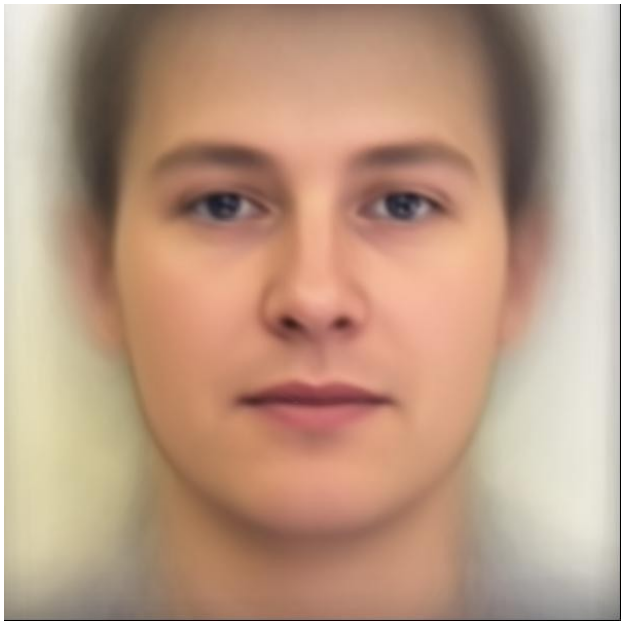
學號：B03201031 系級： 數學四 姓名：王楷

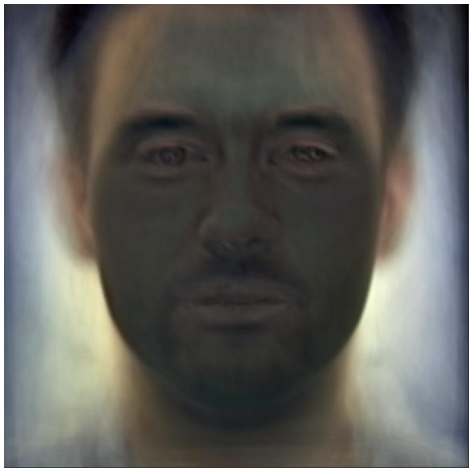
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

