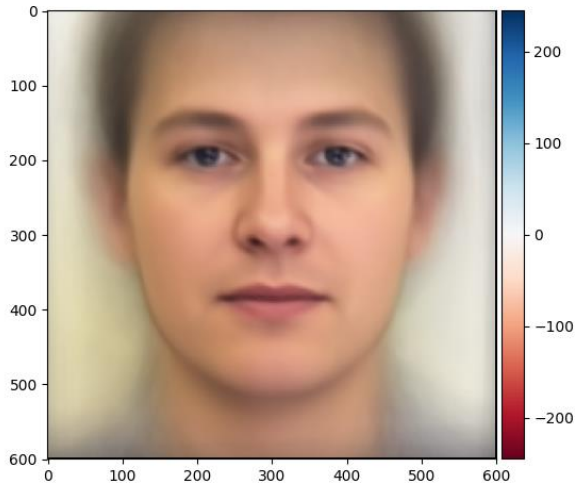
學號：B04902089 系級： 資工三 姓名：林政豪

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。

依序為前四大 eigenface(左至右、上至下)

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。

依序為 index 0 25 129 333 重建之後的結果(左至右、上至下)

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。

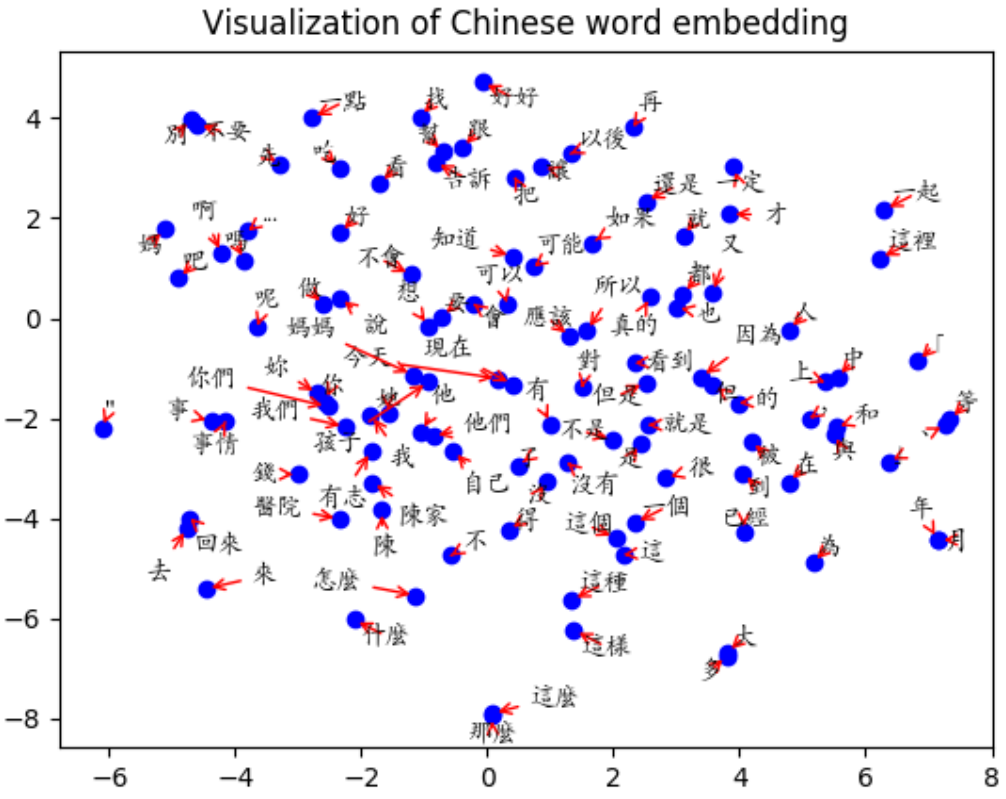
依序為 4.1%, 3.0%, 2.4%, 2.2%

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。

我使用了 gensim 套件，參數如下:
以下參數為在 final project 的 TV 中大致使用的參數，直接沿用於此小題。
vector_size = 96，每個詞的維度
window_size = 3，看到一句話時，每次會一次看三個詞，分析詞之間的關係
negative_size = 3，每個詞反義詞的數量
min_count = 4000，出現次數不足 4000 次的詞會被去掉
iteration = 35，會執行 35 個 cycle
此外還有使用 jieba github 上的 stopwords.txt，將不重要的詞篩掉

