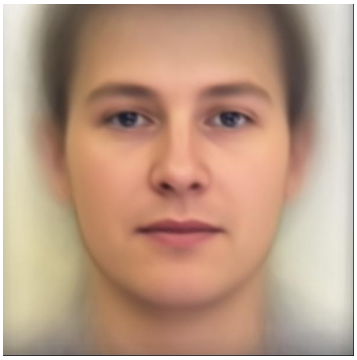
HW6 Report

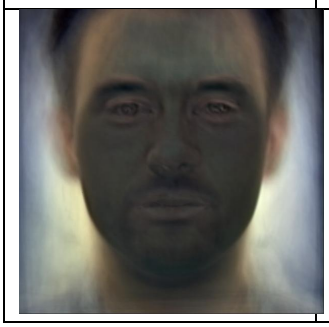
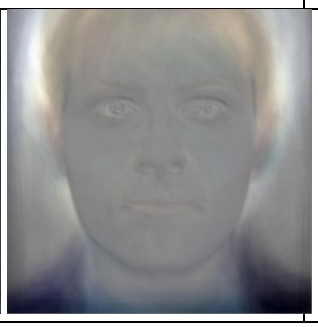
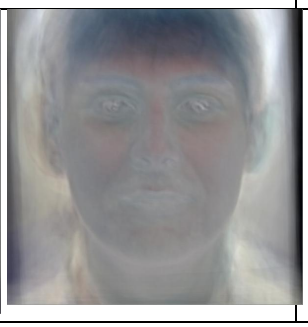
學號：R06942074 系級：電信所碩一 姓名：李宇哲

A. PCA of colored faces：

臉的平均：



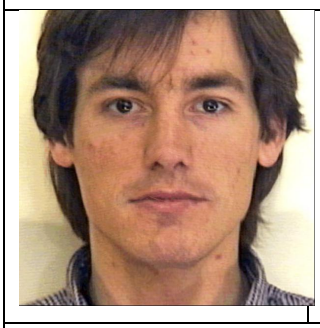
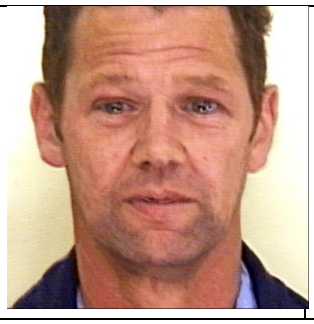
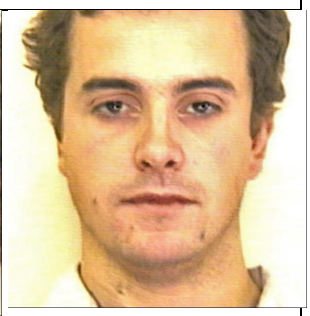
前四大 Eigenfaces：

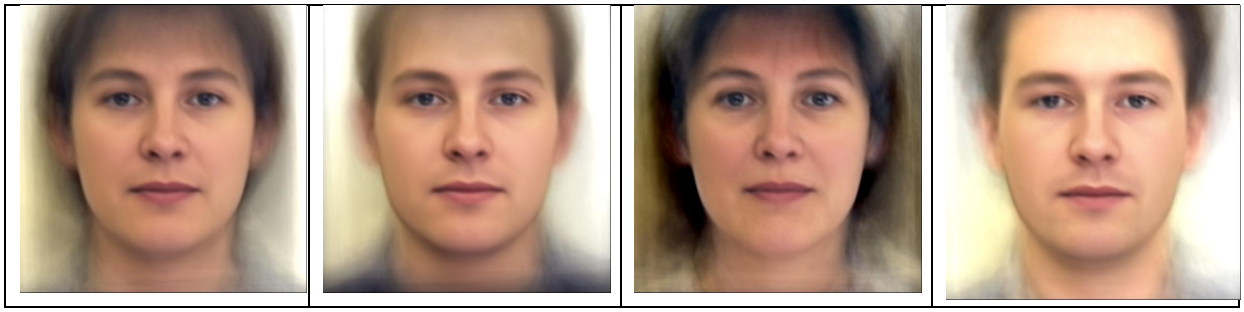
1st	2nd	3rd	4th
			

每張圖片寫出前都做了 standardization:

```
M -= np.min(M)
M /= np.max(M)
M = (M * 255).astype(np.uint8)
```

挑出四個圖片用前四大 Eigenfaces **進行** reconstruction：

Image 0	Image 3	Image 4	Image 9
原圖			
			
Reconstruction with top 4 Eigenfaces			



前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)

1st	2nd	3rd	4th
4.1%	3%	2.4%	2.2 %

B. Visualization of Chinese word embedding

我使用的是 **gensim** 的套件

gensim.models.word2vec.Word2Vec(sentences=<f結巴字串>, size=128, window=5)

Size： the dimensionality of the feature vectors.

