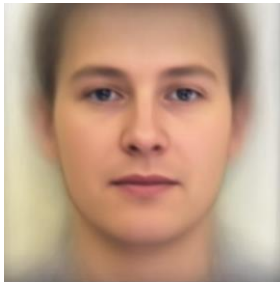


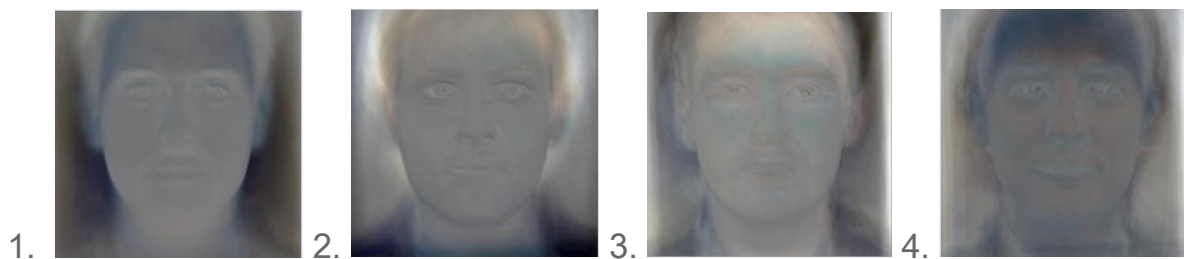
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。







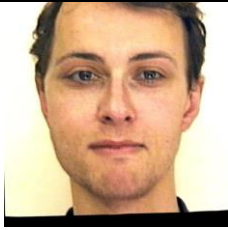


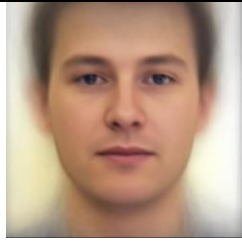
A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

前四大 Eigenfaces 如下



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。

178		265	
原圖	只取前四個 eigenvector	原圖	只取前四個 eigenvector
			

282		302	
原圖	只取前四個 eigenvector	原圖	只取前四個 eigenvector
			

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。

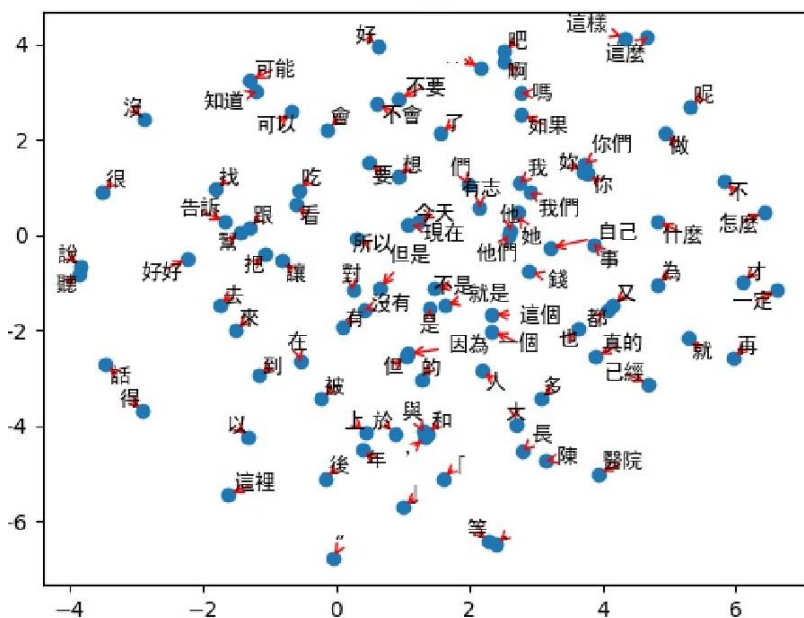
前四大 Eigenfaces 所佔的比重為：1. 0.04、2. 0.03、3. 0.02、4. 0.02

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。

我調整了 word2vec 中的 size=300 和 mincount=4000，size 為 vector 的 dimension，mincount 是字出現的次數。

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



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

用法相近的詞彙較靠近，例如

你、我、他們，

吧、啊、嗎，

來、去，

聽、說。

都是在圖上較接近的幾組字詞。

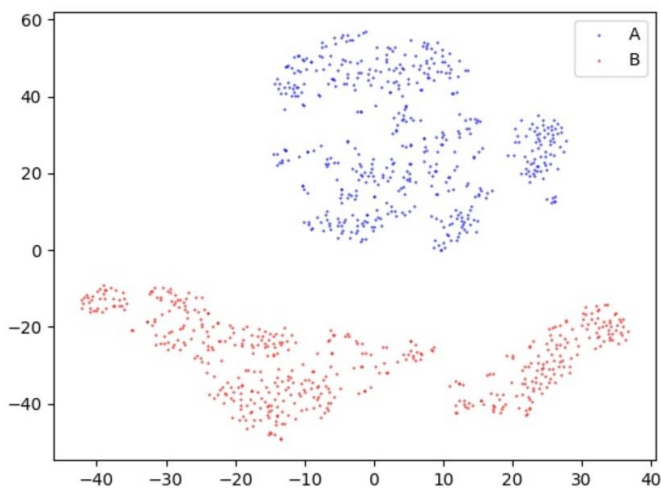
C. Image clustering

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)

PCA 準確率為 0.04，PCA 希望降為後的數據能以損失最小的方式代表原來的那組數據。

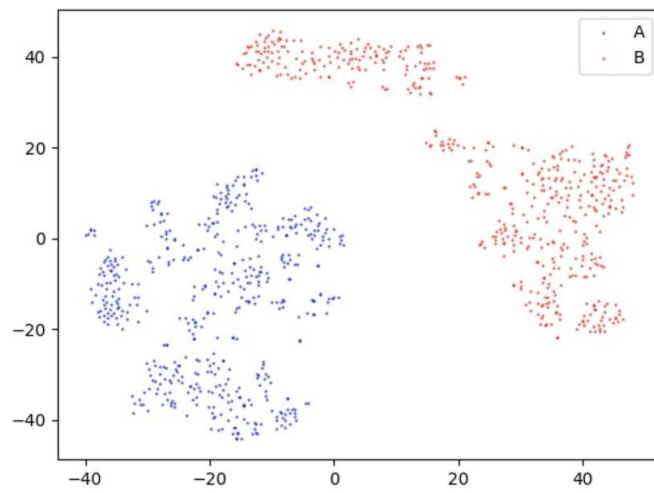
Autoencoder 的準確率為 0.94775，autoencoder 是希望輸入值等於目標值，在降維後還能還原原本的數據，表示找到有代表性的數據集合。

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。



兩者幾乎無差別，只是 TSNE 的降維方式是 random 的。