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。

可以看到他把相似的詞放在一起，如圖中中央偏左的部分有:(你、妳、我們、孩子、我、他、她、他們....)等等稱謂的詞都相距不遠，還有左上角有一些:(阿、嗎、吧)等語助詞，反義詞的部分也稍微看的出來，在圖中可以看出(是、不是)、(有、沒有)、(會、不會)等等大致上都是左上、右下的關係，但有的是正向的部分在左上、有的是負向的部分在左上，這部分看起來做得不太好，應該是因為我使用針對 final project 所 tune 好的參數對其訓練導致。

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)

我使用了 autoencoder 降到 32 維，再使用 Kmeans 分成兩群。

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 392)	307720
dense_2 (Dense)	(None, 128)	50304
dense_3 (Dense)	(None, 64)	8256
dense_4 (Dense)	(None, 32)	2080
dense_5 (Dense)	(None, 64)	2112
dense_6 (Dense)	(None, 128)	8320
dense_7 (Dense)	(None, 392)	50568
dense_8 (Dense)	(None, 784)	308112

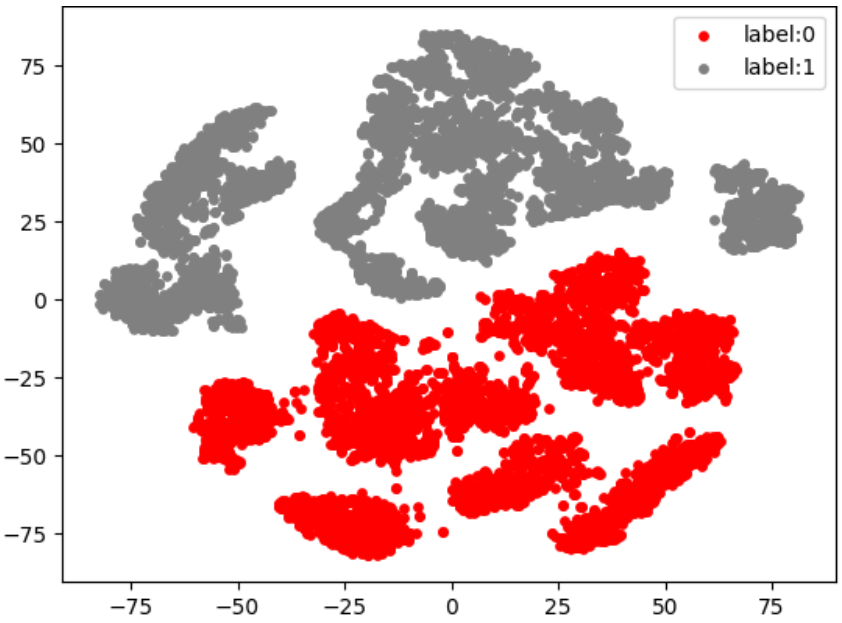
Total params: 737,472
Trainable params: 737,472
Non-trainable params: 0

以及先使用 PCA 降到 64 維，TSNE 再降到 3 維，再使用 Kmeans 分成兩群。

方法	Autoencoder	PCA+TSNE
分群個數	70000,70000	72026,67974
Kaggle(Public +Private/2)	1	0.004302

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

我使用上小題的 autoencoder 降維之後使用 Kmeans 分類，且使用 TSNE 將降維後的結果再降成兩維，並畫圖。



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。

我的 autoencoder 在 Kaggle 上的表現非常好(Public 和 Private 分數皆為 1)，在這次的分類上看起來也完全正確，我檢查結果時發現，他把前五千個分在一群，後五千個分在一群。