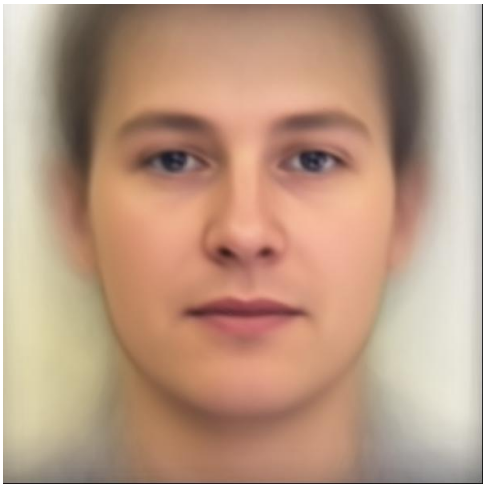
學號：R05922156 系級： 資工碩二 姓名：黃子瑋

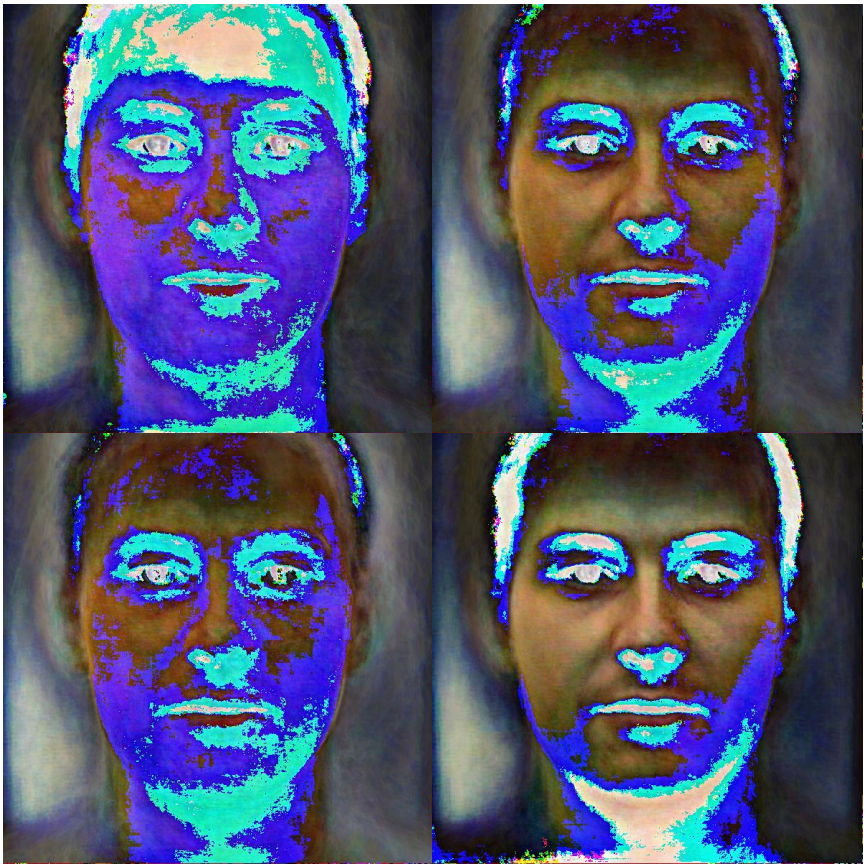
A. PCA of colored faces

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。



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. 7.7%
2. 1.7%
3. 1.4%
4. 1.1%

```
0.07679847139 19
0.01732939131 07
0.01411074172 81
0.01060168133 24
```

