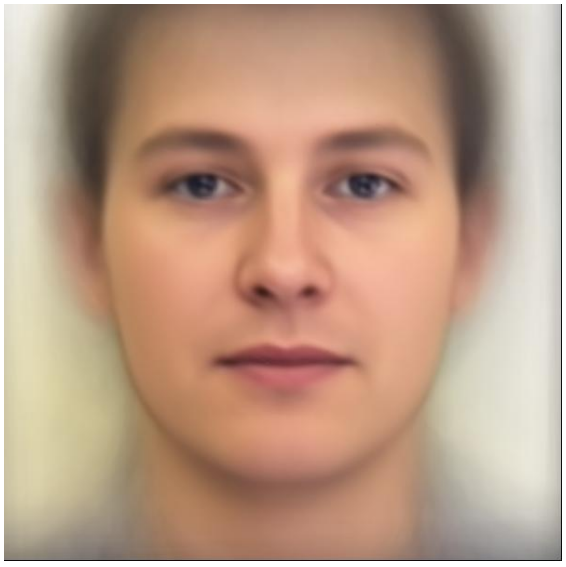
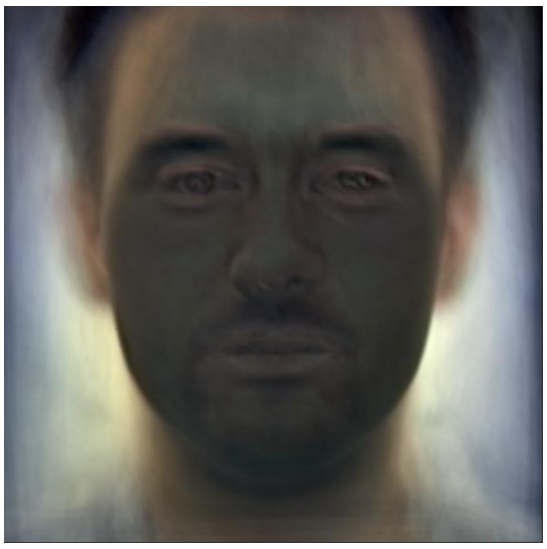
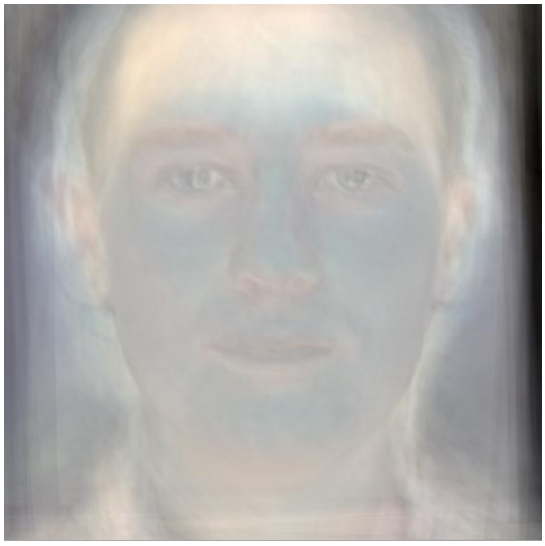
學號：R05944041 系級： 網媒所 姓名：戴長昕

1. PCA of colored faces
    1. (.5%) 請畫出所有臉的平均。

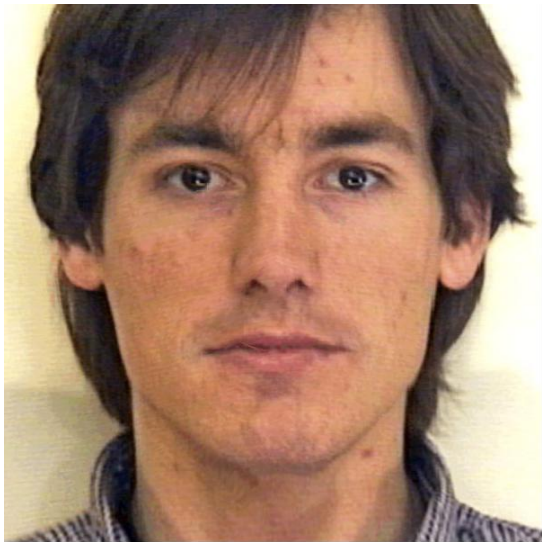


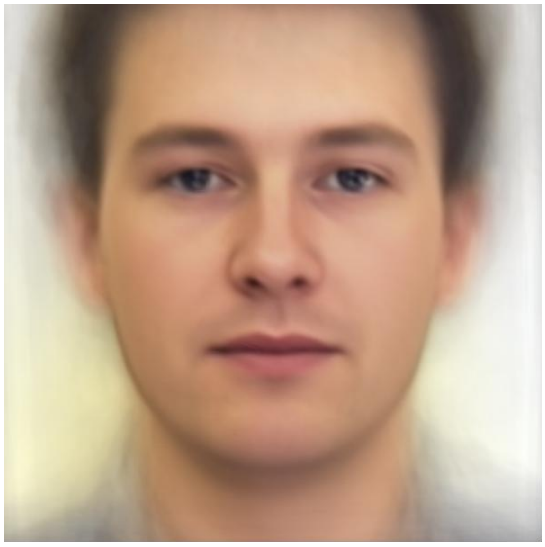
    2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

