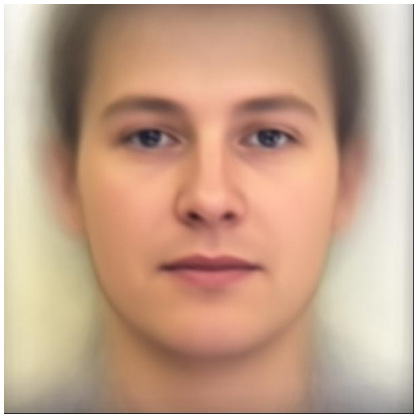
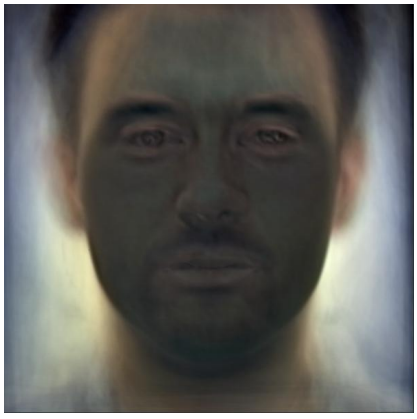
學號：R05525096 系級： 工海碩二 姓名：郭捷

A. PCA of colored faces

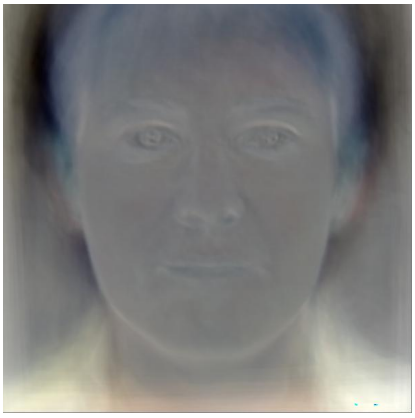
(.5%) 請畫出所有臉的平均。



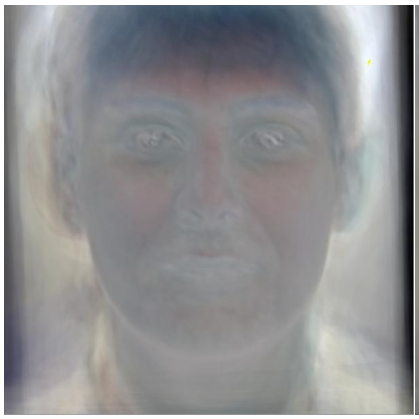
(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



1



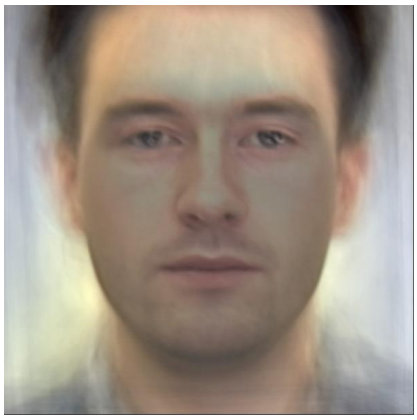
2



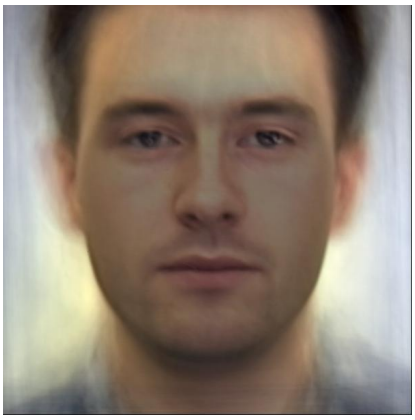
3

4

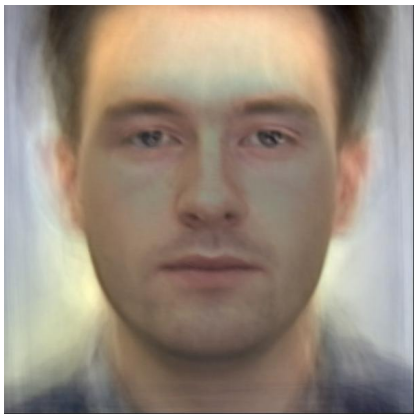
(.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。



