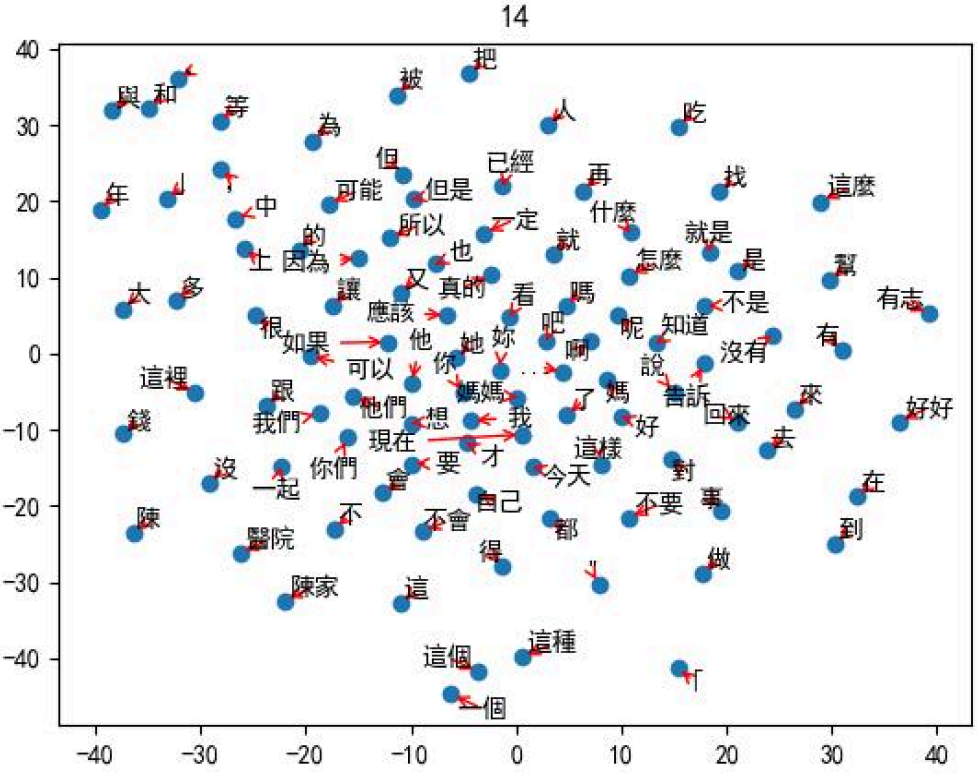
第一大 => 4.1%，第二大 =>3.0%，第三大 => 2.4%，第四大 => 2.2%

2.Visualization of Chinese word embedding

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。

我使用的是 gensim 這個套件，利用這個套件裡頭所提供的 word2vec 物件和方法去做中文句子的 training ，裡面我所調整的參數包括了： size ，也就是輸出的 word embedding的維度，我設定為256 ; window，也就是針對每一個詞在做 training 會往前往後看幾個字的意思，因為裡面的算法是用 cbow ，所以我設定為比較小的6 ; min_count ，也就是一個詞出現次數小於幾次就不會被視為是訓練對象，我把它設為 3。

