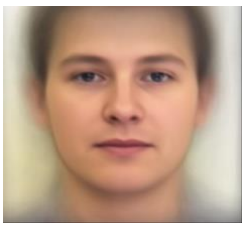
學號：R05921016 系級： 電機碩二 姓名：傅鈞笙

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

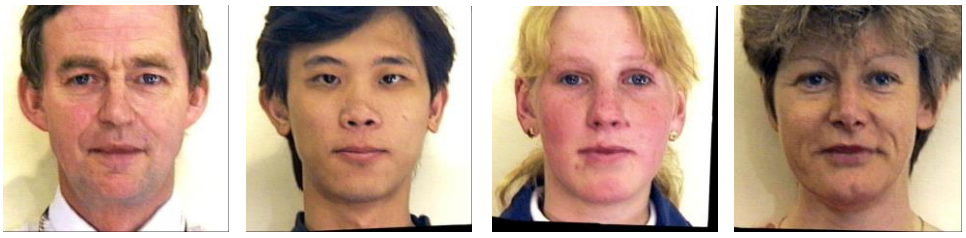


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

1st 2nd 3rd 4th

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。

1st 2nd 3rd 4th



↓ ↓ ↓ ↓

僅使用四個 eigenface 的重建效果其實不大，都跟 averageface 不多。

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。

[4.1% 3.0% 2.4% 2.2%]

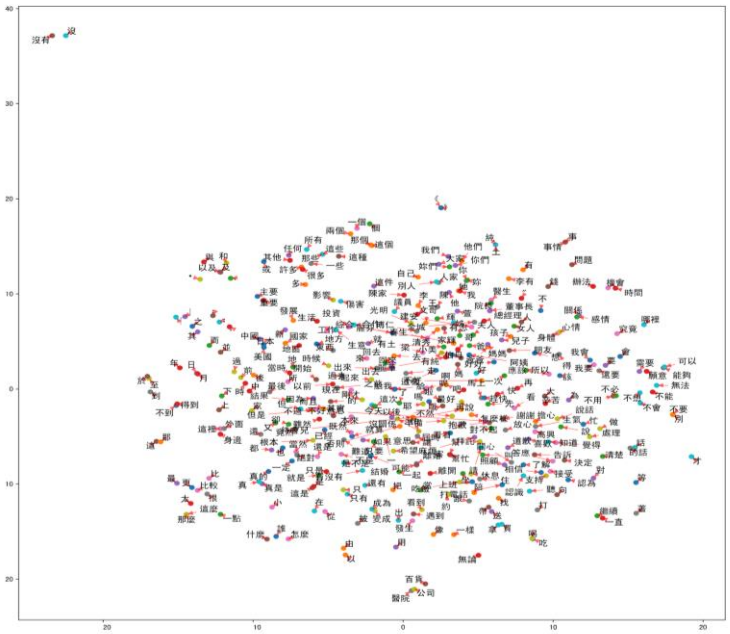
B. Visualization of Chinese word embedding

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 套件計算 word2vec (先經 jieba 斷詞)，調整的參數為 size 以及 min_count。Size 代表 embedding size，決定將一個詞轉成多長的向量；min_count 代表該詞出現次數的 threshold，出現次數<min_count 則不將該詞加入字典。本次實驗取 size = 400, min_count = 1。

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

只取出現次數>1200 的詞進行顯示。



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

可以觀察到語意相近的詞都出現在附近。

不知為何，沒和沒有和其他詞距離很遠 (B.2 圖)。

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 : NN auto encoder，double layer (784→600, 600→400)。

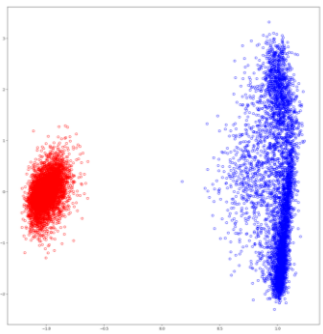
方法二 : PCA 降維 (784→400)。

使用上列兩方法降至 400 維後，用 K-means cluster 成 2 群。

準確率	testing private acc.	testing public acc.
NN auto encoder	0.85692	0.85692
PCA	1.00000	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。

使用的是 PCA 方法，取得 K-means 估測的 label 後，將 400 維的資料再度使用 PCA 降至兩維顯示，如下圖。



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。

由於 PCA 方法其實已經有 100% accuracy，所以其實做出來跟 ground-truth 應該是一樣的，下圖維 ground-truth label。