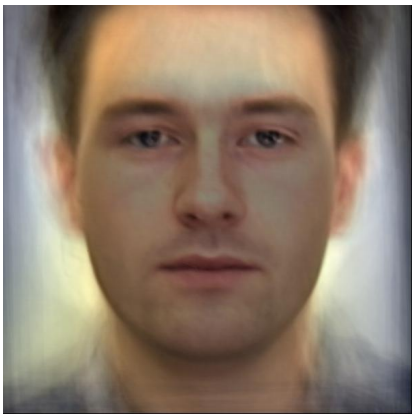
19



28



66



93

(.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

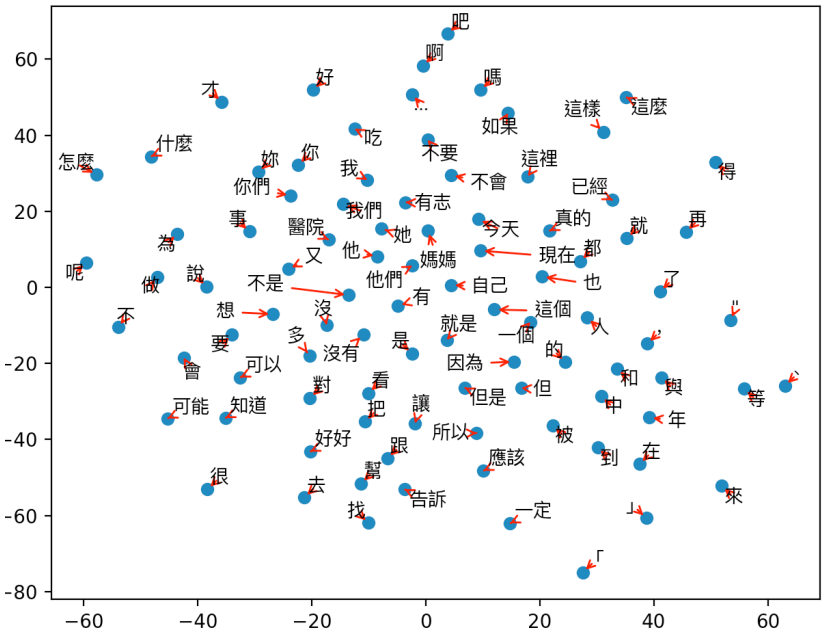
1	2	3	4
4.1%	2.9%	2.4%	2.2%

B. 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。

我使用了 Gensim 的套件。Min_count 設 6000，因為如果對全部的字都做 visualization 的話會太多。Size 設 250 用來定義每個字要用多少維度表示。

(.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。

可以從結果看到，‘今天’和‘現在’距離很近，‘沒’和‘沒有’距離很近，‘你’、‘我’、‘你們’和‘我們’等詞距離很近。表示詞意相近的詞，距離就很小，同時‘不要’和‘要’的距離相似與‘不會’和‘會’的距離，代表詞和詞直接的距離和方向可以表示他們之間的關係。

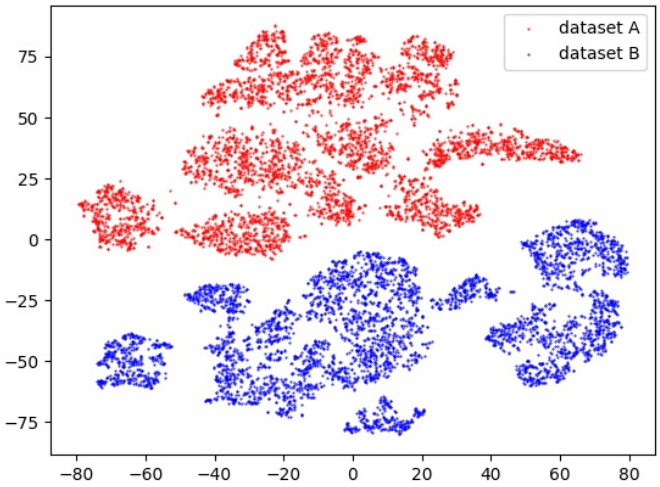
C. Image clustering

(.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)

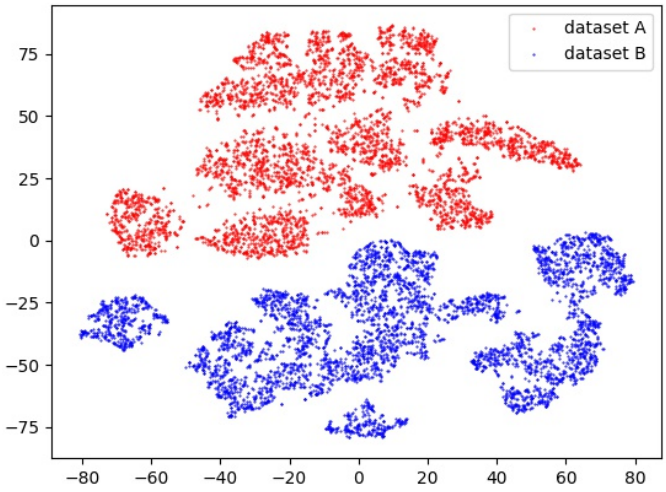
PCA	AutoEncoder
0.02645	0.99935

本實驗比較了相同的 cluster(Kmeans)情況下，分別使用 PCA 和 AutoEncoder 的方法做降維的結果，結果顯示 AntoEncoder 的方法遠好於 PCA。

(.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。



根據這個資訊，在二維平面上視覺化 label 的分佈，與我自己預測的 label 之間幾乎沒有差別。除了少許特徵在二維平面上有些許差異，但是其預測結果幾乎全部正確。