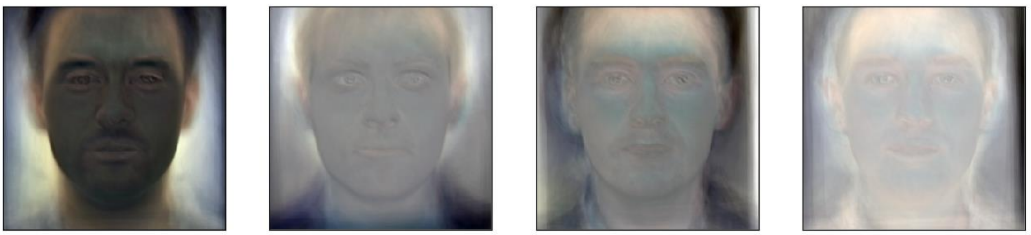
學號：R06922152 系級： 資工碩一 姓名：袁晟峻

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 (explained variance ratio)，請四捨五入到小數點後一位。

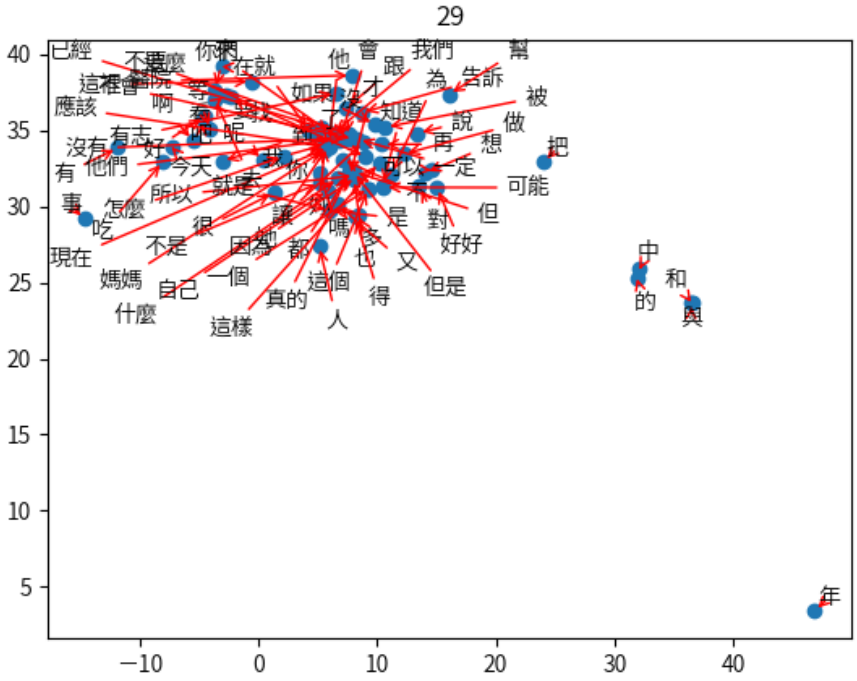
4.2%、3.0%、2.4%、2.2%

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。

Word2vec 套件使用 gensim，調整了 min_count 和 window ，min_count 是決定重複出現次數不少於多少才拿去訓練，window 決定 model 在 training 過程中會參考前後各多少詞。

