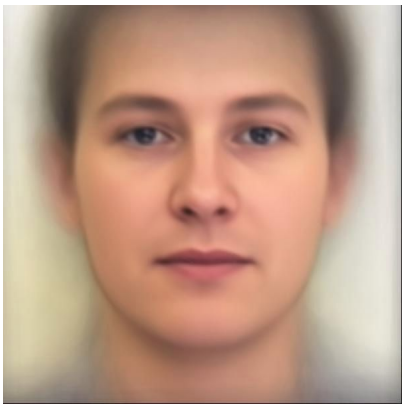
學號：r05922145 系級：資工碩二 姓名：郁錦濤

A. PCA of colored faces

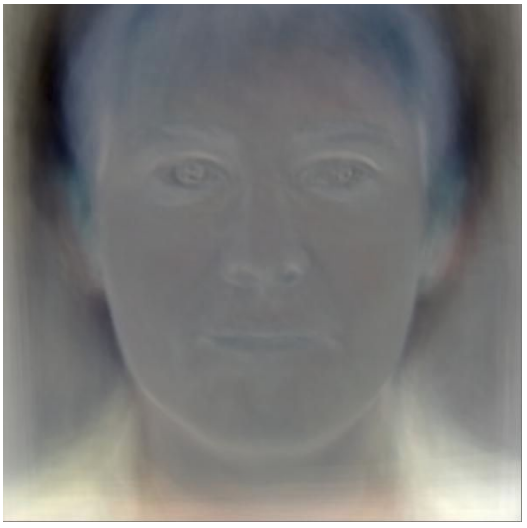
A.1. (.5%) 請畫出所有臉的平均。

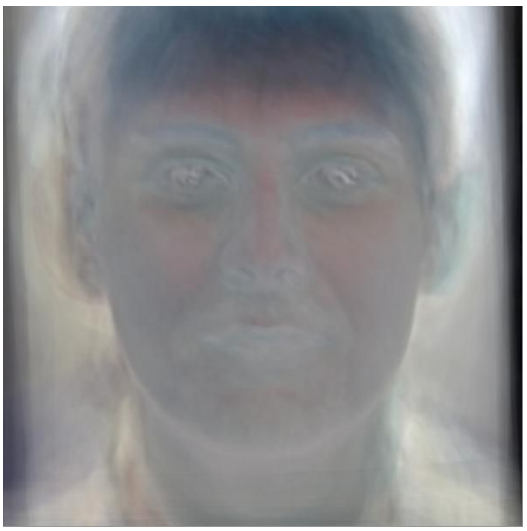
答：



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

答：

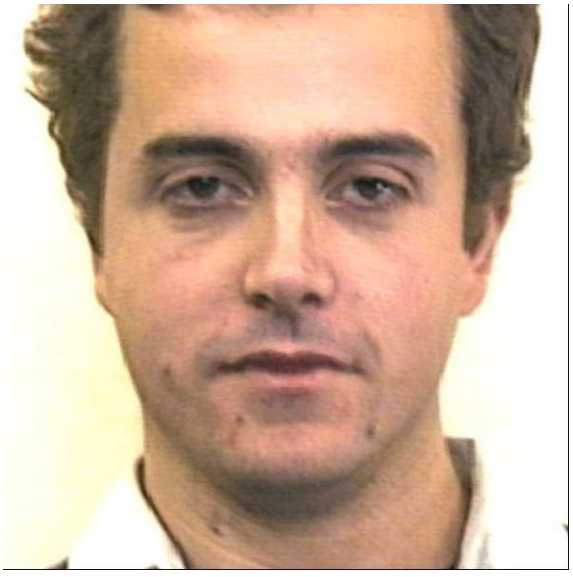




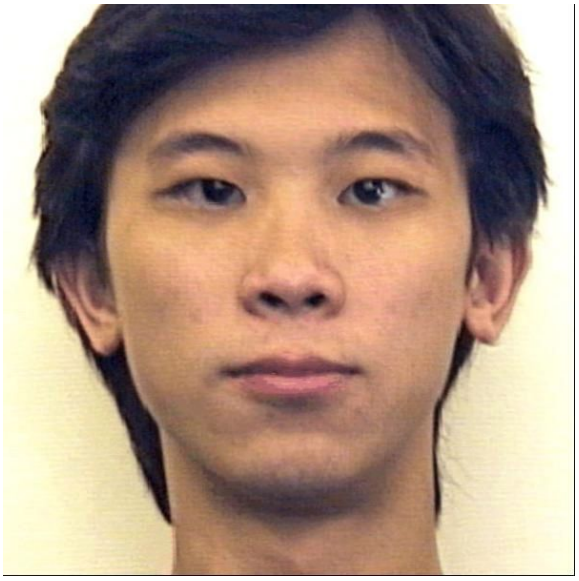
