學號：R06922116 系級： 資工碩一 姓名：賴柏恩

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。

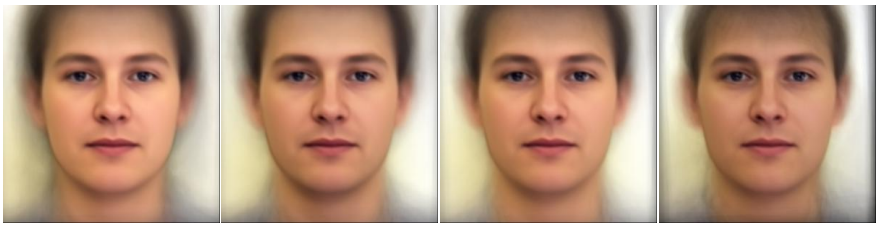
A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

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。

所有臉的平均：

eigenfaces(2*2 順序為左至右)

face reconstruction(face1,10,100,400 順序為左至右)



eigenfaces ratio

7.5% 3.1% 2.8% 2.2%

B. Visualization of Chinese word embedding

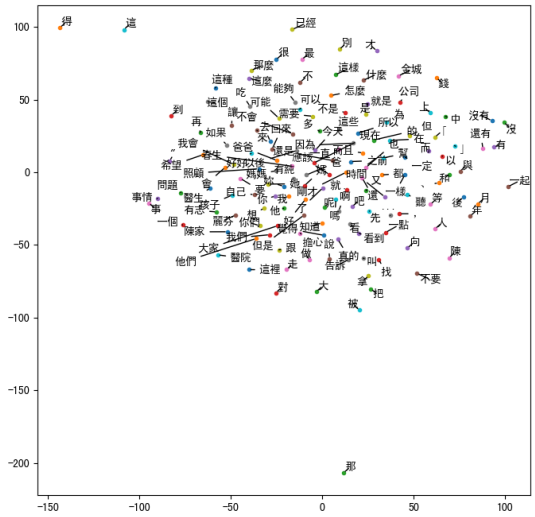
B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

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。

我使用的是 genism 套件裡面的 w2v，因為 final 也是用這個去做，所以直接選擇這個，有調整的參數為 size, 以及 min_count, size 是 feature vector 的維度，而 min_count 是指如果字出現次數低於 min_count 就不會考慮進去。

Visualization result:



從圖中可以看出一些小東西，例如他們我們大家你們這四個詞的位置就相當靠近，還有很多語助詞例如呢啊吧啊嗎也都很靠近，還有很多地方可看出一些詞是用在很類似的地方，他們之間的位置也會較為靠近。

C. Image clustering

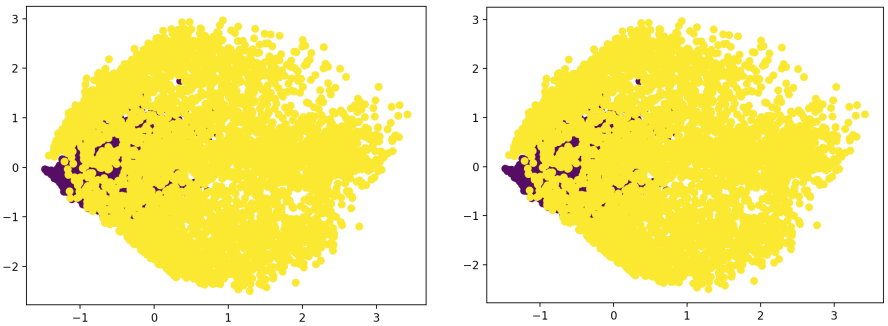
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

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。

我用了三種降維的方法，分別是 autoencoder, t-sne, PCA, 而這三種方法中其實剛開始試出來結果最好的是 autoencoder, 只有他有到 0.5 分, 但其實我覺得很奇怪，因為看到大家的結果都很好，應該是有什麼錯誤，後來我覺得可能是因為我只取了 20 維的 feature，但可能 20 維會造成太多圖片特徵的流失，導致效果很差，所以我試著將 feature dimension 提高，結果確實是變好，但我提高只有使用 PCA，我提高到 250 時 kaggle 分數高達 0.994 因此我就使用這個當作我最後的答案。但我想說不定用 autoencoder 若調到差不多的 dimension 也可以有相同的效果。



左圖為正確 label 畫出的圖，右圖為我預測之 label 畫出的圖
x, y 為 pca 250 維的前兩維
可能是因為 feature 取得蠻好的。所以預測的很準，沒有差異。
檢查了 label 準確率，為 100%。