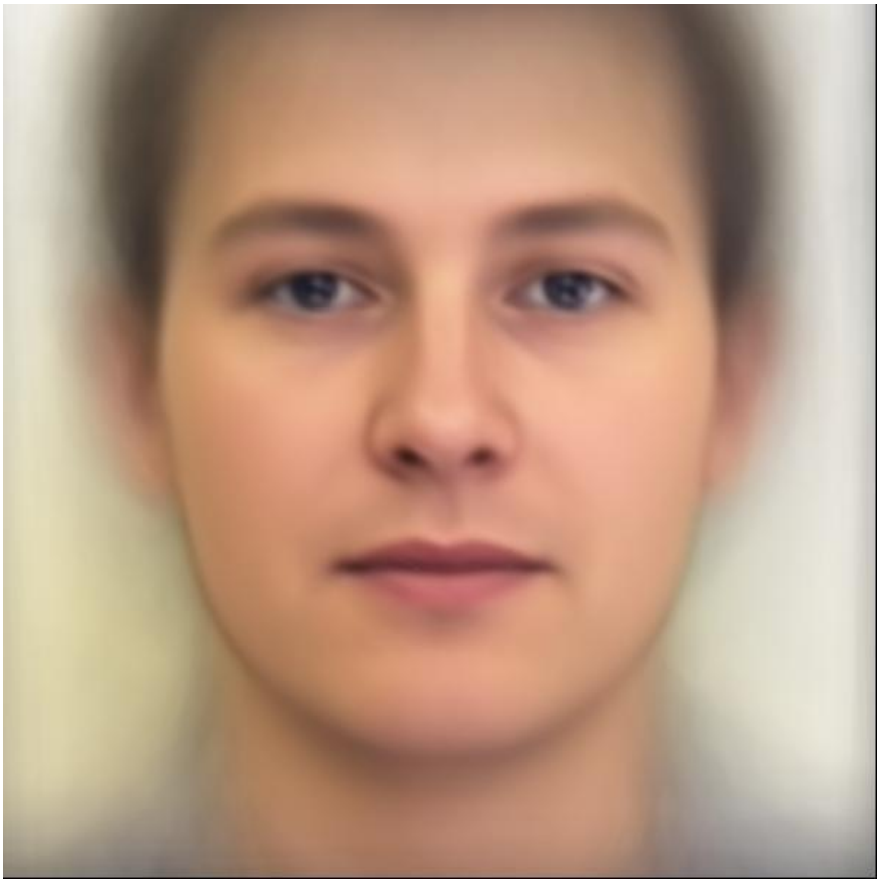
學號：R05546014 系級： 工工碩二 姓名：洪紹綺

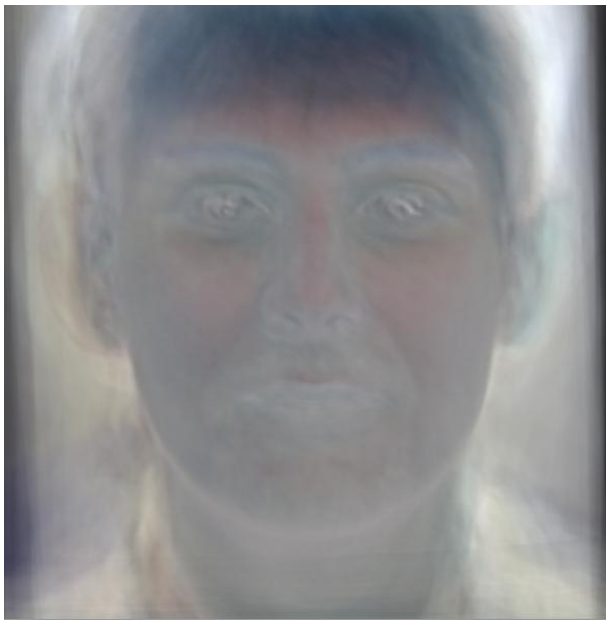
A. PCA of colored faces

Collaborator : 林冠廷 R05546016

A.1. (.5%) 請畫出所有臉的平均。

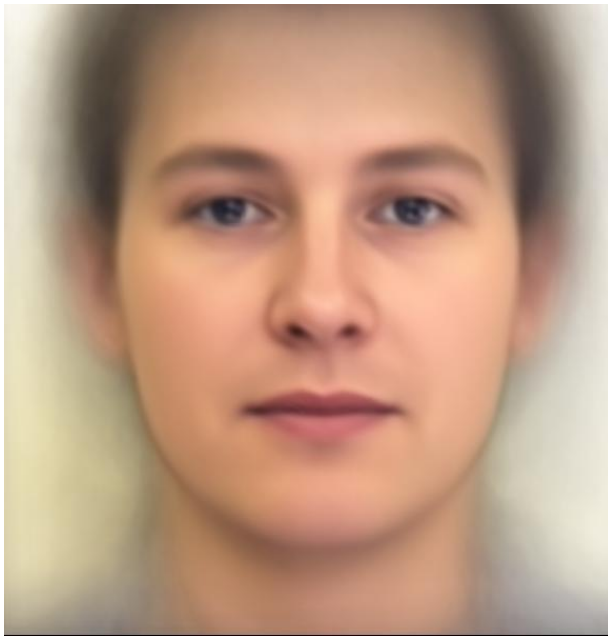


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

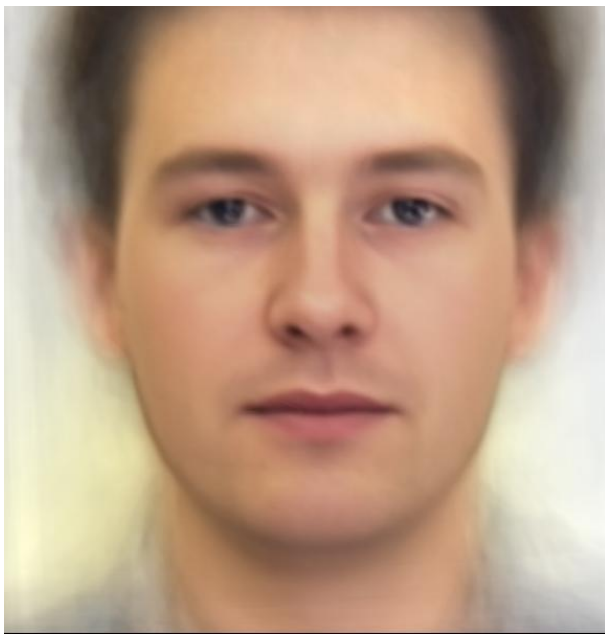
