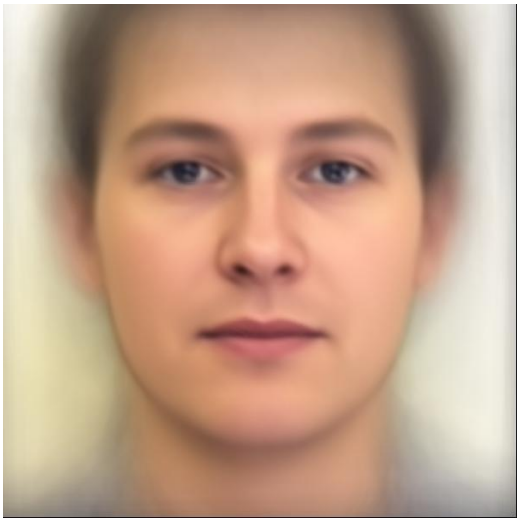
學號：r06921011 系級： 電機碩一 姓名：劉逸霖

A. PCA of colored faces

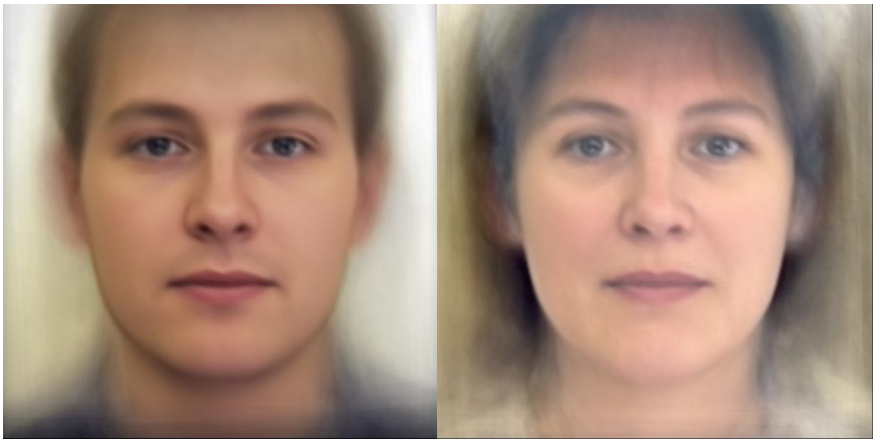
A.1. (.5%) 請畫出所有臉的平均。



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。

Reconstruction_face1 Reconstruction_face2



Reconstruction_face3 Reconstruction_face4

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。