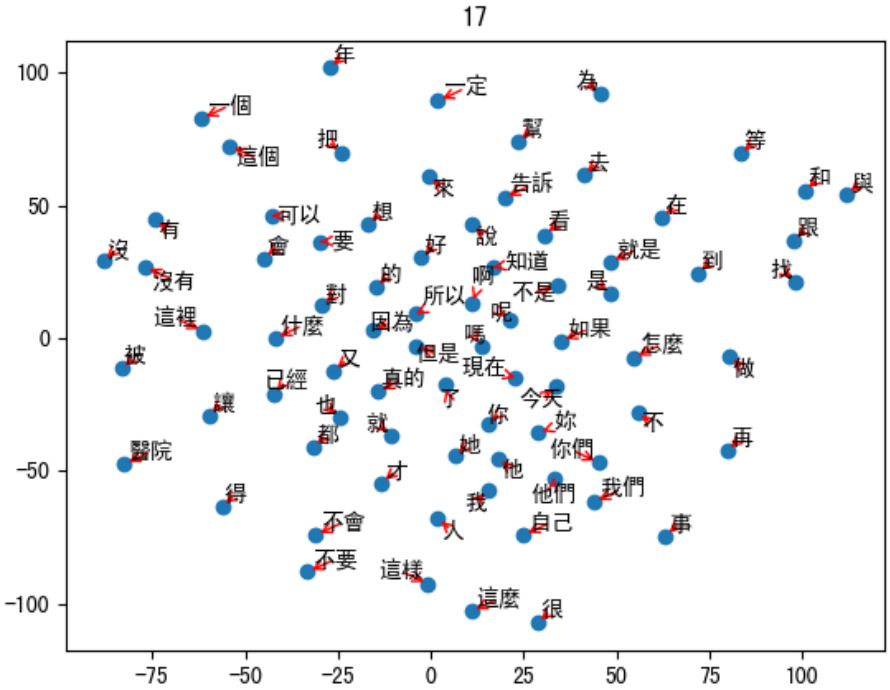
希望 encode 的 feature dimension 為 128 維

Window： the maximum distance between the current and predicted word within a sentence.

