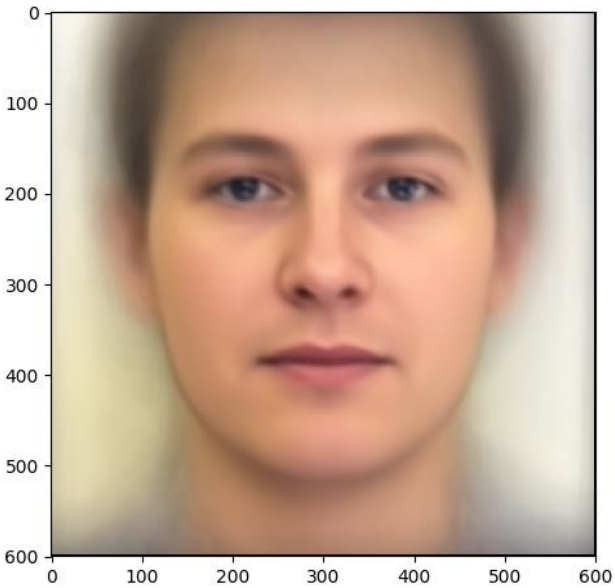
學號：R06922093 系級： 資工碩一 姓名：陳禹齊

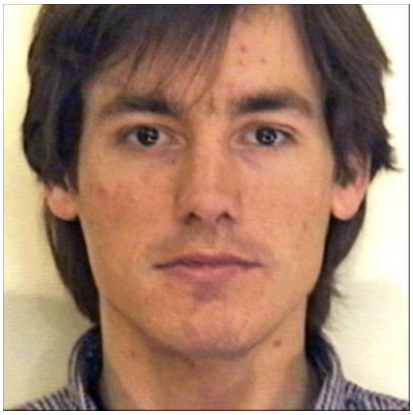
A. PCA of colored faces

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。

原圖: 

重建: 

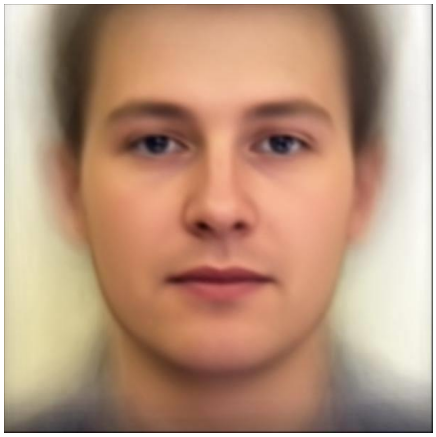
原 圖 : 

重建: 

原圖: 

重建: 

原圖: 

重建: 