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

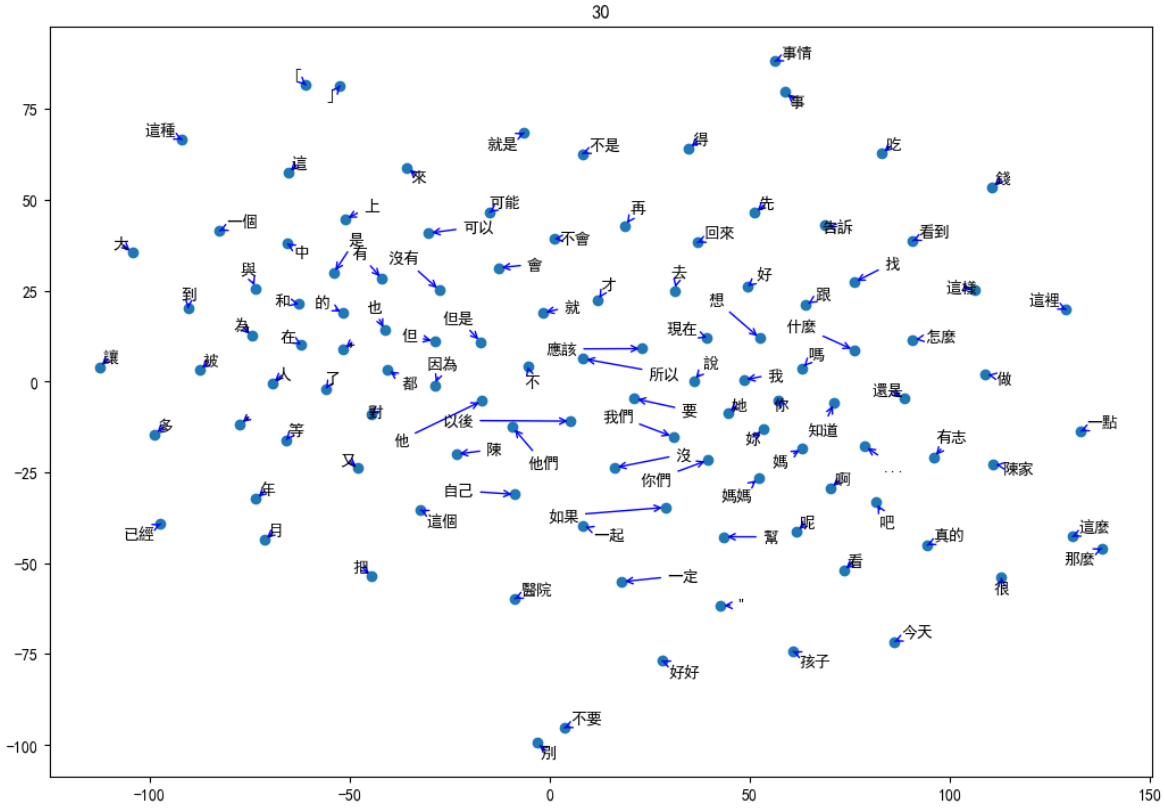
B. Visualization of Chinese word embedding

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。

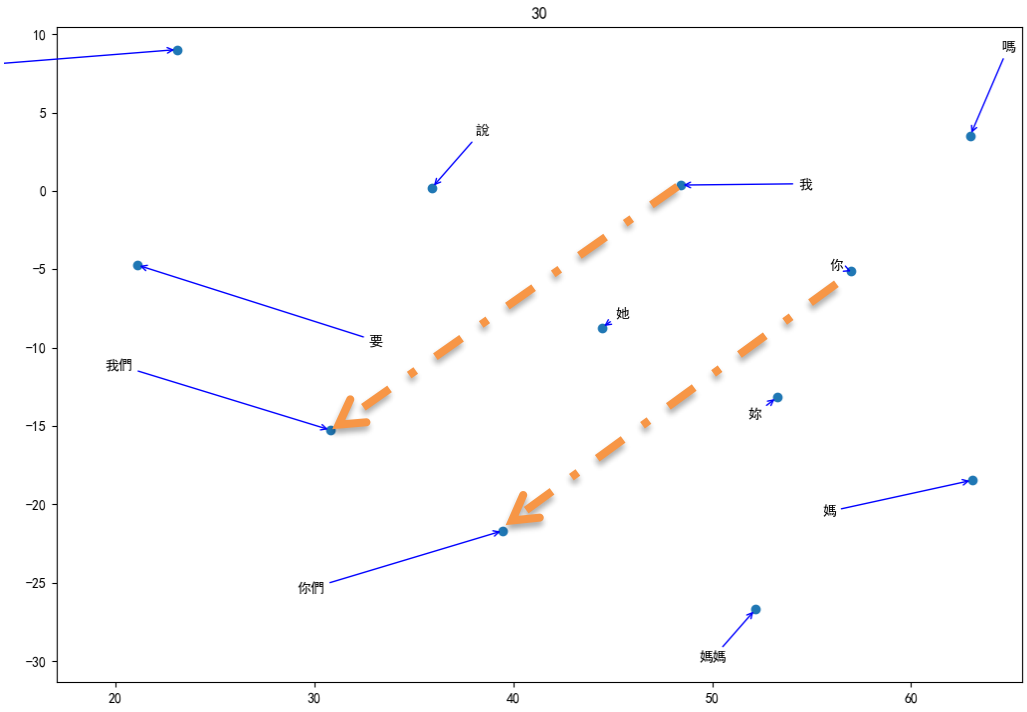
Genism, size=256, window=7, alpha=0.015, sg=1, hs=1, min_count=5, workers=7, iter=20

