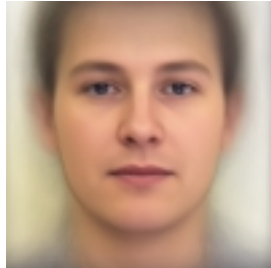


## A. PCA of colored faces

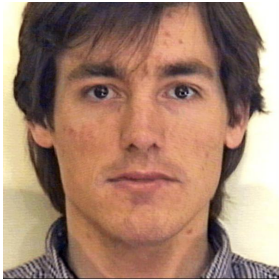

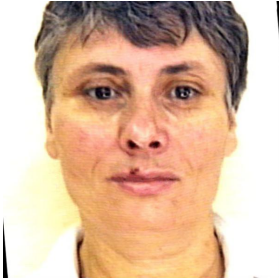
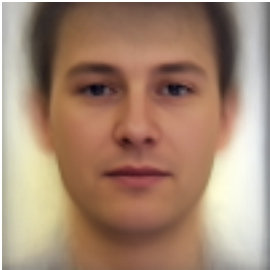
1. (.5%) 請畫出所有臉的平均。


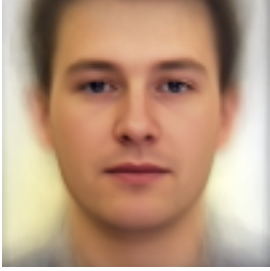

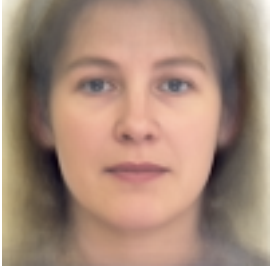


2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



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。

| num | origin  | reconstruction   |
|-----|---|--|
| 1   |  |  |
| 50  |  |  |

|     |   |  |
|-----|---|--|
| 101 |  |  |
| 201 |  |  |

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。

|   |      |
|---|------|
|   |      |
| 1 | 4.2% |
| 2 | 3.0% |
| 3 | 2.4% |
| 4 | 2.2% |

## B. 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。

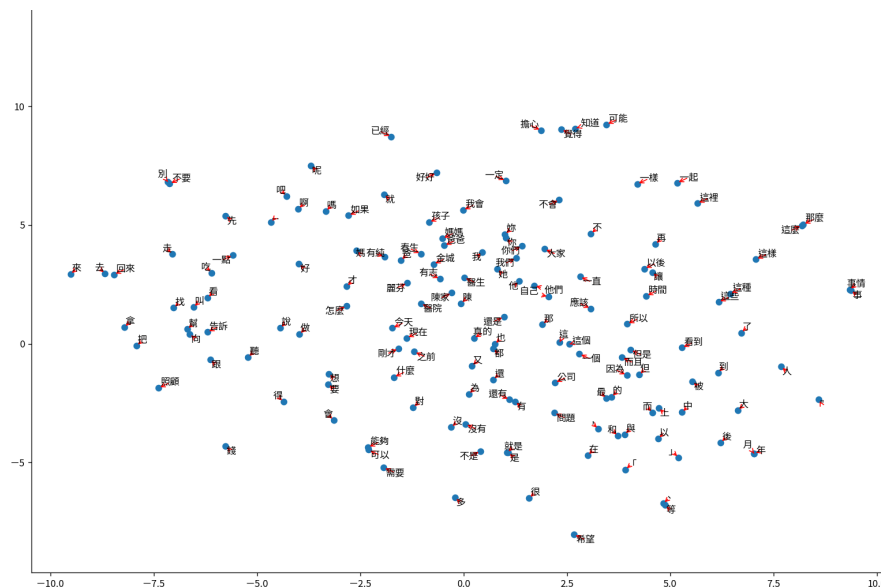
套件：gensim

參數只有調 size 和 min\_count

size 是指產生的向量維度

min\_count 最少出現次數，超過才會放進 model

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。

屬性類似的字彙會聚集在一起，像是中間有一群偏向稱呼，爸爸、媽媽、孩子；而旁邊一群，你、你們、他們...此類代名詞密集分佈；右下角有一群是連結詞，「但是」、「而且」等詞密集分佈。

