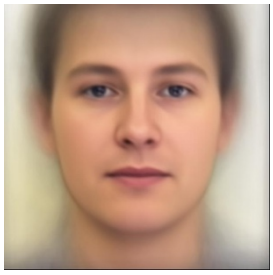
學號：B03505031 系級： 工海四 姓名：邱昱軒

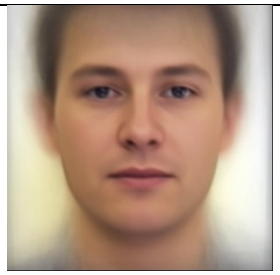
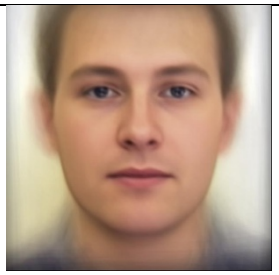
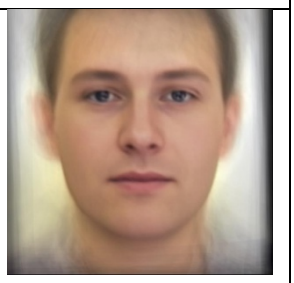
A. PCA of colored faces

(.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。

(.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。

ID	23	160	200	371
原圖				
重構				

(.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

4.1% 2.9% 2.4% 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。

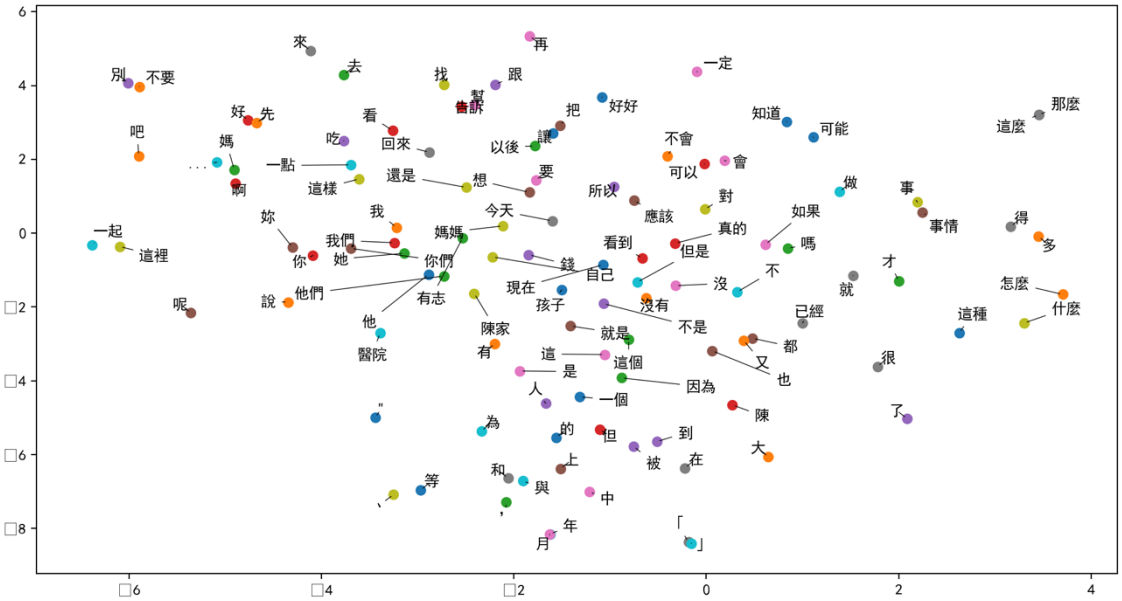
我使用 Gensim Word2Vec，調整的參數有

size（output 向量的維度）

min_count（出現次數大於此值的字詞才會被拿來計算，可以忽略噪聲字詞）

alpha（初始的 learning rate）

(.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。

經過 T-SNE 降維繪圖後，可以看到有類似意義的字詞會在附近。比如

右上角的「這麼」「那麼」、「事」「事情」

左上的「別」「不要」、「來」「去」

左邊的「我」「她」「他」「你」「妳」「你們」「他們」「我們」

下面的「年」「月」

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)

方法一：降維- T-SNE 降至 2 維，分群- KMeans。 Kaggle 0.28159
方法二：降維- T-SNE 降至 2 維，分群- cos distance。 Kaggle 0.07891
方法三：降維- PCA 降至 100 維，分群- KMeans。 Kaggle 0.03024
方法四：降維- Autoencoder 降至 32 維，分群- KMeans。Kaggle 0.88684
方法五：降維- Autoencoder 降至 64 維，分群- KMeans。Kaggle 1.0000

