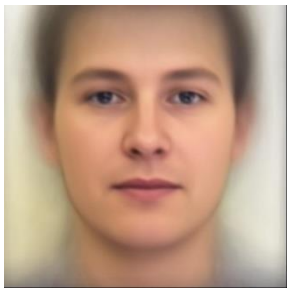
學號：B03901149 系級： 電機四 姓名：陳咸嘉

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。

由左至右是 Eigenface 1~4。

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。

針對 100.jpg、200.jpg、300.jpg、400.jpg 進行重建。

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。

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。

使用套件：genism.model 裡的 word2vec 訓練模型。

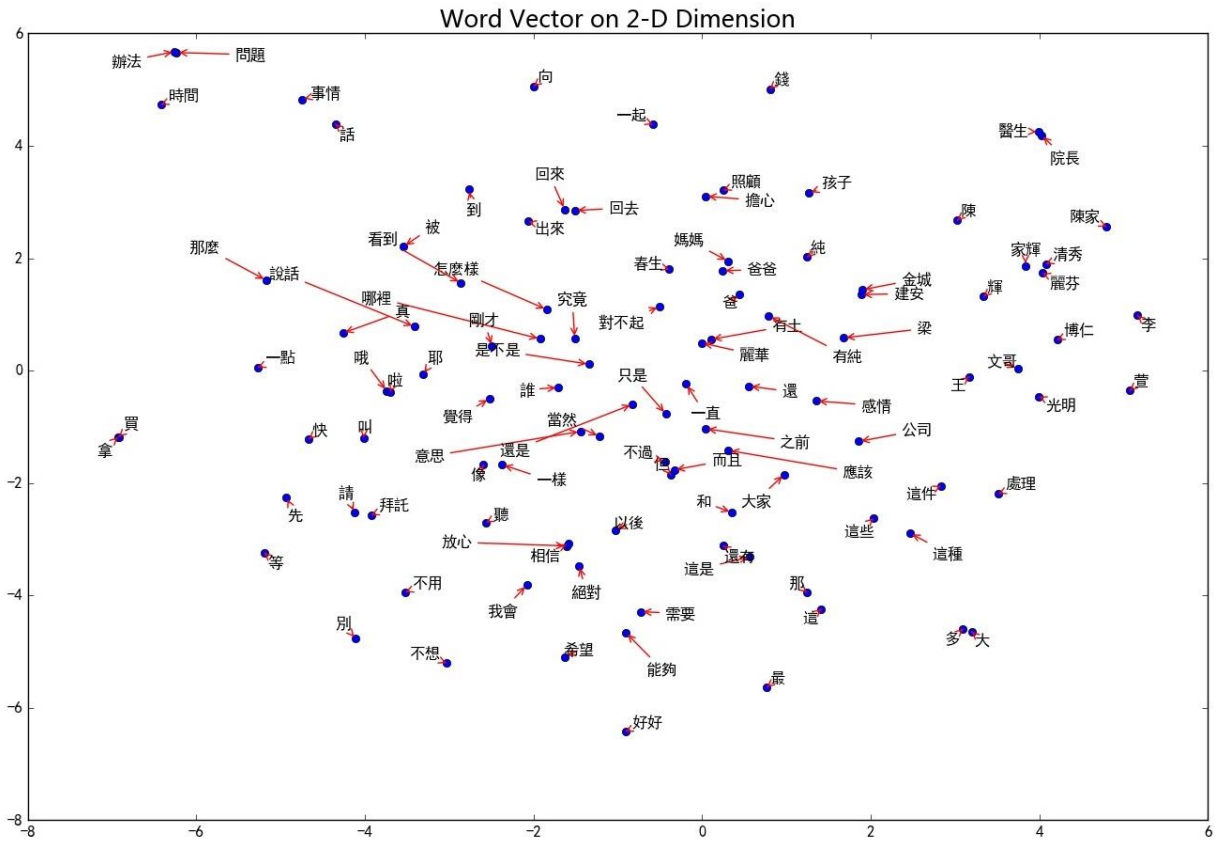
size：訓練出來每個詞向量的維度大小。

min_count：訓練語料庫中，出現次數大於 min_count 次的字詞。

window：訓練過程中一個詞參考的左右相鄰字詞數。

iter：訓練語料庫的次數。

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。

詞義相近的會被投影到相近的位置：如(爸爸/媽媽/爸)、(辦法/問題)、(照顧/擔心)、(回來/出來/回去)、還有幾個人名被歸類為相近的位置。

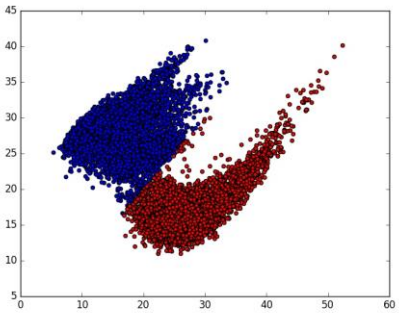
C. Image clustering

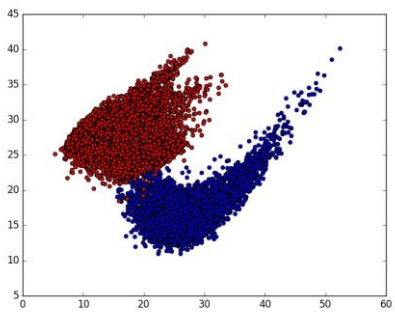
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

Auto-Encoder：使用 DNN，將 784 維降到 254、64 維，再升到 784 維。Loss function 使用 mse，optimizer 使用 adam，降到 64 維後將其結果取出，使用 KMeans 進行 Cluster。Kaggle 分數為 0.94。人工比對前 100 筆資料，觀察到 100 筆皆分群正確。

直接使用 PCA 降維至 32 維，再由 KMeans 進行 Cluster，此法效果不彰，Kaggle 分數只有 0.03。人工比對前 100 筆資料，觀察到只有 80 筆分群正確。

C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。(下左圖)





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。(上右圖)

顏色相反是因為標示 0/1 的不同造成，大致上來看是分類的滿準確的。僅有在兩個顏色邊界處會有標示錯誤的情形發生。