以上參數是參考 final project 時我們在 train word vector 時使用的參數，在這個參數下會有最好的效果。

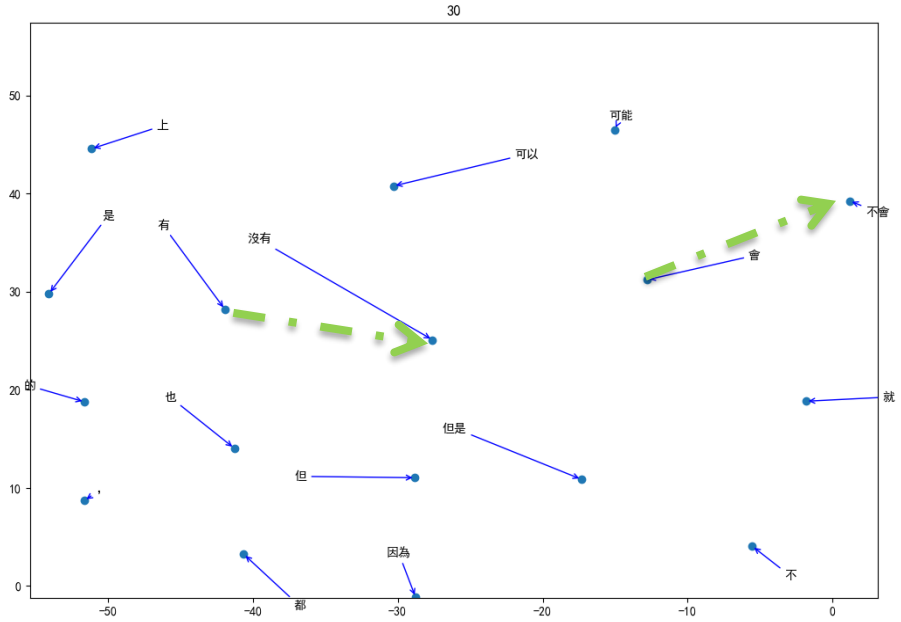
B.2. (.5%) 請在 Report 上放上你 visualization 的結果。



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



可以發現<u>我</u>跟<u>我們</u>的向量與<u>你</u>跟<u>你們</u>的向量類似(橘色)



有跟沒有的向量與會跟不會的向量類似(綠色)

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)

第一種參考助教手把手的 code, 使用 200 個 epochs,得到 kaggle 結果為 0.74576

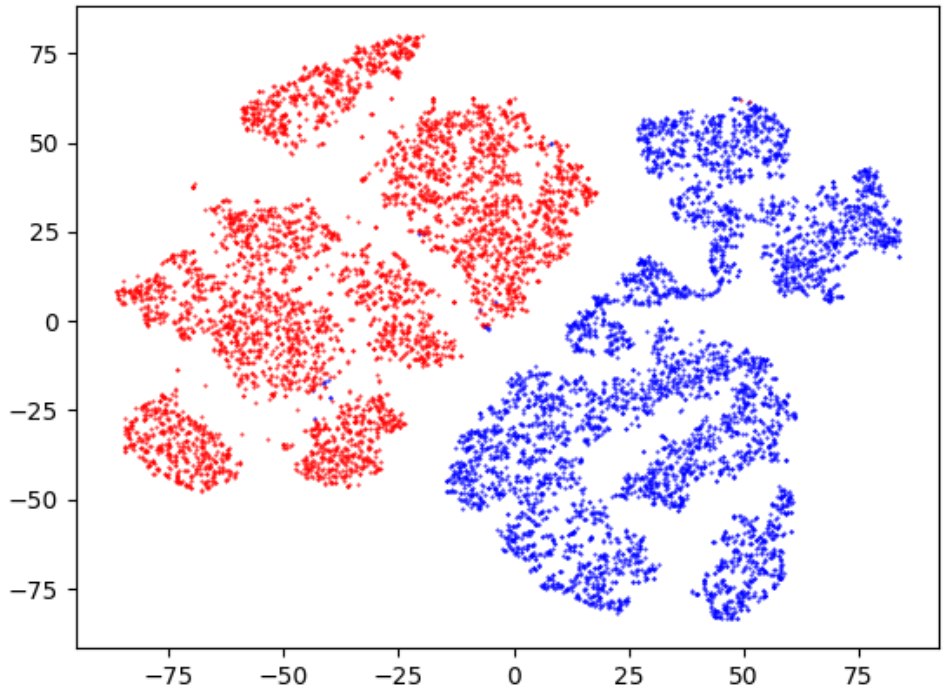
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 256)	200960
dense_2 (Dense)	(None, 64)	16448
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 64)	2112
dense_5 (Dense)	(None, 256)	16640
dense_6 (Dense)	(None, 784)	201488

