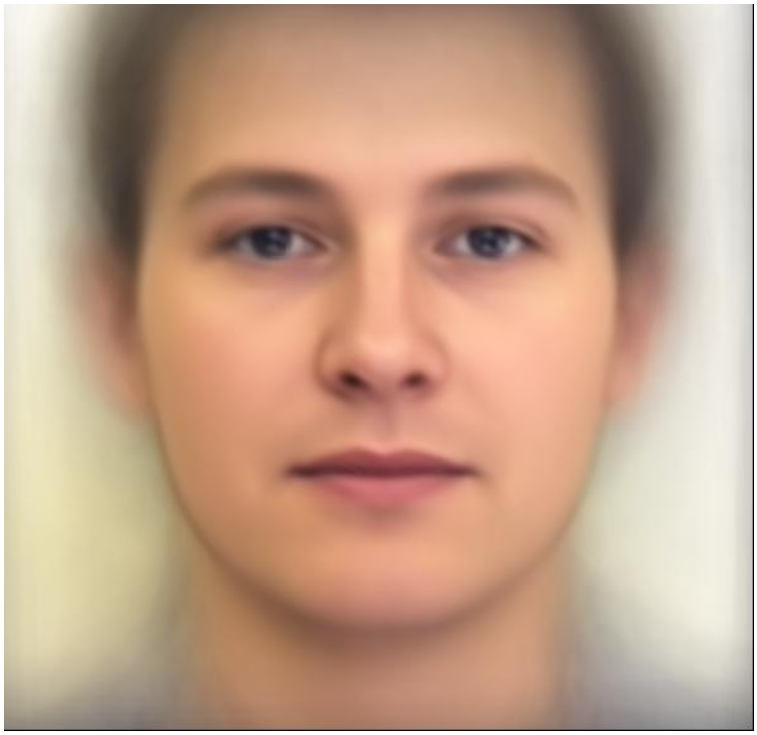
學號：B03902125 系級： 資工三 姓名：林映廷

A. PCA of colored faces

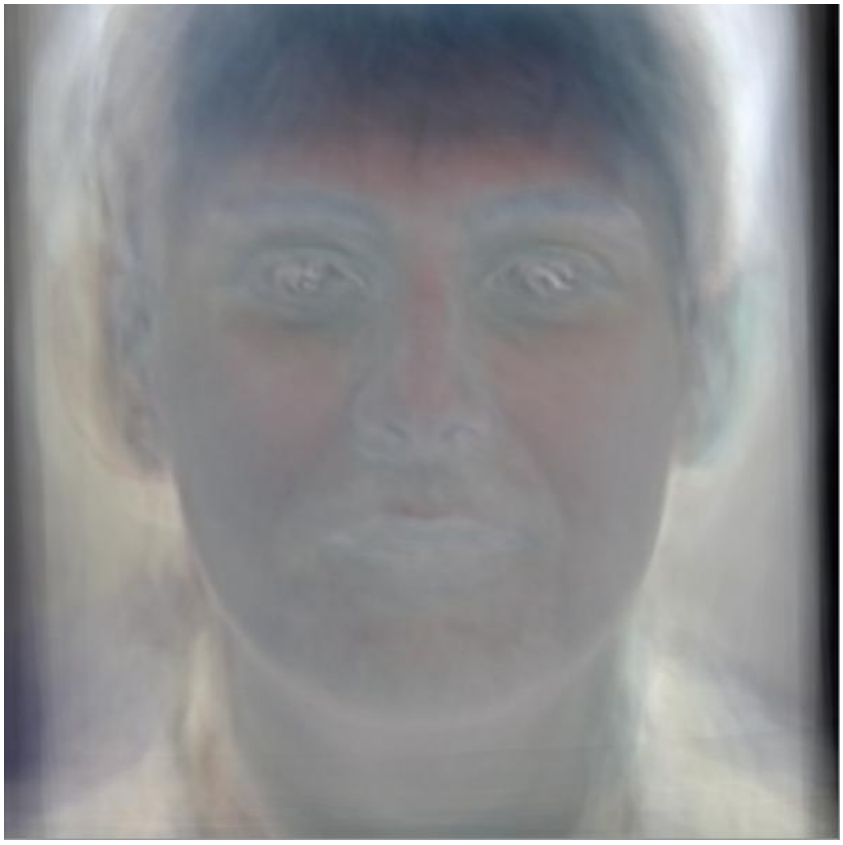
