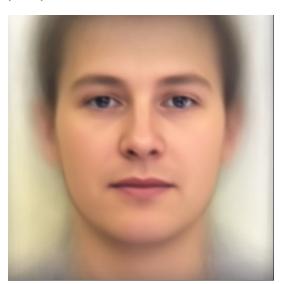
學號:B04901136 系級: 電機四 姓名:張家銘

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。

第2張圖

第4張圖





第6張圖

第8張圖





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。

eigenface1: 4.1%

eigenface2: 3.0%

eigenface3: 2.4%

eigenface4: 2.2%

B. Visualization of Chinese word embedding

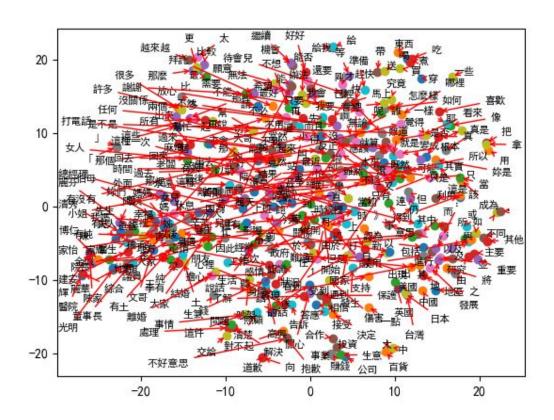
B.1. (.5%) 請說明你用哪一個 word2vec 套件,並針對你有調整的參數 說明那個參數的意義。

使用 gensim 的 word2vec後。

min_count=1, 代表 會把sequence 中每一句話都拿去做word embedding。

其餘參數都使用預設參數,沒有做調整。

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。

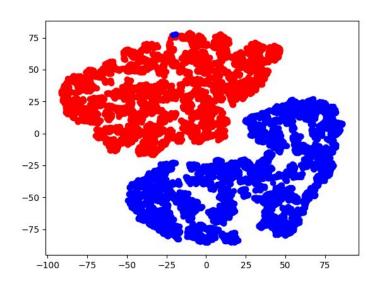
我選擇的是:所有大於出現次數大於3000的字

圖形上的左半部份,出現的都是職位,或人名,然後國家會都出現 在右下角。

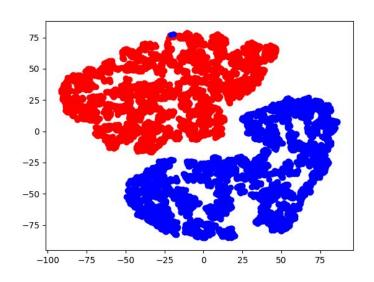
C. Image clustering

- C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的 的降維方法或不同的 cluster 方法都可以算是不同的方法)
 - (1) 用pca 降維到400,然後直接用kmeans 做 cluster,F1準確率: 100%
 - (2) 用autoencoder降維到64維,然後用kmeans 做 cluster, F1準確率: 100%

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。



跟我預測的幾乎一樣了,可是還是有一點點不太一樣,誤差很低。 我印出錯誤的data,看不出為什麼會錯。