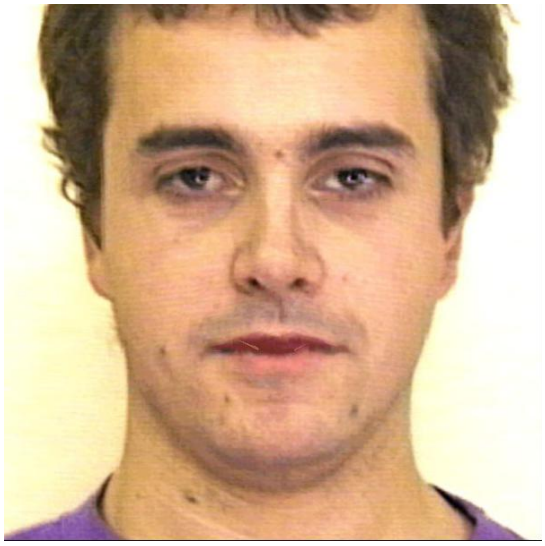


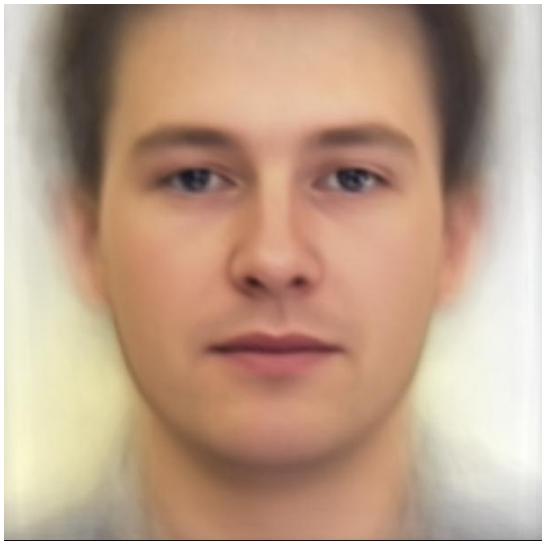


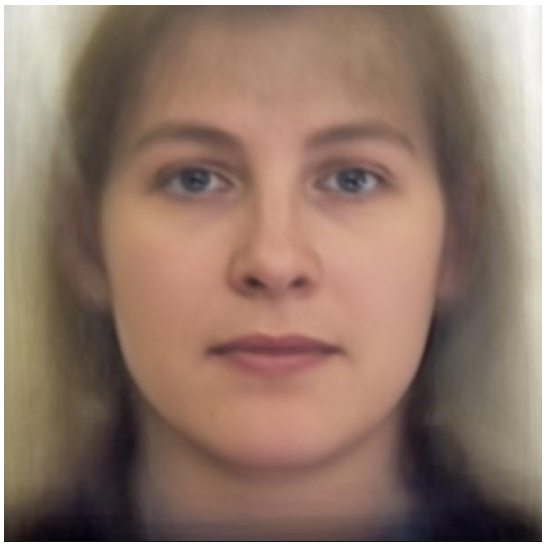
3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