(.5%) 請畫出所有臉的平均。



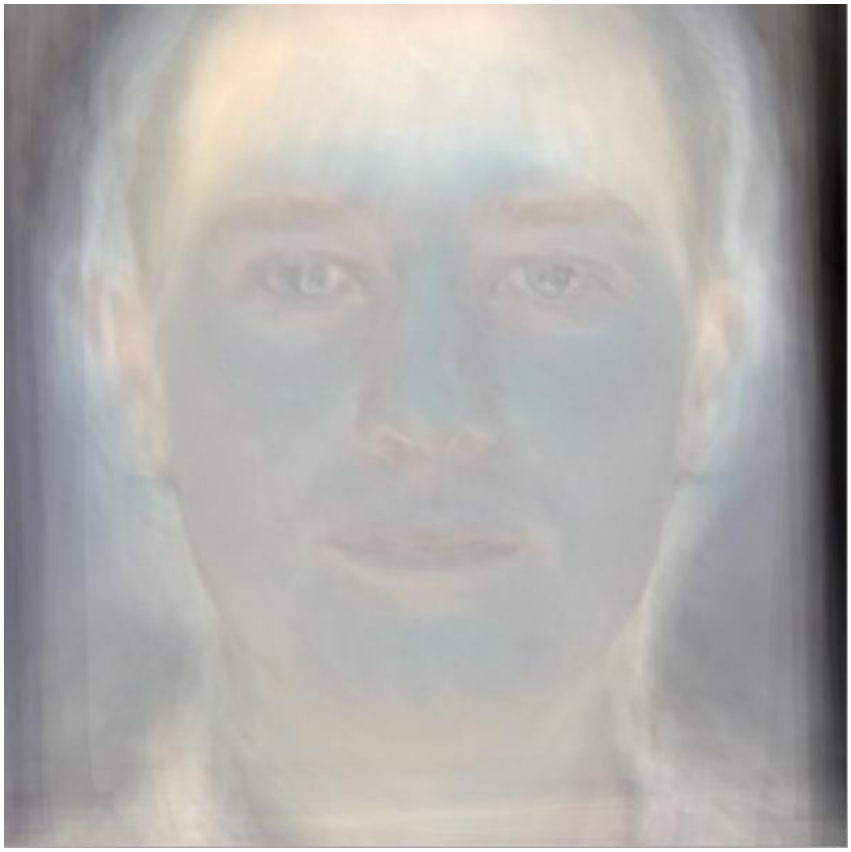
(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

