A. 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。









 (.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。
我挑選 index = 66, 67, 77, 87 重建



Index=66



index=67





Index=77

index=87

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。

四大: (1) 4.1% (2)3.0% (3)2.4% (4)2.2%

B. Visualization of Chinese word embedding

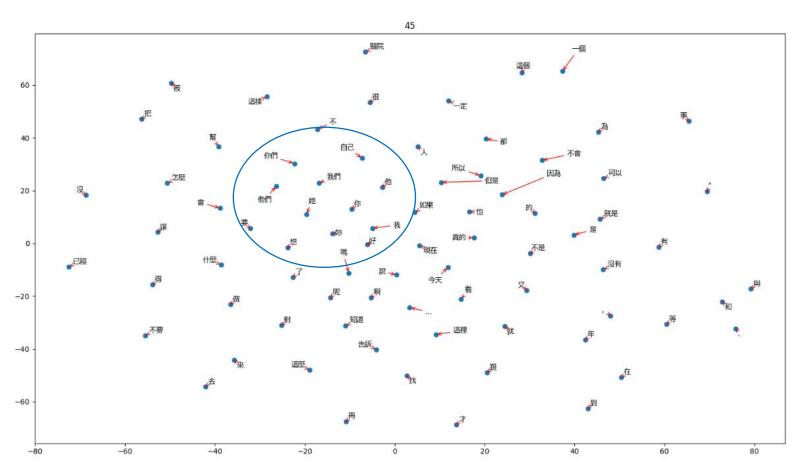
1. (.5%) 請說明你用哪一個 word2vec 套件,並針對你有調整的參數說明那個參數的意義。

我使用 gensim, Size=250 代表把每個詞都轉成 250 維

model = word2vec.Word2Vec(sentences, size=250)

而在 visualize 時候只針對出現 K=6000 次以上的詞

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



3. (.5%) 請討論你從 visualization 的結果觀察到什麼。 我發現我用藍色圈起來的部分,裡面有你、你們、他們、 我、我們等等的詞,在二維空間上的距離都蠻近的! 和、與 相似詞也在隔壁而已,

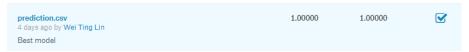
C. Image clustering

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)

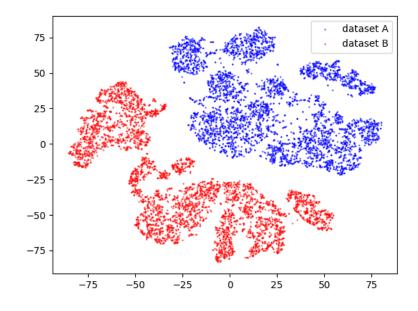
我一開始使用 pca 降到 32 維,然後使用 Kmeans 分群,丟到 kaggle 後結果很差。



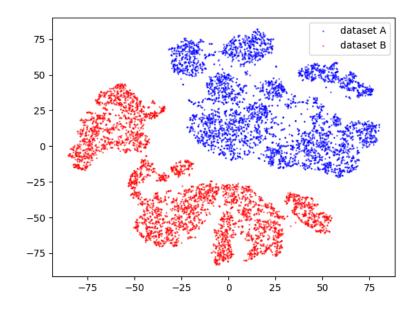
後來改用多層 autoencoder,train 兩百個 epoch 後得到我最好的 model



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



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。



比較結果發現一樣,於是印出全部的 label 看看。發現前五千個都被分類到 1,後五千個都分類到 0。因此完全預測正確!