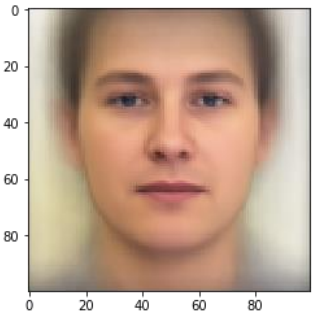
學號：r05229016 系級： 大氣碩二 姓名：羅章碩

# PCA of colored faces(collaborator：r05229014 鄒適文)

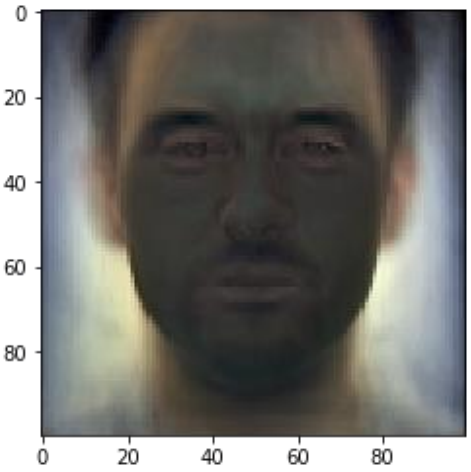
※此題我將圖片 reshape 成 100x100x3，運算較快

A.1.(.5%) 請畫出所有臉的平均。

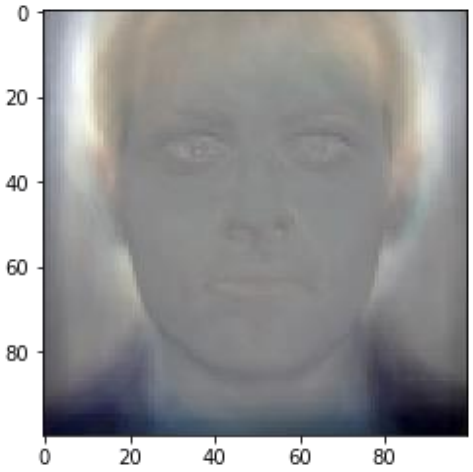



A.2.(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

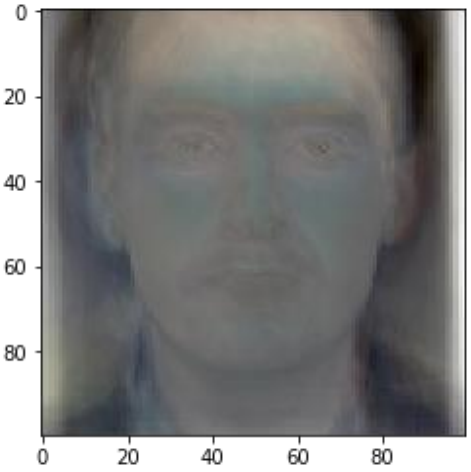
1.



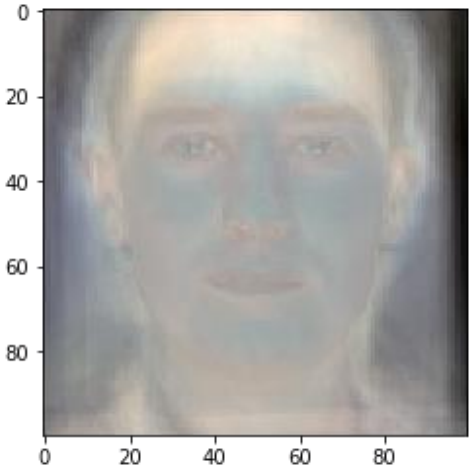

2.




3.

