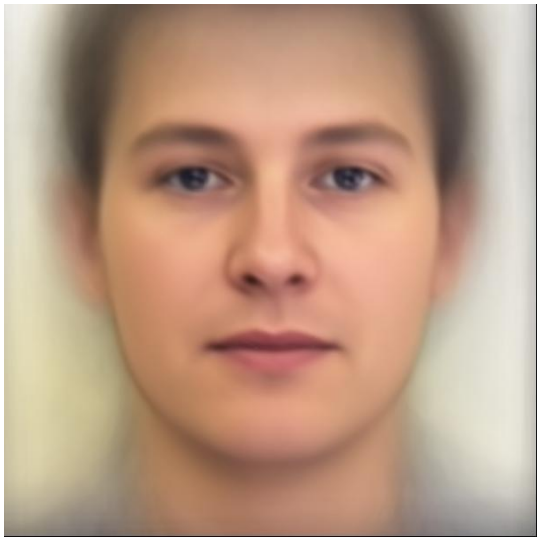
學號：R06945003 系級： 生醫電資一 姓名：林鈺盛

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。

(依序為前四大 eigenfaces)

(.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。

以上四張左圖為用前四大 eigenface 進行重建，右圖為用全部 eigenface 進行重建，可以發現只用四個進行重建的結果看起來幾乎都一樣。

(.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。

依序為 4.1%，3.0%，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。

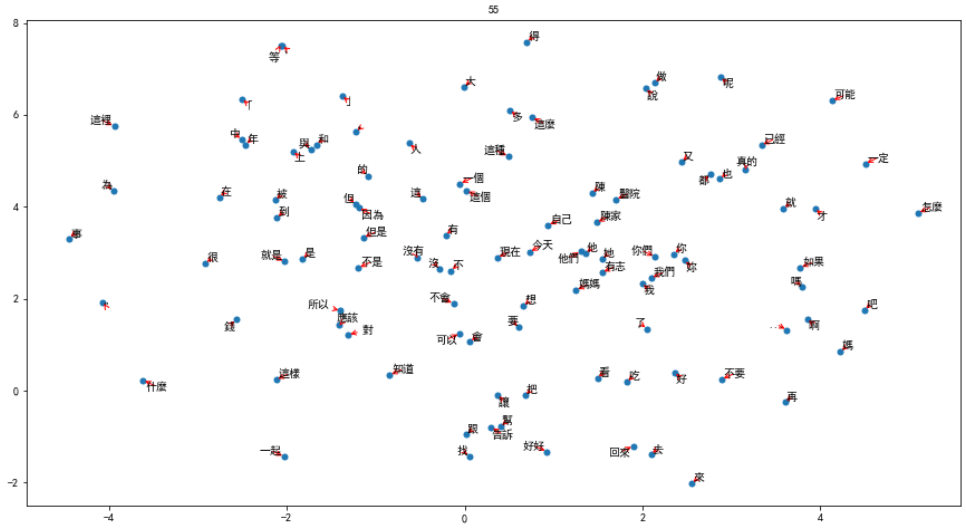
使用 gensim 的 Word2Vec，其中調整的參數有：

min_count = 4000：針對出現次數大於 4000 的詞作訓練

sg=1：使用 skip_gram 算法

size=256：output 的 wor2v 有 256 維度

(.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。