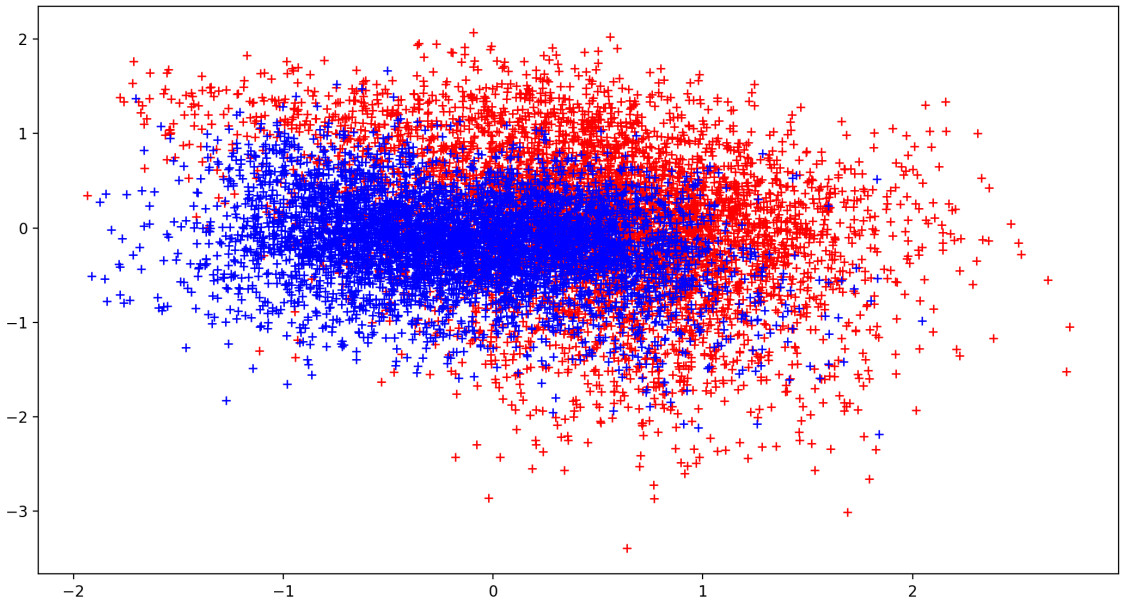
降維部分比較：

TSNE 因筆數較多所以跑很久，且只能降到 3 維以下，測試後發現 2 維效果最好。
PCA 不管降至幾維效果都不好
AutoEncoder 訓練快，而且效果不錯，第一層 Dense 的大小也對結果影響很大。

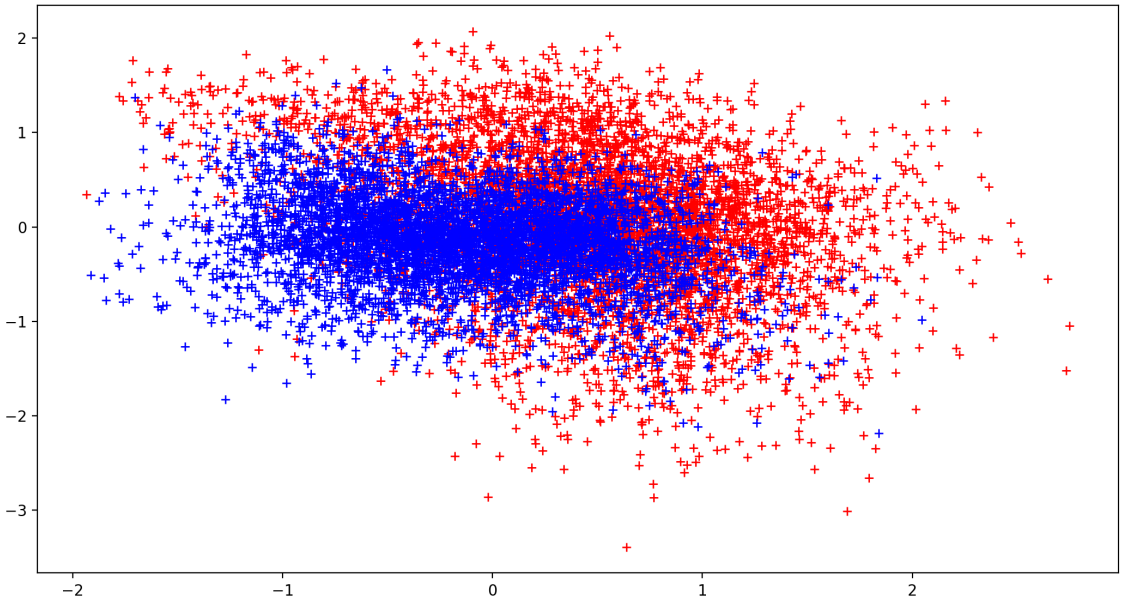
分群部分比較：

Scipy.spatial.distance.cosine 效果比自己定義的 cos distance 還差
KMeans 則在三者中最好

(.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。



兩者幾乎一樣。