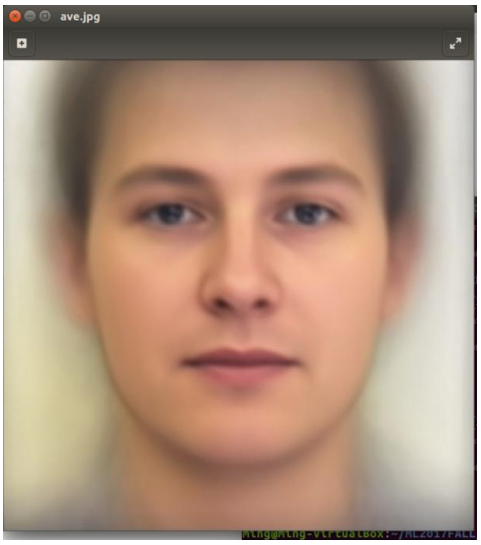
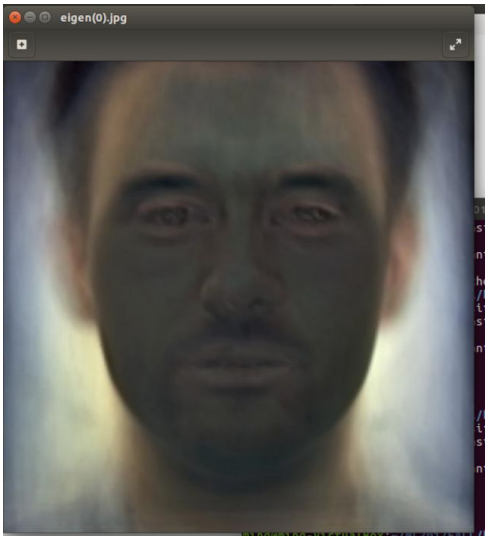
學號：B04901102 系級： 電機三 姓名：簡仲明

