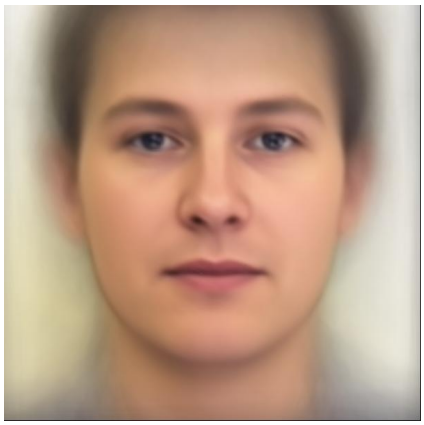
學號：R06922118 系級： 資工碩一 姓名：吳政軒

以下題目 Collaborators 皆為：R06944032 倪溥辰、R06944049 黃敬庭

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。



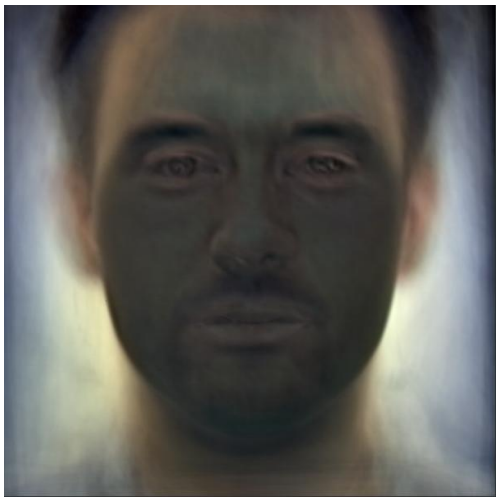
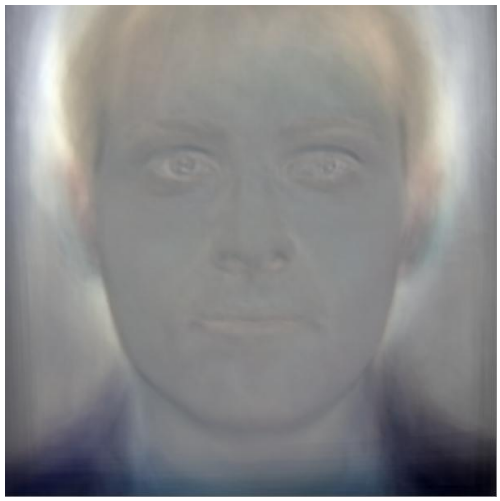
A.2.

A.3.

A.4. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

A.5.

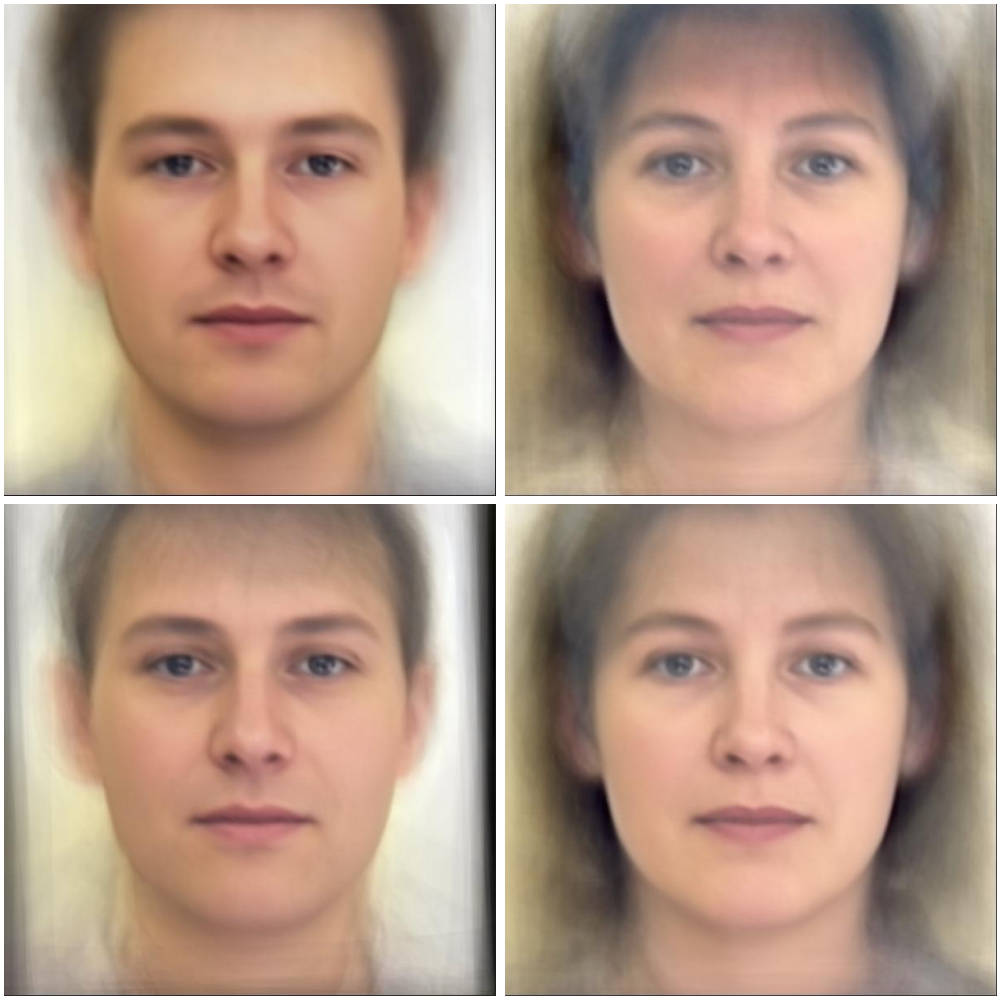
A.6.