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。

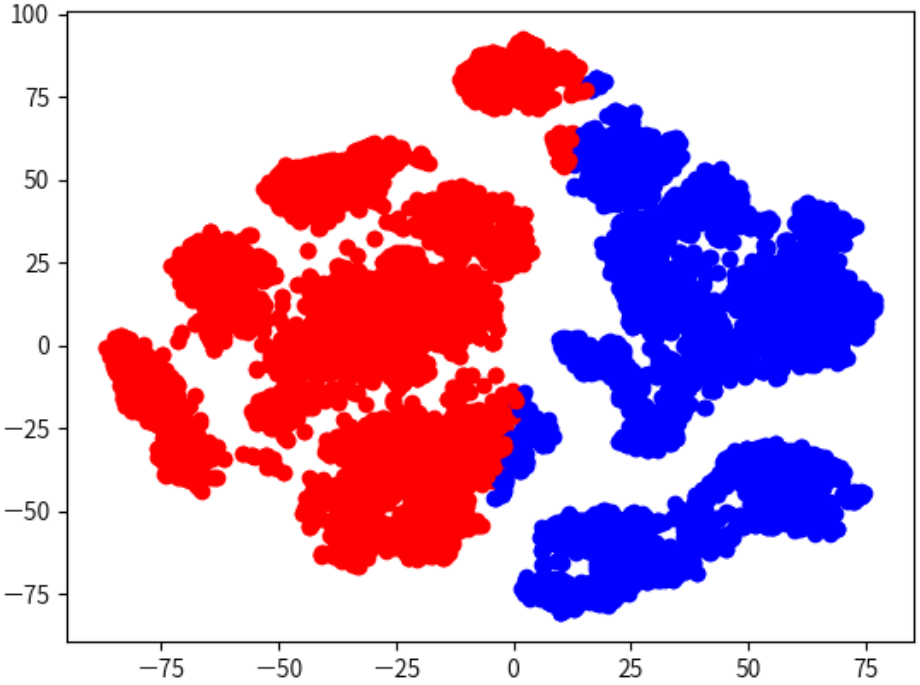
年被分得很遠，沒有跟他相近的詞，中、和、的、與等連接詞有被分出來，剩下的則幾乎聚成一團，中間還有一點分界，但看不太出來分界的標準以及每團之間為什麼相似。

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)

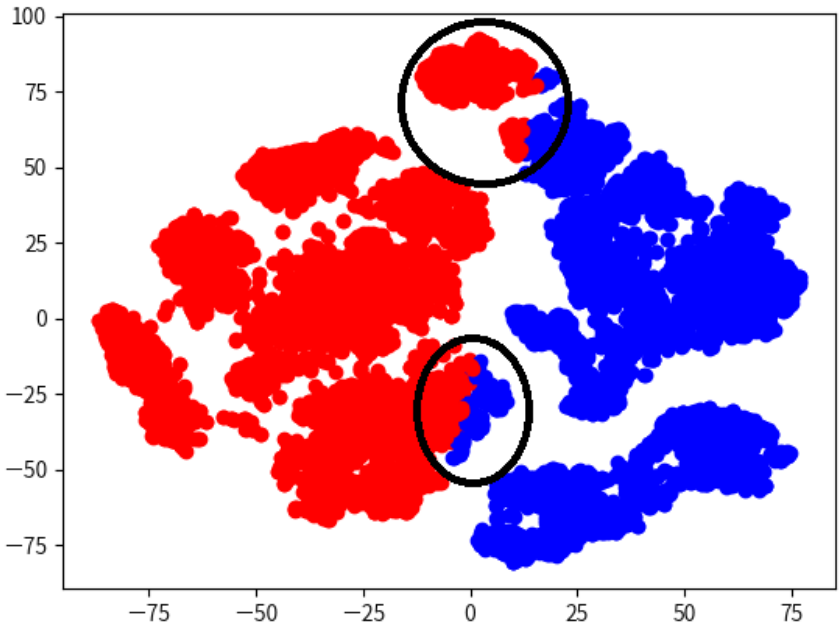
1. 使用 auto-encoder + Kmeans：降維到 128 維後用 Kmeans 分群，在 kaggle public 上得到 1.0 的分數。
2. 使用 TSNE + Kmeans：直接降到二維後用 Kmeans 分群，但是效果很差，在 kaggle public 上只得到 0.025 的分數，而且花的時間還很久。

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。

比較之後發現預測錯誤的有 647 個，集中在下圖用黑色圓圈框出來的部分，由於數量有點多，如果分別標出是第幾筆 data 錯誤整個畫面會變得很亂，所以就用框出來代替。


100
75
50
25
0
−25
−50
−75
−75
−50
−25
0
25
50
75