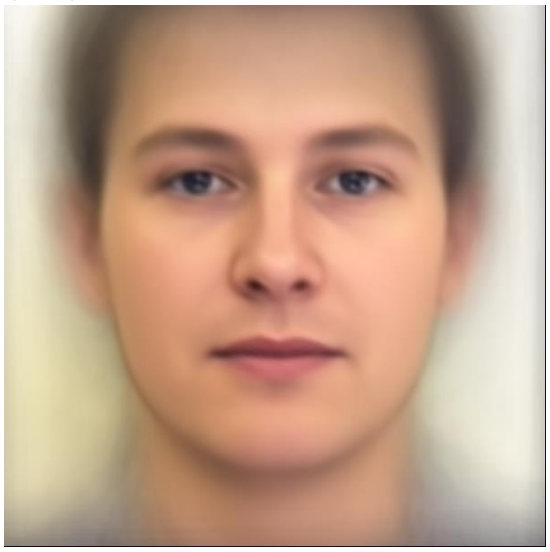
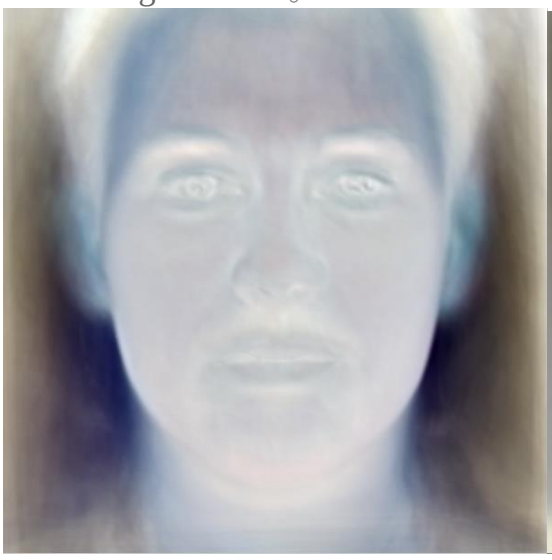
學號：B04902021 系級：資工三 姓名：陳弘梵

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。

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。

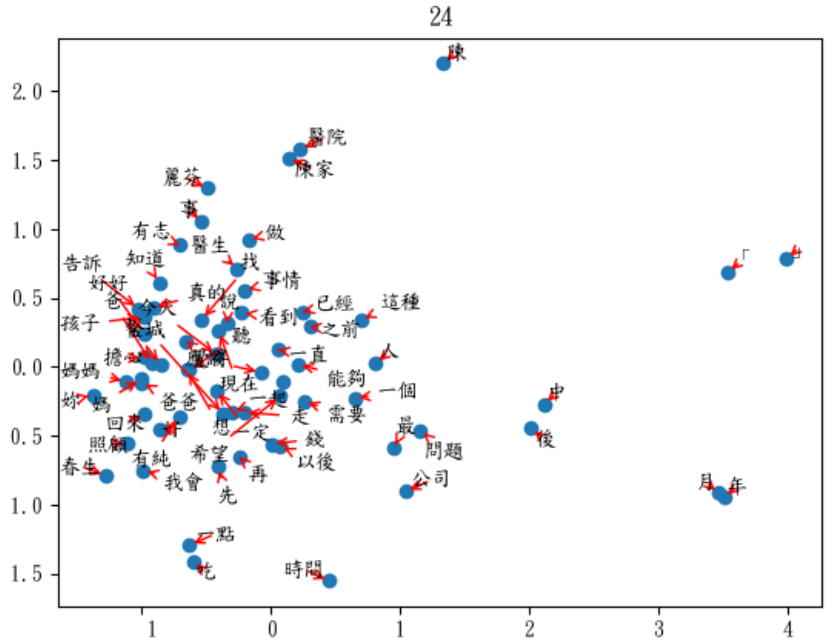
B. Visualization of Chinese word embedding

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。*

我使用的是 gensim.models 的 word2vec，參數有 size(代表每個 vector 的維度大小)=64、min_count(代表這個要被計算的單字至少出現次數)=3000。

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

用 PCA 降維之後，呈現結果如下圖:



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

我覺得 vector 訓練得很好，家庭身分都在一起，「月年」、標點符號、介係詞也都被投影在附近的點上。而左邊中間投影出來的字密度較高，比較難觀察出字之間的直接關係。這次在使用 stopwords 之後的結果，比原先的圖片來的清楚、乾淨非常多。

