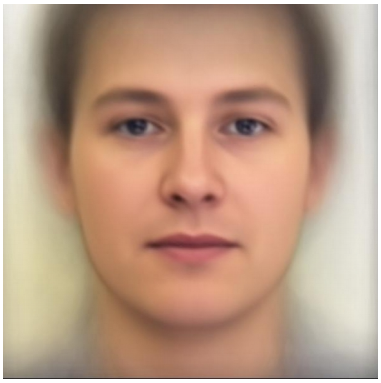
學號：B04901155 系級： 電機三 姓名：鄭閎

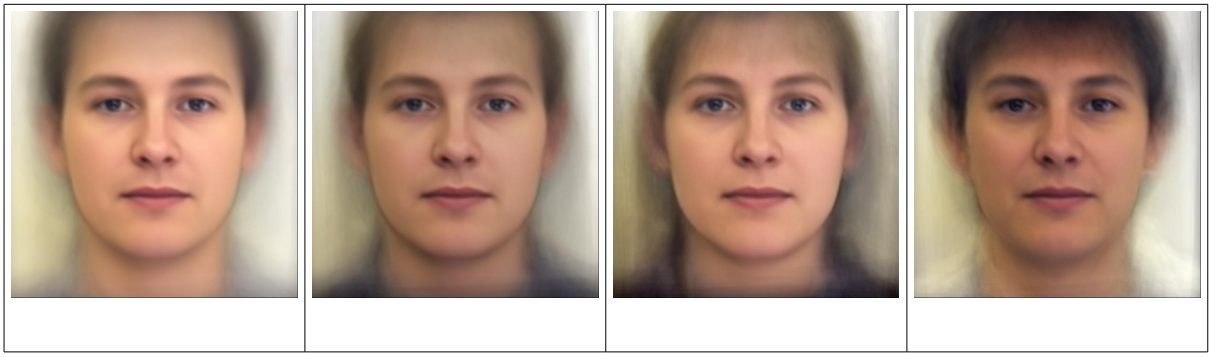
A. PCA of colored faces

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。



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。

4.1%	2.9%	2.4%	2.2%

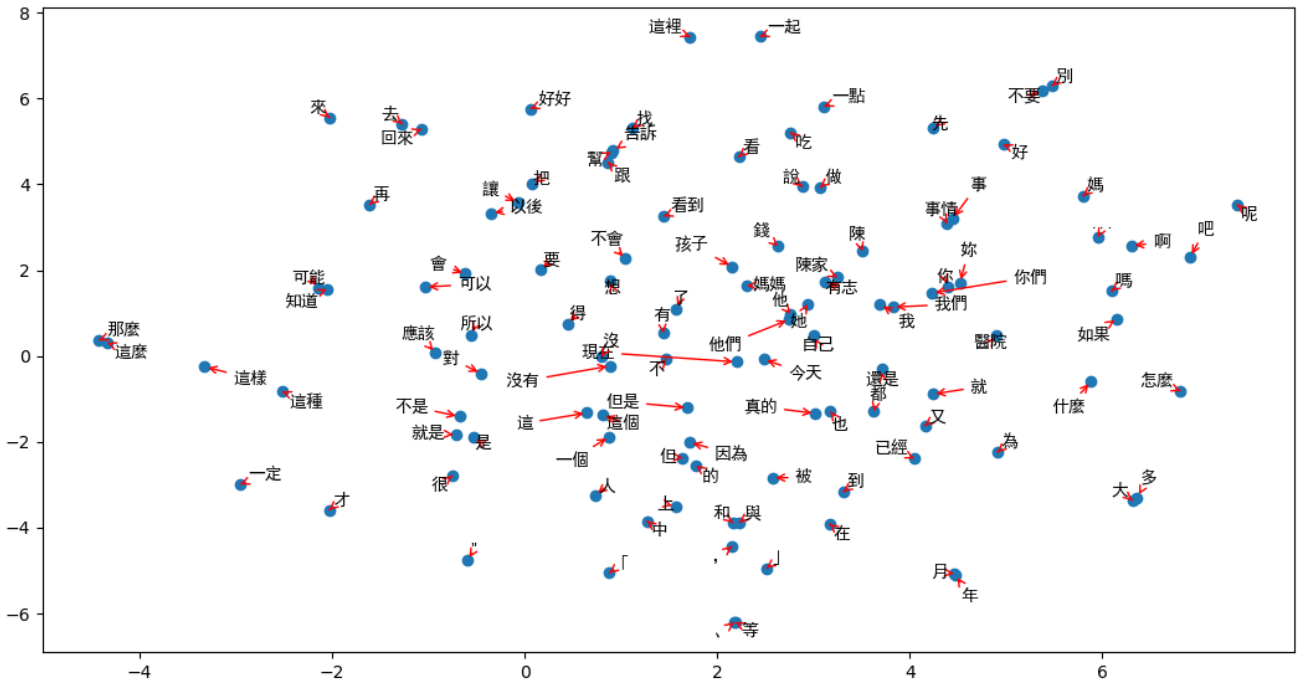
B. Visualization of Chinese word embedding

