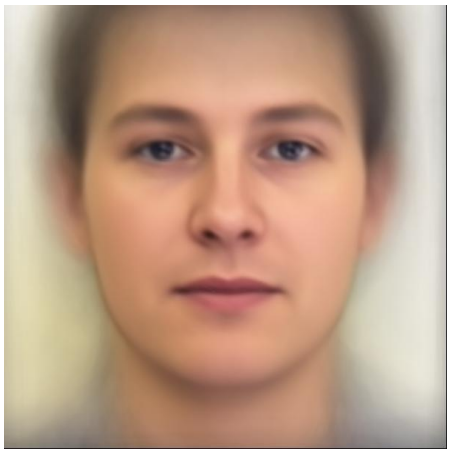
A. PCA of colored faces

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。

Picture22 Picture200 Picture331 Picture354

由於僅從四個 eigenfaces 做出圖片，因此輪廓都差不多，比較有區別性的則是在頭髮上。

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。

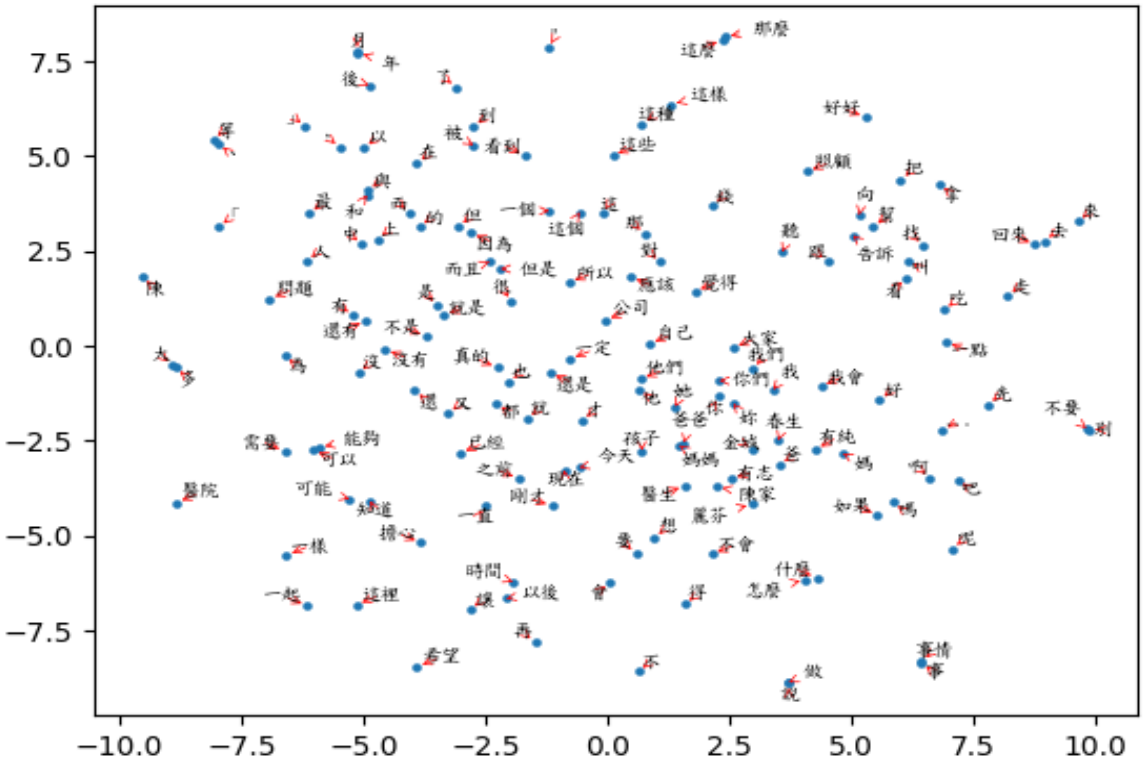
1:4.1% 2:2.9% 3:2.4% 4:2.2%

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。**

使用 gensim，將所有的資料塞入，並將
iteration 調為 10，意義是 word2vector 中 train10 次；
size 調為 50，做出來是 50 維，再用 tsne 降維；
mincount 設 50，至少要出現 50 次的字才會被放上來 train

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。**

中下附近有許多人名，應該是某些 data 有密集重複的出現這些人名的故事，因此放在一起；而右上角幾乎都是動詞，左上角幾乎都是副詞，代表詞性相近的經過 word2vecor 之後，會擺在一起。

