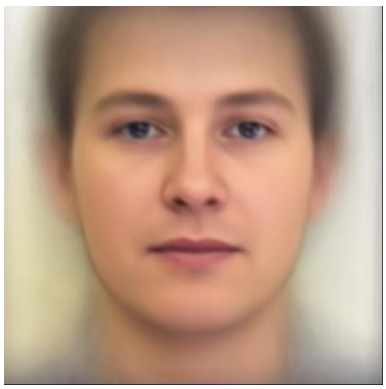
學號：B03901145 系級： 電機四 姓名：郭恆成

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。

依序是最大的 Eigenface 到第四大的 Eigenface。

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。

上排原圖，下排是用前四大 eigenface reconstruct 的結果。

5.jpg 82.jpg 172.jpg 217.jpg

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

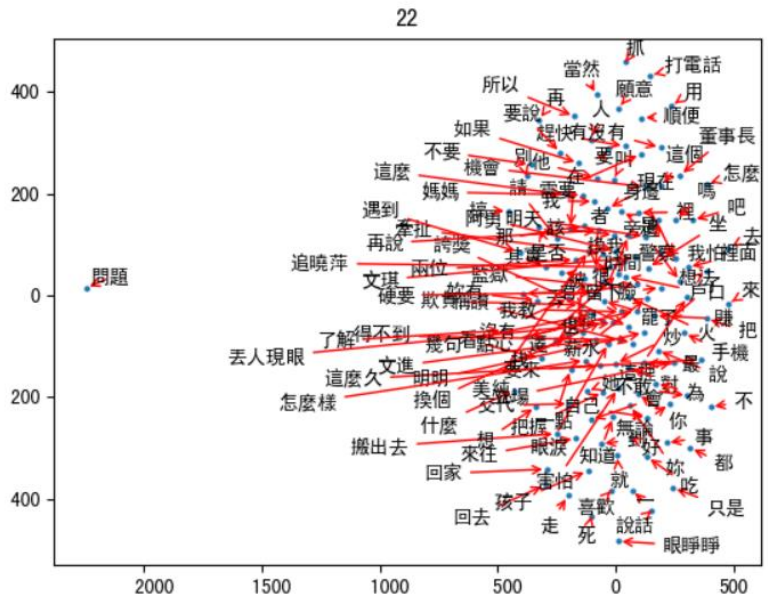
B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

使用 gensim 的 word2vec。然後調整 word vector 的 size 與 window 與 minimum count，word vector 的 size 越大理論上可以包含的資訊越多，但相對的運算量也越大，且 performance 不保證提升，故用適當的 size 保證 performance，且不會耗盡運算資源。

而 minimum count 的話是取至少出現一定次數的 word 較具有代表性，不常出現的字視為雜訊。

Windows 則是要考慮相鄰距離多少距離的字之間的關係。

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。

從圖片看不出來什麼特別的結果，我猜是因為用的 training data 是對話而非文章，所以詞語間的相對關係沒這麼明顯而且因為為了在圖片上方便觀察，只取

了 150 個左右的字，沒有觀察到有特別關聯的字除了幾個人名例如美純、文琪、追曉明等，大致成同一個方向。

C. Image clustering

C.1. (.5%) 請比較至少兩種不同的 feat

ure 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)

方法一:

利用 DNN 的 autoencoder 把 training data 降成 32 維的 feature，再利用 Kmeans 把降維的 feature 分成兩類。

方法二:

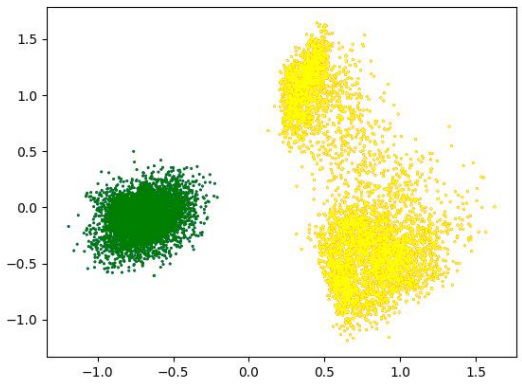
先用 PCA 降到 128 維，再利用 Kmeans 把 feature 分成兩類。

	Public score	Private score
DNN + KMeans	0.99925	0.99934
PCA + KMeans	0.03024	0.03048

可以看到在這個 case 中 DNN 降為可以保留更多 data 中的資訊，有效的去做分類。

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

綠色是 predict 屬於 classes 0, 黃色是屬於 classes 1。然後用



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。

可以看到用正確的的 label 畫出來的圖根原本的答案一模一樣，有鑑於在 kaggle 上的準確率維 0.99 以上，所以 predict 出來的正確率很高。