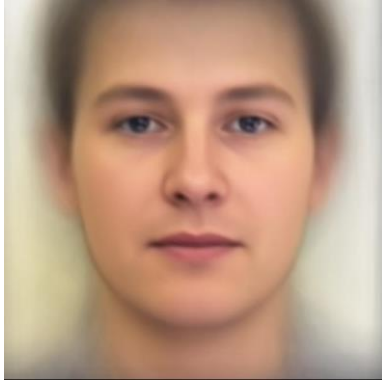


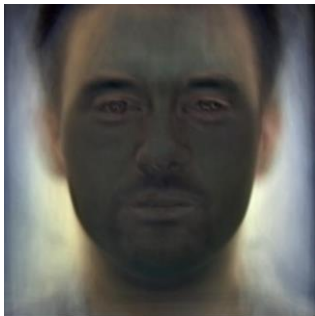
A. PCA of colored faces (Collaborators: R06942010 蘇建翰)

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。

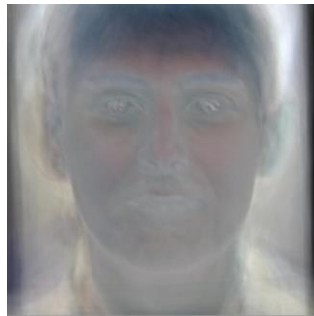
Eigenvector1:



Eigenvector2:











Eigenvector3:



Eigenvector4:



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。

	37.jpg	137.jpg	237.jpg	337.jpg
original pictures				
4 eigenvec to reconstruct				

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

Ans: 分別是 4.1%, 2.9%, 2.4%, 2.2%。

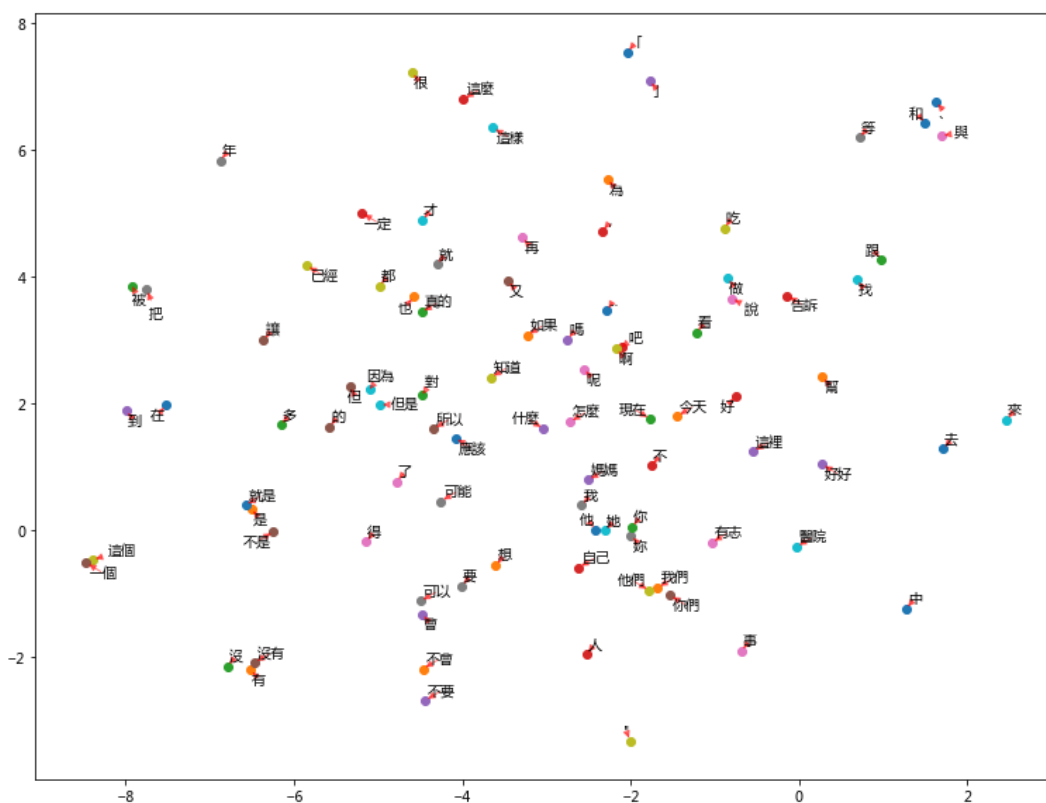
B. Visualization of Chinese word embedding

(Collaborators: R06942010 蘇建翰)

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。

Ans:我所使用的套件為 gensim 的 word2vec，所配置的有 size(每一個詞要用幾個維度的向量表示，這裡設 128)，window(每一個字要往前往後看幾個字，這裡設 5)，min_count(出現次數低於 min_count 將不會被考慮，這裡設 1)。

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。

Ans:

Word2vec 會盡量地把概念類似的詞給聚集在一起，可以看到比方說(與 和 、)這種連接詞，(你、妳、我、他、她、自己)，(我們、你們、他們)這種代名詞，(怎麼、什麼、呢)疑問詞，(有、沒、沒有)諸如此類的被歸類在一起，表示他們在句子裡面的腳色關係非常相近。

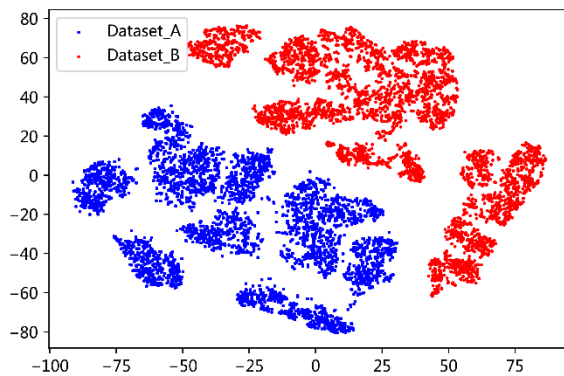
C. Image clustering (Collaborators: R06942010 蘇建翰)

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)

Ans: 使用了兩種分類方法，一種是使用 PCA 將圖片降為 50 維，接著用 tsne 降到 10 維做 kmeans，kaggle 上面成績為 0.03288(單純對圖片做降維處理似乎對於圖片的分類有困難)。另一種方式，利用 autoencoder，疊了 784、256、32、256、784 五層，利用中間那層 32 維的結果作 kmeans，kaggle 上面分數為 0.99930。

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

Ans: 利用第三題 autoencoder 的 model 降到 32 維並且分類，再用 tsne 把圖片從 32 維降到 2 維畫出圖形。



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。

Ans: 計算了自己預測的 label 共 5000 個 Dataset_A 和 5000 個 Dataset_B，此題所給予也是 5000 個 Dataset_A 和 5000 個 Dataset_B。從兩張圖片看來，預測的結果看起來應該和正確答案是一致的。