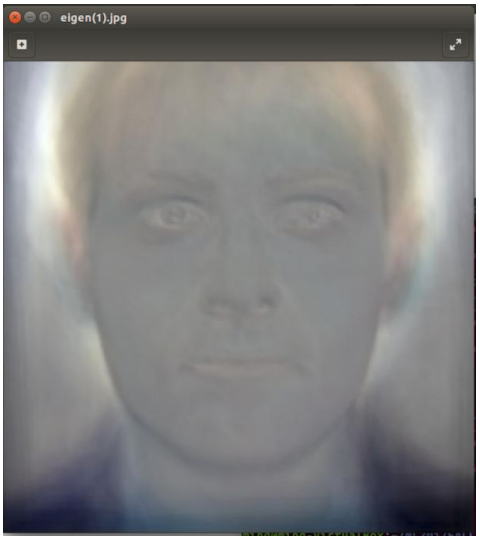
A. PCA of colored faces

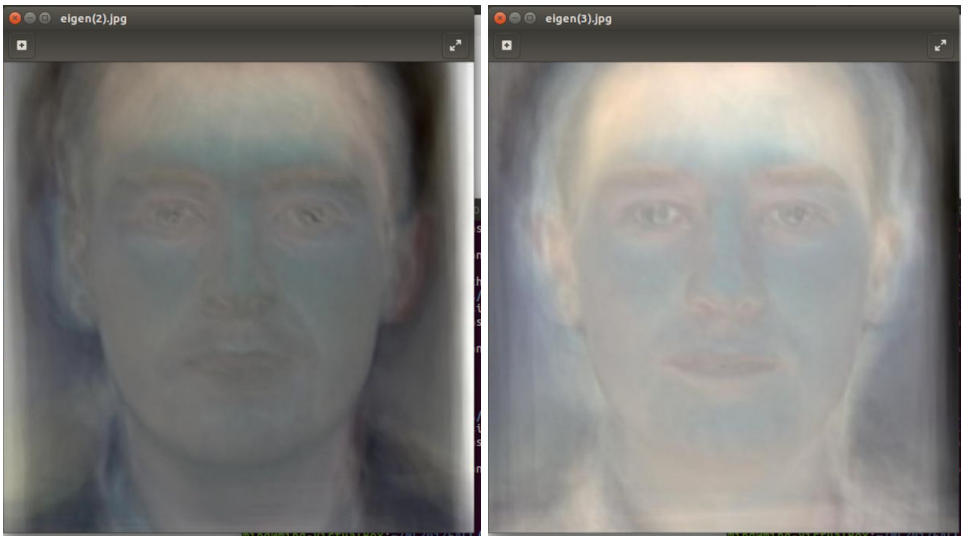
A.1. (.5%) 請畫出所有臉的平均。



A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

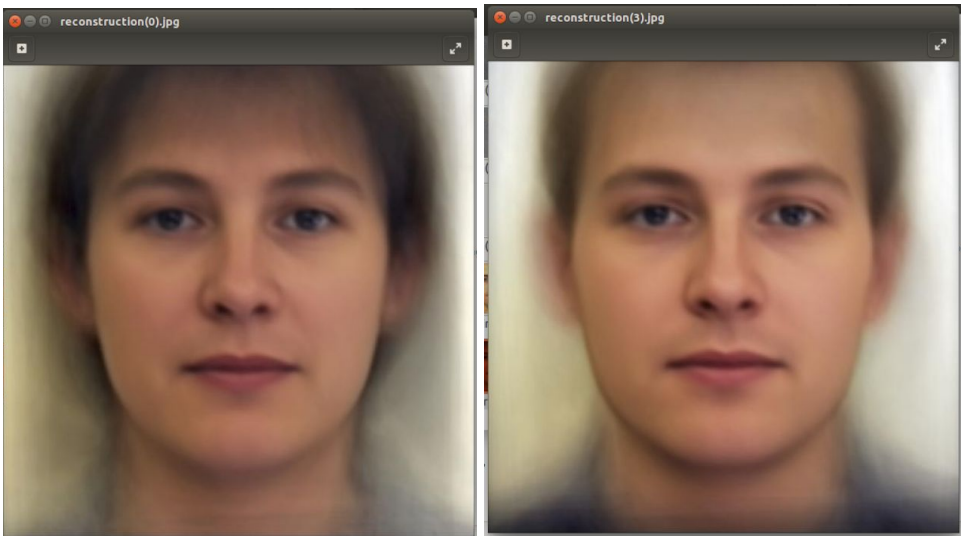


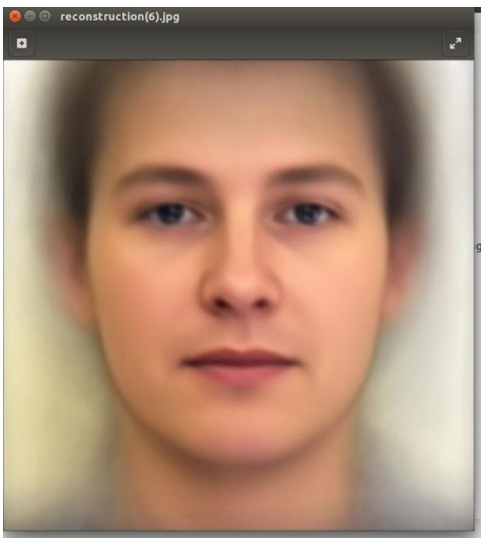


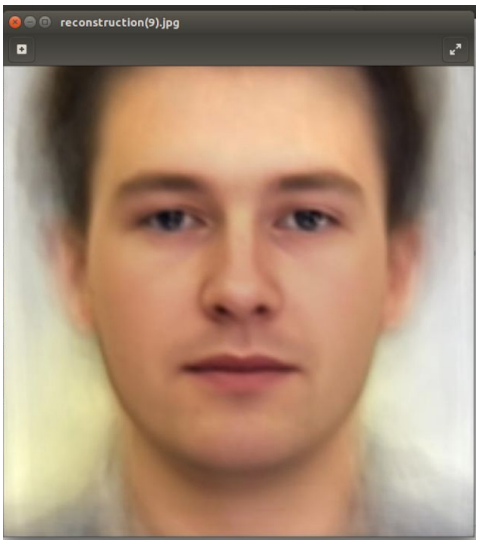


A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

以下四張分別為0.jpg、3.jpg、6.jpg、9.jpg的reconstruction結果。







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。

本題resize為200*200。

eigenface1：4.1%

eigenface2：3.0%

eigenface3：2.4%

eigenface4：2.2%