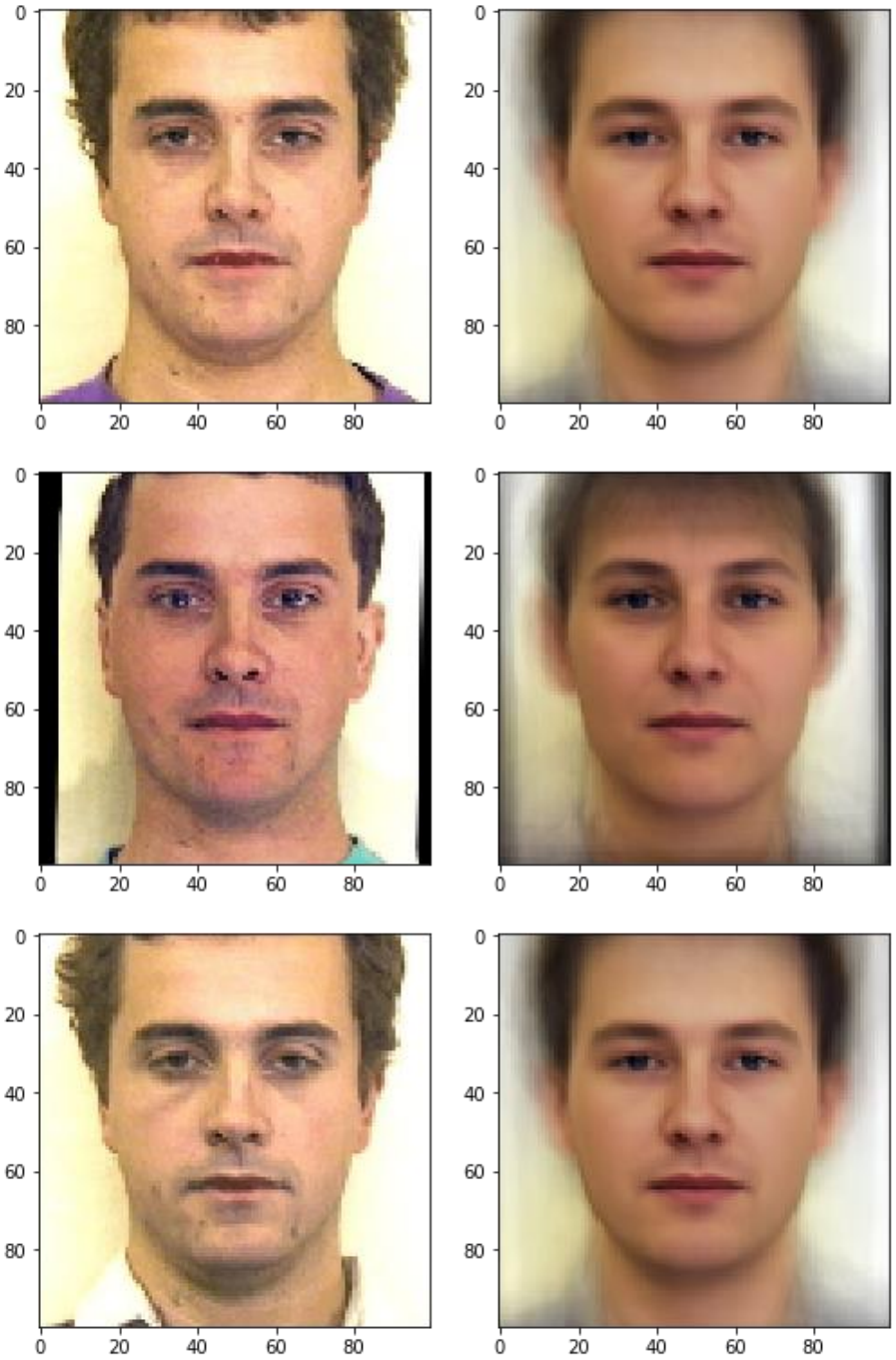


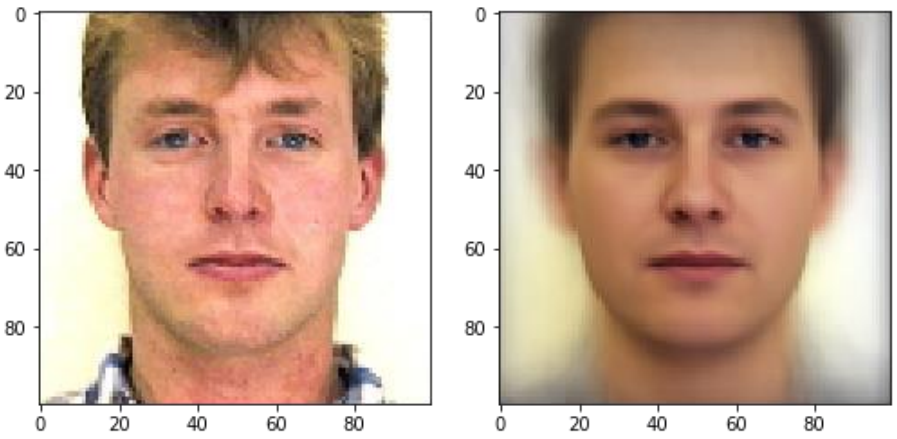

4.



A.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。

True　　Reconstruct(4 Eigenfaces)






A.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。

1. 4.2 %
2. 3.0 %
3. 2.4 %
4. 2.2 %

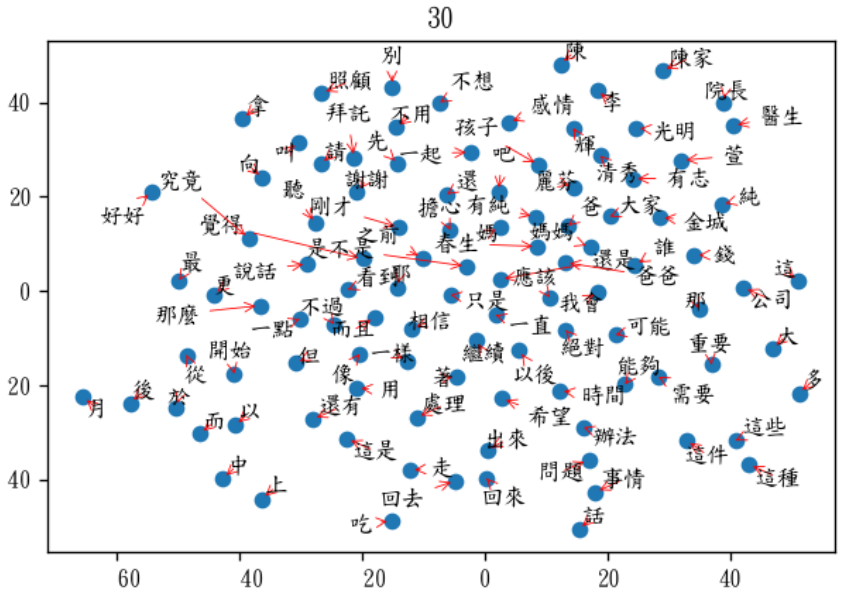
## B. Visualization of Chinese word embedding (collaborator：r05229014 鄒適文)

B.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 中的 Word2vec 來 train data，我調了 size 到 256，此參數代表的意義是訓練出來的詞向量會有幾維，而調到 256 代表用 256 維來表示一個詞，除此之外還有調 min_count=1，這個參數的意義是若這個詞出現的次數小於 min_count，那他就不會被視為訓練對象。