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

第一 Eigenface:21.7%

第二 Eigenface:2.7%

第三 Eigenface:2.4%

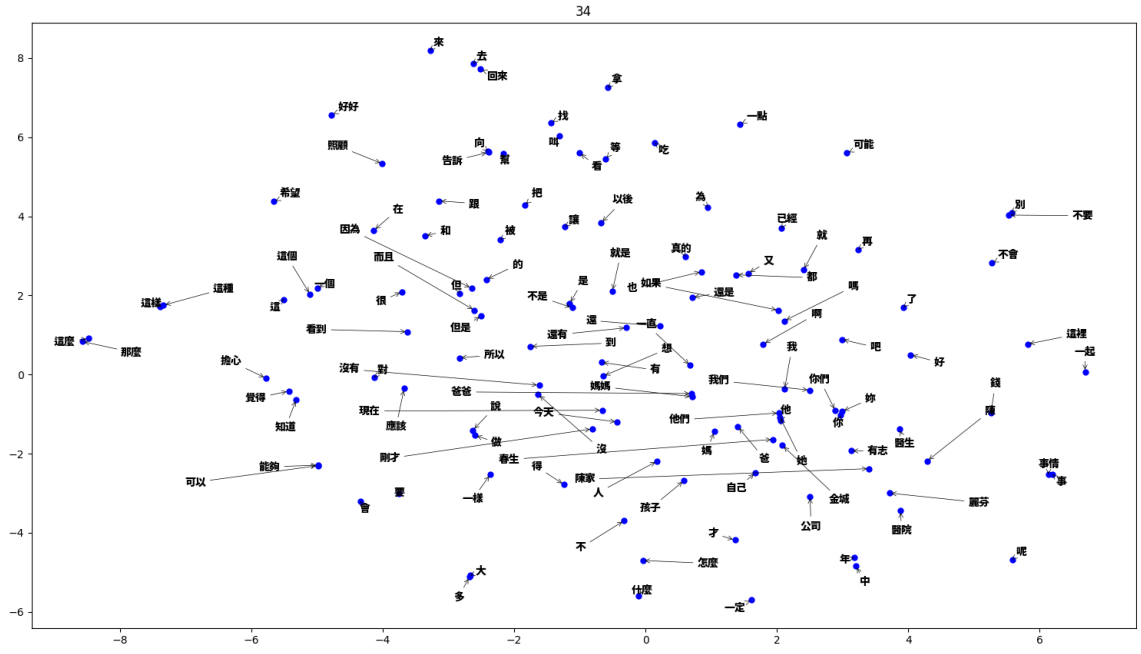
第四 Eigenface:1.8%

B. Visualization of Chinese word embedding

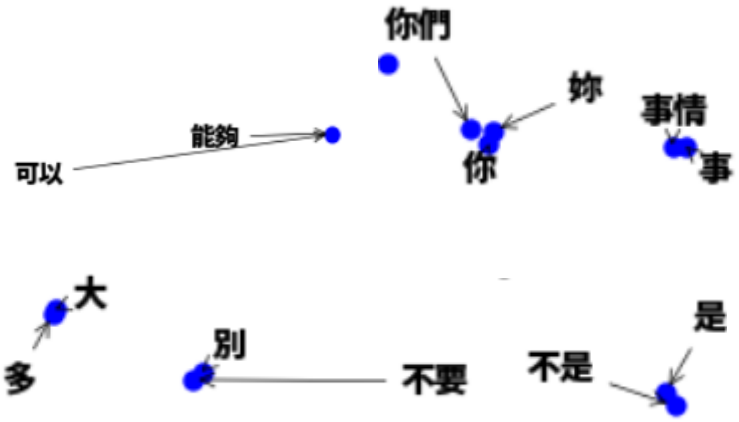
B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

我是使用 gensim 的 Word2Vec,參數的部分我調了 size、min_count 和 iter，這些分別代表了向量的維度、單字至少要出現的次數和迭代次數。

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。



相似的詞性會在圖上的距離會非常的相近,或是直接疊在一起。

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)

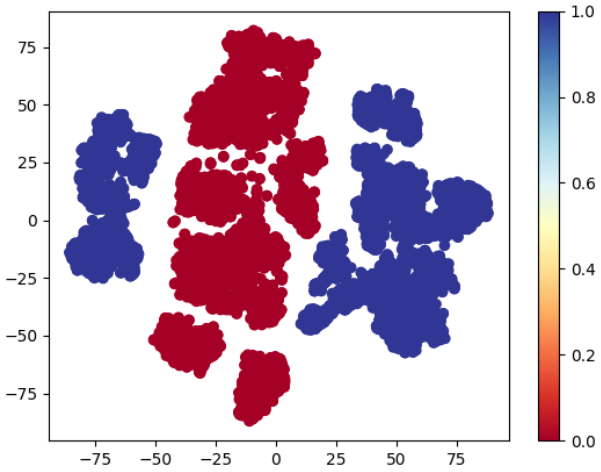
一、 使用 PCA 降維,只使用前 10 大的 eigenvector,並且使用 k-means 來分群

Kaggle public score:0.01949

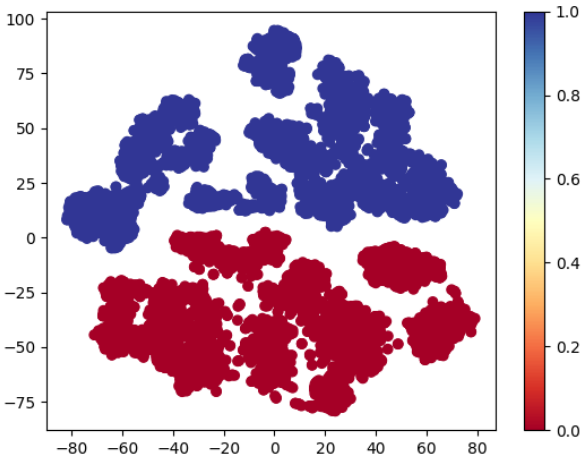
二、 使用 auto-encoder 來做降維,使用 k-means 來做分群

Kaggle public score:1.0

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。



從這兩張圖來看有幾個區塊預測的結果跟實際答案是相同的,像是預測圖的最左邊的藍色長條,右上角的小藍色區塊,都可以在實際答案圖的上面看到不過有些小小的地方看起來不樣,應該就是預測錯誤的地方。