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 word2vec來實作，有調整過的參數分別為：

size=10，為word轉成的vec的維度(因為要做visualization所以不宜太高，跟一般實作不同)

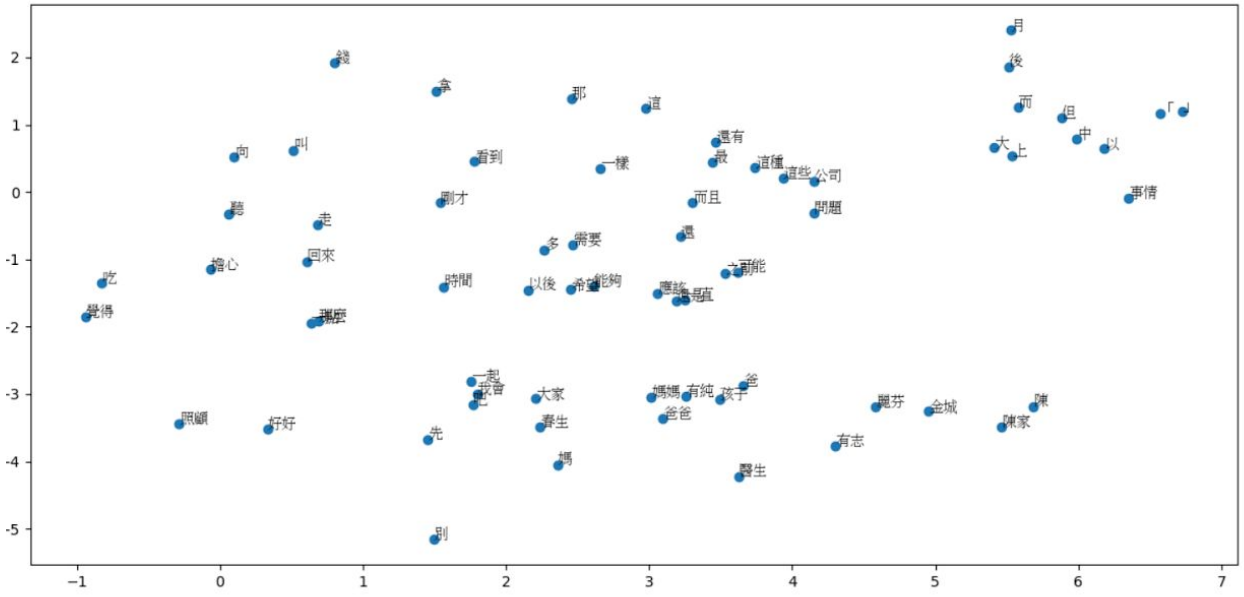
window=3，以目前字為中心，往兩邊看的寬度

min_count=10，出現次數小於此數字的字將被視為OOV

sg=1，使用skip-gram

iter=20，在corpus上面train的迴圈次數

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。

途中可以看出，相似的詞幾乎都已經被擺在一起，例如右下角到芳邦的"春生"、"媽""、"大家"、"媽媽"、"爸爸"、"醫生"、"爸"、"有純"、"孩子"、"有志"、"麗芬"、"金城"、"陳家"、"陳"皆為人物，中間有"希望"、"能夠、"需要"的助動詞組合，在他們右邊則有"應該"和"可能"兩個同義詞'，右上角則是一系列順接連接詞，至於整個左上角到左邊則是動詞的天下。

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)

