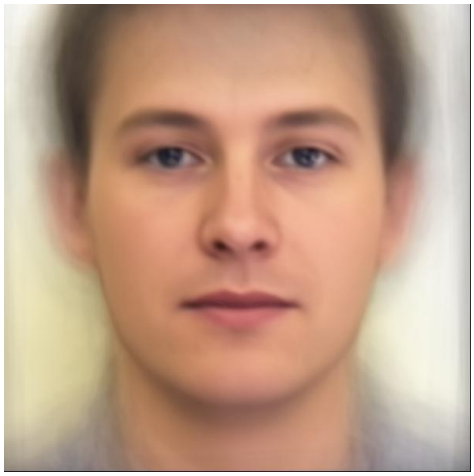
學號：R06922113 系級： 資工所碩一 姓名：陳宣伯

A. PCA of colored faces

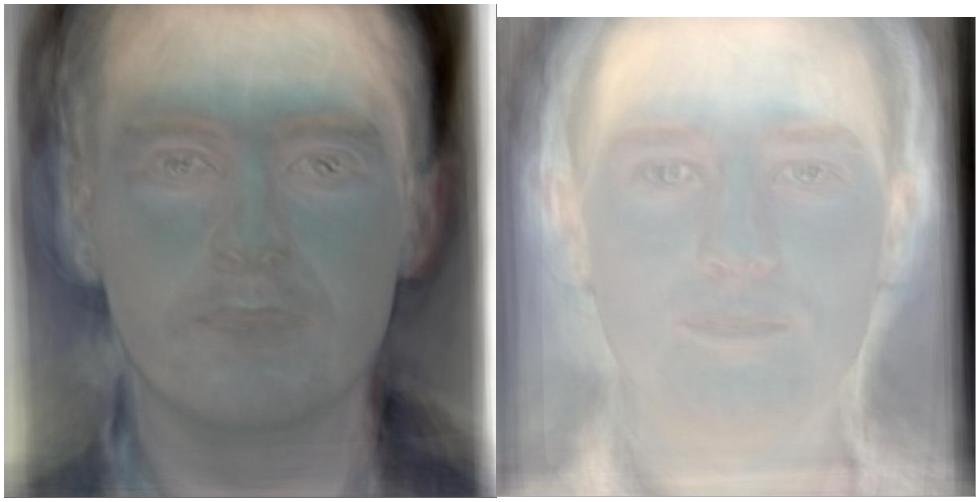
A.1. (.5%) 請畫出所有臉的平均。

A.2. 

A.3. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

A.4. 前四大 eigenvectors 由大到小是，由上而下、由左而右

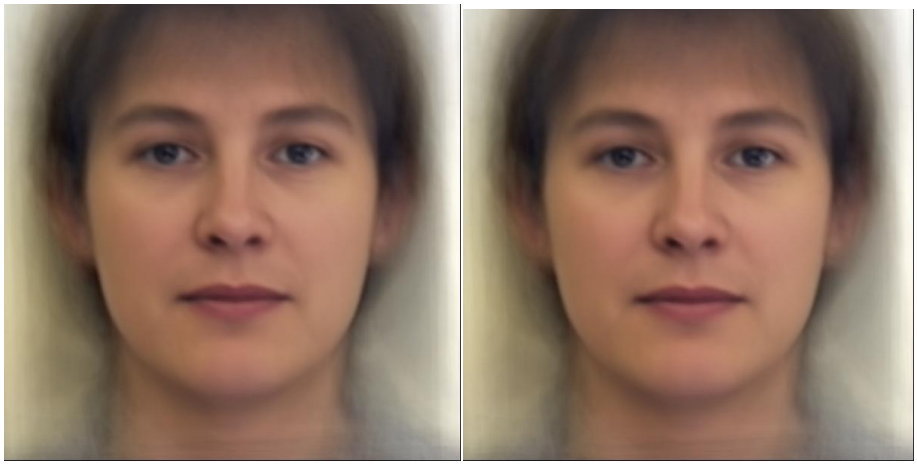
A.5. 

