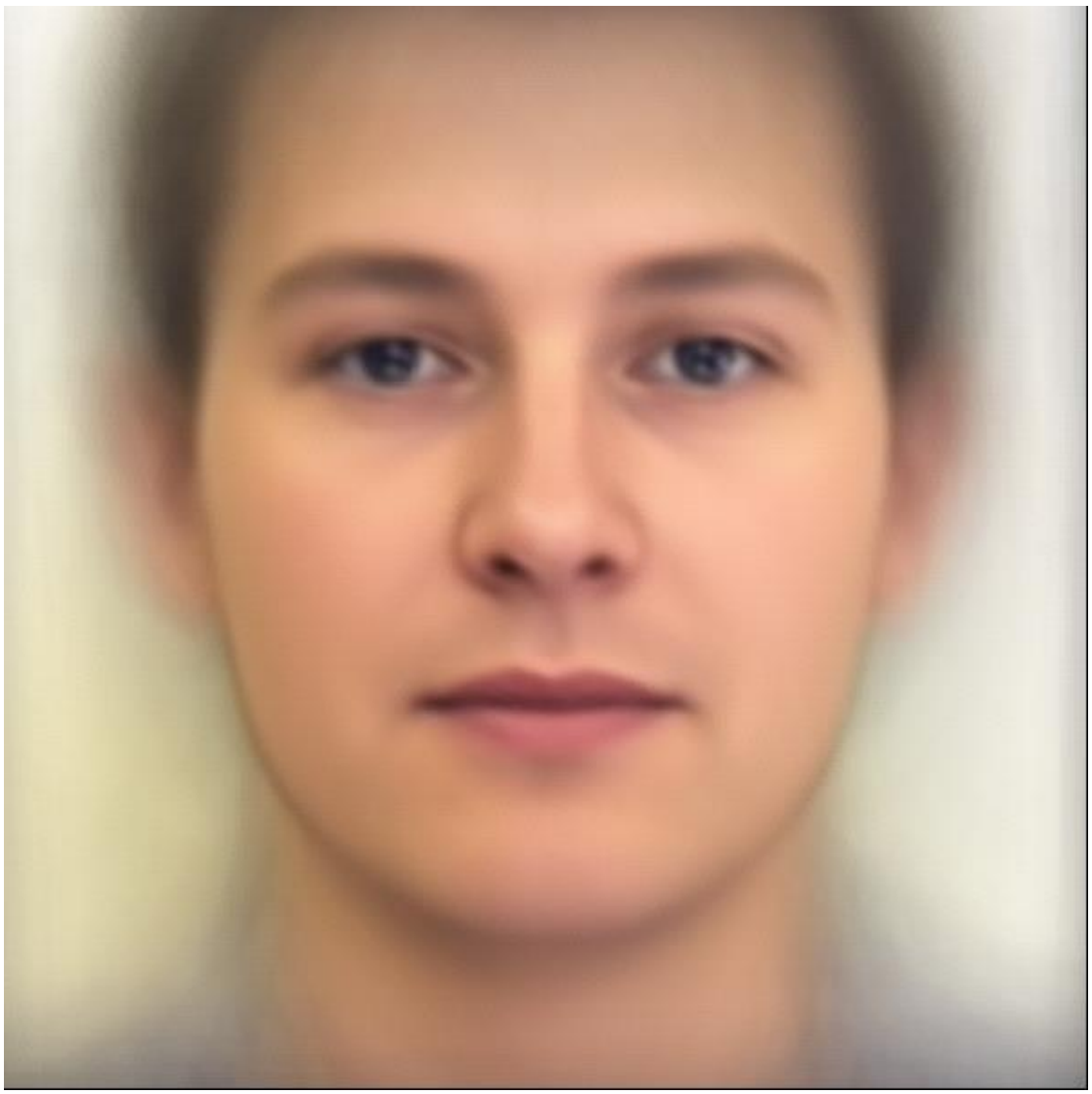
學號：r06942095 系級： 電信碩一 姓名：劉翔瑜

- 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。


- (.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。

前四大的比重分別占 4.1% , 2.9% , 2.4% , 2.2%

- 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。

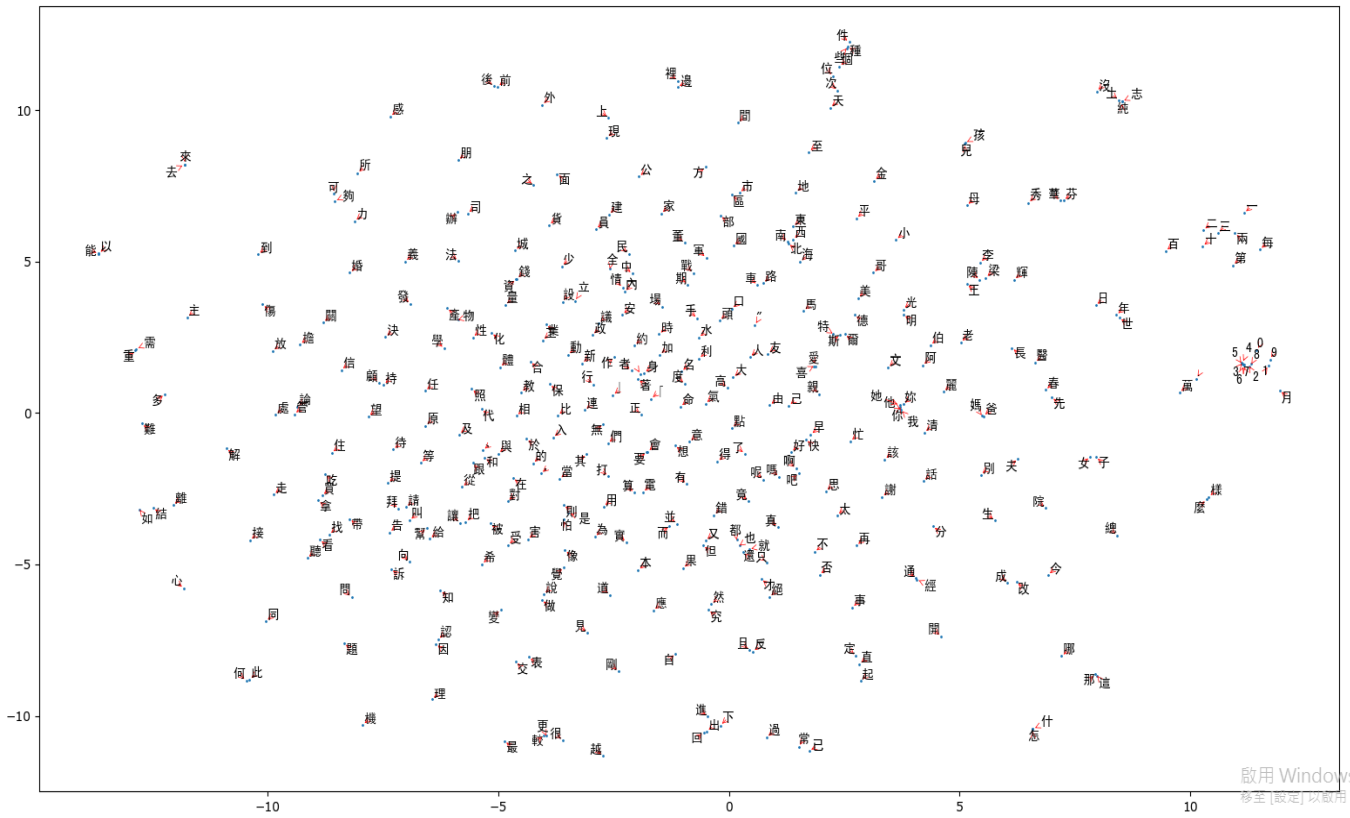
我使用 Word2Vec 來訓練 word embedding，其中

Word2Vec(data, size=150, window=3, min_count=3500, workers=4)

size 為 embedding 輸出的維度，window=3 即為左右 3 個單字內的單字皆會被考慮，min_count=3500 即為出現超過 3500 次的單字才會列入考慮，workers=4 代表同時使用 4 個線程處理 word

embedding(加速用)

- (.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。

由圖可知，wordembedding 會將具有高度相關的字分在一起，如右邊的 0.1.2.3.4...等數字明顯的被分在同一區，還有一些姓氏(李,陳,王..等等)也會被分在同一區，代表 word embedding 可以有效地將詞的特性訓練出來

- 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)

