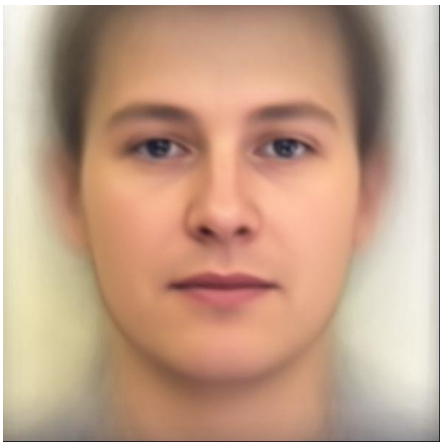
學號：B04901056 系級： 電機三 姓名：張承洋

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

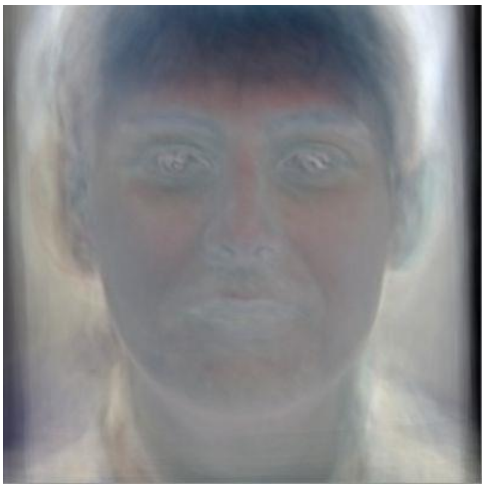


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

Eiganface_0 :

Eiganface_1 :

Eiganface_2 :



Eiganface_3 :

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。

reconstruction_1 :

reconstruction _10 :

reconstruction _22 :

reconstruction _37 :

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。

Eiganvalue_0 : 4.2%

Eiganvalue_1 : 3.0%

Eiganvalue_2 : 2.4%

Eiganvalue_3 : 2.2%

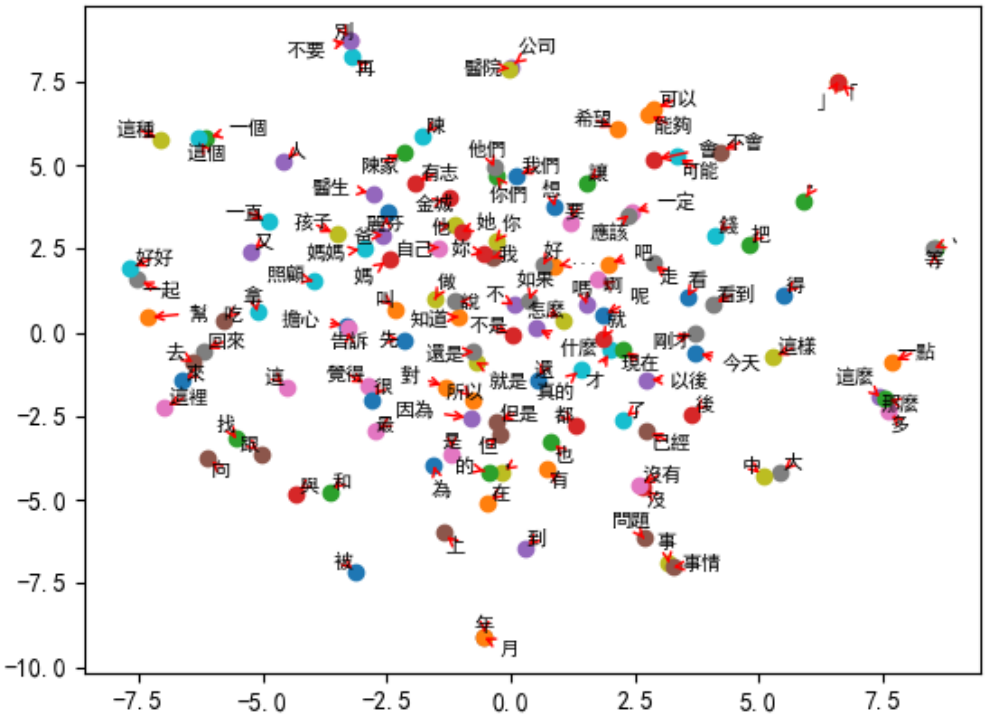
B. Visualization of Chinese word embedding