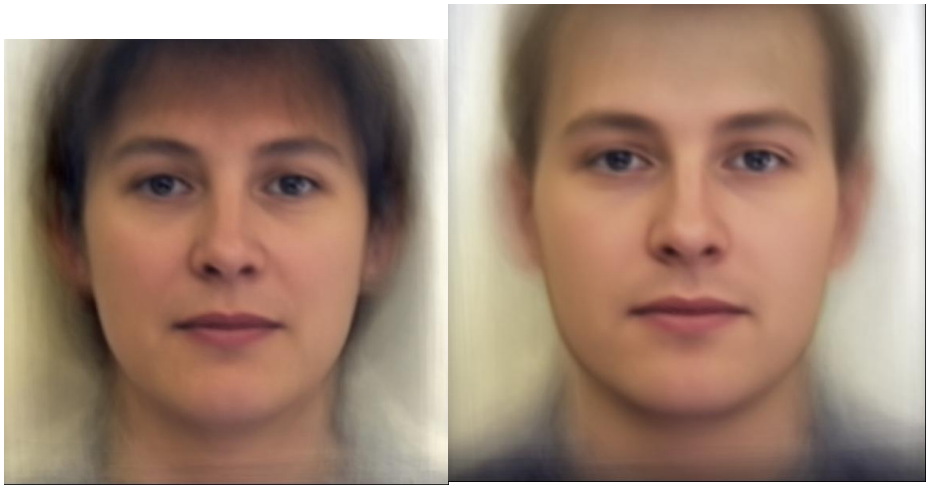
A.6. 

A.7. (.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。

A.8. 我挑圖片裡的：0.jpg, 1.jpg, 2.jpg, 3.jpg 四張圖片來做 reconstruction,結果分別是左上、右上、左下、右下

A.9. 

A.10. 

A.11. (.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。

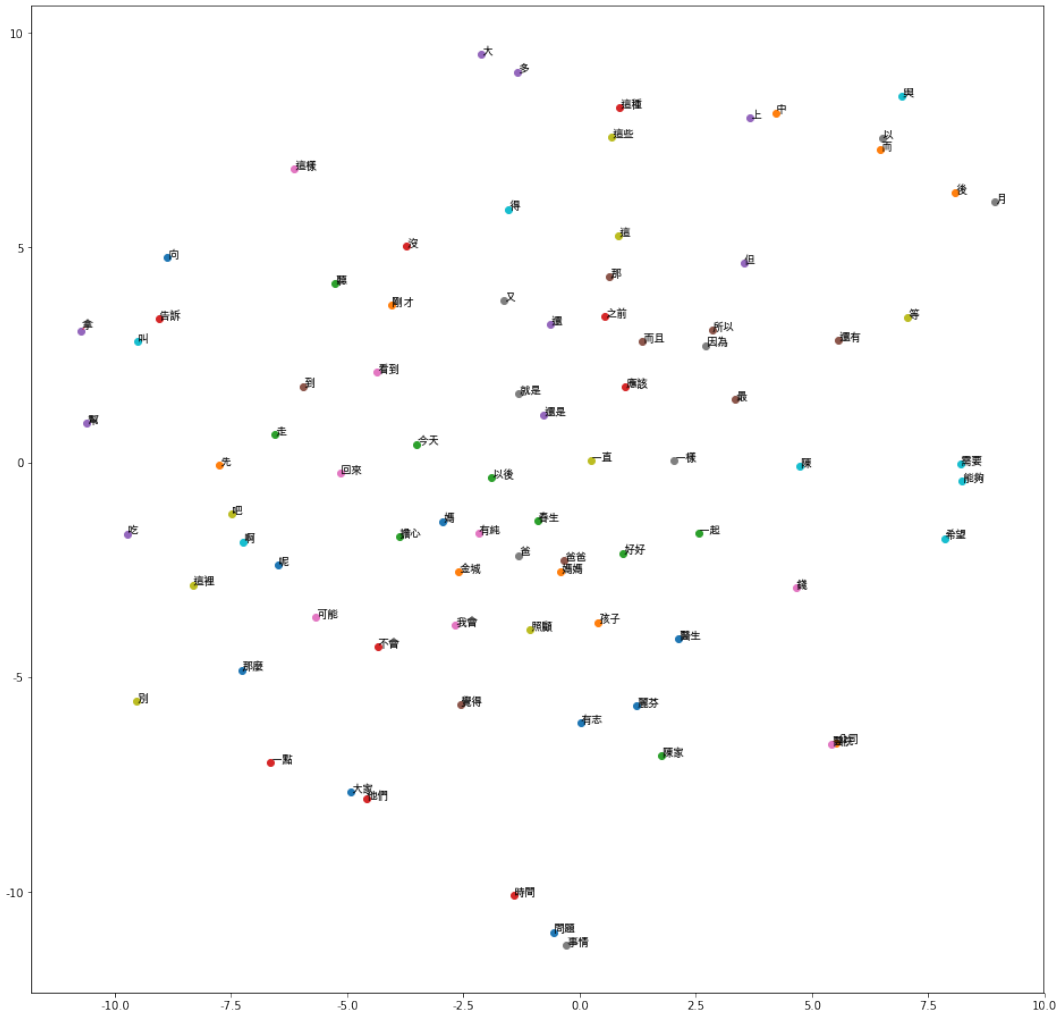
A.12. [第一大, 第二大, 第三大, 第四大] =
[4.1%, 3.0%, 2.4%, 2.2%]

B. Visualization of Chinese word embedding

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。

B.1.1. 我使用 gensim 套件來實現 word2vec，我總共嘗試調整了以下參數 learning rate(alpha)=0.03、window=5、min_count=10。learning rate 就是模型學習 word 分群的速度，window 為每次最多看幾個字，調太大會模糊一些字詞的關聯，調太小會失去一些長字詞的辨識，min_count 則是任何字數統計低於此值，即不參與建模

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



B.3.

