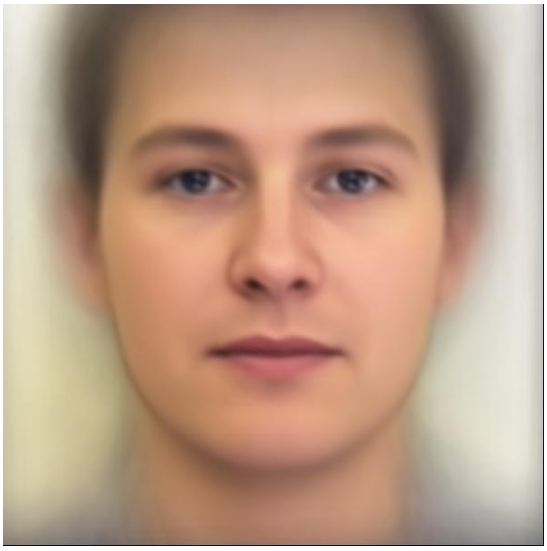
學號：R06946009 系級： 資料科學碩一 姓名：林庭宇

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，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

由左到右依序是第一大、第二大、第三大和第四大 Eigenvalues 的 Eigenvectors

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。

上面是四張原圖，下面是對應的四張用前四大 Eigenfaces 進行 reconstruction 的結果

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。

第一大 Eigenface 比重	第二大 Eigenface 比重	第三大 Eigenface 比重	第四大 Eigenface 比重
4.1%	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。

使用 gensim 套件，調整 gensim 套件中 Word2Vec 的兩個參數為 size=150 和 min_count=1

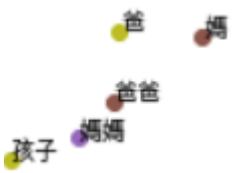
size 參數為 word vector 的維度

min_count 是一個最小次數，若一個詞出現的次數小於 min_count，那他就不會被視為訓練對象

原本直接設定 min_count=3000，但效果不太好，因此改設 min_count=1，將所有字詞都拿來訓練，之後才根據出現次數決定要來拿做圖的字詞

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。



親屬類

單人代名詞類

多人代名詞

日期

轉折語

語助詞

這

沒

有

是

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)

分群都使用 kmeans 分成兩群，降維則試了 Autoencoder 和 PCA

一開始嘗試使用 Autoencoder

Encode 過程：784 維→600 維→400 維→200 維→100 維→50 維

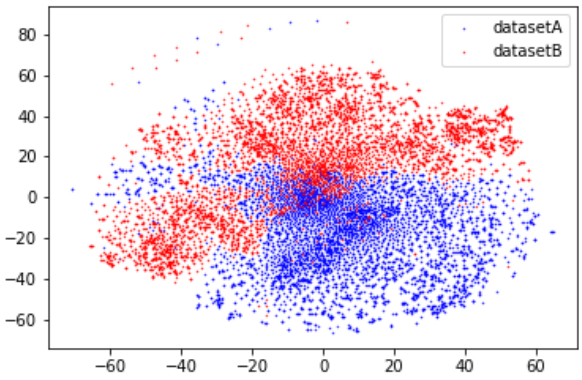
Decode 過程：50 維→100 維→200 維→400 維→600 維→784 維

但 Autoencoder 的結果不太好，因此改試 PCA

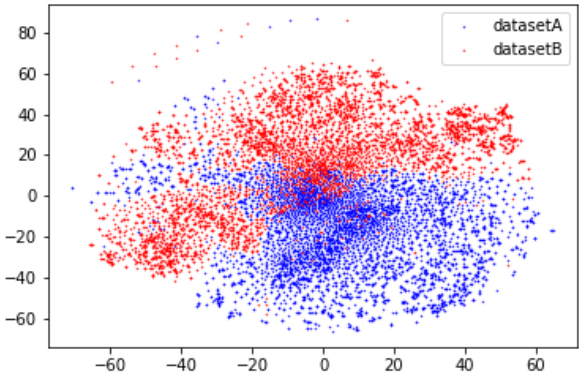
利用 PCA 將維度降到 400，並且把參數 whiten 設成 True，最後的結果很不錯

	Autoencoder	PCA
Kaggle Public Score	0.45424	1.00000
Kaggle Private Score	0.45409	1.00000

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。



最後算出來的分布看起來幾乎一模一樣，推測應該是 PCA 降維後做分群的準確度很高，因此預測出來幾乎跟答案一樣，所以分布也長得一樣