在一個句子裡面字的距離不希望超過 5

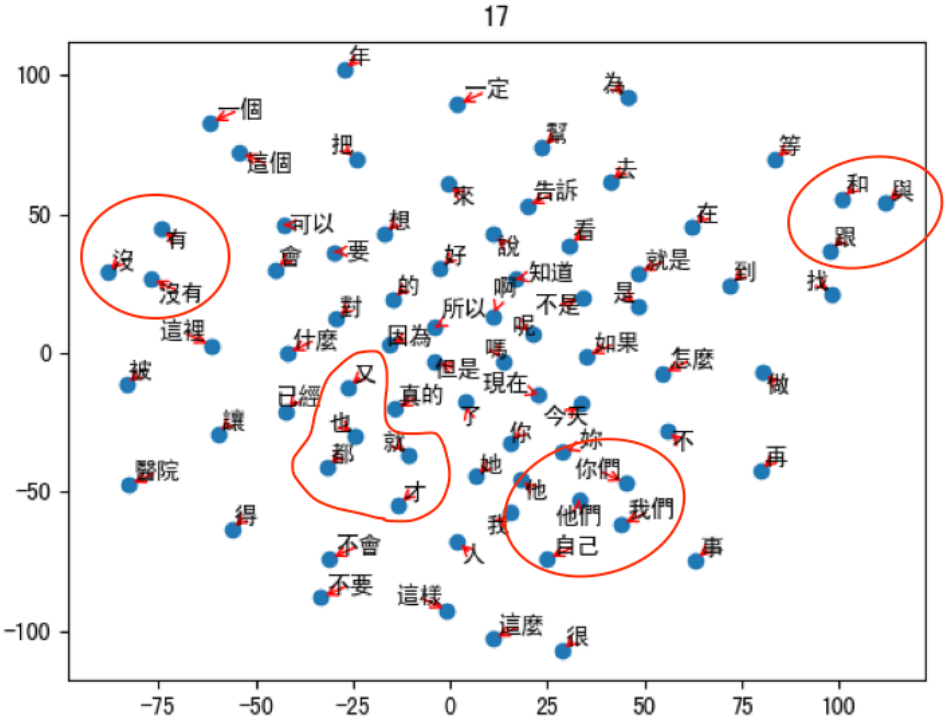
K 的字數：6000

Visualization 的結果：



我從 visualization 的結果觀察到：

可以發現字的意思比較有關係的會在附近，像是”有”、 “沒”、”沒有”基本上都在附近一起，以下只圈出個人覺得很明顯的字團，像是代名詞還連接詞等等。



C. Image clustering

兩種不同的 feature extraction 及其結果。

1. Autoencoder→32 → Kmeans→2

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

我的 Autoencoder 主要以 dense 為主，原本試過 CNN 但是效果反而不彰，最後再用 Kmeans binary cluster 分類

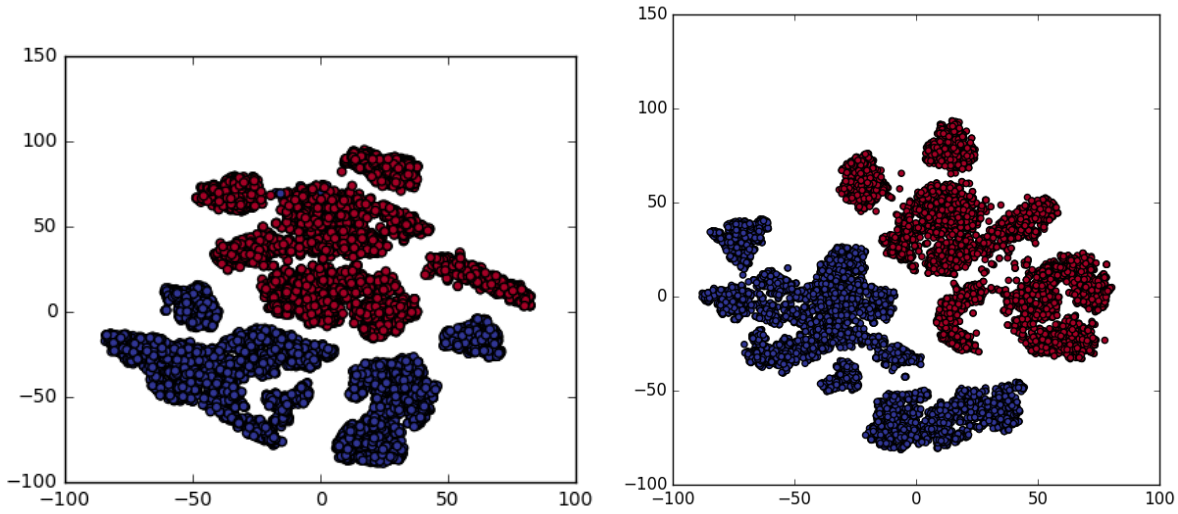
prediction.csv 3 days ago by Jack add submission details	0.86708	0.86800	☑

2. PCA →128→ Kmeans→2
   PCA 的方法就是先將 image 直接用 pca 降為

prediction_pca_2.csv a minute ago by Jack add submission details	0.03048	0.03024	☐

發現直接用 PCA 降為的準確率極低，畢竟 PCA 可以看成是一層的 Autoencoder，多層的 Autoencoder 的狀況比較好的確比較合理。

預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈
用 TSNE 降為後，左圖為我的 model 的預測狀況然後用 tsne 降為，右邊為 ground true label(前 5000 個 images 跟後 5000 個 images 來自不同 dataset)的狀況



可以發現我的圖(左圖)會有一些藍色的點混在上方紅色團裡面，畢竟我的準確率只有 0.88 左右，然後這裡可以發現用同樣的 model 做 TSNE 降為後會有不同的結果。