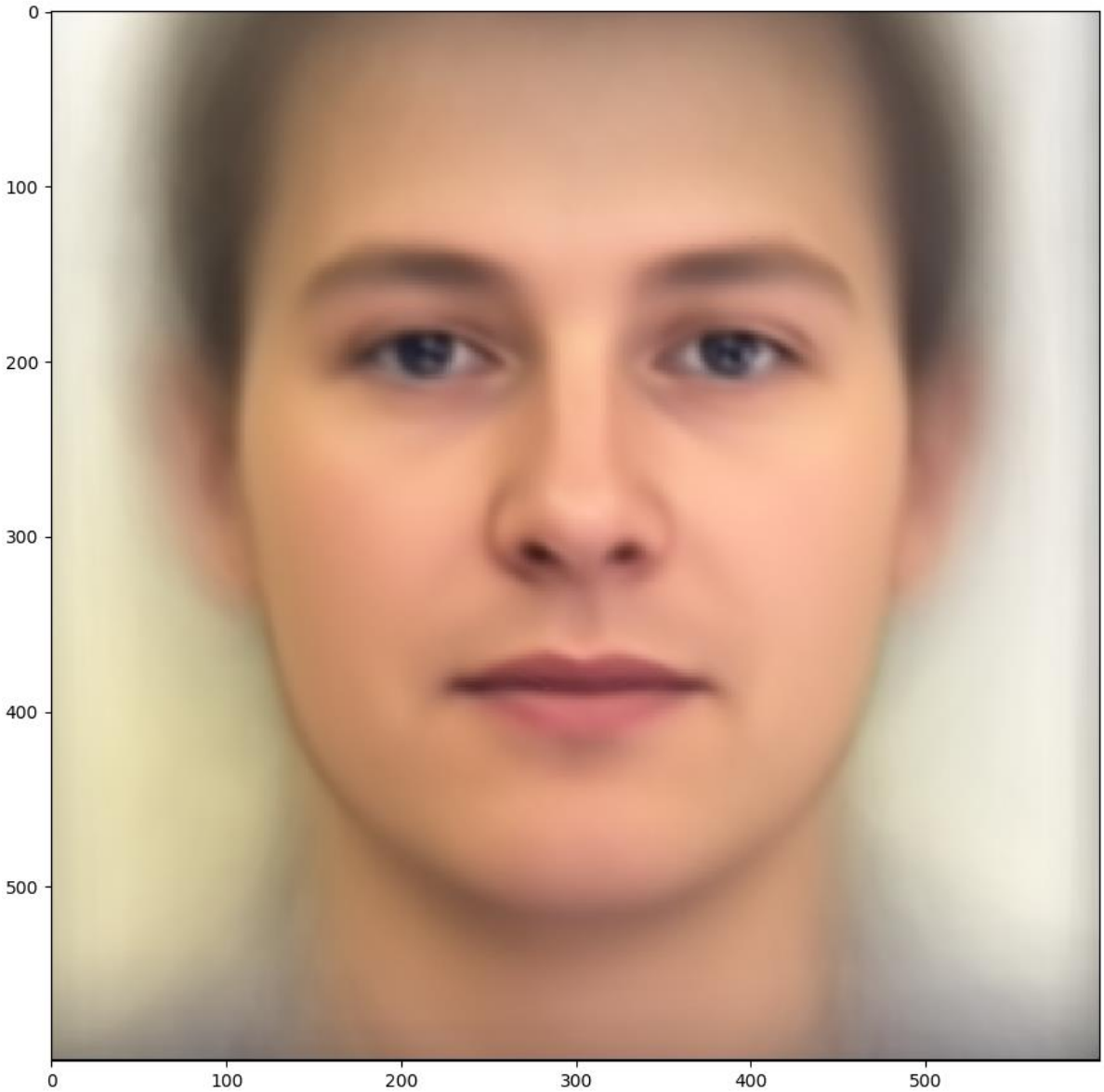
學號：R05921120 系級： 電機碩二 姓名：黃浩恩

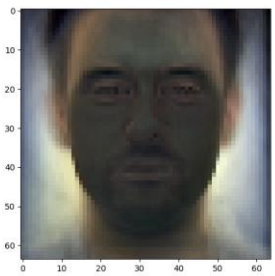
合作人：歐靖 R06921089、賴棹沅 D05921011

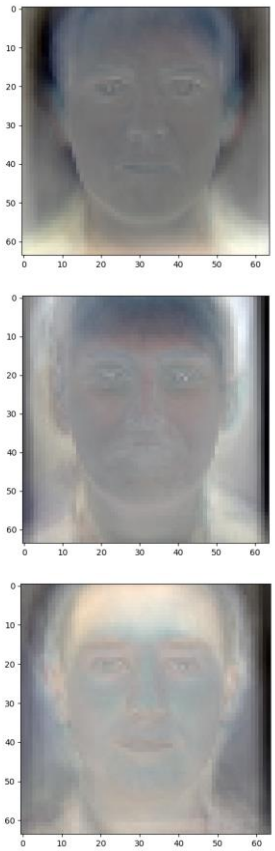
A. PCA of colored faces

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。

選取照片：108、140、177、216

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。

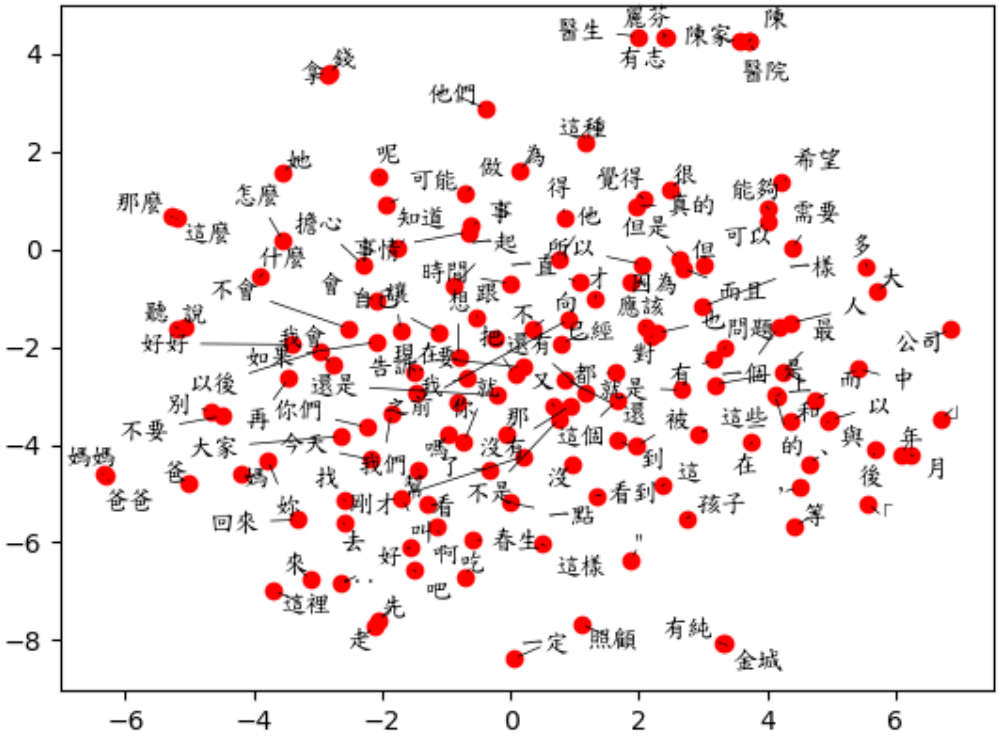
B. Visualization of Chinese word embedding

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。

```
model = word2vec.Word2Vec(sentences, size=250, min_count=2000)
```

於此使用 gensim，調整參數有 size 與 min_count，設定為 250 與 3000，分別表示 vector_size 之值、篩選出現次數大於 2000 的字。

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。

發現當字詞意思相近之字詞(不要、別...)在圖形上很靠近，表達能力狀態(能夠、希望、需要、可以...)也可以看出集中，也可以看出反義字的距離都較遠，特別的地方是像爸爸媽媽這種詞彙，雖然中