0	1
	
2	3
	

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。

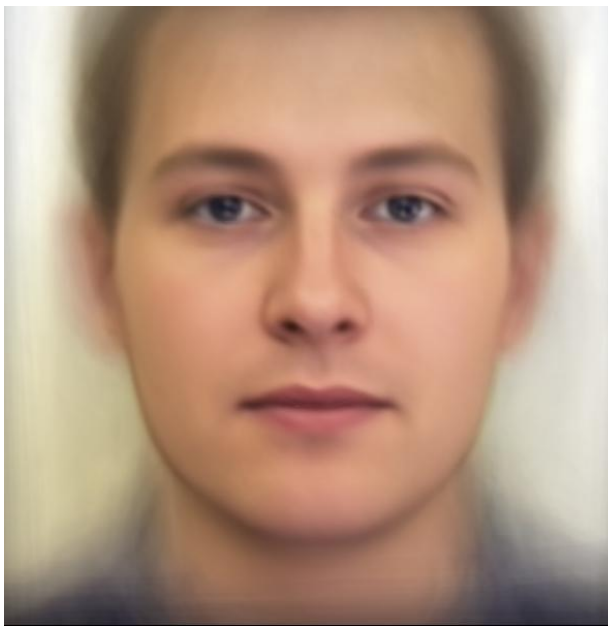
7



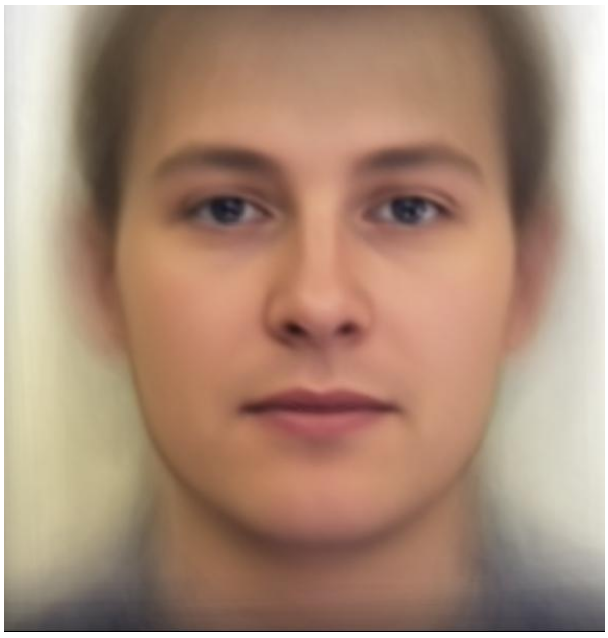
17



70



77



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。

```
0.0414462483826
0.0294873222511
0.0238771129321
0.022078415569
```