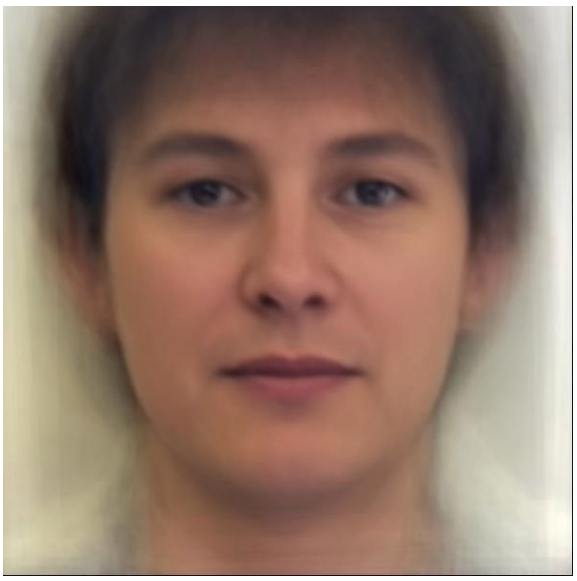

A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

答：我從數據集當中分別按照順序抽出序號爲21,142,221,321的四張照片，分別用前四大Eigenfaces進行reconstruction。以下圖片從上往下依次爲index爲21,142,221,321的照片，從左往右依次爲原圖和reconstruction的圖片。

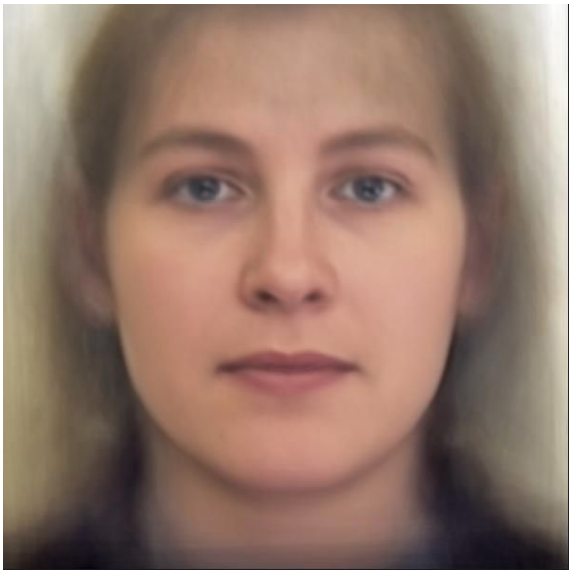
index=21:



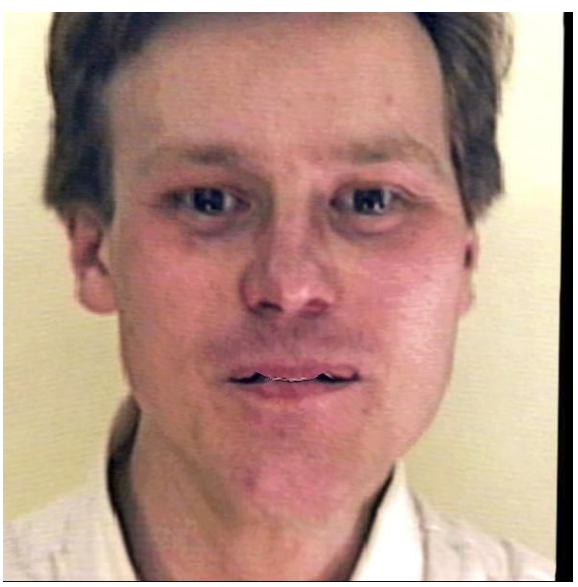
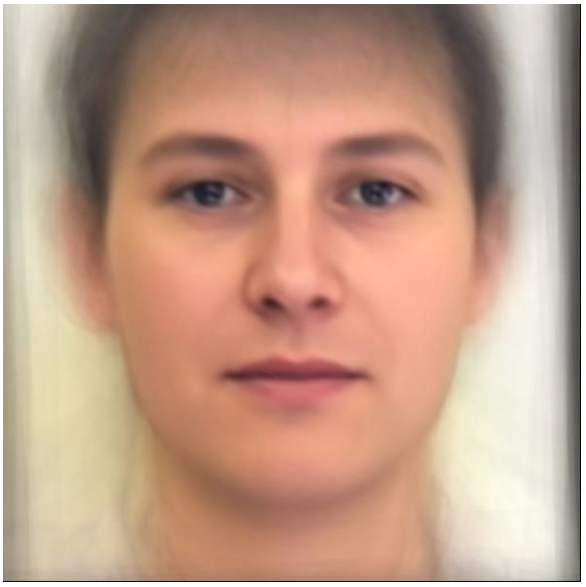

