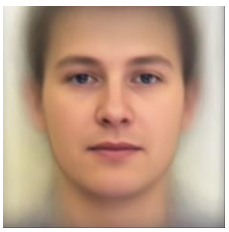
學號：R06942082　系級：　電信碩一　姓名：黃釋平

A. PCA of colored faces

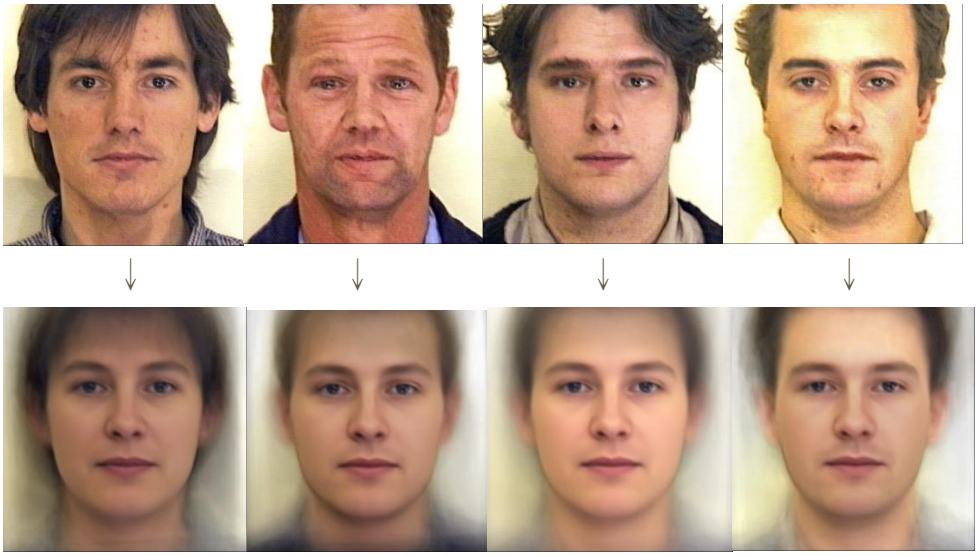
A.1. (.5%) 請畫出所有臉的平均。



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

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。



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	2	3	4
4.2%	2.9%	2.4%	2.2%

B. Visualization of Chinese word embedding

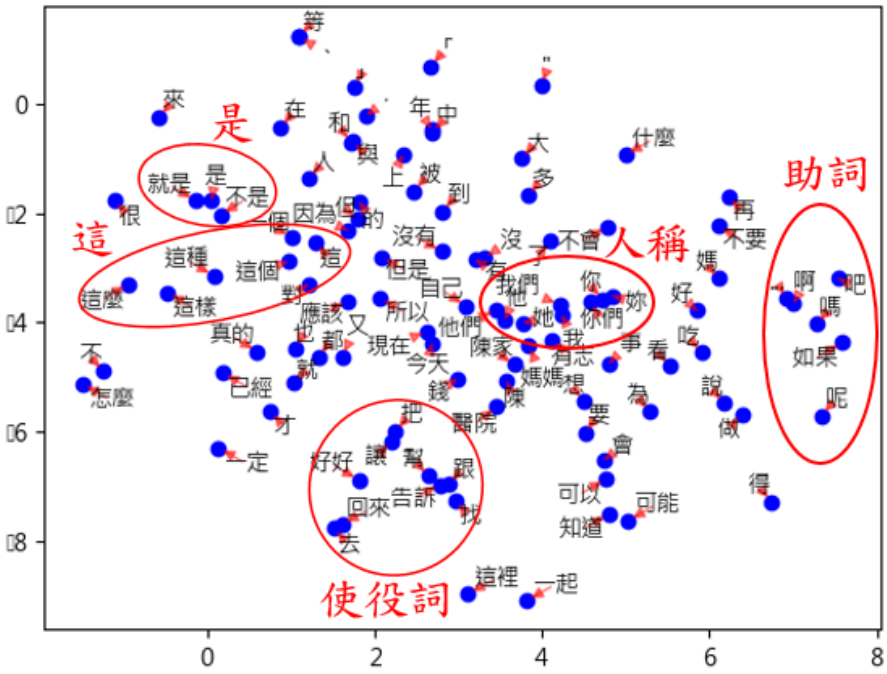
B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

我用 gensim.Word2vec，將 min_count 設為 4500，即出現 4500 次以上的詞才納入字典，可以找出較常用、易於分辨性質的詞。

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

見下題

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



有看到相同性質的詞、類似字詞都會聚在一起

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)

使用:

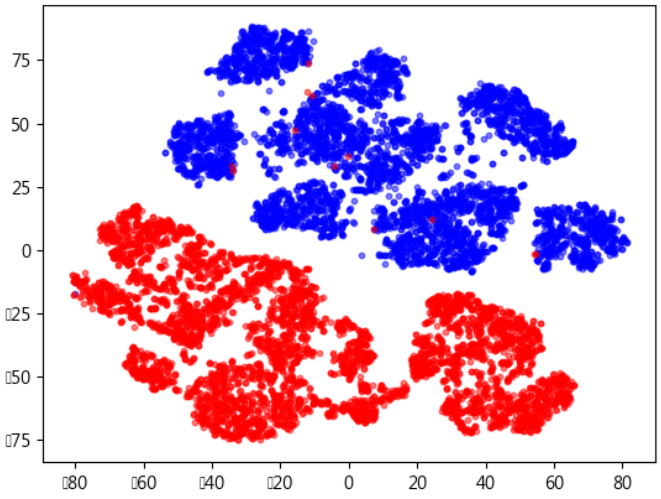
先由 Autoencoder 轉到 32 維，之後用下面兩種方法分類:

	sklearn.cluster.Kmeans	Cosine similarity (thres=0.707)
Kaggle	**0.89038**	**0.00331**

Cosine similarity 很難找到一個好的 threshold 衝高

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

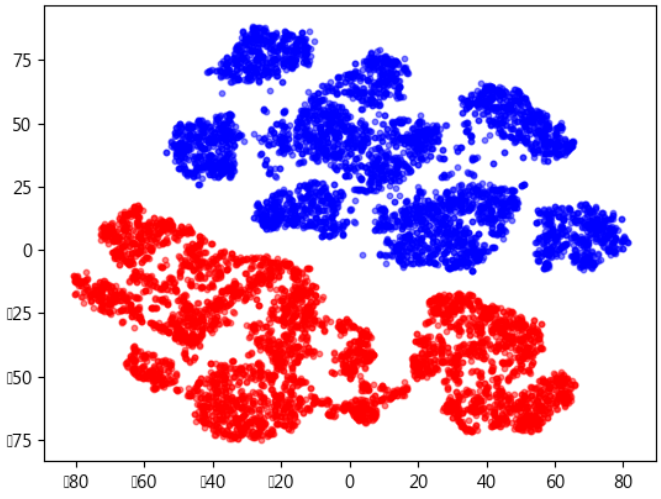
藍色為前 5000 張，紅色為後 5000 張



（我的預測）

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。

藍色為前 5000 張，紅色為後 5000 張



（正確 label）

我預測的約有 12 處錯誤，大部分是將第一類(藍)錯認為第二類(紅)，可能有較容易混淆的地方。