eigenface0

eigenface1



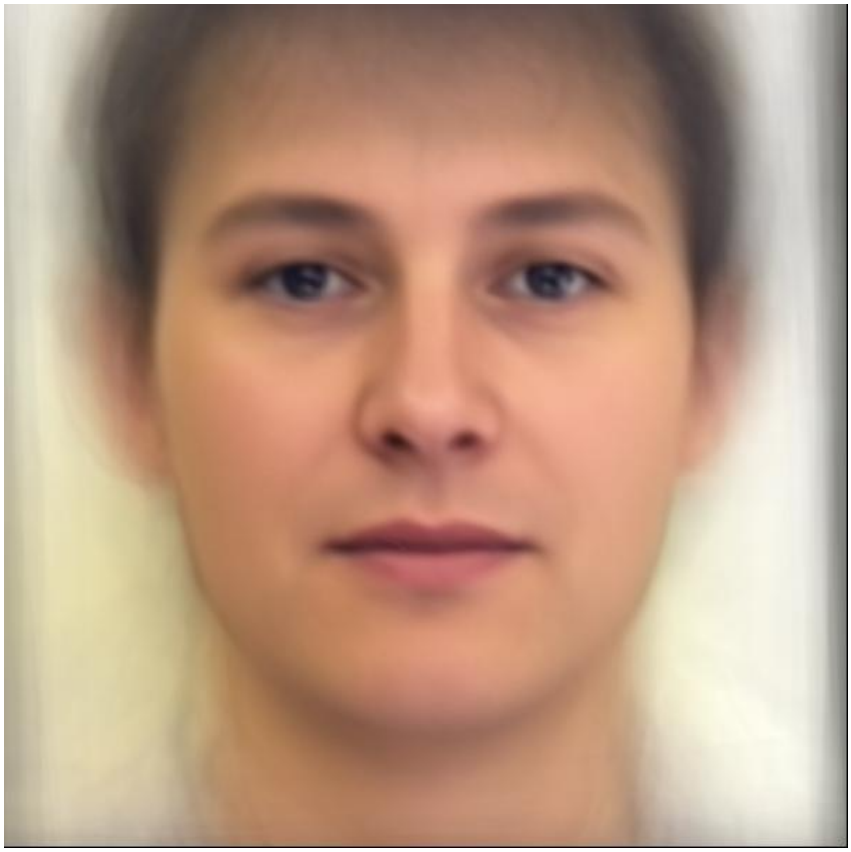
eigenface2



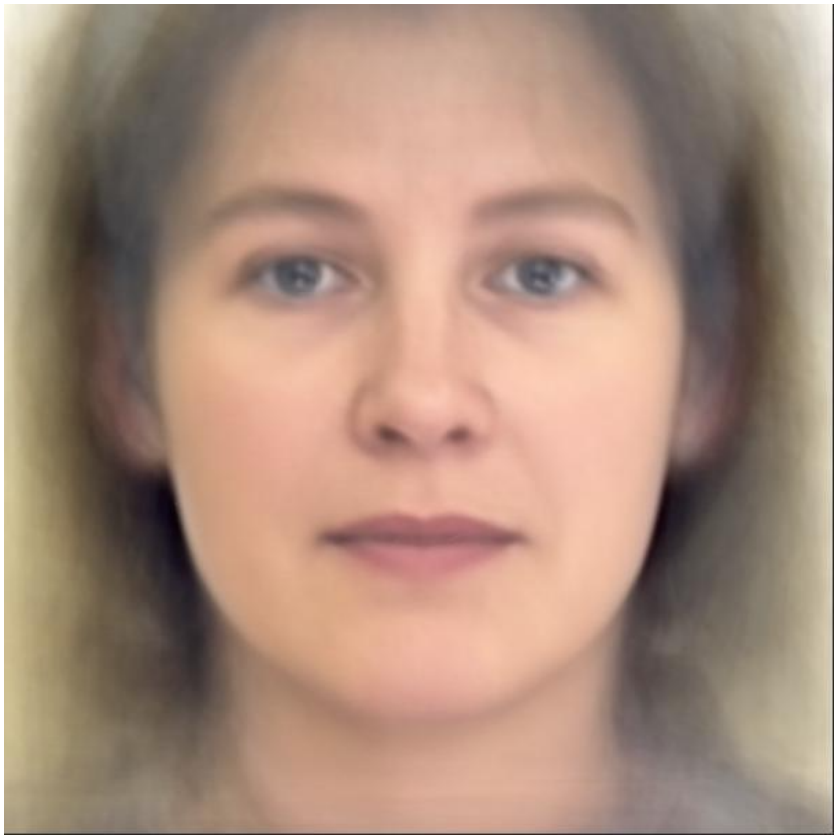
eigenface3

(.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。

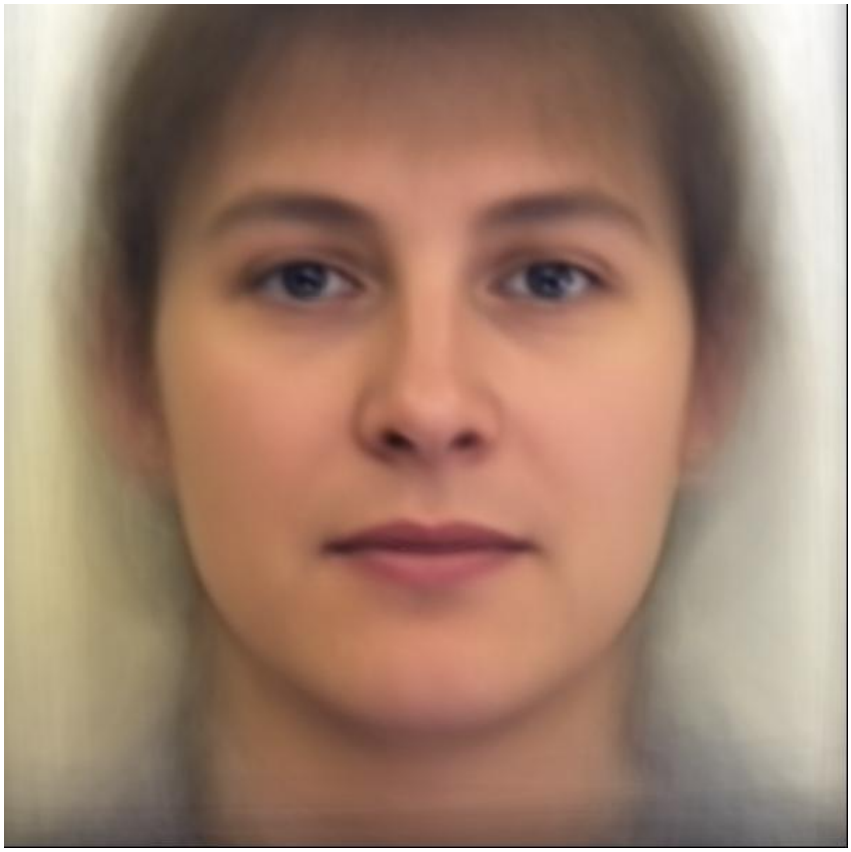
reconstruct9



reconstruct100



reconstruct200



reconstruct300