index=142:

index=221:

index=321:

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。

答：

```
[ 4.14747808e-02  2.95084914e-02  2.38935984e-02  2.20932885e-02
  2.07111886e-02  1.85183718e-02  1.61430734e-02  1.48297575e-02
  1.43741055e-02  1.23772648e-02  1.05063762e-02  9.94618250e-03
  9.70274125e-03  9.26535019e-03  8.91389674e-03  8.76030584e-03
```

No.	1	2	3	4
Ratio	4.1%	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。

答：我用的是gensim中的word2vec套件。我調整的參數有size和window，其他參數都是default。

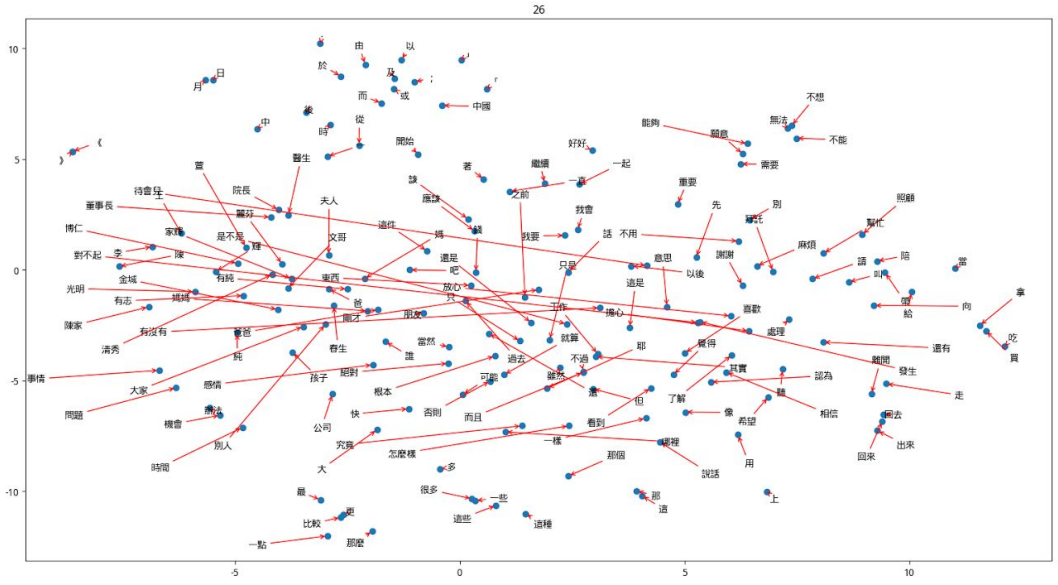
word2vec的參數有很多，下面我將說明幾個重要參數的意義。

- size: 250，用來表示word vector的維度。
- windiw: 5，用來表示訓練窗口的大小，數值5表示每個詞會考慮前5個詞和後5個詞給出機率。
- hs: 指定是否使用Hierarchical Softmax。因爲在默認情況下，只使用softmax，輸出時要計算的參數會很多， hs會先將單詞分類，一層一層輸出，會顯著減少計算量。