C. Image clustering

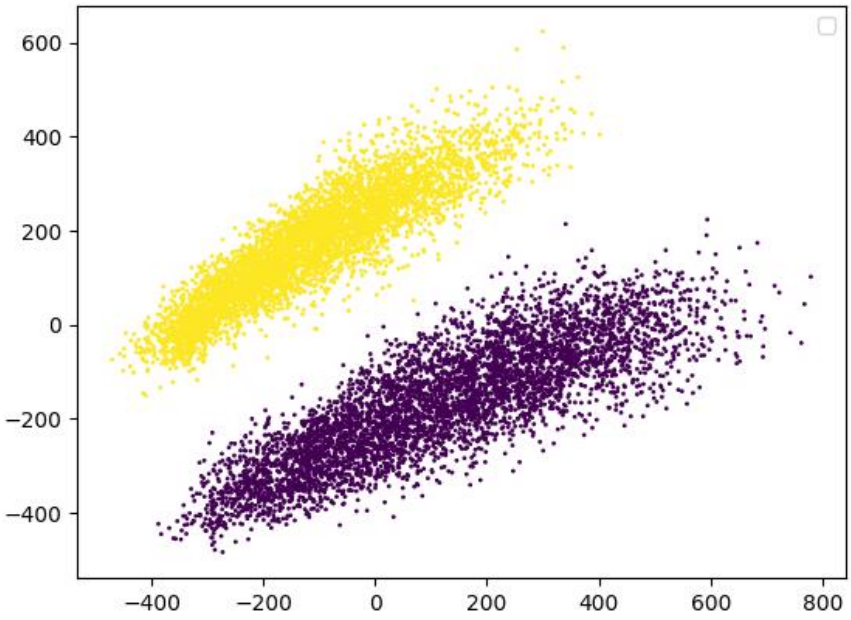
C.1. **(.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)**

方法一：利用 pca 以及 tsne 降維,再利用 kmeans clustering 分類
方法二：利用 DNN autoencoder 降維，再利用 kmeans clustering 分類
結果：

	Public score	Private score
方法一	0.15742	0.15728
方法二	0.96987	0.96844

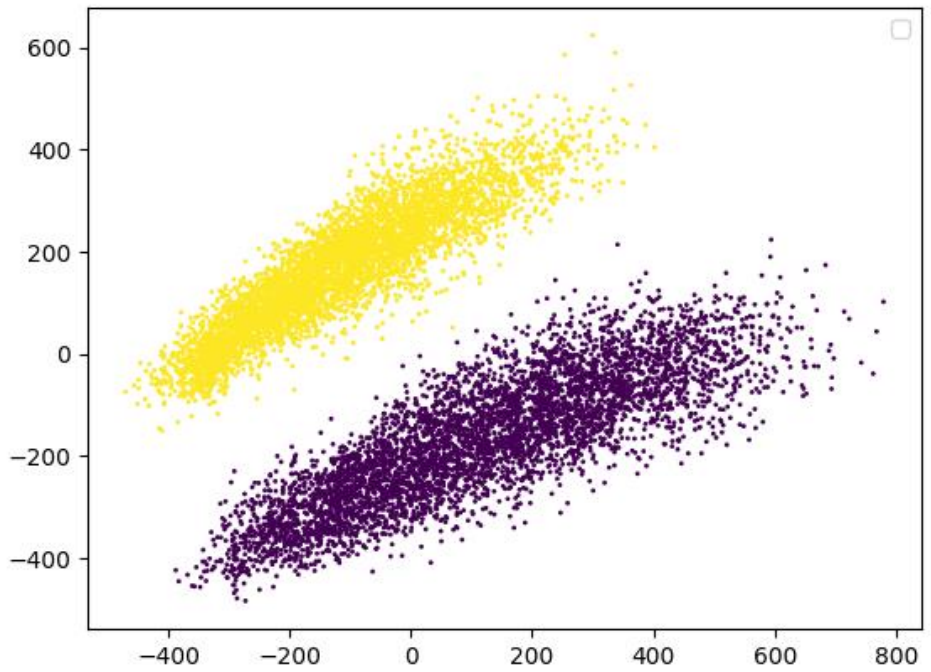
原因推斷：利用 pca 跟 tsne 降維時，降維太快且訓練量不足，導致降至兩維時的代表性不足，因此無法明確的分類。

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



黃色的部分是 Dataset A, 紫色的部分是 Dataset B，經過 PCA 降維成兩維
由於原本的 model train 的不錯，因此可發現明顯分為兩邊，且集中。

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。**



這是原本的 label 分布，可發現做出來結果幾乎跟自己 predict 出來的 label 一致

代表這次的 model train 的還不錯。