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。

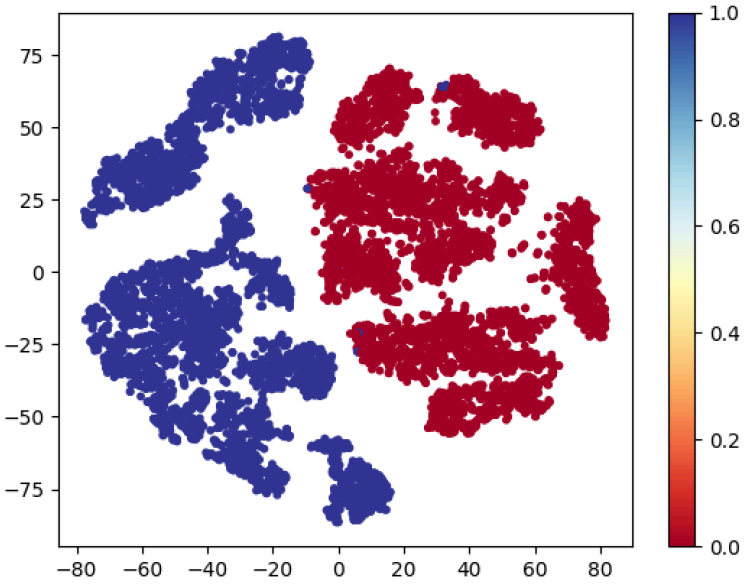
由視覺化的結果可以觀察到一些特徵，相類似的字詞會有接近的分佈位置，比如說：與、和，這兩個字由於意思相近，都分佈在左上角的位置；我們、你們、他們，這三個相似的主詞也分佈在中央偏左下鄰近的位置；嗎、吧、啊、呢，這四個常見語助詞也一起分佈在中央偏右上相鄰近的位置 。相反的字詞也會有接近的位置，比如說：有、沒有，會、不會，兩組都會有接近的位置跟相似的相對位置。最後也可以看到某些特定關係的字詞會有一樣的相對位置，你、妳，他、她，這兩組有男生女生的相對關係的字詞，位置上都分佈在中央附近的位置，並且這兩組各別會有一樣的相對位置，男女關係都是從左下到右上的方向。

3.Image clustering

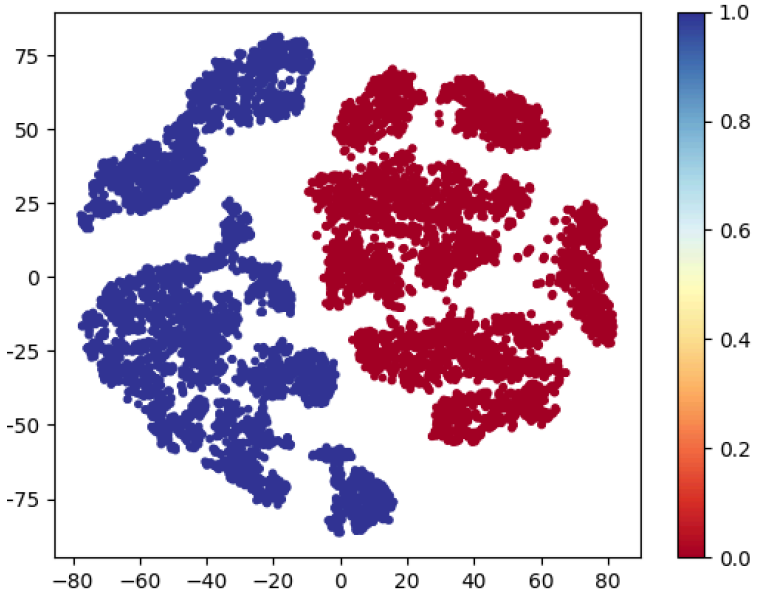
1.(.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

我使用兩種不同的降維方法做比較，一種是用sklearn 做 PCA的降維方法，另一種則是用 keras 實作 autoencoder 的降維方法，兩個方法都下降到低維度，然後把降維後的 image data 用 sklearn 裡面的 Kmeans 做分類分成兩個 cluster，並且兩者都有用 data augmentation 增加 data 數量。結果是，用 dnn 的 autoencoder 的降維方法，搭配 100 epochs 跟 64 batch-size 的 training 降到 64維，得到 cluster的結果是 F1-score = 0.93304，而另一種 PCA 的降維方法降到16 維，得到 cluster 的結果是 F1-score = 0.03，相較之下 PCA 的降維方法得到的結果明顯比 autoencoder 要來得低很多，由此可見 audoencoder 的降維是一種比較好的降維方法。

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



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。



答：
兩者之間比較起來大致上的分類都是一樣的，代表我的 model 分類的結果跟標準答案比起來是非常接近的，但仔細觀察還是可以看到我預測的 label 分佈圖在紅點的那區還是有零星的藍點，代表原本應該是 0 卻判斷成 1 ，那些應該就是誤判的數據。