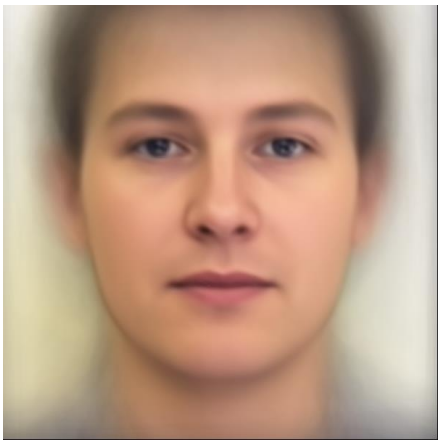
學號：R06942039 系級：電信所碩一 姓名：何明倩

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。

第一個 eigenface

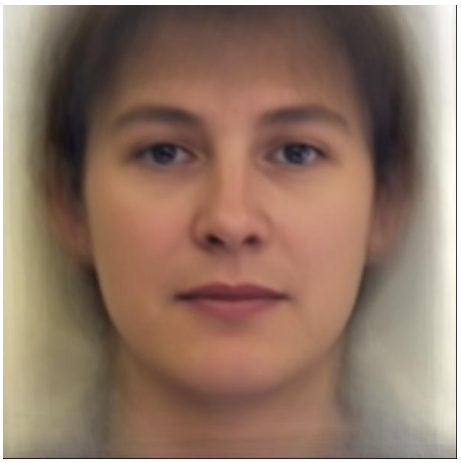
第二個 eigenface

第三個 eigenface

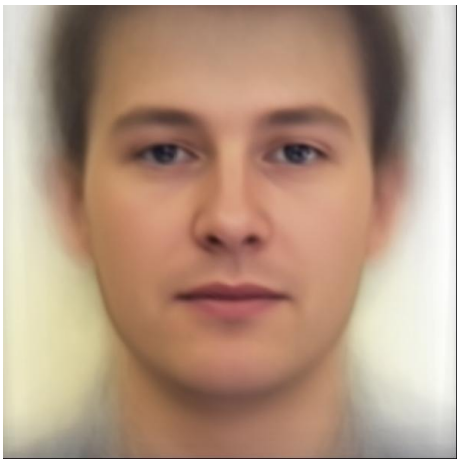
第四個 eigenface

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。

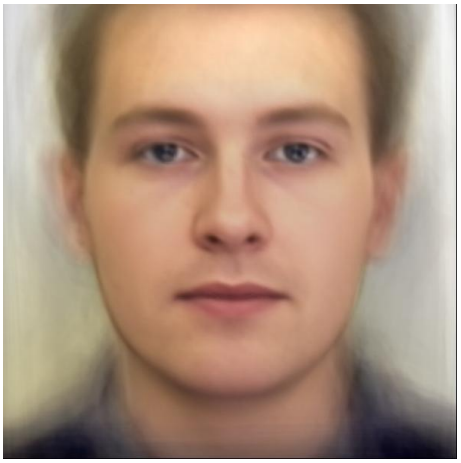
1.jpg



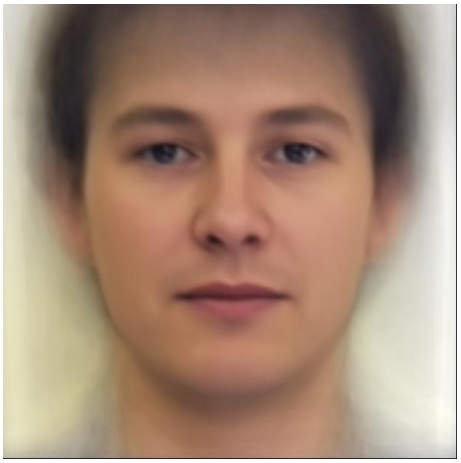
11.jpg



45.jpg



414.jpg



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。

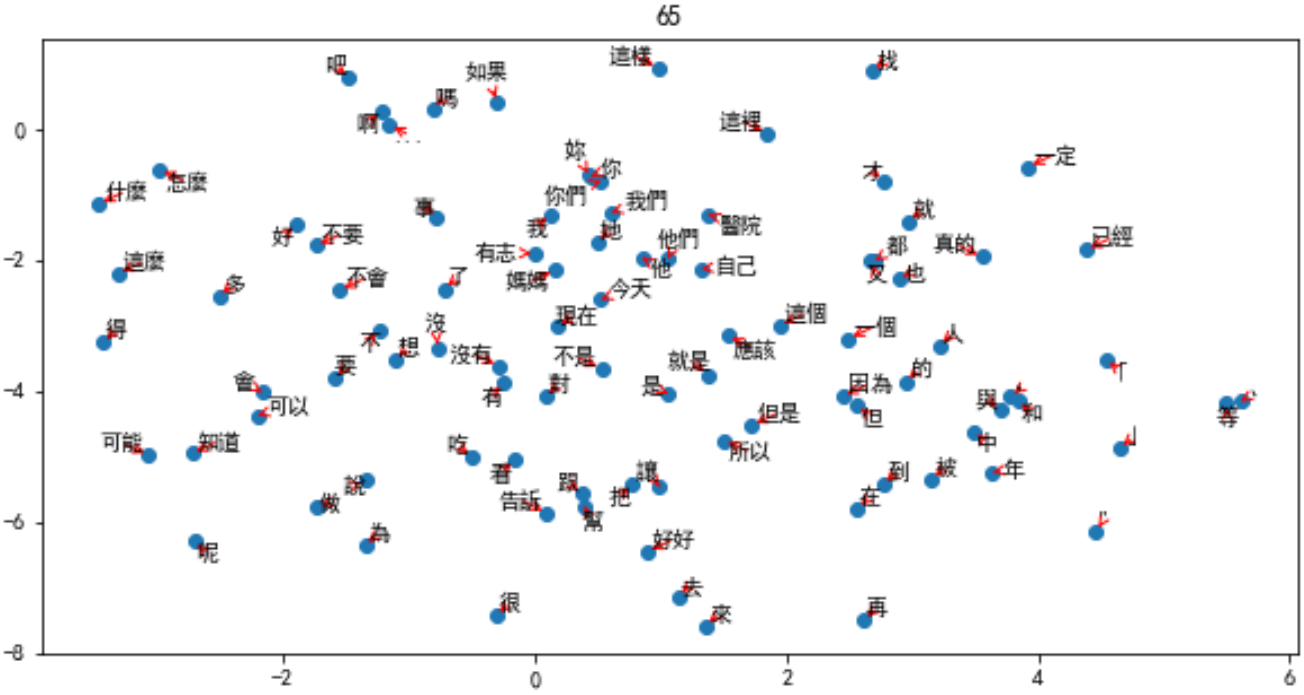
	Eigenface1	Eigenface 2	Eigenface 3	Eigenface 4
weight	4.2%	3.0%	2.4%	2.2%