- sg:　用來表示訓練算法，sg=0（默認）表示使用CBOW，sg=1表示使用skip-gram。
- alpha: 初始化的learning rate。

B.2.　(.5%) 請在 Report 上放上你 visualization 的結果。

答：



B.3.　(.5%) 請討論你從 visualization 的結果觀察到什麼。

答：做Visualization時，我的K選取值爲大於等於2000並且小於等於6000。(與助教所給的有些區別，從我的這個可以很容易看出很多信息。)

從上圖中，我看出，人名和名詞（如麗芬，金城，爸，媽等）在圖片中聚集在左側中央附近，一些介詞（如由，以，及，而，或等）會聚集在圖片上方區域。而一些描述程度的詞語（如最，比較，更等）則會聚集在圖片下方區域，而一些否定的詞語（如無法，不想，不能）則聚集在圖片右上方區域。

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)

答：降維方法我採用了autoencoder和PCA降維。

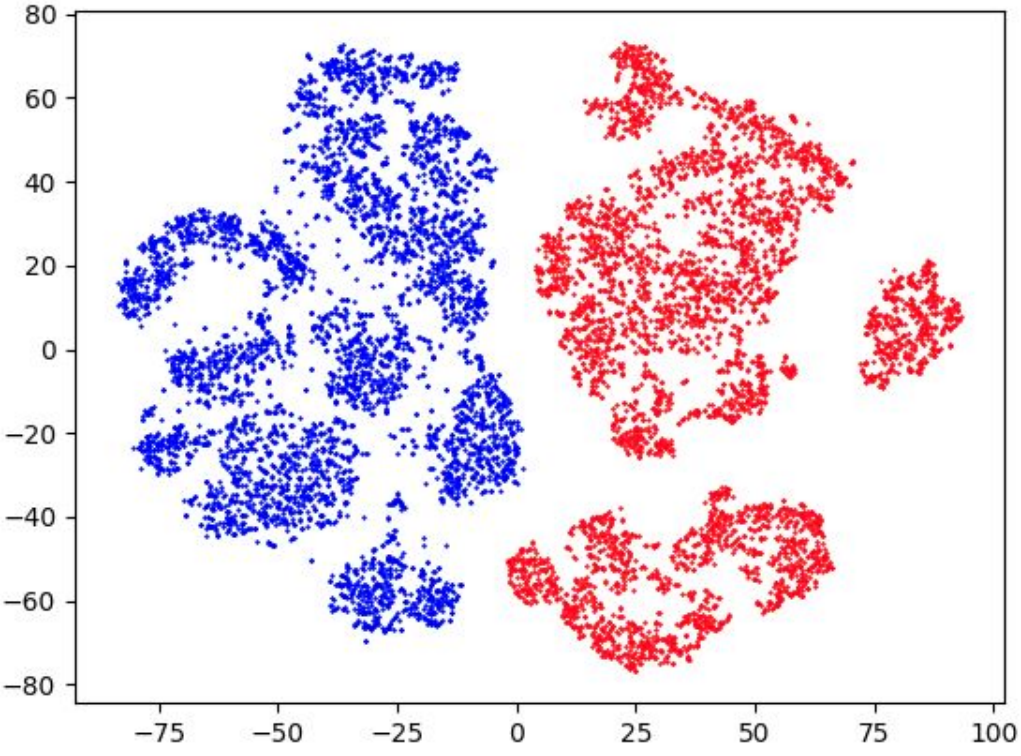
autoencoder降維其實就是使用DNN降維減小feature的維度至64維，然後採用kmeans進行分類。

而PCA是通過sklearn的套件實現降維到64維，然後也採用sklearn進行分類，但是最後performance比較不好，因此得出在此例子中PCA降維效果不太好。

降維	Public Set	Private Set
Autoencoder(DNN)	1.00000	1.00000
PCA	0.03024	0.03048

C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。

答：



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。

答：我用同一個model進行train的時候，分別用我自己用kmeans生成的label和助教所給的label做tsne進行視覺化分析。

我自己用kmeans的方法是將predict的值按照0-1分成兩個list進行繪圖。

助教所給的data則是將前5000個和後5000個進行繪圖。

因爲我的model在kaggle上的得分一般爲0.99~1，所以按照正常情況來看兩個圖的顏色和形狀應該是相同的。

比較兩個圖，我發現我預測的label還是與正確值非常接近的。