Total params: 439,728
Trainable params: 439,728
Non-trainable params: 0

第二種透過觀察，發現手寫數字通常會在圖片的中間，只要找出周邊是黑色的圖片即可，得到 kaggle 結果為 0.86838，比 autoencoder 效果還要好

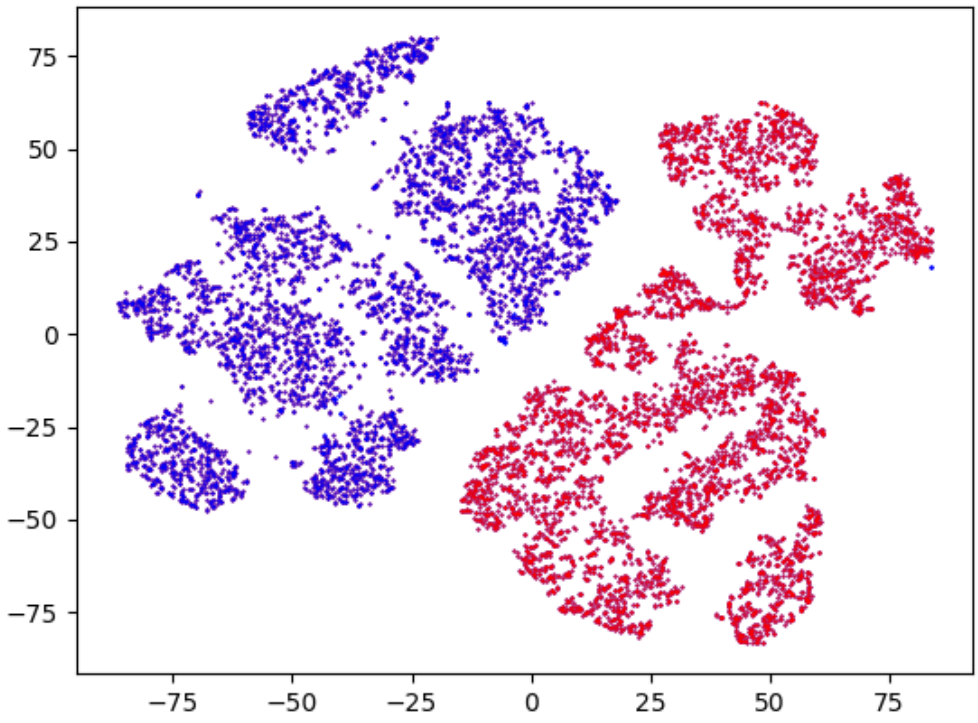
```
label = []
for i in range(data.shape[0]):
    num = 0
    nz = np.count_nonzero(data[i,:])
    image = data[i,:].reshape(28,28)
    col = np.count_nonzero(image[0,:])
    row = np.count_nonzero(image[:,0])
    row1 = np.count_nonzero(image[:,27])

    if col < 2 and row < 2 and row1 < 2 and nz < 300 : label.append(0)
    else : label.append(1)
    if (100*i/data.shape[0]) % 10 == 0 : print('i=',100*i/data.shape[0])
```

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



跟 ground truth 滿接近的，只有少數幾個點分類錯誤

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。


75
50
25
0
−25
−50
−75
−75
−50
−25
0
25
50
75