A.7. (.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.8. 左上、右上、左下、右下分別為 10.jpg、5.jpg、118.jpg、200.jpg

A.9.



A.10. (.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。

A.11.

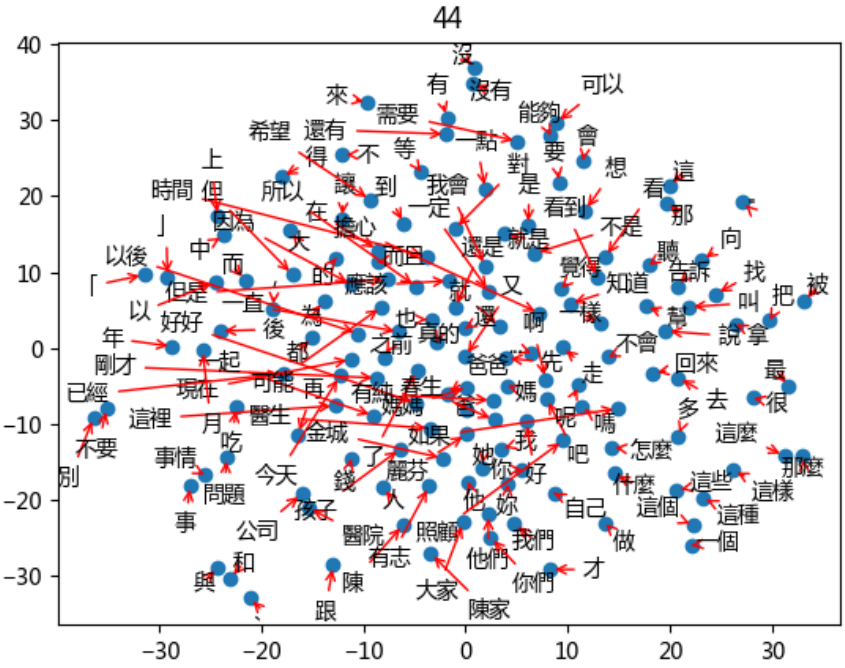
A.12. 前四大比重按照順序為：4.1%, 3.0%, 2.4%, 2.2%

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 gensim 的 word2vec 套件，有特別調整的參數是 min_count，預設是 5，我有調成 1，min_count 的意義是會忽略出現次數小於 min_count 的字，也就是只有大於等於 min_count 的字會在 word2vec 的字典裡。

B.2. (.5%) 請在 Report 上放上你 visualization 的結果。



B.3.

B.4. (.5%) 請討論你從 visualization 的結果觀察到什麼。

B.5.