(.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。

Eigenface0 : 4.1%

Eigenface1 : 2.9%

Eigenface2 : 2.4%

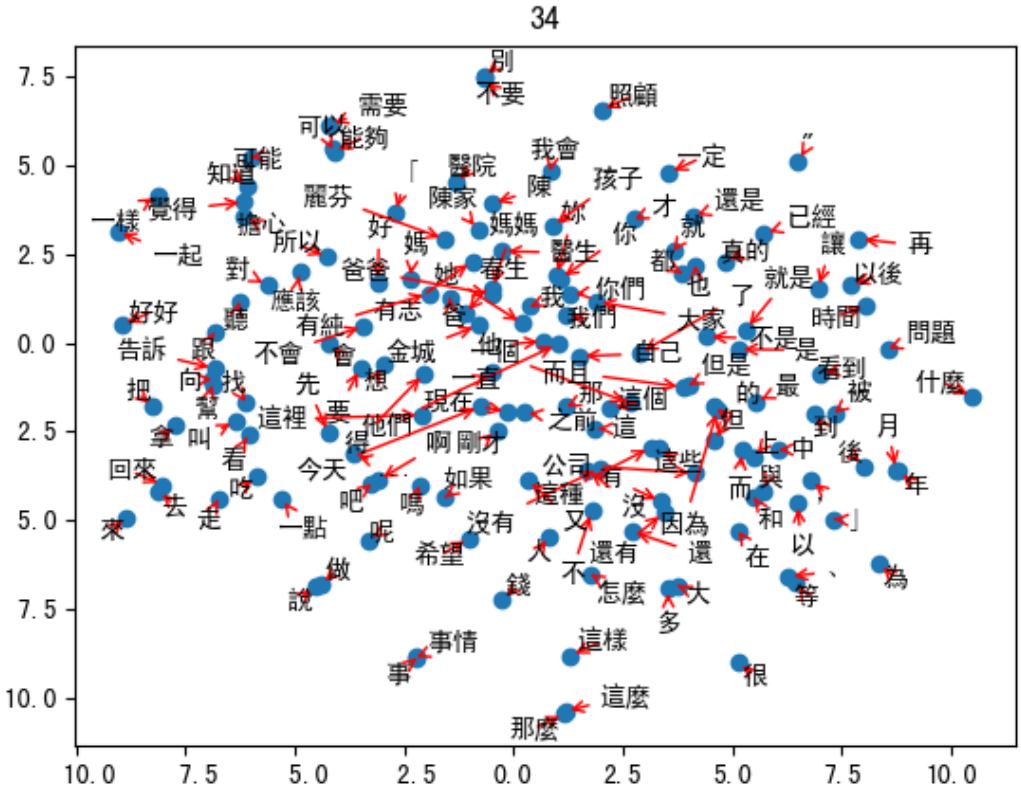
Eigenface3 : 2.2%

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。

gensim 的 word2vec，我調整了 size = 50，讓每一個 word2vec 的維度為 50，min_count=3000，從 text 裡面挑出出現頻率大於或等於 3000 次的字詞，作為訓練對象。