## C. 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)

TSNE (兩個維度) + KMeans (分兩群) (並沒有 normalization)

F1 score:

|   |                |                |
|---|----------------|----------------|
| <b>tsne_1.csv</b>                         | <b>0.17433</b> | <b>0.17446</b> |
| 10 days ago by <a href="#">kelly wang</a> |                |                |
| <a href="#">add submission details</a>    |                |                |

Autoencoder + KMeans (分兩類) (有經過 normalization)

encoder 疊了 4 層，分別是 512 256 128 32

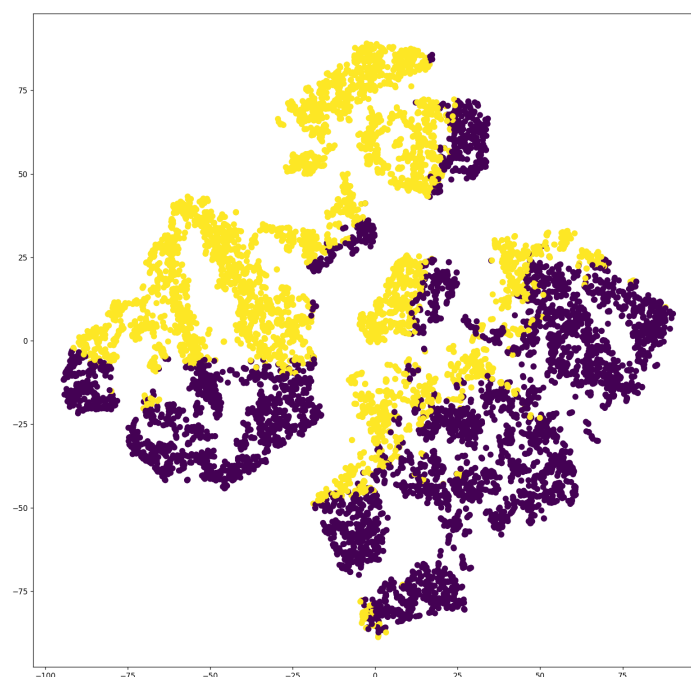
decoder 也疊了 4 層，依序分別是 64,128,256,512

batch 為 128，epoch 100

F1 score:

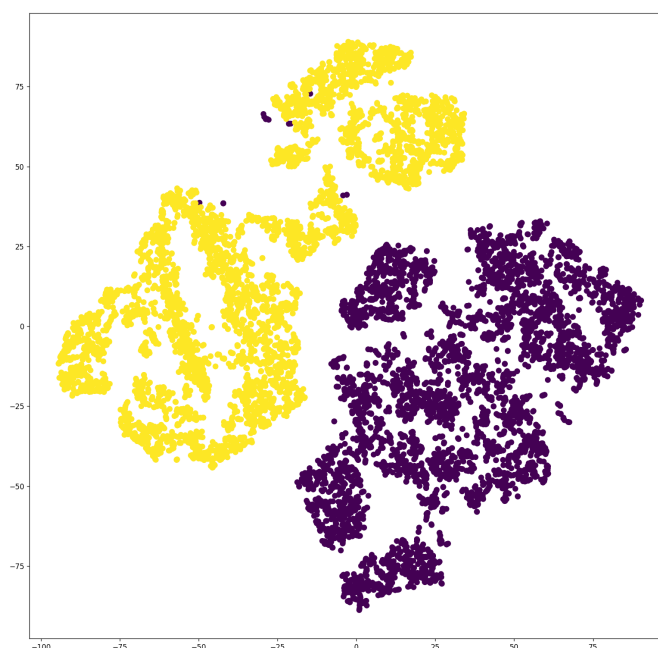
|   |                |                |
|---|----------------|----------------|
| <b>auto_5.csv</b>                       | <b>0.97991</b> | <b>0.98108</b> |
| a day ago by <a href="#">kelly wang</a> |                |                |
| <a href="#">add submission details</a>  |                |                |

D. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化



label 的分佈。

1. (.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。



比較上下兩圖（兩圖的座標相同），在左邊分佈，較右邊的分佈出現誤判，而右邊的分佈，較左邊的部分，則出現誤判。