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。

我是用 gensim 這個套件

他的 word2vec 有蠻多參數可以調的

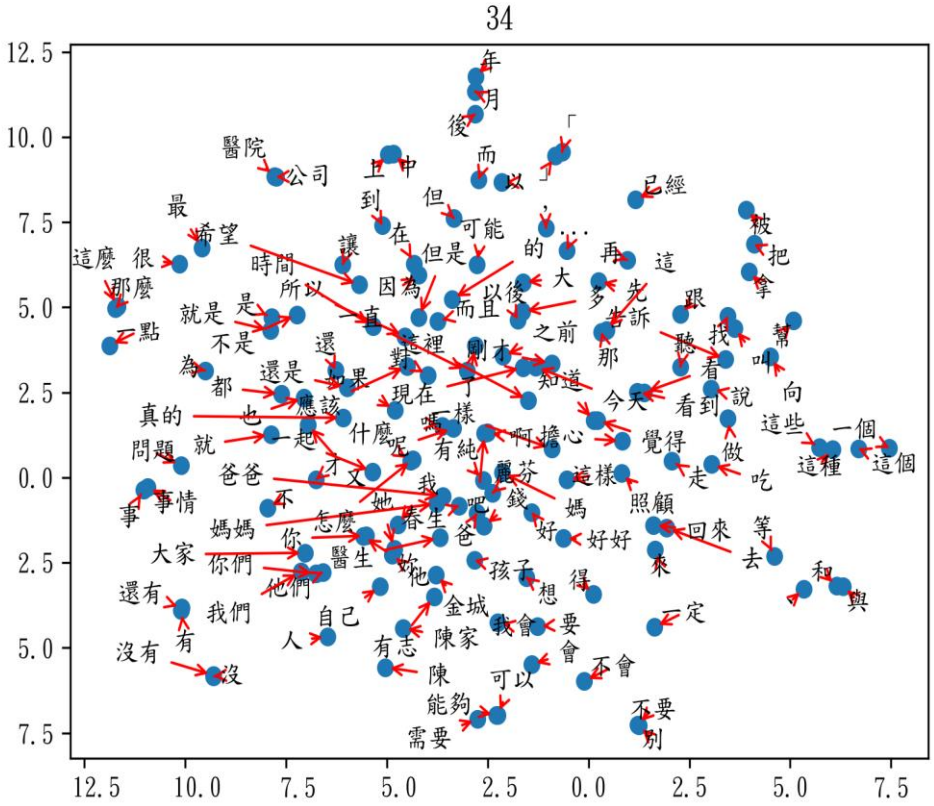
我有嘗試過調整 “ size ” “ window ” “ min_count ”這幾個參數

size 是指最後 train 出來每個 word 對應到的 vector 的維度

window 是指在 train 的過程中 “當下詞彙” 與 “預測詞彙”在一個句子中的最大距離

min_count 則是指 word 出現的次數小於”min_count”次會被捨棄

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。

從上圖可以發現

比較相似的詞在位置上會比較相近

例如 “不要” “別” 就很靠近

而在比較中間的詞感覺上是意義比較普通或是較沒意義的

例如 “一樣” ”甚麼”意思就比較 general

而”有純” ”春生” 是人名 感覺沒什麼意義

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)

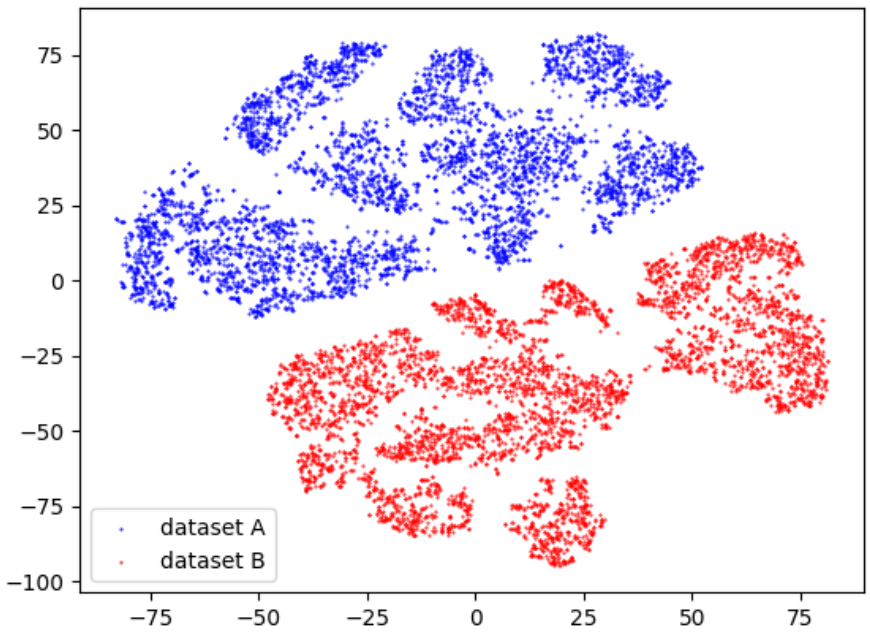
	Autoencoder	pca
Kaggle 上的 score	0.95	0.43

我在 Autoencoder 上有嘗試過 encode 部份加上 CNN 或是 encode 成較多維的 feature

結果後來大概得出 0.95 的成績

然後 pca 的方法跑出蠻爛的分數所以最後還是選 Autoencoder

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。

我發現預測出來的結果跟真實結果

在二維圖片上來說超級接近

Encoder predict 出來的結果很接近 100%(大概 95%左右吧)