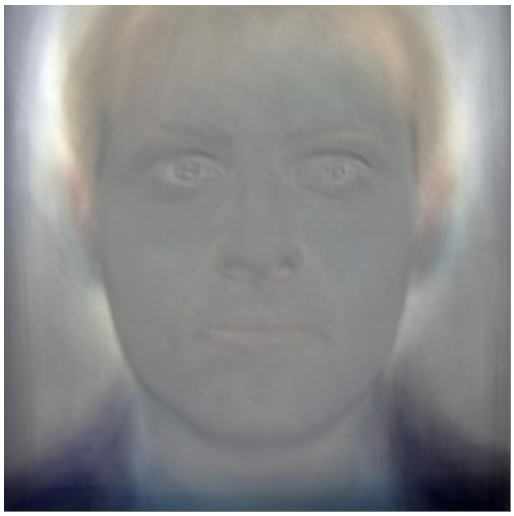


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

1

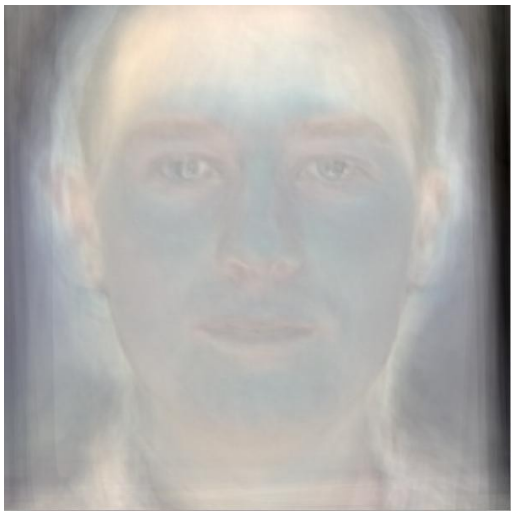


2



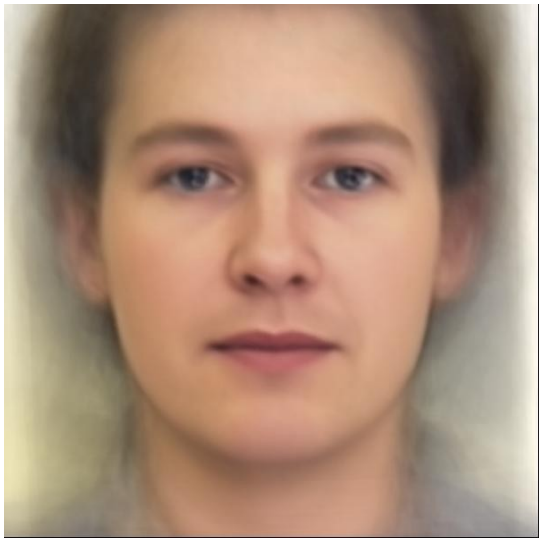
3

4

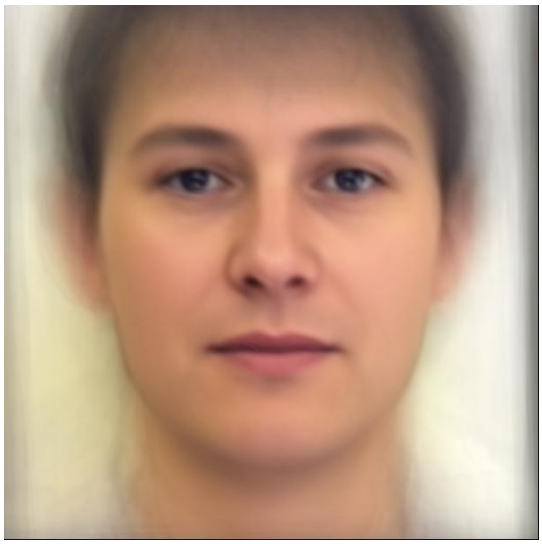


A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

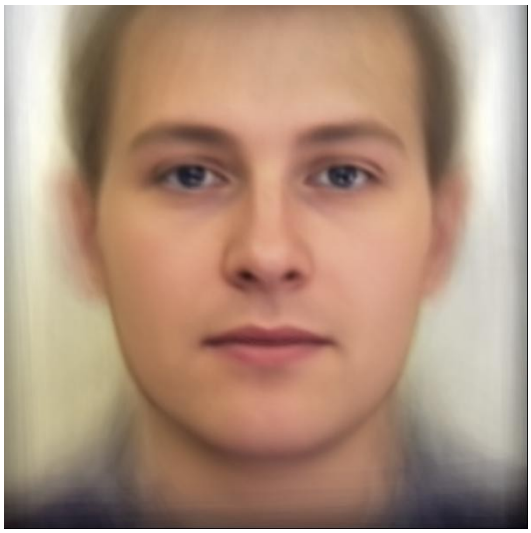
35



100



211



242

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。

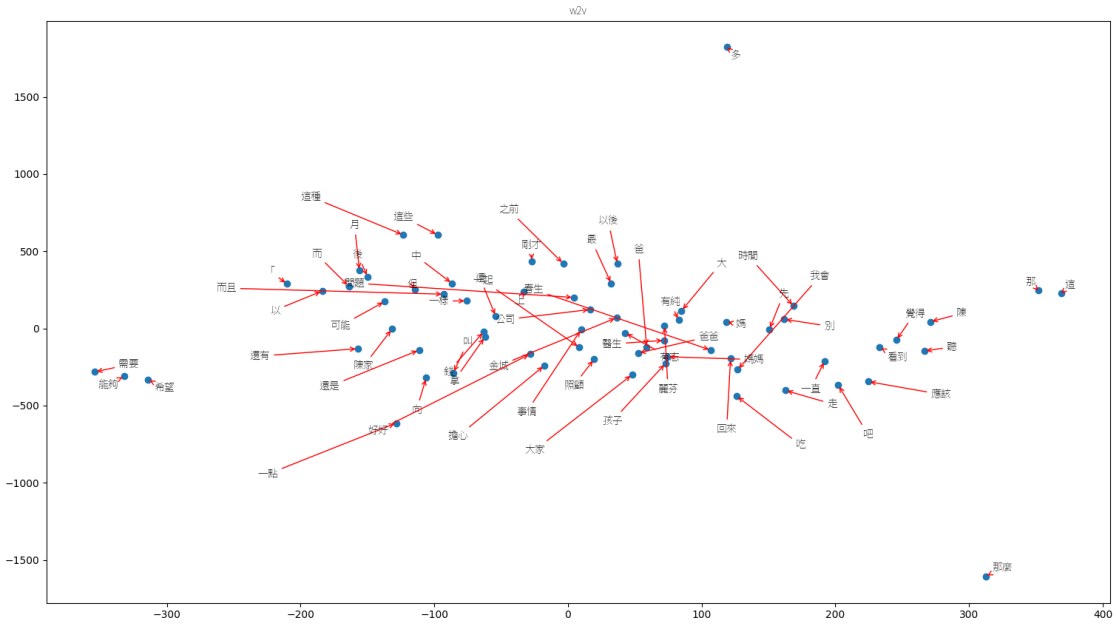
21.6%
10.9%
7.2%
6.1%

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。