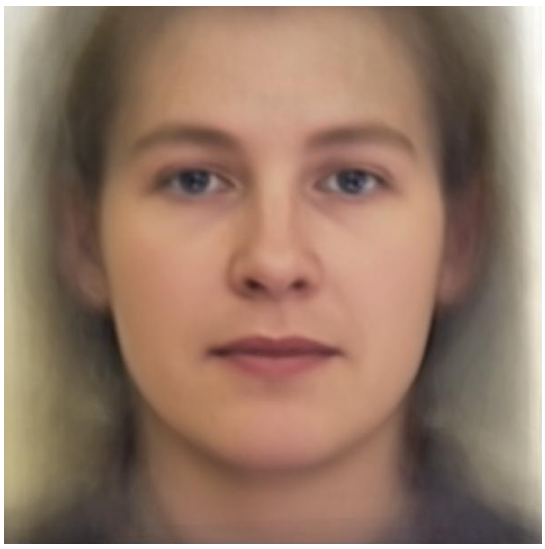










4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

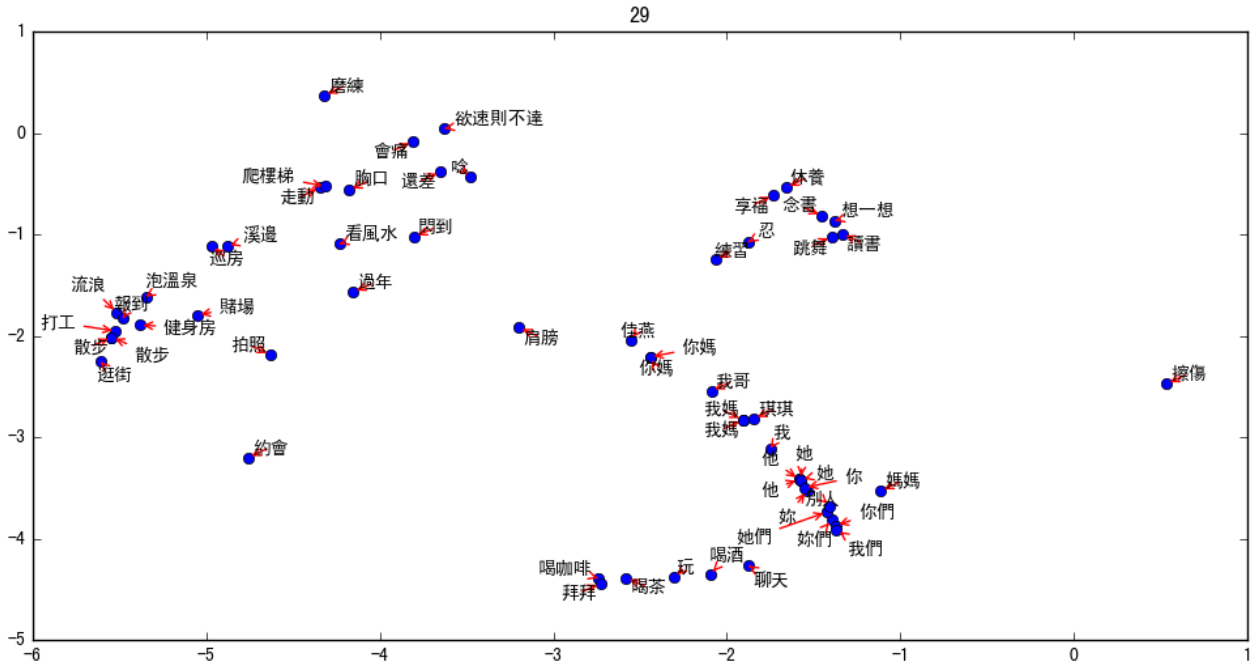
4.1%	3.0%	2.4%	2.2%

2. Visualization of Chinese word embedding

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 word2vec，調整參數有：
   - window = 5 : 左右滑動的窗口，一次看幾個字。
   - size = 256 : embedding的維度。

- min_count = 20 : 只看出限字數大於20的字。
- workers = 8 : 使用8個cpu核心。

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



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

語意相似的字詞成群聚出現在圖中，如：「你、我、他」等代名詞，「喝茶、喝咖啡、喝酒、聊天」等動詞。詞性相同也會有所區分成群，如「散步、逛街、泡溫泉」等休閒屬性的成一群。

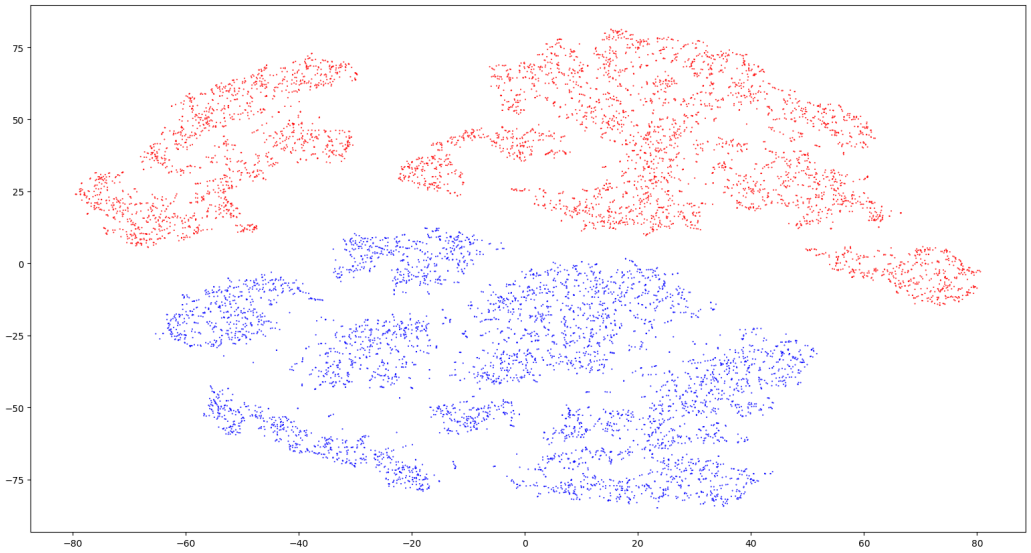
3. Image clustering

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)

	Kaggle Public Score
GMM	
Auto Encoder + Kmeans	0.91333

Encoder將原始pixels值降維到32維，再用Kmeans分群。

2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



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。

沒有太大不同。