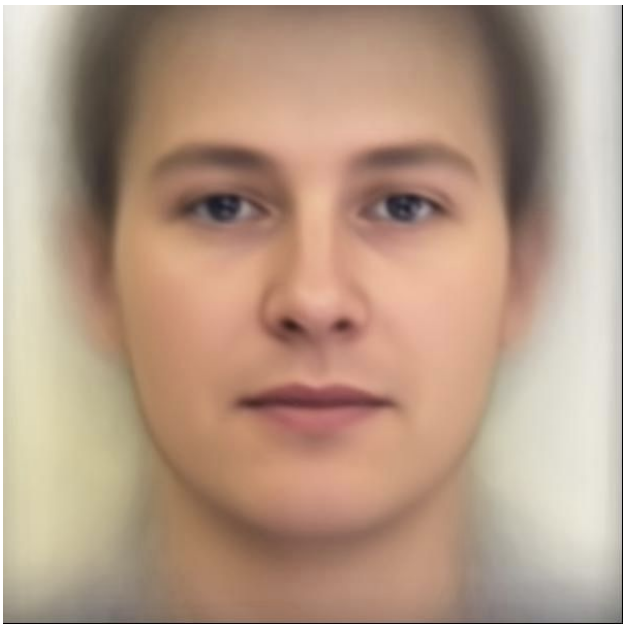
學號：R06922129 系級：資工碩一 姓名：丁縉楷

(collaborator: r06922130 葉韋辰、 r06944034 黃禹程)

A. PCA of colored faces

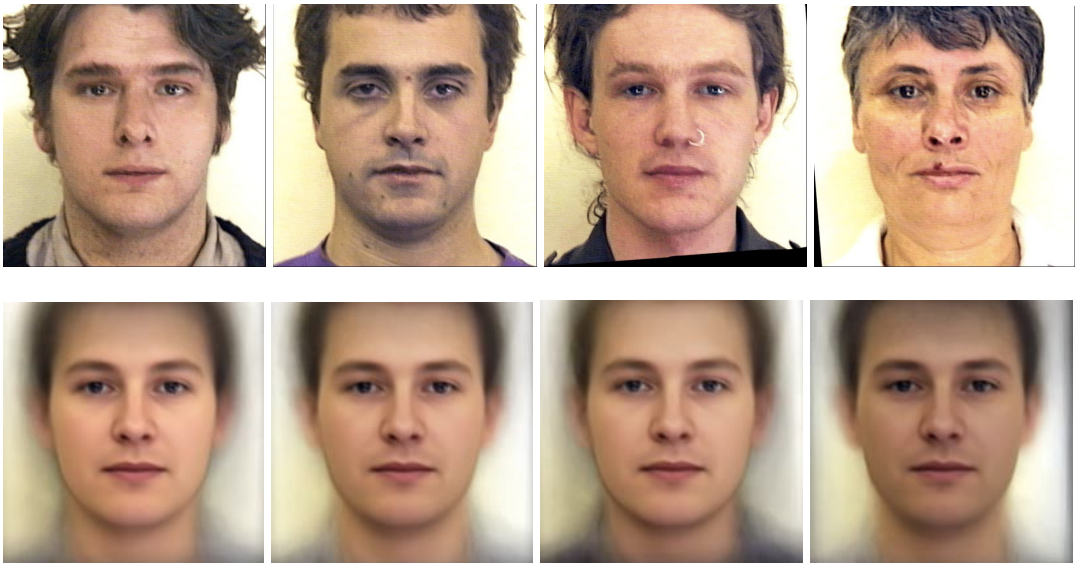
A.1. (.5%) 請畫出所有臉的平均。



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

加負號

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。



我只取了4個eigenvector，所以圖片都大同小異

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。

(resize成200)

