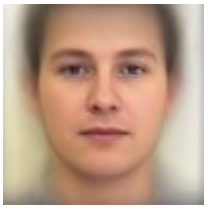
學號：B04902105 系級：資工三 姓名：戴培倫

A. PCA of colored faces

由於記憶體不夠，我將照片resize成100*100。

A.1.



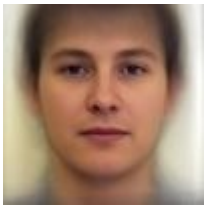
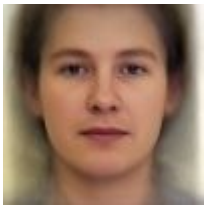
A.2. 由左至右分別是第0~3個eigenface




A.3. 由左至右分別是10,20,30,40.jpg，下面是reconstruct的結果





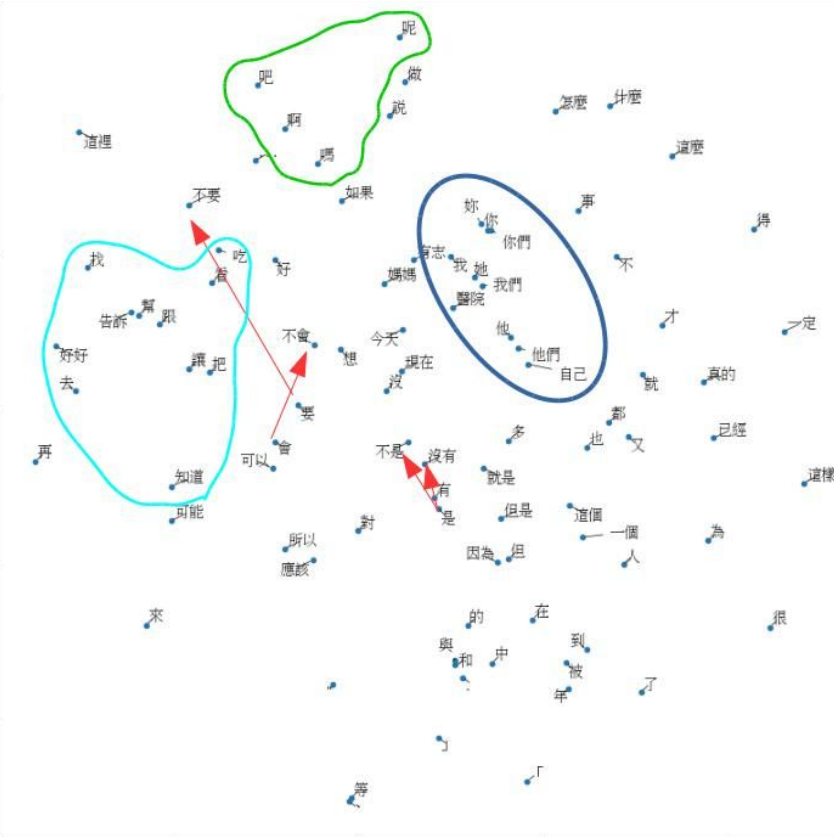




A.4. 前四大eigenface在所有eigenface中的weight佔的比例分別為 4.2%, 3.0%, 2.4%, 2.3%

B. Visualization of Chinese word embedding

B.1. 我用的是gensim.models.Word2Vec(size=100, iter=100)，用出現超過5000次的詞，共92個詞做embedding，我有試過dimension調到200, 300, 並調整iteration，但發現從embedding中看出來的意義都沒有這組參數好。

B.2.



B.3. 深藍色橢圓中可以看出(我,我們)、(你,你們)、(他,他們)有很明顯的左上->右下的關係，而(你,妳)、(他,她)也有很明顯右下->左上的關係。紅色箭頭則是正面->負面，除了(會,不會)指向上偏右之外，其他都指向左上。從綠色區塊可發現語尾助詞都集中在同一區。而淺藍色區域則包含了幾乎全部的動詞。

C. Image clustering

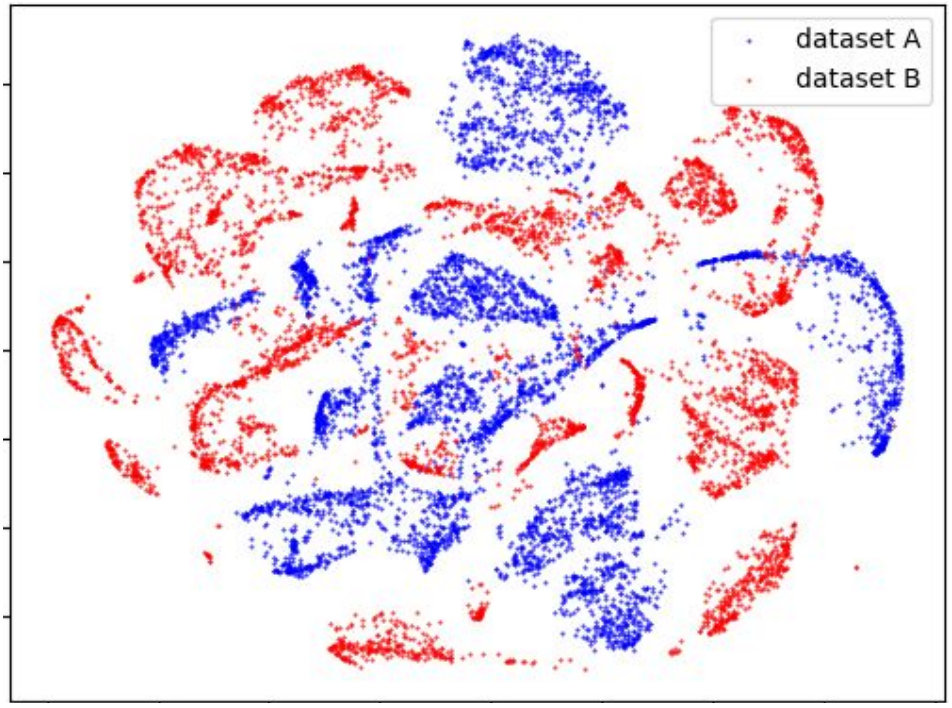
C.1.

C.1.1. PCA降到100維，用kmeans分兩個cluster
=> public : 0.03024

C.1.2. PCA降到100維，用kmeans分4個cluster，[0,2,3]為一組，[1]自己一組 => public : 0.08047

C.1.3. Deep autoencoder降到32維，用kmeans分兩個cluster => public : 0.98326

C.2. 這是我將data用autoencoder降到32維後，再用TSNE降到2維，並做了kmeans clustering的結果。



C.3. 兩張圖看不出什麼差別檢查後發現10000筆資料中只有將4個dataset A的判成dataset B的，其他判斷都正確。