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。

我使用的 word2vec 是利用 gensim 裡面的 word2vec 套件，我選的 size 為 95，size:表示訓練出的詞向量會有幾維。Min_count 設為 5000，min_count:若這個詞出現的次數小於 min_count，那他就不會被視為訓練對象。

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。

我們可以從 visualization 的結果發現，類似意思(像是'與'、'和')、相似的字(像是'妳'、'你'、'你們')或是相反的詞彙(像是'沒有'、'有')會聚集在一起。我覺得這可能是因為這些字在在類似的句子有所相關，所以 train 出來的結果才會很靠近。

C. Image clustering

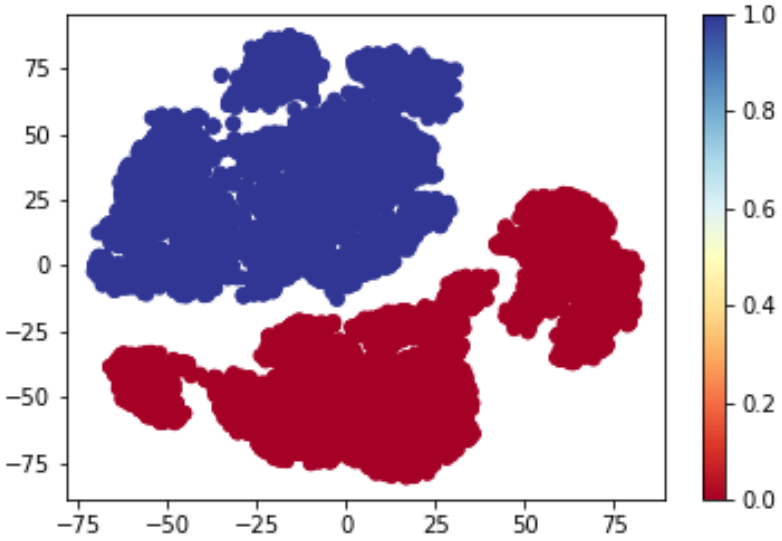
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

我使用了兩種降維方式: (1)PCA (2)Auto-encoder

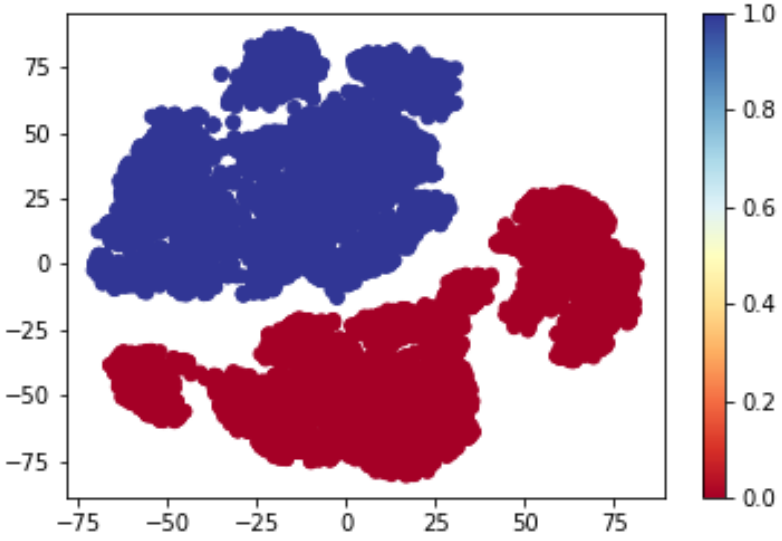
accuracy	PCA	Auto-encoder
private	0.02945	1
public	0.02925	1

從結果可以很明顯的看出 PCA 降維所預測出來的結果很差，而用 auto-encoder 的方法就好非常的多。我覺得可能是因為 PCA 在線性降維的過程中損失太多資訊導致再做 kmeans 的時候無法學到相似的 feature 來做準確的區別。

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。



由於我預測來的 label 和 ground truth label 是一模一樣的，所以做出來的二維分布是一模一樣的。