1. 7.35%
2. 3.64%
3. 2.75%
4. 2.23%

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。

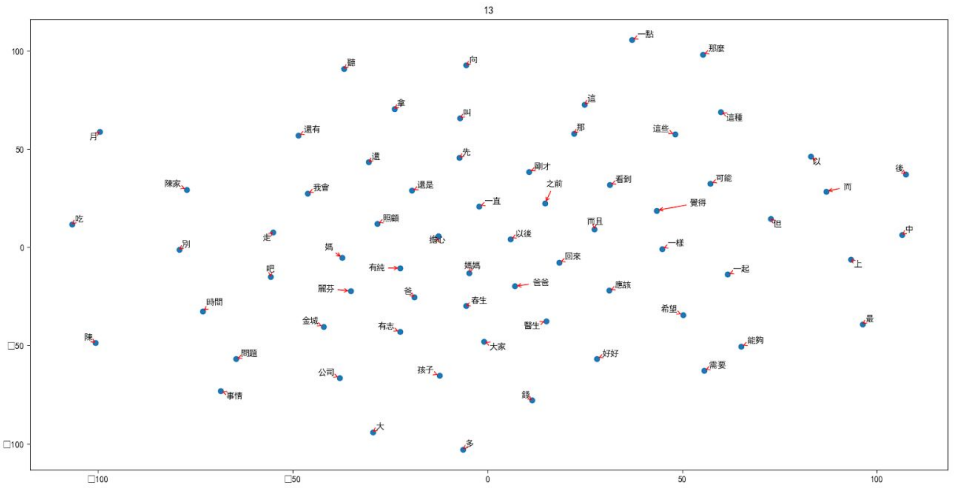
使用gensim的word2vec

window: 建vector時，考慮這個包含這個word的sentences大小

size: vector 的 dimension大小

min_count: 只考慮出現次數大於某個數的word

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。

可以看到人名分成一群，動詞分成一群，連接詞分成一群

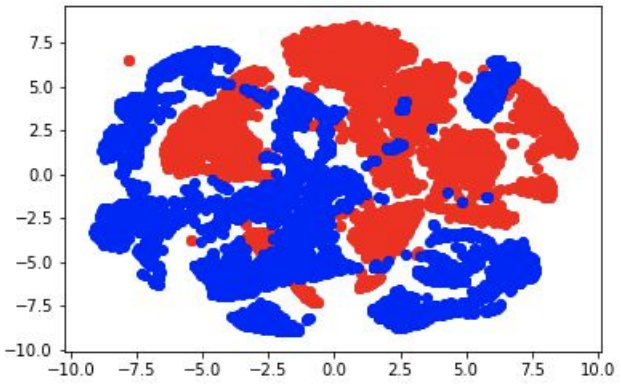
C. Image clustering

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)

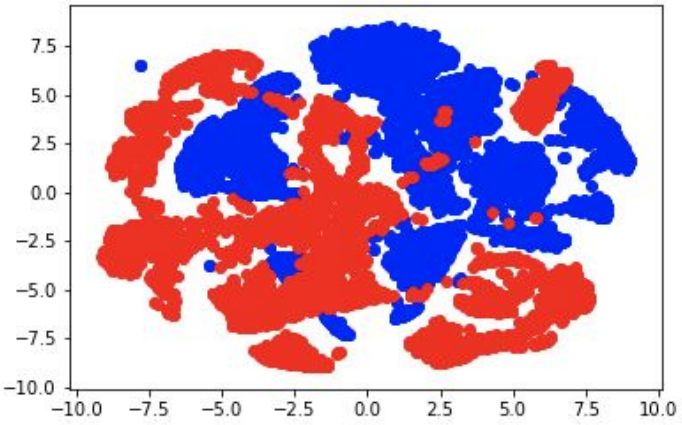
	private	public
PCA(dim = 300)	0.03048	0.03023
DNN(dim = 64)	0.93538	0.93620

表格為兩種降維方法，分群方法皆為kmeans，可以看到deep auto encoder 明顯比PCA還要好

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。



可以看到跟自己預測的幾乎一樣，只是label相反，但是降維之後有些點有重疊，雖然預測的結果幾乎全對，不過可能是降成兩維視覺化的方法不好。