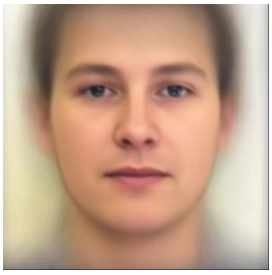
學號：R05522621 系級：機械碩二 姓名：李哲銘

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。

由左而右分別是第 1、2、3、4 個 Eigenfaces。

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。

由左而右分別是 0.jpg、1.jpg、2.jpg、3.jpg。

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。

top 1: 4.1%、top 2: 2.9%、top 3: 2.4%、top 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。

我使用的是 gensim.models 的 word2vec 套件，以下為我有使用到的參數:

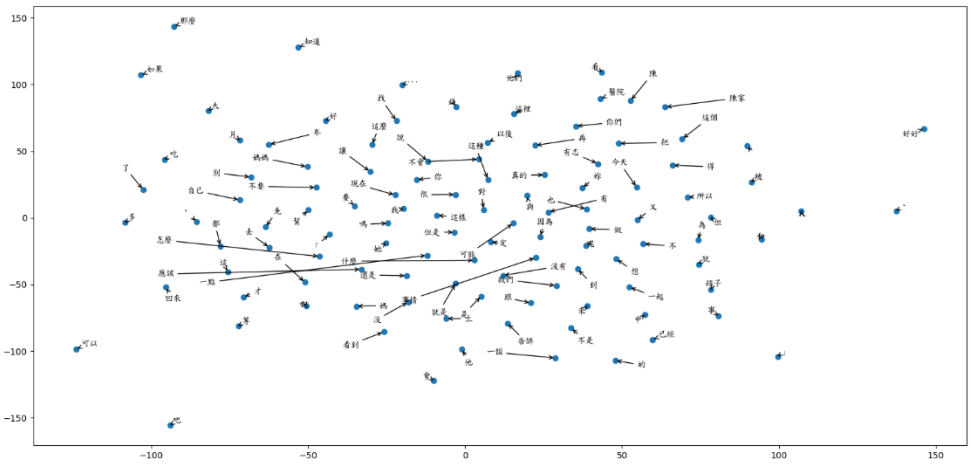
```
size=200, window=5, min_count=4000
```

size = 200 : 代表使用 200 維的向量來描述一個詞。

window =5 : 表示一個字被相鄰 5 個字影響。

min_count=4000 : 表示出現 4000 次以下的詞不加入訓練。

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。

因為已經濾掉出現次數低於 4000 次的中文辭彙，所以剩下的便幾乎是口語跟文章中的常用辭彙，像是你、我、他等代名詞，說、吃、幫等動詞，還有常見的地點名詞等等。由圖中可以很輕易看出，常一起使用的兩個辭彙的距離會較近，像是要跟不要，你跟很之類的。比較有趣的是陳、陳家跟醫院的距離非常接近，我想應該是劇本裡面醫院是陳家的，且甚至就是叫陳醫院或是陳家醫院。

C. Image clustering

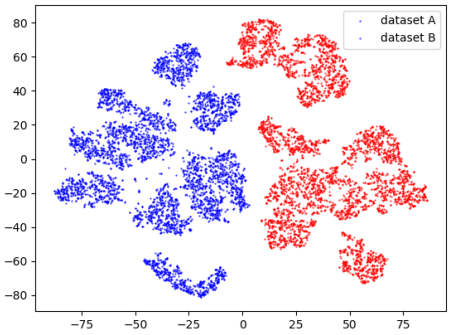
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

這題我比較的是兩種不同的 cluster 方法，一種是 K-means、另一種是直接計算 euclidean distance，兩者皆使用 DNN autoencoder 來降維，以下為兩者的結果:

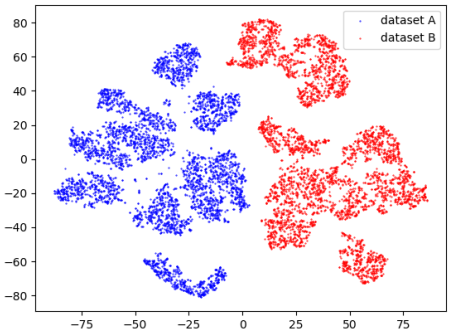
CLUSTER	MEAN F1-SCORE
K-MEANS	1.00000
EUCLIDEAN DISTANCE	0.82573

以這次的降維結果來說，使用 K-means 便能直接在 public 拿到 1.0 的成績，不過如果是降維結果沒那麼好的 model 來說，直接調整降維後兩向量的歐氏距離當作分群標準，其效果會比直接放入 K-means 計算好，且更能夠自由調整。

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。



與前一題比較可以看出兩張圖一模一樣，由此可知此預測結果與實際分群相同。