=> [4.1%, 2.9%, 2.4%, 2.2%] #

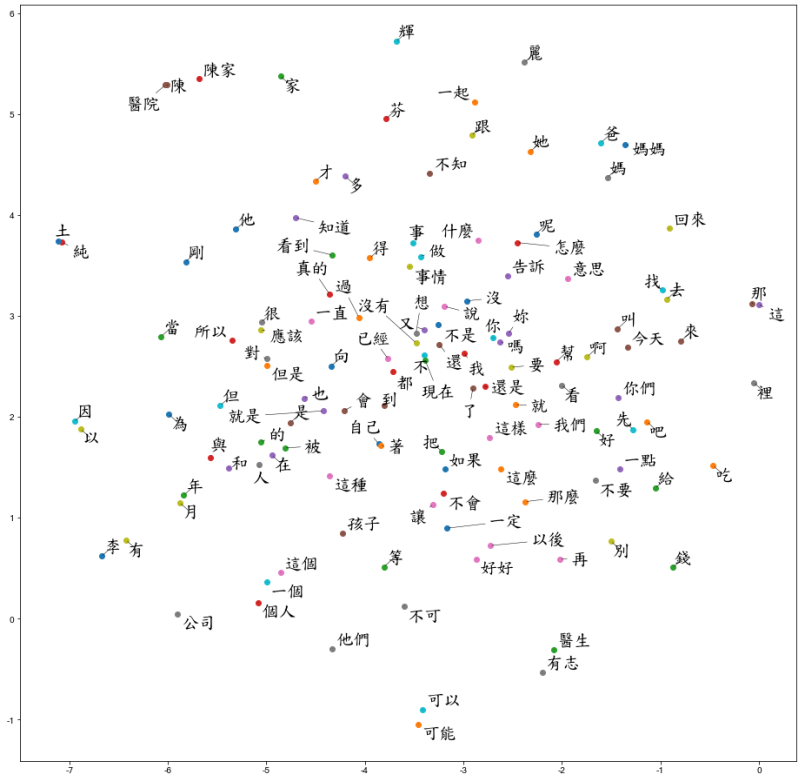
B. Visualization of Chinese word embedding

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。

word_model = word2vec.Word2Vec(sentences, size=300, min_count=4000, sg = 0, window = 5, iter=10)

使用 gensim 函數庫的 Word2Vec 建立 300 維的詞向量，不計出現次數小於 min_count=4000 次的詞，每一個詞在句子中與預測詞最大距離為 window=5，迭代 iter=10 次。

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。

使用 Jieba 搭配助教提供的字典，設定全模式斷詞，所以”陳家醫院”，會被切成”陳”、”陳家”、”家”、”醫院”等這幾種詞。其實斷詞效果滿有趣的，醫生與醫院的距離比我想像的遠，但也有可能是因為一個代表地點，一個代表人物，李有志是醫生，所以有志跟醫生距離最近，非常合理。

C. Image clustering

Collaborator : 張少豪 R05546038

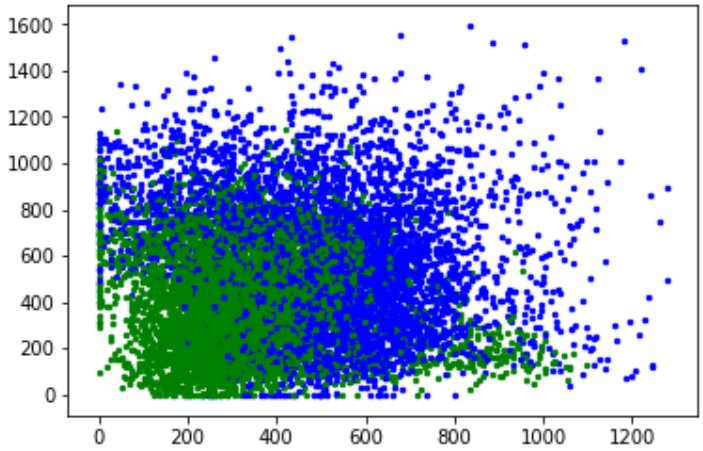
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

降維方法	Kaggle Public
autoencoder(維度=128)	0.99929
autoencoder(維度=512)	0.02634

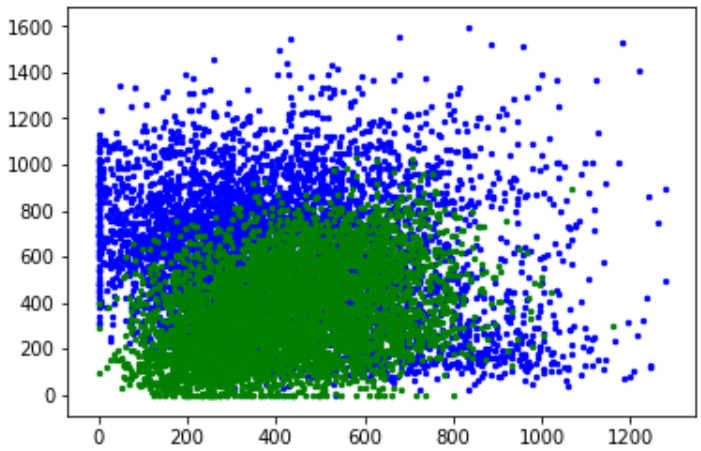
clf = cluster.KMeans(init='k-means++', n_clusters=2)

cluster 皆為 K-Means；維度較低的模型準確率較高。

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。



隨意取兩個維度作圖，我的 model 預測準確率大概 80%，並非完全正確。