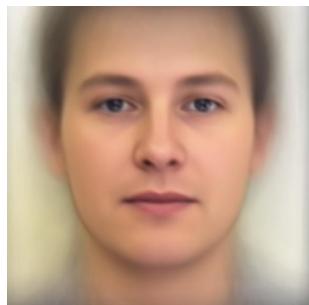


學號：B03902101 系級：資工四 姓名：楊力權

1. PCA of colored faces

1. (.5%) 請畫出所有臉的平均。



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



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。



從左至右分別是200.jpg, 250.jpg, 300.jpg, 350.jpg的reconstruction

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。
- 4.1% 2.9% 2.4% 2.2%

2. Visualization of Chinese word embedding

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。

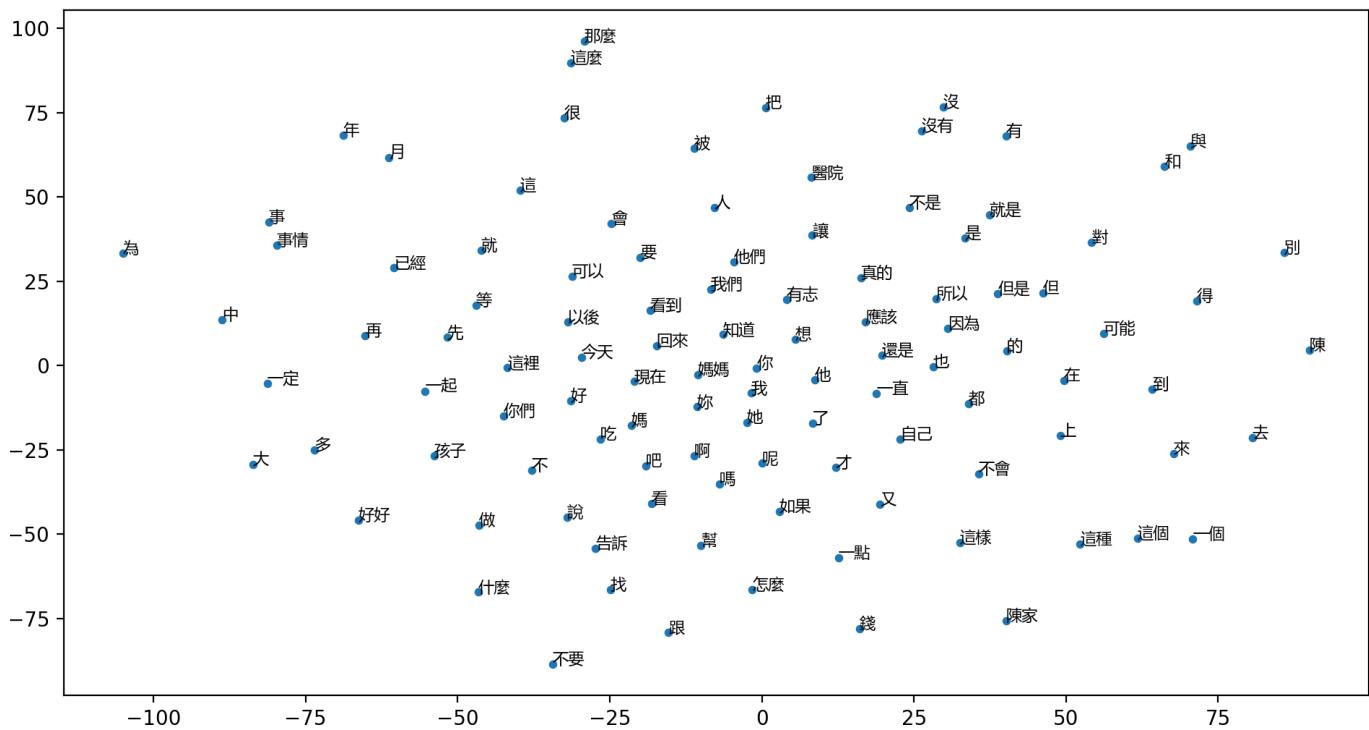
使用Gensim套件，並設定

`size=250`：每個字以250個維度表示。

`min_word=3`：有出現3次以上的字才算有效字。

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

取出現頻率大於4000的詞用tsne降至兩維的結果：



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

可以發現主詞如我、你、妳、他、她、媽媽會聚在一起(圖中間區域)

語氣詞如吧、嗎、啊、呢等也聚在一起(圖中間偏下)

這麼、那麼；和、與；但是、但、因為、所以；有、沒有、沒；不是、

就是、是...等都聚在一起。表示用法類似或辭意相近的詞之word

embedding是比較相近的。

3. Image clustering

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)

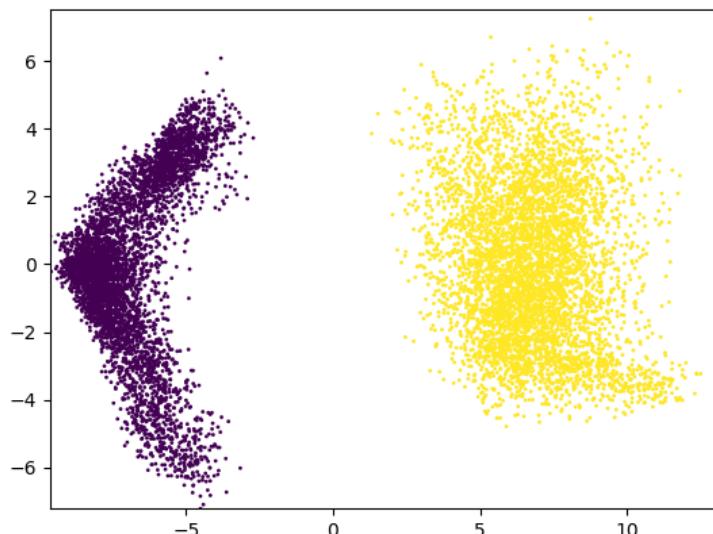
我使用了四種方式：

1. 展開圖片，用PCA降至2維，再做k-means
2. 展開圖片，用PCA降至64維並用tsne降至2維，再做k-means
3. 展開圖片，用DNN autoencoder，取出中層layer 64維，再做k-means
4. 用CNN autoencoder，取出中層layer，再做k-means

方法	kaggle public score	kaggle private score
PCA 2維	0.03180	0.03151
PCA 64維 tsne 2維	0.08824	0.08833
DNN autoencoder	0.94505	0.94623
CNN autoencoder	0.37533	0.37363

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

做完DNN autoencoder 再用PCA降到2維後，k-means的結果。



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。

發現圖形長得與上題一模一樣，表示2.的分類方法已經以準確率100%的表現，成功把data分成兩群。