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。

我使用的是 gensim.Word2vec 套件，我使用的參數有：
Min_count = 1 代表經過 jieba 切出來的詞都會被加到 dictionary 裡

Size = 300 代表一個詞是由一個 300 維的向量表示

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。

我是選擇出現次數 3500 以上的詞。從上圖來看可以發現，雖然分類的效果沒有很好，但還是可以看初一些細節，例如：底下的年、月(時間詞)相當接近，中間的你、我、她、你們、他們人稱詞聚在一起，上方的公司跟醫院在一起，左邊的去、來、回來在一起…等，顯示一些常見而且詞意明顯的詞有不錯的 embedding 效果。

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)

方法一: DNN autoencoder + Kmeans

方法二:CNN autoencoder(中間有接 flatten+DNN) + Kmeans

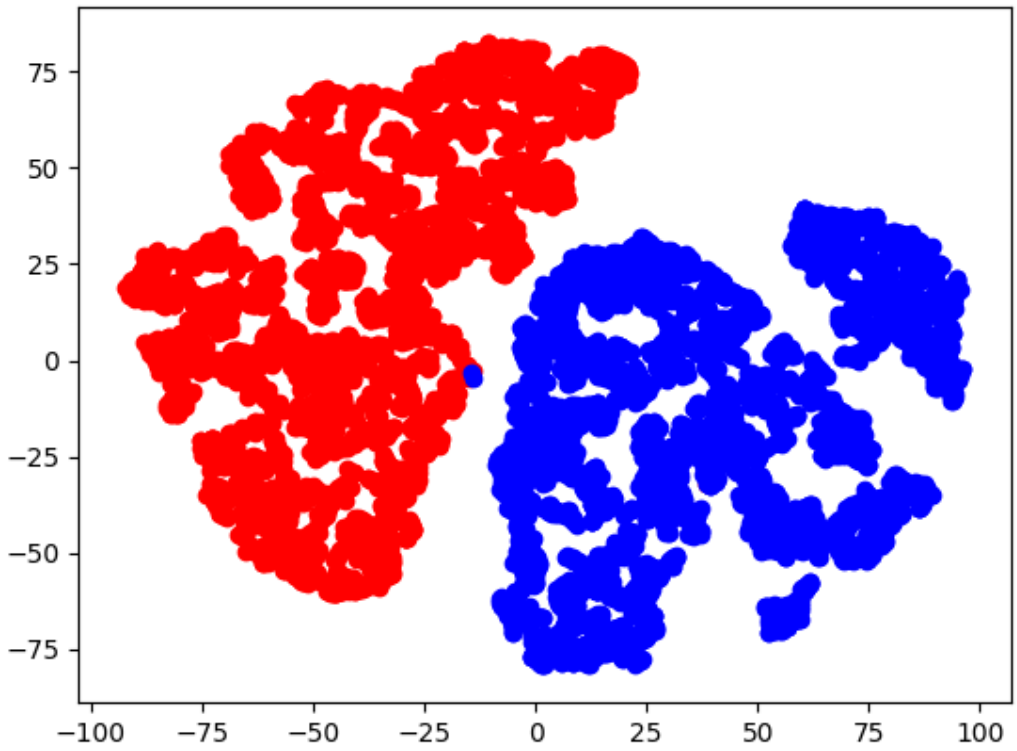
從結果來看我覺得這次作業主題比較單純，只是要分辨圖片來自哪一個 dataset，所以單純 DNN 就能有很好的效果，但如果要觀察

細節的話，或許將 CNN 加到 model 裡面會是一個比較好的選擇。

其他方法(PCA、TSNE): 表現都不好，可能是降維太多資訊流失

	方法一	方法二	PCA	TSNE
Score	1	0.03524	0.03091	0.03925

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。

從上圖可以發現我的 model 將資料投影到二維的分類能力還不錯，基本上用肉眼就可以分辨出來，除了有一些 dataset B(藍色)的資料跟 dataset A 容易搞混，而且 dataset B 內部也有一部分 data 跟其他性質不太一樣(右上)。