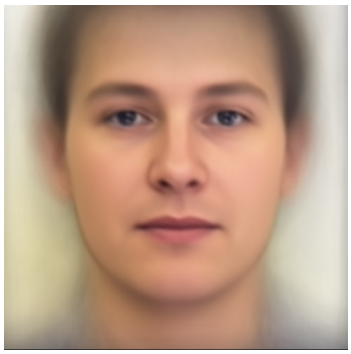
機器學習 hw6 report
b04507009 電機三 何吉瑞

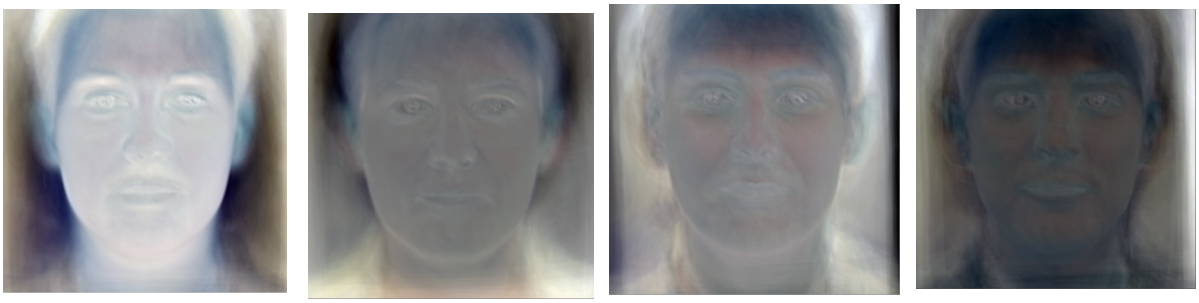
Part A: PCA of colored faces

(.5%) 請畫出所有臉的平均。



看起來還像是大衆臉，但是周圍可能隨每個人髮型等不同而邊界有些模糊

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



左起依序為前四大 eigenfaces，都有蠻明顯的偏白或黑，但是沒有全部都偏白或偏黑，此外第一大特徵臉的邊緣對比相對高

(.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。

臉部都接近大衆臉，沒有太多個人特徵，但是周圍部份像髮型或是黑框邊界，可以觀察出差別

(.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

第 1: 4.2%

第 2: 3.0%

第 3: 2.4%

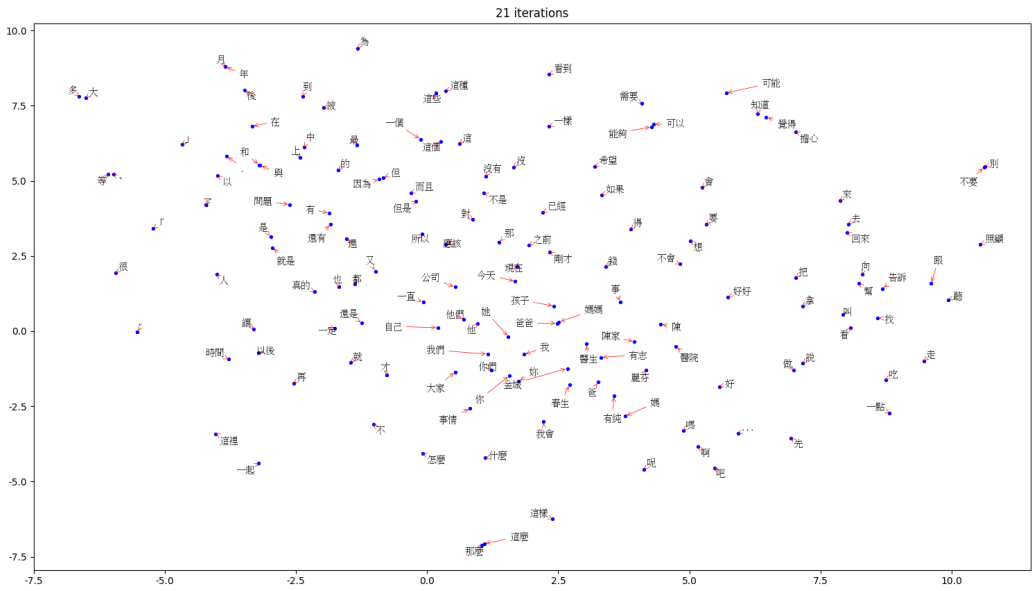
第 4: 2.2%

Part 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。

我用 gensim 的 Word2vec，有調整過 embedding 的 size 調整 embedding 後的大小，也就是用幾個維度去描述每個單詞，128 附近效果不錯，以及 max_vocab_size 和 min_count 去調出現的次數，

