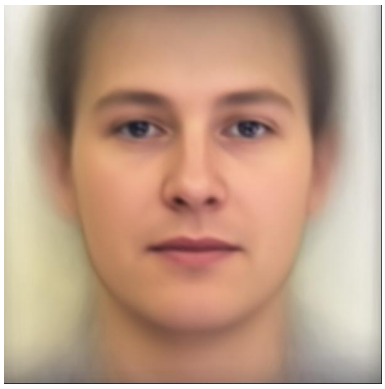
學號：R06922117 系級： 資工碩一 姓名：李岳庭
(collaborators:R06922113)

A. PCA of colored faces

(.5%) 請畫出所有臉的平均。



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

(.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。

0.jpg 1.jpg 2.jpg 3.jpg

(.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。

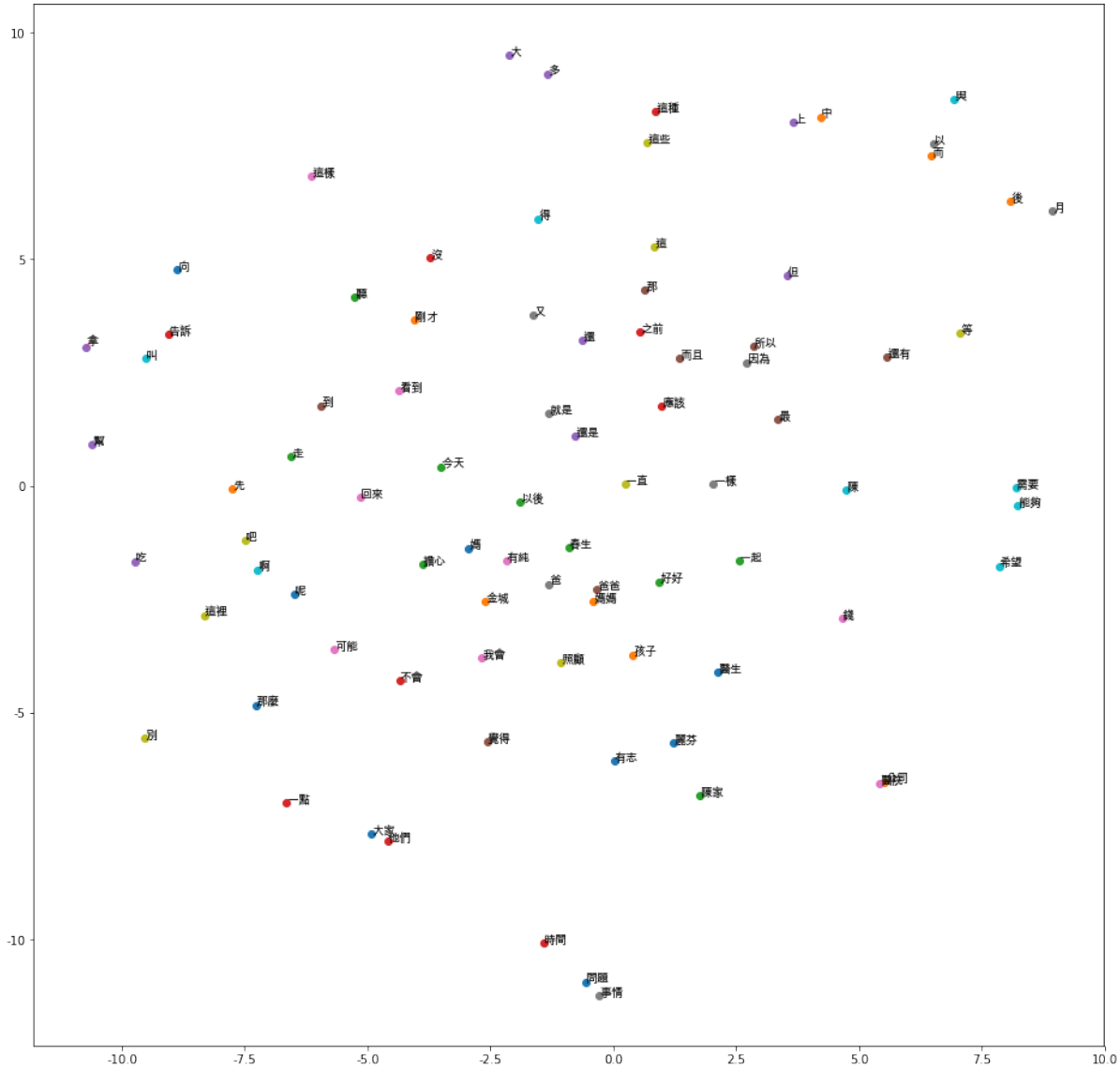
4.2%、2.9%、2.4%、2.2%

B. Visualization of Chinese word embedding

(.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 gensim 的 word2vec 套件，我調整的參數有 size、window 和 alpha，size 就是將每個詞轉成幾維的 vector，window 是句子中前後看幾個詞，alpha 是 learning rate。