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)

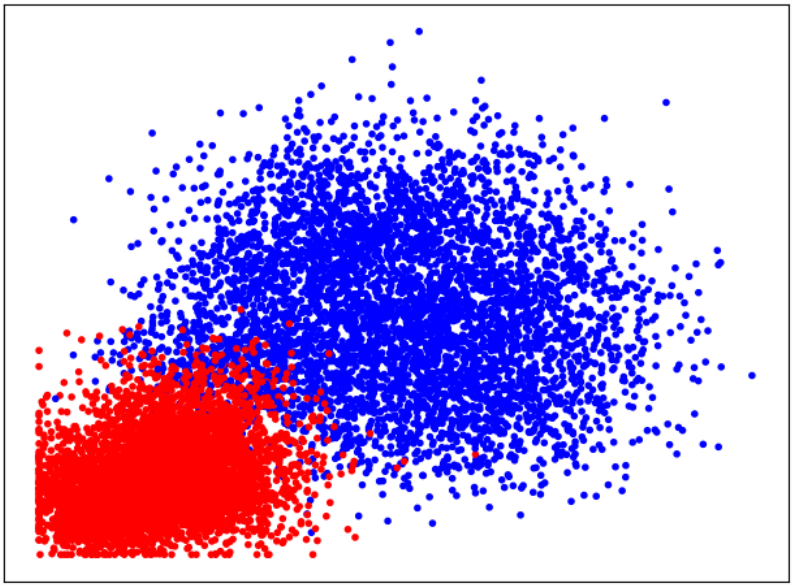
我做了 auto encoding(1)以及 pca(2)的方法，cluster 部分都是採用 sklearn 的 KMeans，其表現成績以及程式如下:

```
input_img = Input(shape=(784,))
encoded = Dense(392, activation='relu')(input_img)
encoded = Dense(196, activation='relu')(encoded)
encoded = Dense(98, activation='relu')(encoded)
encoded = Dense(32, activation='relu')(encoded)
```

```
pca = PCA(n_components=2).fit_transform(x_train)
```

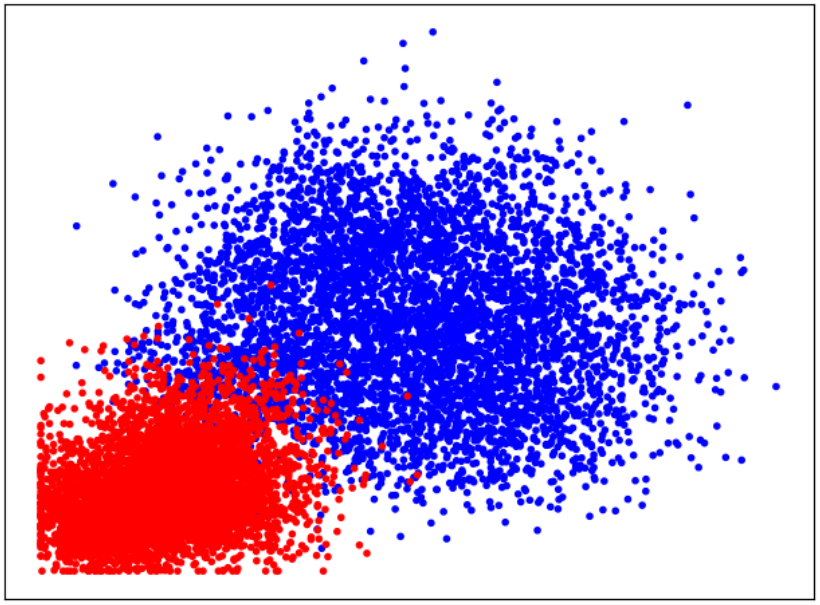
	Private Score	Public Score
PCA	0.03012	0.02984
Auto encode	0.88315	0.87932

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



(紅色=Dataset 1，藍色=Dataset 2)

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。



(紅色=Dataset 1，藍色=Dataset 2)

由於我的 model 在 kaggle 的表現是 public: 0.88315 (在第一小題的表格中) ，因此最後區分出來的兩張圖片，差異並沒有很多，唯一的差別大約都是散佈在教遙遠的紅、藍色點點，若是訓練的抽 feature 效果更好，在這邊投影出來的兩張圖片，會更吻合甚至沒有差異。