學號: R06921083 系級: 電機碩一 姓名: 鄭克宣

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。

** 第一列為原本算出的 eigenface,第二列是第一列取負值

















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。

分別使用 0.jpg, 4.jpg, 10.jpg, 14.jpg, 第一列為原圖,第二列為 reconstruction















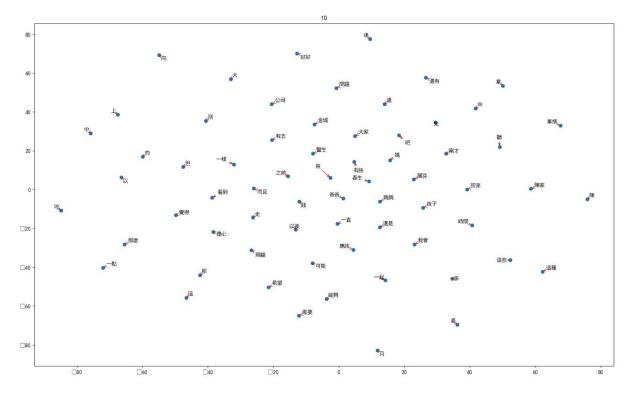


A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重,請用百分比表示並四捨五入到小數點後一位。

7.3% 3.6%	2.7%	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。
 - 1. gensim 的 word2vec
 - 2. min_count=6,至少出現6次才加入計算 word vector
- 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。

可以發現 "爸" 之於 "媽" 就如同 "爸爸" 之於 "媽媽" ,其他實在沒能觀察出什麼

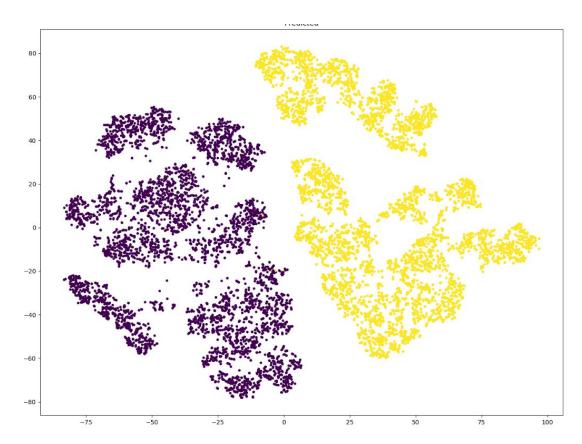
C. Image clustering

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)

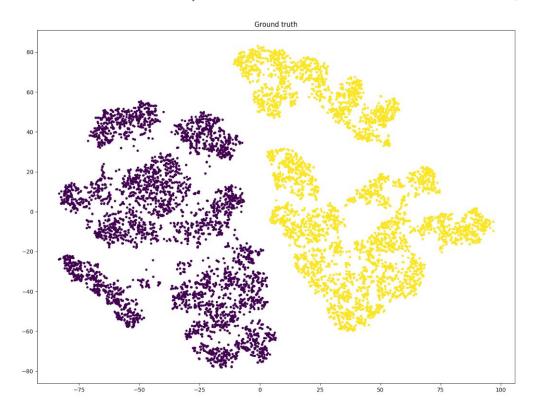
	DNN encoder	CNN encoder
f1_score	0.93158	0.03246

可以發現 DNN 遠比 CNN好很多,個人認為是因為 CNN 在 max pooling 時流失了不少資訊

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。



稍微可以發現 dataset B 在預測的結果有些會混入 dataset A 中,但dataset A 幾乎不會混到 dataset B