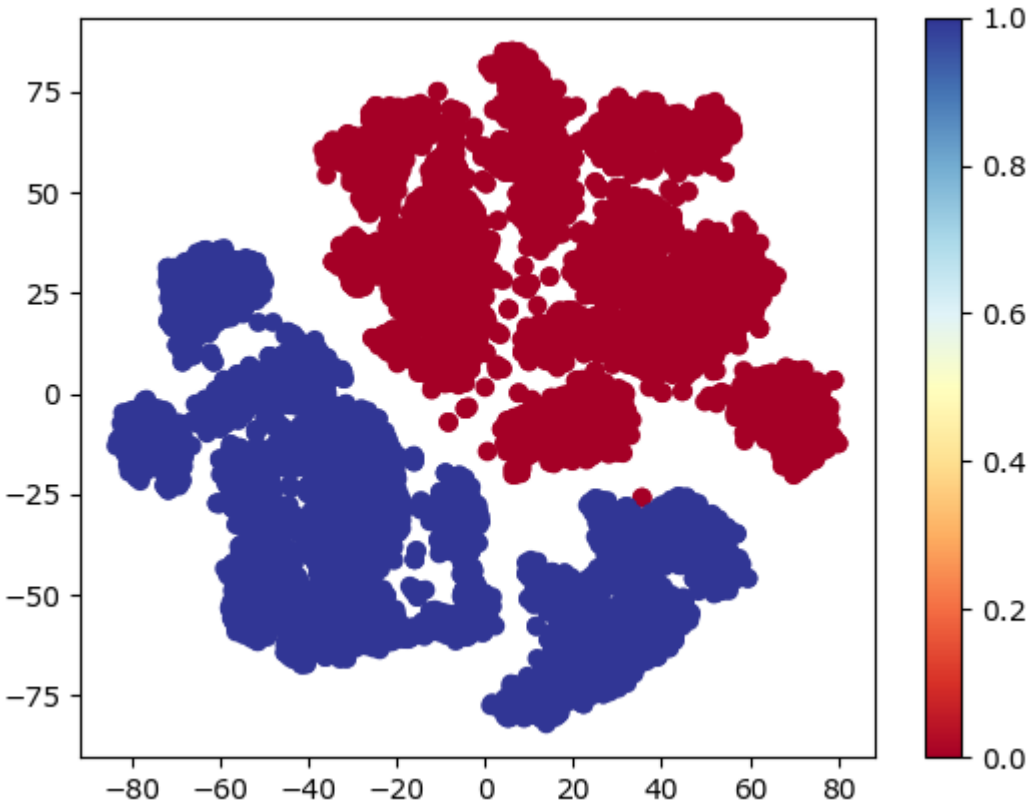
A: 第一次使用 auto encoder :先用 128 64 32 64 128 output_dim 的 DNN 來訓練，練好之後 32 那層的 output 為 auto_encoder 的輸出，

再使用 kmean(n_cluster=6)，最後在 kaggle 上的分數為 0.63(意外 的是，若 n_cluster=2，分數為 0.37)

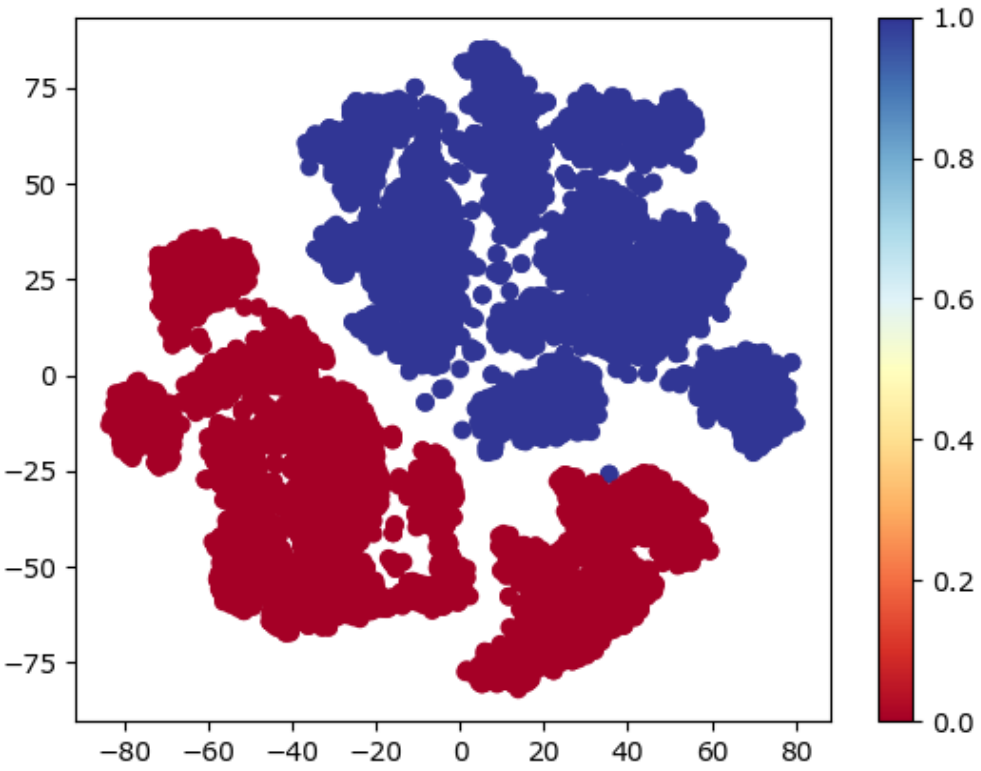
另一次是用 PCA(n_components300,svd_solver='full',whiten=True,copy=True)

練好後再使用 kmean(n_cluster=2)，kaggle 上的分數為 0.99，可 得 PCA 在此類型的分類問題上也許較有效

- (.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。



由圖可知，預測的結果幾乎一模一樣