B.2.(.5%) 請在 Report 上放上你 visualization 的結果。

我只把出現次數大於 2500、小於 6000 的詞給挑出來並做 visualization。以下為做出來的結果。




B.3.(.5%) 請討論你從 visualization 的結果觀察到什麼。

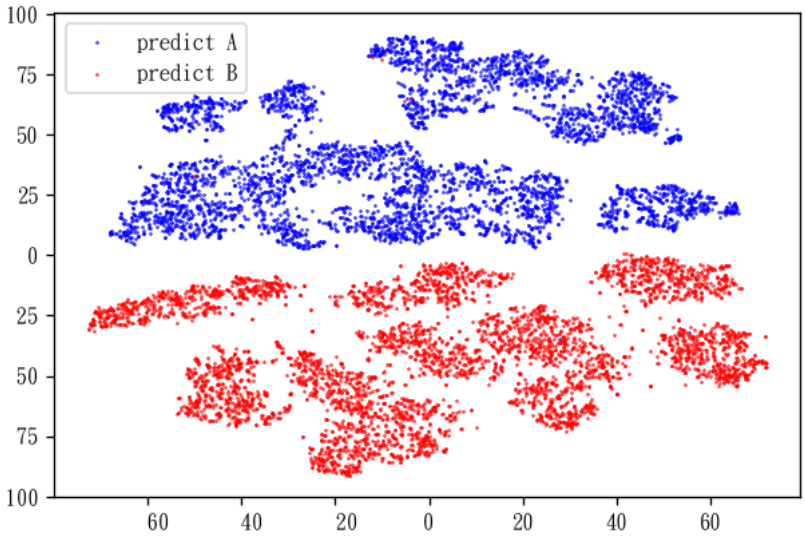
從上圖結果可以觀察到像是爸爸、媽媽或是人名的詞都會比較聚在一塊(圖偏右上的地方)，而圖的下方則是看到「走、回去、回來、出來」這幾個詞比較靠在一塊，其他地方也可以看到相似的詞會有聚在一塊的情形(像是繼續、一直、以後……)，但是對於只有一個字的單詞好像就不是分類的很好，像是"吃”這個詞的位置靠近回去、回來，或是拿的位置是靠近聽跟叫，或是“還”這個位置也感覺不太對。

## C. Image clustering(collaborator：r05229014 鄒適文)

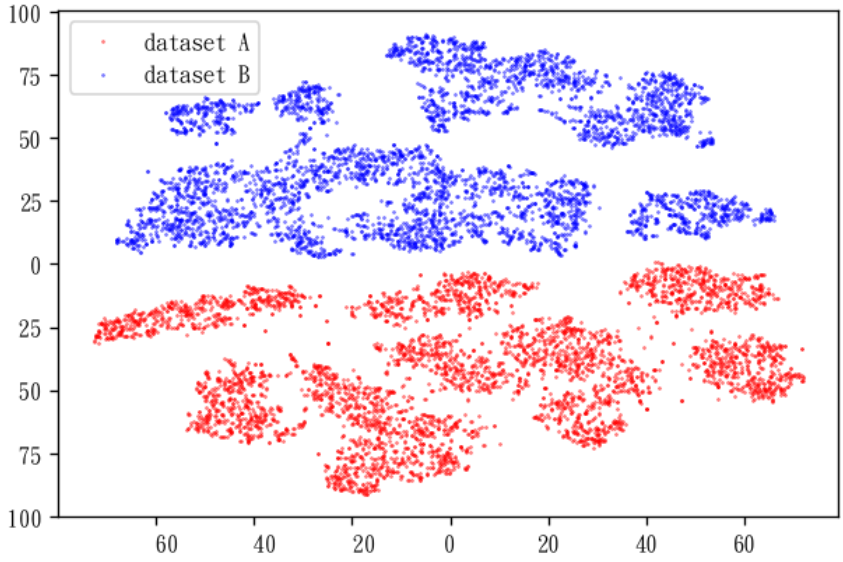
C.1.(.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

第一種使用了 KMeans 的方法做預測，得到的結果在 kaggle 上的分數大約為 0.876，而第二種使用了 MinibatchKmeans 得到的結果在 kaggle 上分數大約是 0.83。

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




C.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。




從兩張圖來看，可以看到 predict 的結果還不錯，只有些微點是預測錯的，藍色預測成紅色的比紅色預測成藍色的還多。