發現可以湊成一句句子的詞通常比較靠近，而詞義相近的詞也會比較靠近，整體來說視覺化的結果還算不錯。

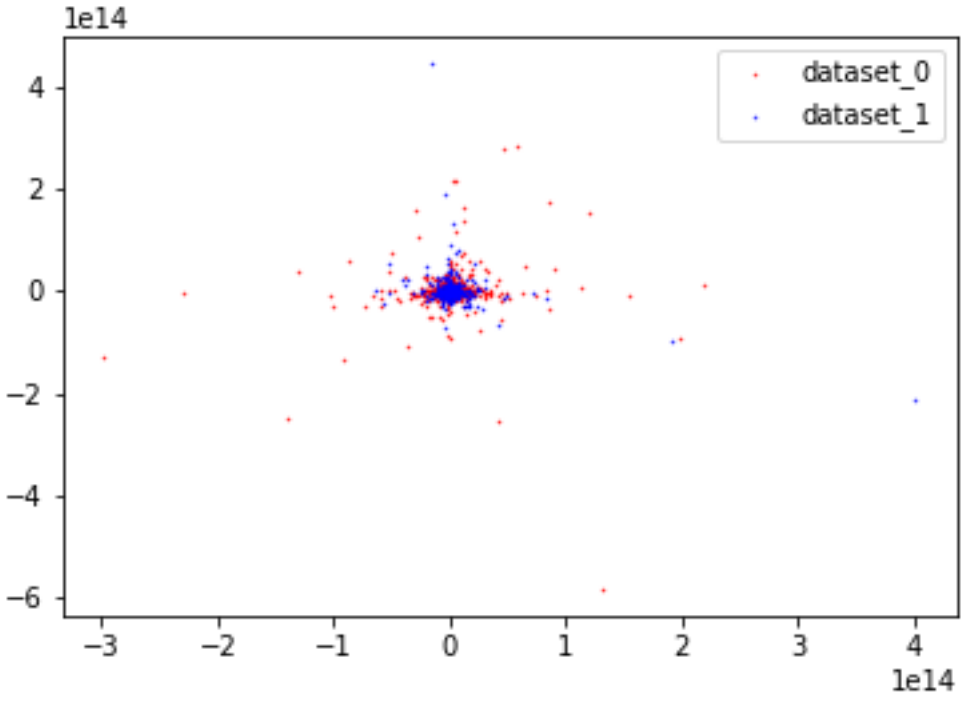
C. Image clustering

(.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)

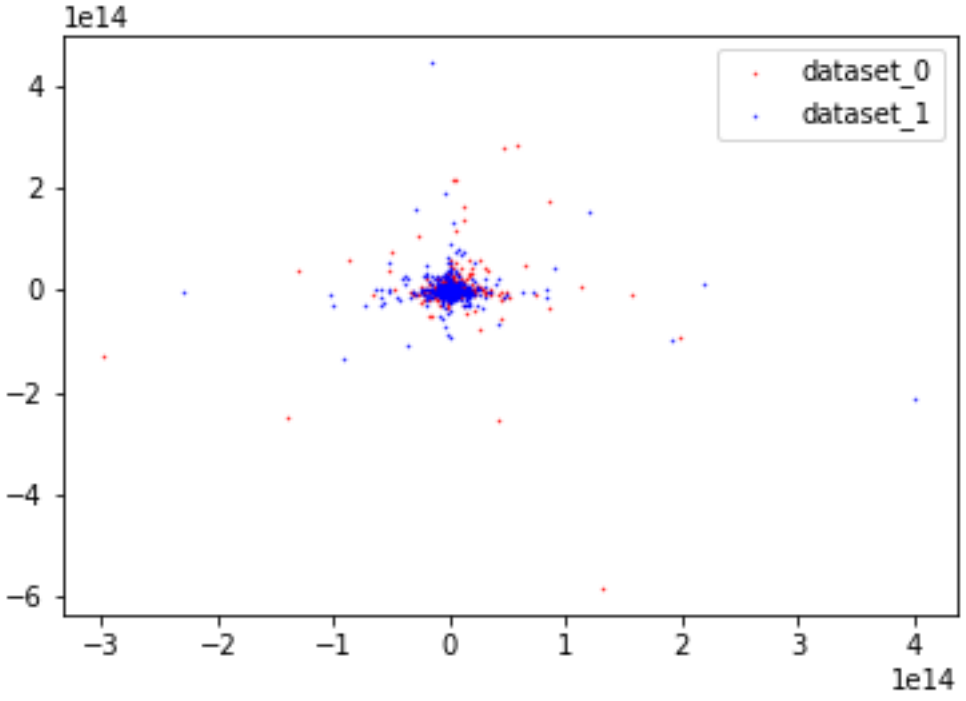
1.auto-encoder 方法將 dim 降到 64 維，再用 kmeans 分成兩群來預測，在 kaggle public/private 皆拿到 1 的分數。

2.一開始是嘗試直接 run TSNE 降到 2 維，再用 kmeans 分成兩群，不過效果很差，kaggle public/private 只有 0.08 左右，應該是直接降到 2 維損失太多的資訊。

(.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。



第二題預測的結果中有 6995 筆 data 被歸類在第一個 data set，兩個圖的視覺化似乎都不是太好，我想原因可能是因為我 train 出來的 encoder 只把 dim 降到 64 維，之後再 TSNE 將維的過程中 loss 掉不少資訊，或許之後可以嘗試直接 train 一個可以降到二維的 encoder 來比較結果。