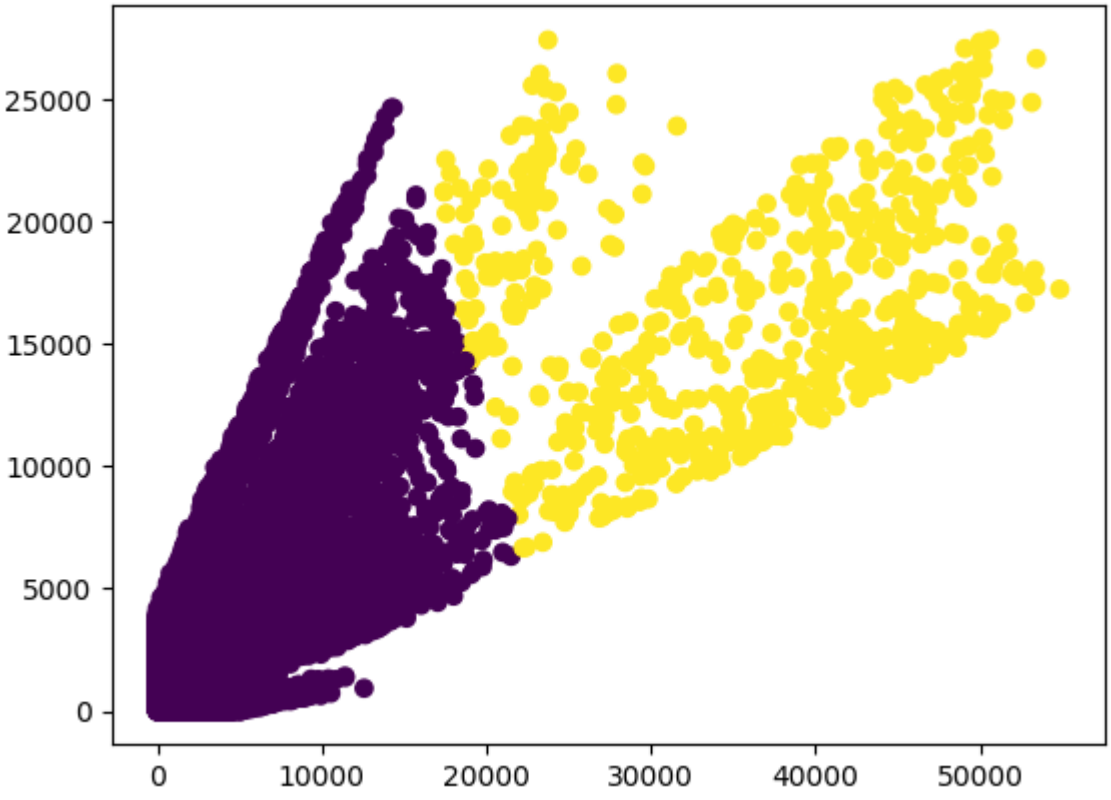
文意思有些差別，但在電腦中的學習可以看出意義是相近的，也蠻合理的。

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)

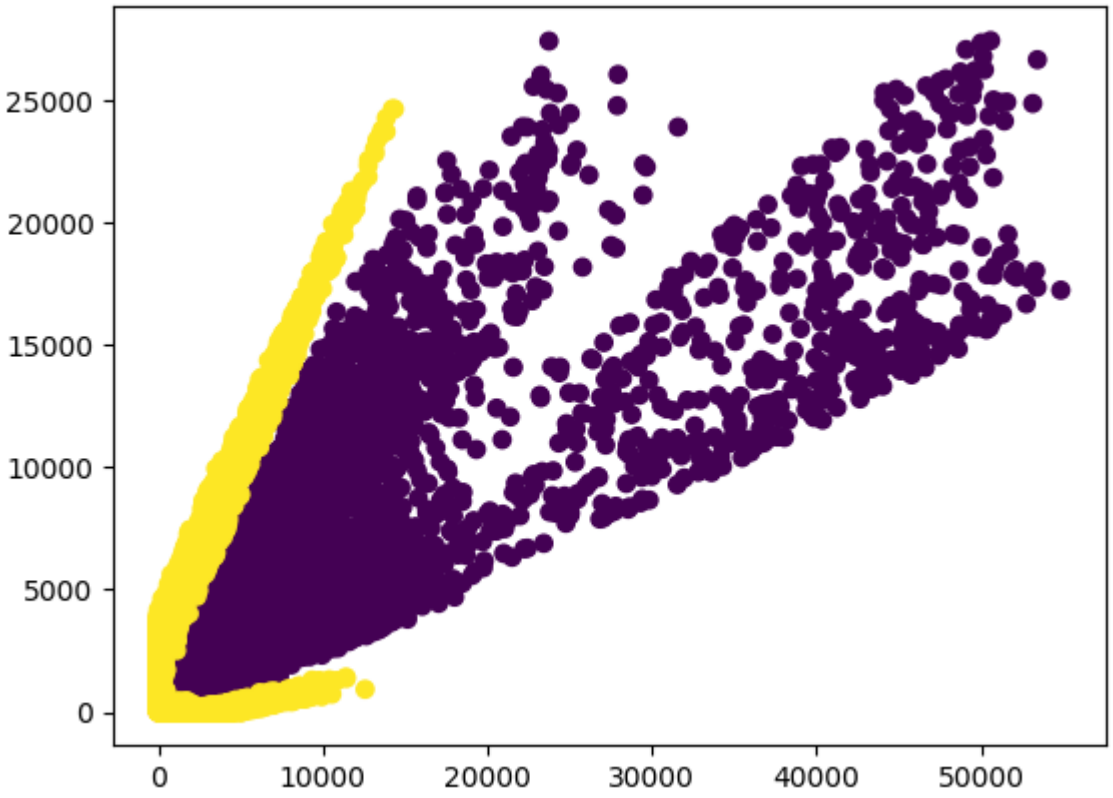
	Public	Private
使用 Tsne 降維至 2 維再透過 k-means 分 2 類	0.92481	0.92184
使用 auto-encoder	0.92369	0.92066

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。

下圖是前 5000 個 images 跟後 5000 個做分類之結果(沒 predict)



可見在分類上有落差，在原本答案上的分類可能不夠強，導致在原始分類中標記不同 label 上在圖形中看不出分類區隔。