B.6. 可以看到一些相關的字在空間中的距離的確比較近，例如：「這個、這些、這種」、「你、妳」、「沒、沒有」等等。

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)

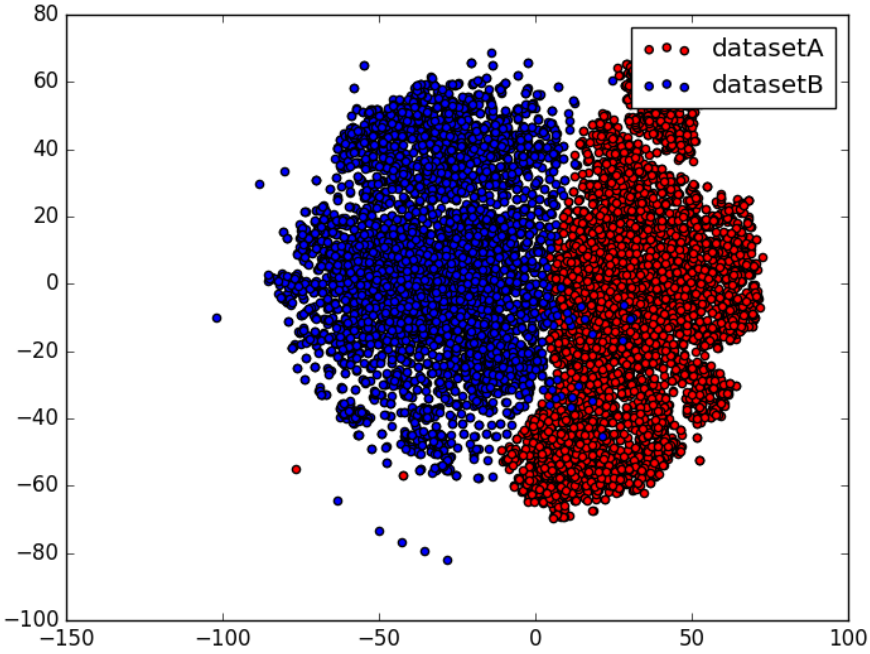
C.2.

C.3. 第一種我使用 PCA 降維成 300 維，再用 KMeans 分成兩群，在 kaggle 上的分數是 1.0（在 deadline 前是 0.99925）

C.4. 第二種我用 DNN train 一個 Autoencoder，拿 encoder 來降維成 32 維，再用 KMeans 分成兩群，在 kaggle 上的分數是 0.76929。不過後來才想到應該要降維成相同維度比較才有意義。

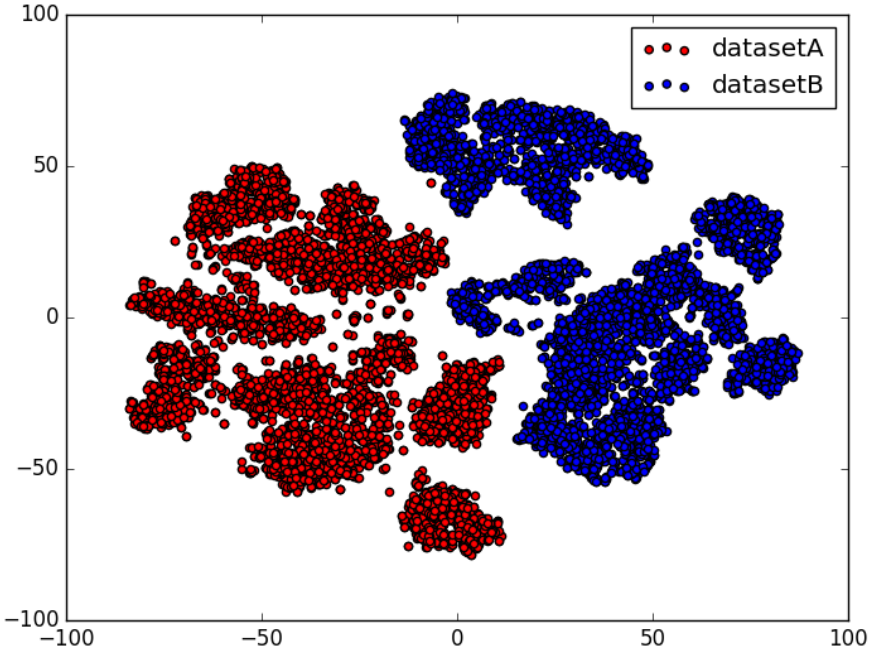
C.5. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

C.6.



C.7.

C.8. (.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。



C.9.

C.10. 我預測的大致上有分開，不過不同群之間沒有隔很開，所以靠近中間的點有時候會預測錯誤，不像第二張圖完全的分開來。