(.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。

意思相近的詞，降維後向量的終點明顯接近，比如『可以』和『能夠』，『和』和『與』，同一個屬性的詞也會接近，比如『爸爸』和『媽媽』，『年』和『月』

Part C: Image clustering

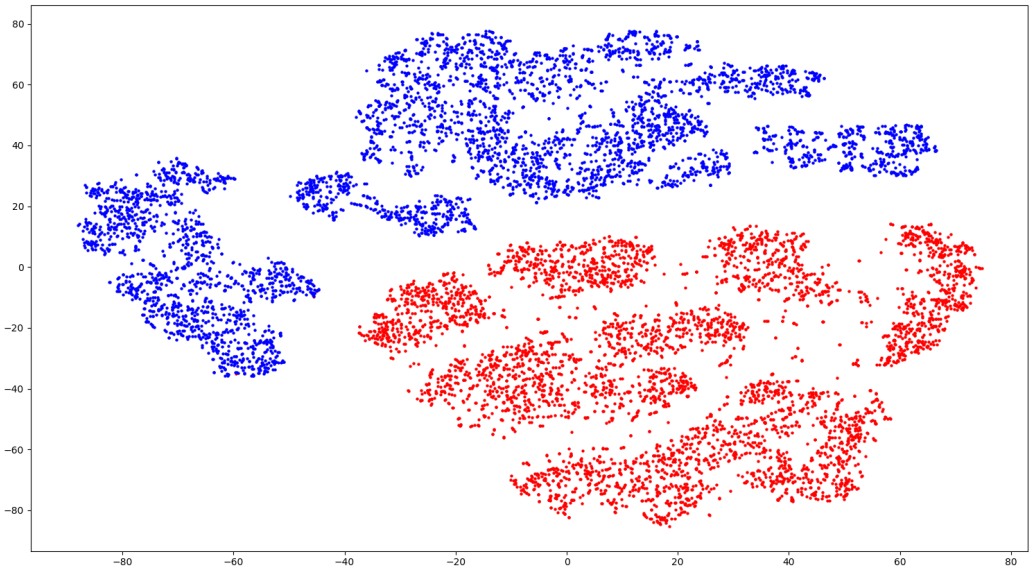
(.5%) 請比較至少兩種不同的 feature extraction 及其結果。

用 DNN 和 PCA 做 autoencoder

DNN 降維到 75 public 0.94239 private 0.94359

PCA 降維到 600 public 1.0000 private 1.0000

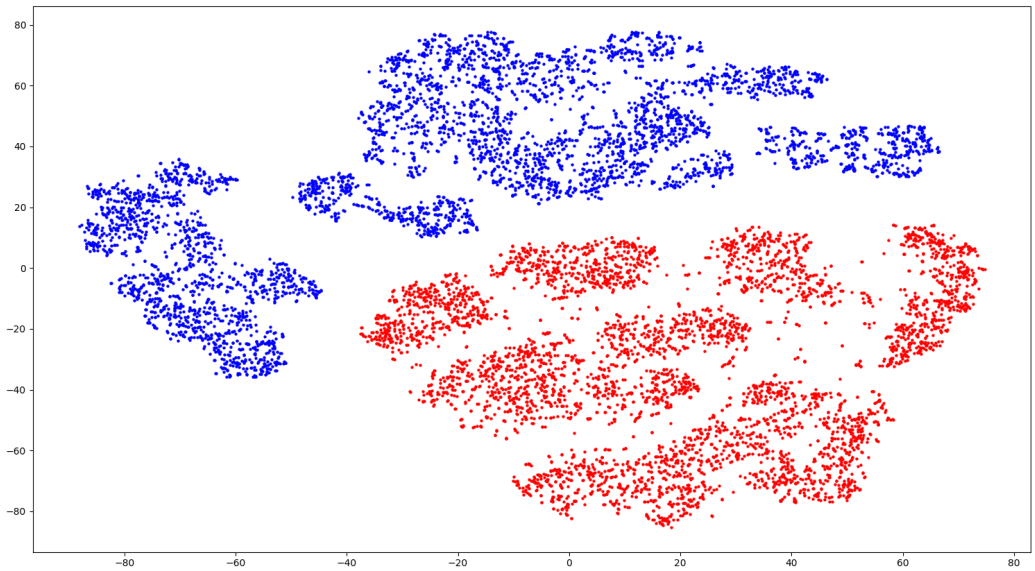
(.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



紅色預測為 datasetA 藍色預測為 datasetB，預測結果是 DNN 做 encoder 降維到 75

(.5%) visualization.npy 中前 5000 個 images 來自 dataset A，後 5000 個 images 來自 dataset B。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。(visualization.npy 將在 Kaggle deadline 之後公布在 Kaggle 上)

紅色為 dataset A 藍色為 dataset B



結果幾乎符合預測