(.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。

有相近的詞意會很集中，甚至剛好在附近，像是「可能」、「需要」、「能夠」這一組彼此就很接近，但像是「爸爸」、「爸」這一組相較之下，詞意相近，但卻被分得沒那麼近。很明顯有些做得很好，但有些沒做好。

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 + kmeans:

Private: 0.03048

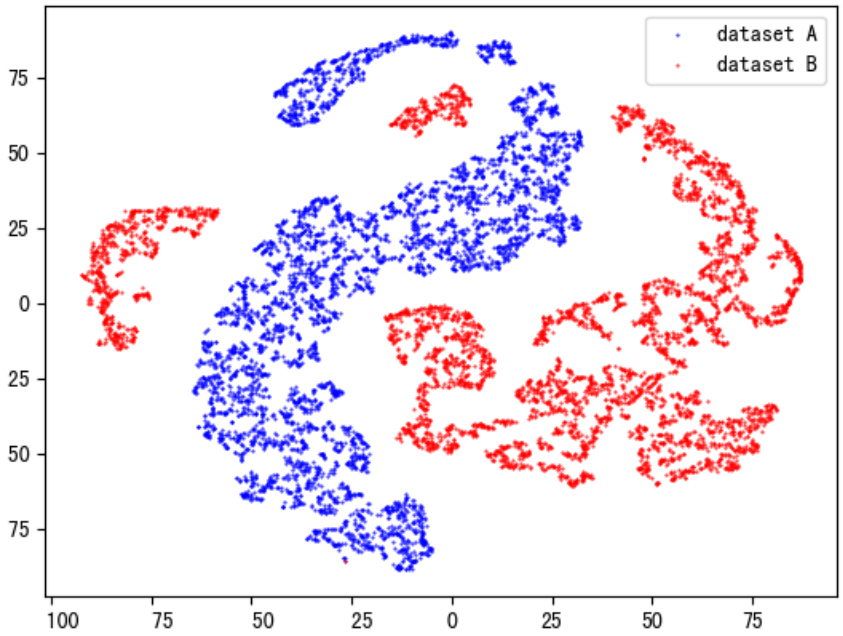
Public:0.03023

Autoencoder+kmeans:

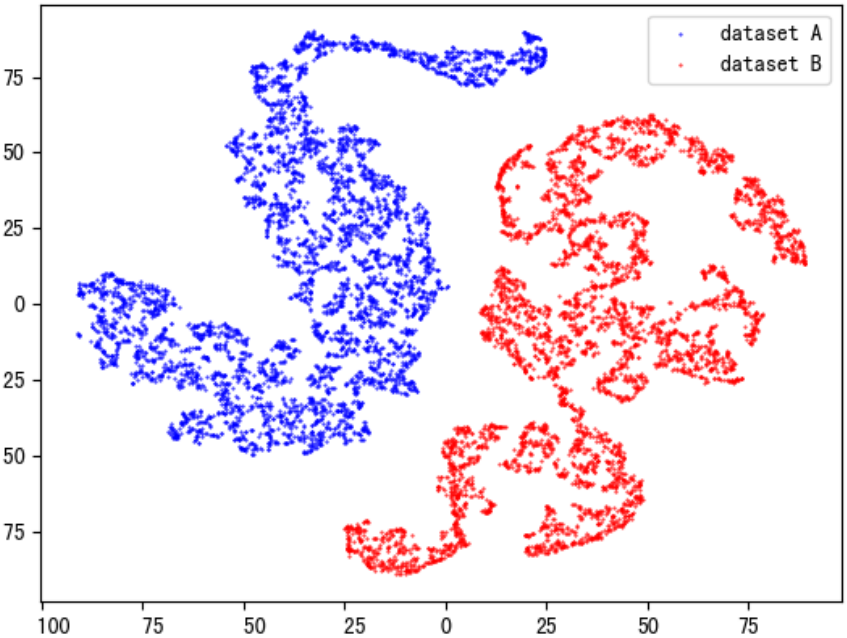
Private:0.97737

Public:0.97835

(.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。



如果有根據這個資訊分群，兩者很明顯是分得很開，而且自己一群的會比較緊

密；反之，如果沒有根據這個資訊分群，會發現有些地方沒分得很好，藍色的那一群出現兩小群紅色的部分。