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。

使用的是 gensim.models.Word2Vec。調整參數：

size	決定向量要幾維
window	訓練的時候會看詞附近多遠的前後文
iter	訓練要跑幾個 iteration

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。

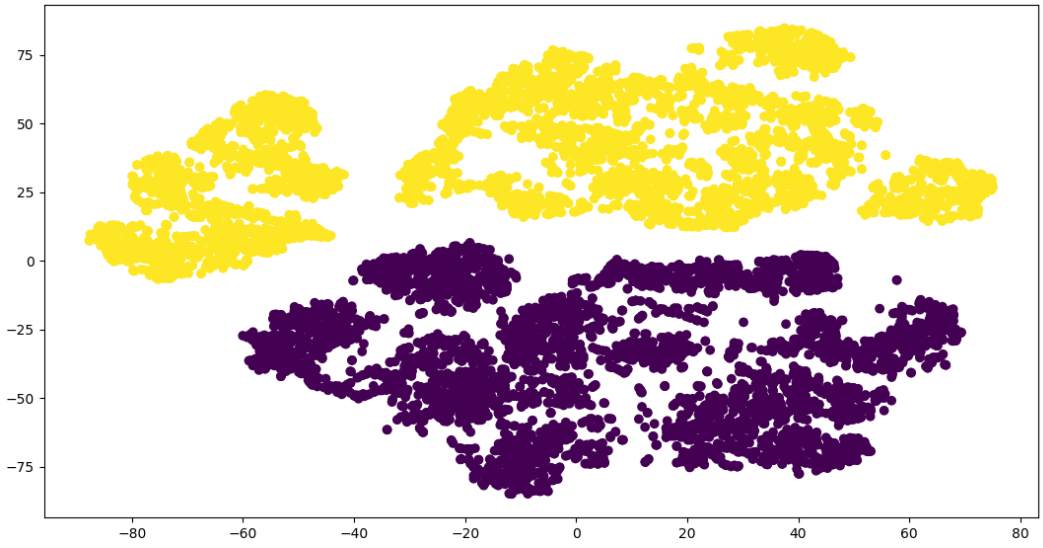
1. 語意或是文法詞性相似的詞會在一起，例如：左邊的那麼、這麼; 右邊的啊、吧、呢; 中間偏右的你、你們、我、我們等。
2. 我找到幾組有相對性的詞：「要 vs 不要」、「會 vs 不會」、「有 vs 沒有」、「是 vs 不是」。前兩組都是否定在右邊肯定在左邊，距離滿遠，但是後兩組的兩個詞沒有分開，聚在一起。

C. Image clustering

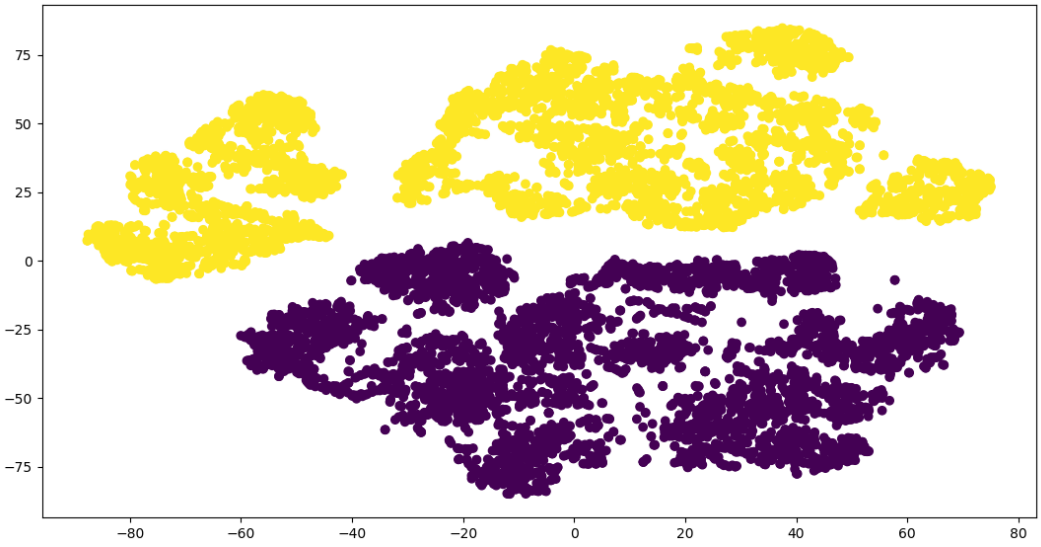
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

法 1：用 autoencoder 降至 32 維後，KMeans cluster
F1 score on Kaggle: 1.0
法 2：用 PCA 降至 32 維後，KMeans cluster
F1 score on Kaggle: 0.03
autoencoder 表現較好

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。



cluster 的結果完全和答案一樣，把兩個 dataset 分開來了。