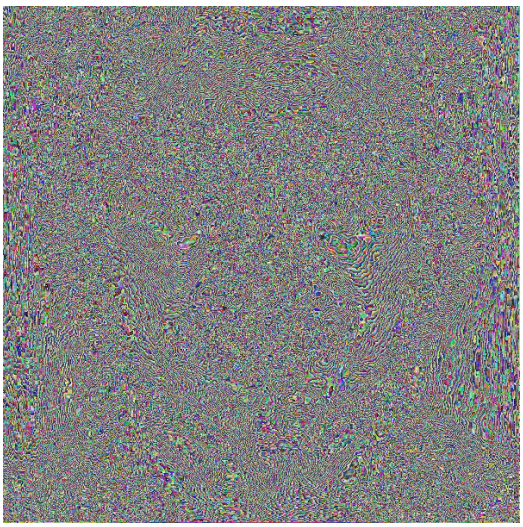
學號：R06521608　系級： 土木碩一　姓名：陳德元

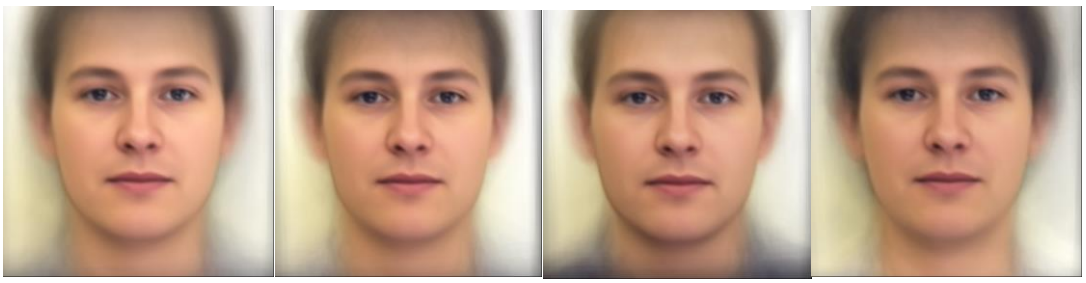
A. PCA of colored faces

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。



26　81　156　281

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。

Ans: [7.5% 3.1% 2.8% 2.2%]

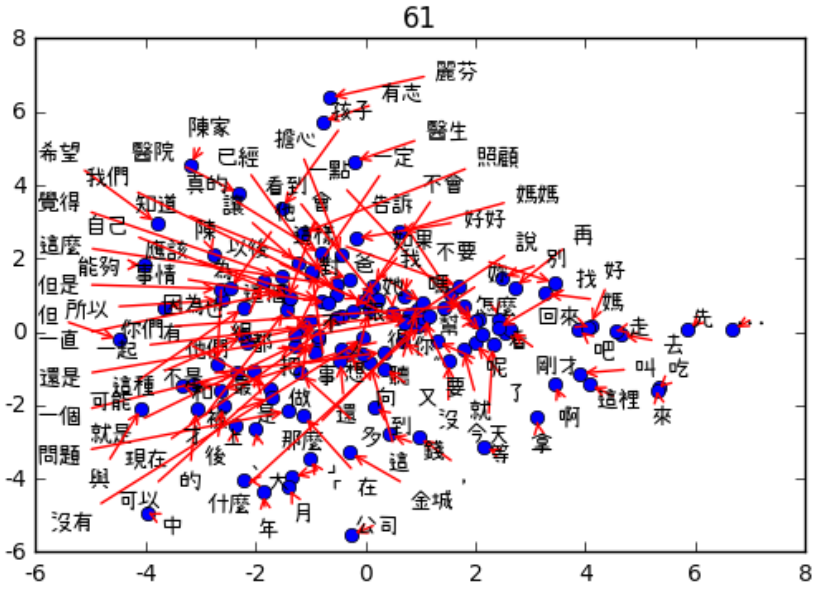
B. Visualization of Chinese word embedding

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。

```
from gensim.models import word2vec
model = word2vec.Word2Vec(sentences, size = 100, min_count = 3500)
```

使用的 word2vec 是 gensim 的套件，而我調整的參數 size 是代表轉換出來的 vector 是 100 維的向量，min_count 則是指要有出現 3500 次以上才會被轉換。

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。

從視覺化後出來的結果，我發現很多人名(麗芬、有志)都被分類在圖表的上方，而許多希望、覺得這類的祈使句的詞彙，則在圖表中的左邊被分為較類似的一類，還有許多語末助詞(啊、吧)，則在圖表的右下方。

C. Image clustering

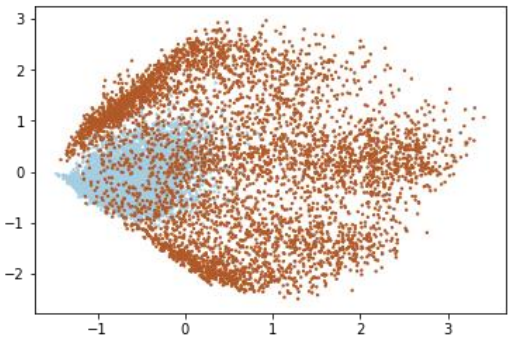
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 784)	0
dense_1 (Dense)	(None, 128)	100480
dense_2 (Dense)	(None, 64)	8256
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 64)	2112
dense_5 (Dense)	(None, 128)	8320
dense_6 (Dense)	(None, 784)	101136

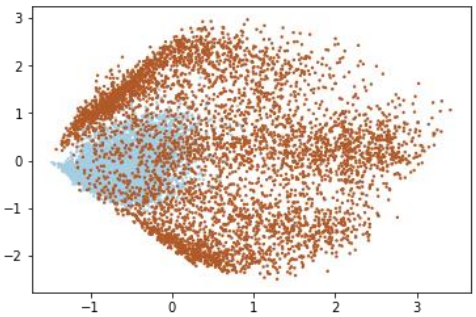
Total params: 222,384
Trainable params: 222,384
Non-trainable params: 0

一種是 DNN 的 autoencoder 的方式，其用 kmeans 分類的方式結果為{0: 68246, 1: 71754}，另一種是用 sklearn 的 PCA 直接降維至 280，而用 kmeans 分類的結果可到{0: 70000, 1: 70000}。

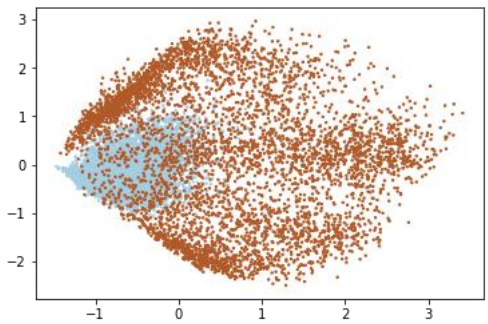
C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



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。



(圖一)為自己預測



(圖二)為直接標註前後 2dataset

兩者長得一模一樣，代表預測力極高。