Genism 套件
Parameter :
Sg = 1 skip-gram 算法
Min_count = 2 取用至少出現兩次的字，因為只有出現一次可能是噪音

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。

Model make sense，因為人工判斷相似的在圖上有聚集再一起，例如:

需要、能夠、希望

C. Image clustering

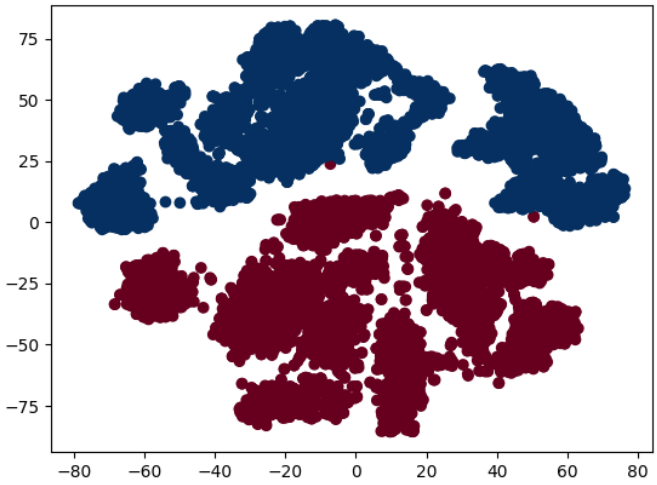
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

第一種: 利用 NN 做 auto encoder kaggle 分數為 0.99

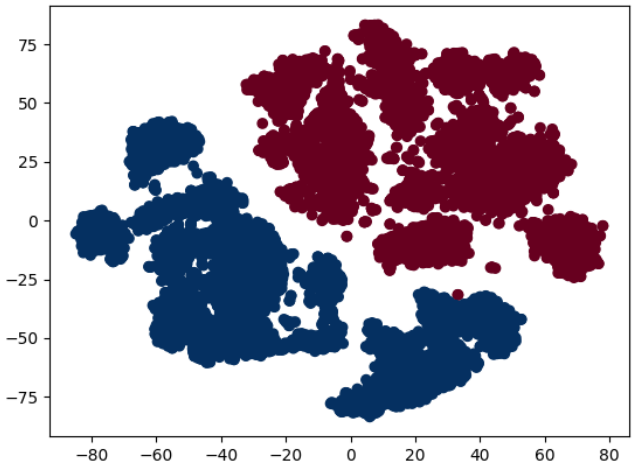
第二種: 利用 CNN 做 auto encoder kaggle 分數為 0.03

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。



因為本身 model performance 還不錯，所以 clustering 的結果也很好