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



(.5%) 請討論你從 visualization 的結果觀察到什麼。

{“大家”,“他們”}、{“問題”,“事情”}、{“爸爸”,“媽媽”}、{“需要”,“能夠”}、{“而”,“以”}、{“因為”,“所以”}這些點在圖上的距離非常接近，而這些詞也常常在同一句話同時出現，可以猜測他們的關係密切。

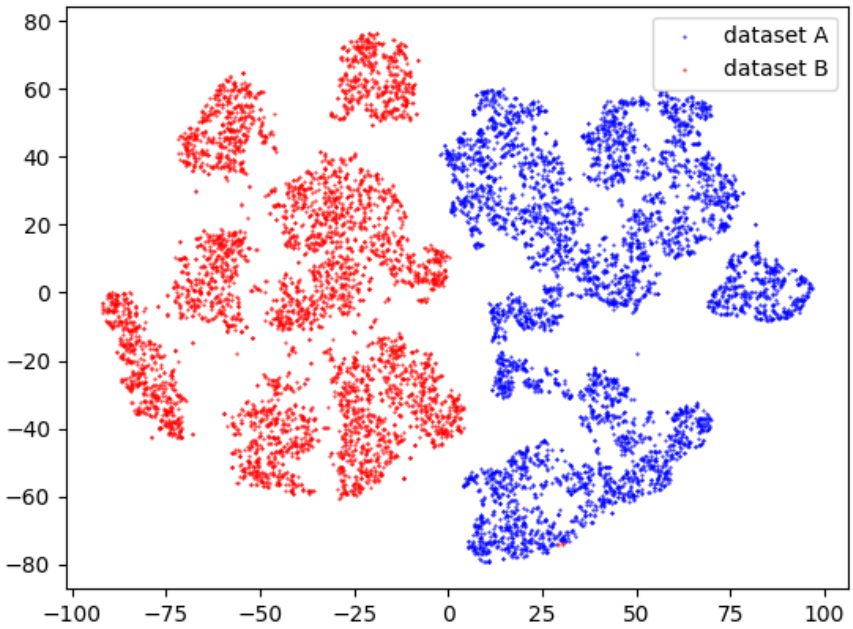
C. Image clustering

(.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)

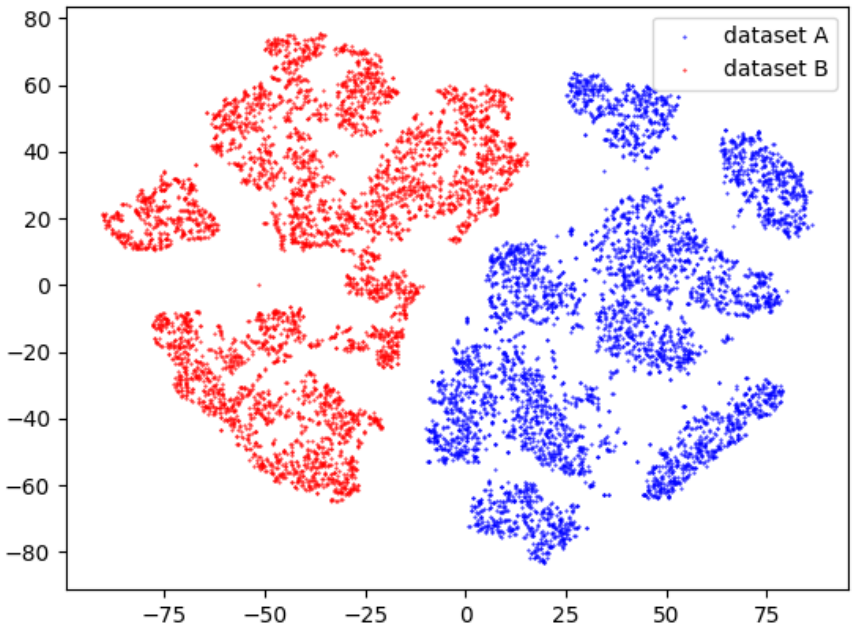
	dim	public	private
PCA	128	0.17558	0.17770
T-SNE	128	0.12285	0.12338
auto encoder	32	0.94467	0.94430
auto encoder	32 (normalized)	0.14799	0.14591
auto encoder	16	0.52726	0.52661
auto encoder	64	0.41120	0.41793

試了上述幾種降維的方式，分群的方式皆為 k-means，降維的 PCA 、t-SNE 的效果都很差，以 auto encoder 效果最好，就多試幾個 auto encoder 的維度，就以 32 最高，而且若對 training data 做 nomalization，結果都超差。另外聽說 PCA 能夠做到 F1 socre = 1.0000，只是沒時間就沒做那些實驗了。

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



(.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。



跟正確答案相比，我的預測結果有兩三個紅色的點落在右下方，這些漏網之魚表示我預測結果沒有百分之百將兩個 dataset 分開，視覺化的結果蠻合理的，預測結果也算很好。