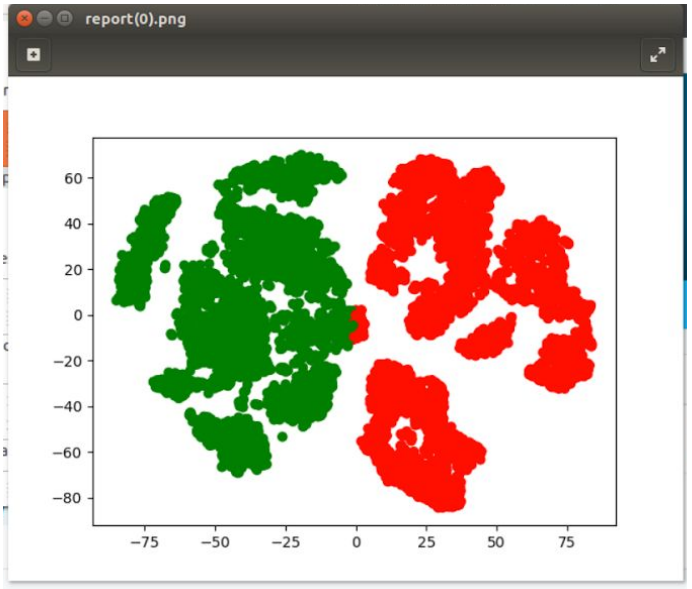
方法一：

首先用PCA降到16微，接著用t-SNE降到8微，最後用AgglomerativeClustering降到2維(kmeans的算法在只有兩個core的情況下常常會將明顯在同一個cluster的兩個點切開，因此用pair-wise distance來計算的本演算法較適合此task，實作上也得到相同的結果，在F1-score上大約可以進步0.1)。由於14萬個點實在太多，會造成t-SNE降維時無法漂亮的切出分界線，因此實驗證實大約將每1000個點分成一個epoch計算，最後再merge起來可以有較佳結果。最佳結果為0.52617。

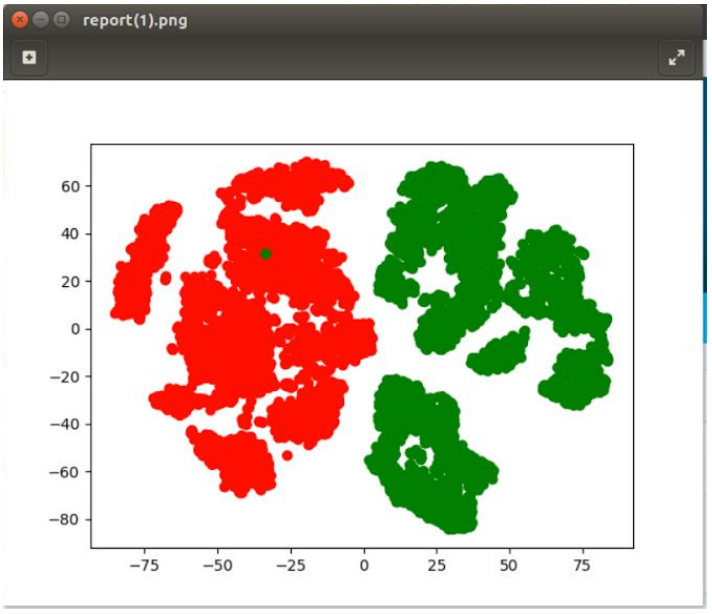
方法二：

首先用PCA降到400維，接著直接使用kmeans降到2維(因為AgglomerativeClustering的pairwise distance對14萬個點的計算量過大，硬體無法負荷)，可以有1.000000的結果。這根本是黑魔法不知道如何解釋。若硬要解釋，可以想像成t-SNE在降維過程中丟失了許多資訊，因此已經不適合用單純的距離演算法來分群。

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。



這兩張圖是使用方法一所得到，可以看到相當不錯的分群結果，明顯切出兩個dataset之間的分界，唯獨正確label中有一個綠色的點落在左半邊的紅色區域，而且是在非常中間的位置，令人有些匪夷所思，不知道是不是助教刻意留下來的?