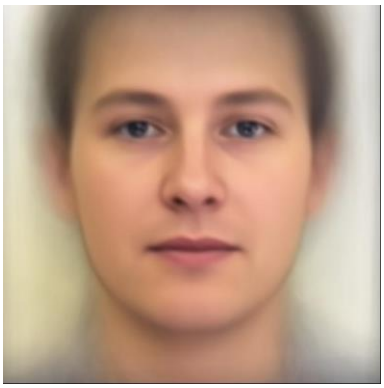
ML2017Fall - HW6

學號：R05945012 系級：生醫電資碩二 姓名：張凱崴

A. PCA of colored faces

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。

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。

Original

Reconstructed

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。

前四大的百分比分別為：4.1%、2.9%、2.4%、2.2%。

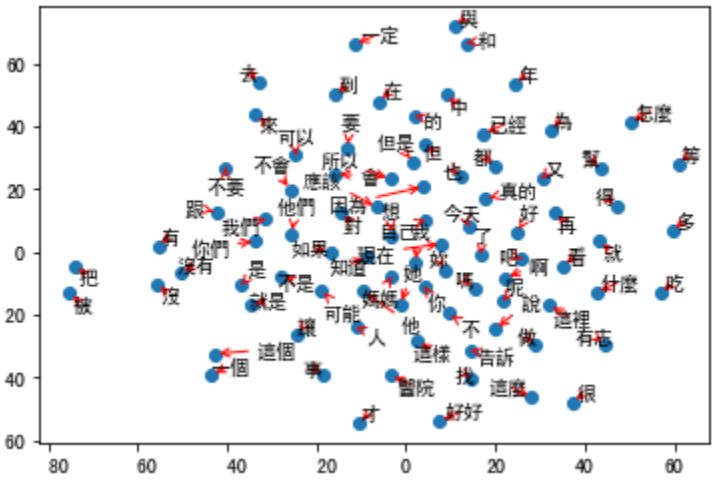
B. Visualization of Chinese word embedding

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。

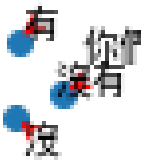
我使用的套件為 gensim，參數調整 size=100、min_count=1，代表 word vector 的 dimension 設為 100，出現過一次以上的詞就會計入訓練。

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

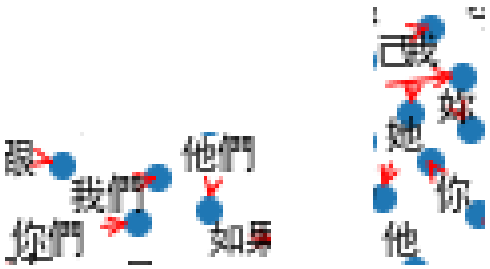
(出現次數超過 5000 次才顯示)



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

- 

“有”、”沒”、”沒有”是相近的，但可以看出”沒有”與”沒”比較近，因為”沒有”與”沒”同義。

- 

”你們我們他們”與”你我他”各自相鄰很近。此外，”你妳”與”他她”，具有男女分別的詞的相對位置是相同的(女性偏右上、男性左下)。

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)

我共使用了兩種方法：DNN 與 CAE (CNN+DNN)，結構如下：

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 256)	200960
dense_2 (Dense)	(None, 128)	32896
dense_3 (Dense)	(None, 64)	8256
dense_4 (Dense)	(None, 32)	2080
dense_5 (Dense)	(None, 64)	2112
dense_6 (Dense)	(None, 128)	8320
dense_7 (Dense)	(None, 256)	33024
dense_8 (Dense)	(None, 784)	201488

Total params: 489,136
Trainable params: 489,136
Non-trainable params: 0

DNN model

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 28, 28, 1)	0
conv2d_1 (Conv2D)	(None, 28, 28, 16)	416
max_pooling2d_1 (MaxPooling2	(None, 7, 28, 16)	0
reshape_1 (Reshape)	(None, 3136)	0
dense_1 (Dense)	(None, 784)	2459408
dense_2 (Dense)	(None, 256)	200960
dense_3 (Dense)	(None, 128)	32896
dense_4 (Dense)	(None, 64)	8256
dense_5 (Dense)	(None, 32)	2080
dense_6 (Dense)	(None, 64)	2112
dense_7 (Dense)	(None, 128)	8320
dense_8 (Dense)	(None, 256)	33024
dense_9 (Dense)	(None, 784)	201488
dense_10 (Dense)	(None, 3136)	2461760
reshape_2 (Reshape)	(None, 7, 28, 16)	0
conv2d_2 (Conv2D)	(None, 7, 28, 16)	6416
up_sampling2d_1 (UpSampling2	(None, 28, 28, 16)	0
conv2d_3 (Conv2D)	(None, 28, 28, 1)	401

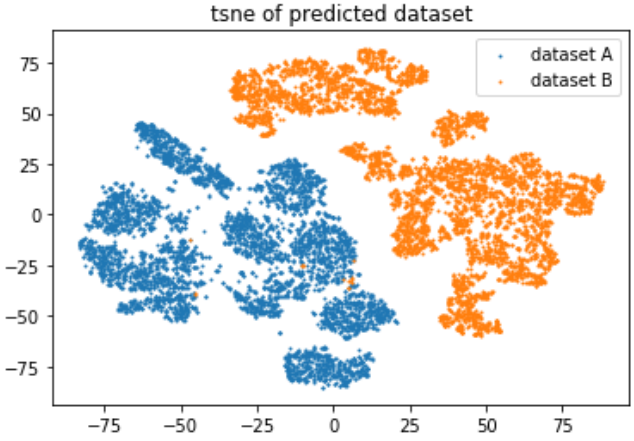
Total params: 5,417,537
Trainable params: 5,417,537
Non-trainable params: 0

CAE model

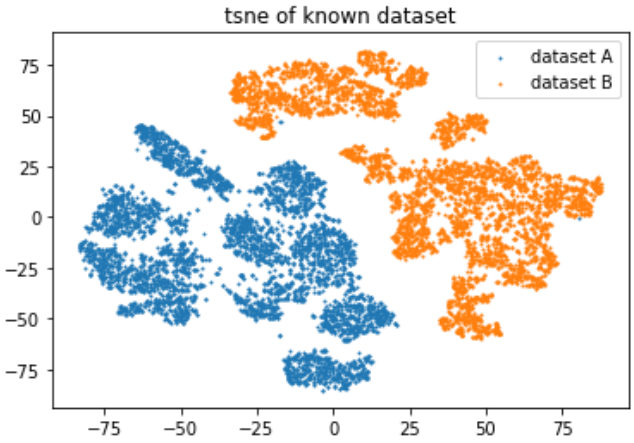
model	DNN	CAE
Valid mse loss	0.0090	0.0100
Public score	0.88400 (K-means)	0.02898 (K-means) 0.31476 (調整 threshold)

由上表可知，DNN model 表現得比 CAE 還好，這有點反直覺，因為圖像處理通常使用 CNN 會比較好。觀察原圖後發現，因兩種分群可以很直接地用眼睛分辨出來，特徵差距很大，所以使用 CNN 擷取特徵，反而不會有太好的結果。

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。



把第二題自己預測的分群與其相比，僅有一些點是有誤的，且在交界處的錯誤率比較大。