B.4. (.5%) 請討論你從 visualization 的結果觀察到什麼。

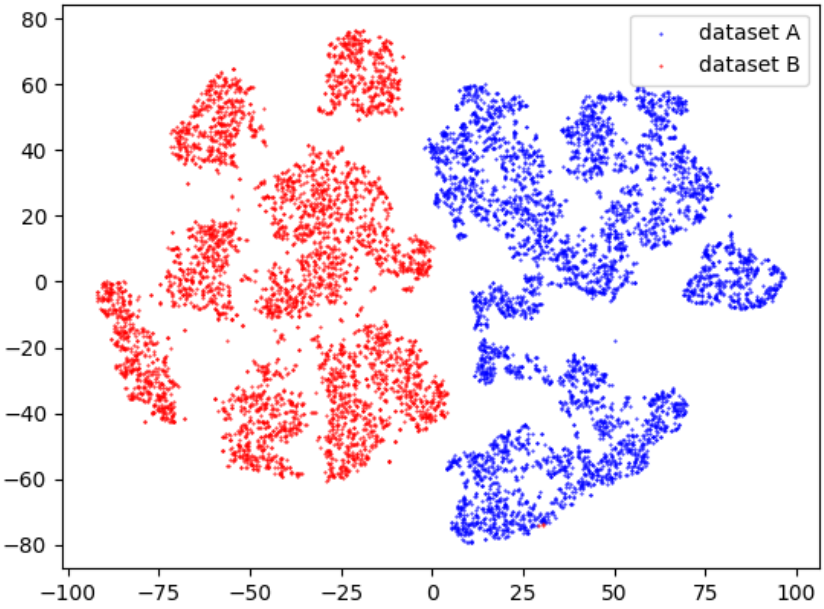
這<->這些的距離與還<->還是的距離非常相似，可以看出如果這個字跟下一個字可以組成常用的詞，這個字跟這個常用的詞距離都會雷同。另外一些動詞單詞例如拿、吃、叫會集中在圖的左邊、一些介系詞如與、以、而會集中在右上角

C. Image clustering

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)

Method	Public score	Private score
PCA(#dim=128)	17.558	17.770
t-SNE(#dim=128)	12.285	12.338
Auto-encoder(#dim=32)	83.946	83.842

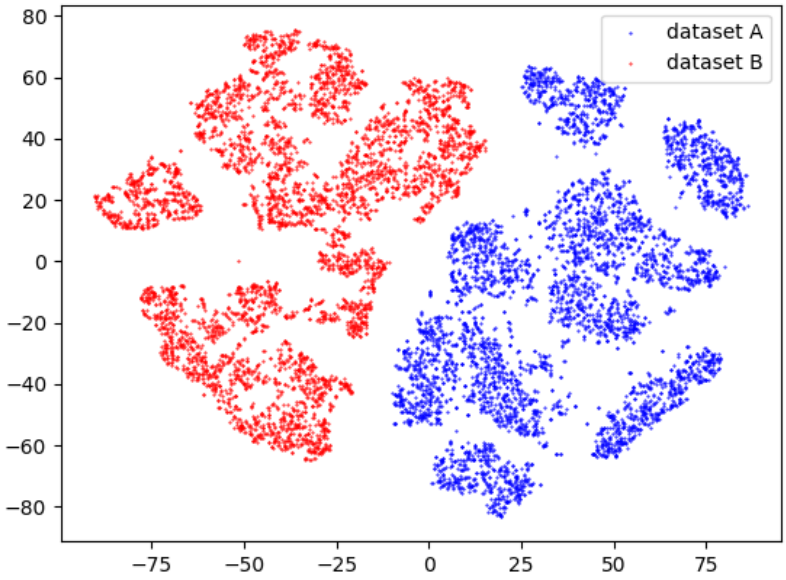
(.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。(Collaborators:李岳庭)



C.2.

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。(Collaborators:李岳庭)

C.4.



C.5.

C.6. 與前一子題預測的標籤相比，前一題的圖右下角有一些 dataset B 的紅色點與 dataset A 的藍色點混在一起，代表沒有百分之百預測，但是加入已知前 5000 與後 5000 分別屬於 dataset A 與 dataset B 後，就再也沒有混在一起的點，能夠很清楚看到分隔兩個 dataset 的界線。