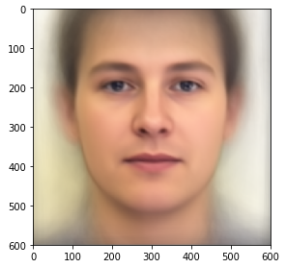
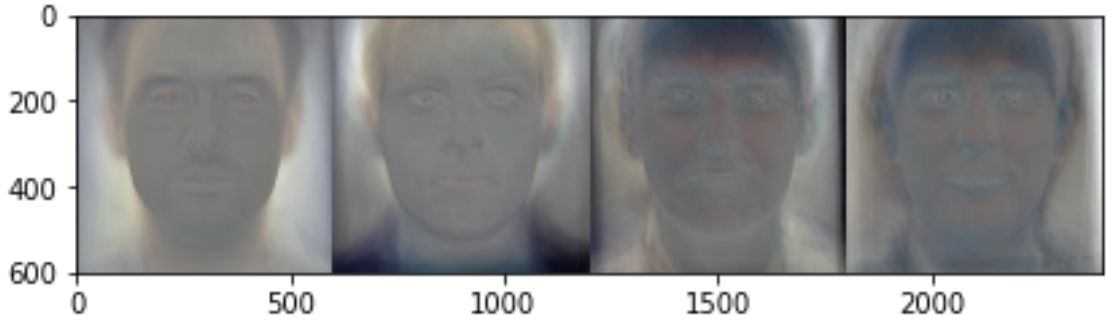
學號：B02123456 系級： 電機四 姓名：涼宮春日

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。



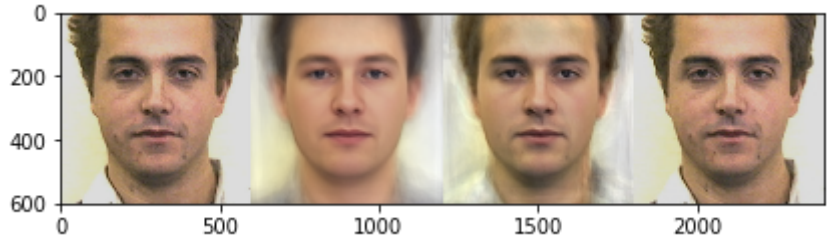
A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



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。

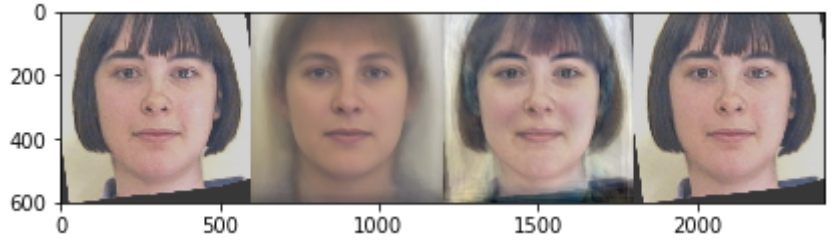
左到右： 原圖, eigenfaces= 4, eigenfaces= 40, eigenfaces= 400

21:



342:

220:



14:

A.4. (.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。

0.041, 0.030, 0.024, 0.022

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。

斷詞：jieba, dict=big5

embedding: gensim.word2vec,

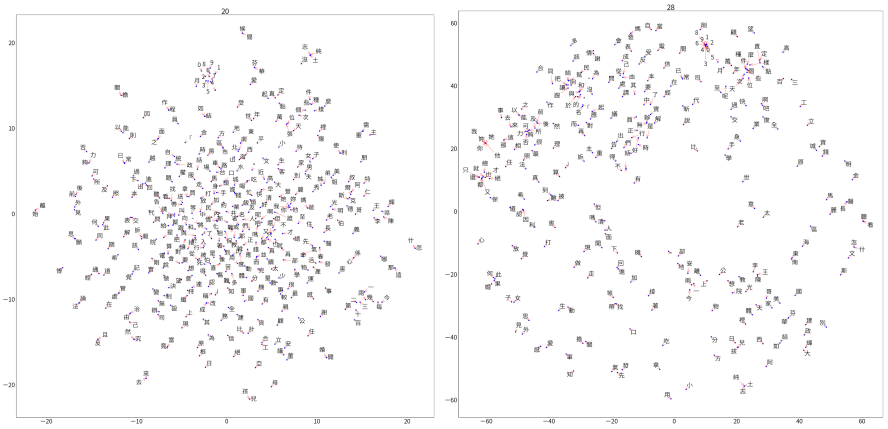
size = 64, embedding vector的大小

n_iter = 20, train 幾個epoch

左圖 有min_count = 3000, 右圖的min count 是在visualization前才切掉的

降維： sklearn.mainfold.TSNE

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



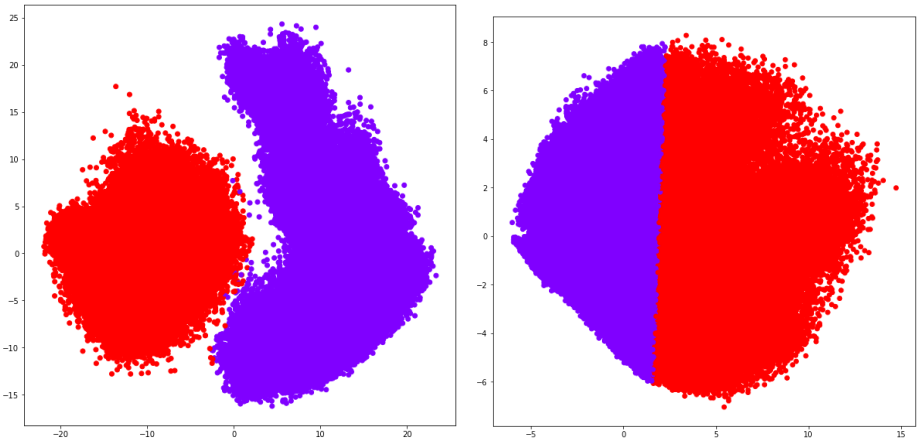
B.3. (.5%) 請討論你從 visualization 的結果觀察到什麼。

如果只選 > k =3000 的字來做降維 它們會變成一大群, 比較明顯的cluster只有數字

如果是在visualization時才切掉 < k 的 就可以看到比較分散的cluster, 數字一樣是很緊的一群但其他字也有自記的群

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)



左：autoencoder 降維, k-means cluster, 0.99 accuracy

右：pca降維, k-means cluster, 0.03 accuracy

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。