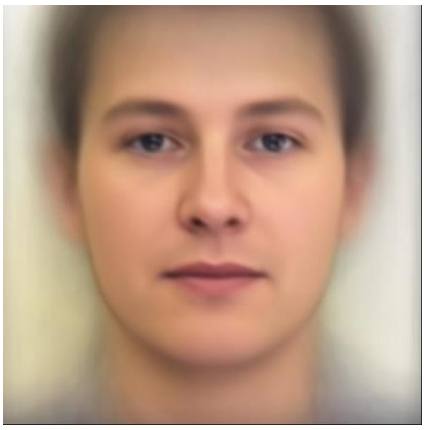
學號：R06944049 系級： 網媒碩一 姓名：黃敬庭

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。

分別是第 7,41,77,414 張圖片

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。

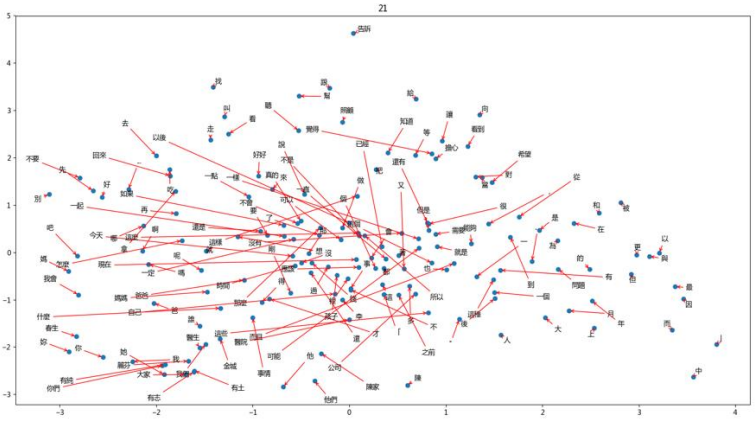
4.1	(0.0414747807674)
3.0	(0.0295084913796)
2.4	(0.0238935983997)
2.2	(0.0220932885439)

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。

我使用 Gensim 裡面的 Word2Vec 套件,我所調整的參數有 min_count 與 size, min_count 代表在整個 dataset 中要出現幾次才會將它放進 word2vec 的 model 中(最後定為 3000),而 size 則代表要將這些詞投射到幾維度的空間中(最後選用 300)

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。

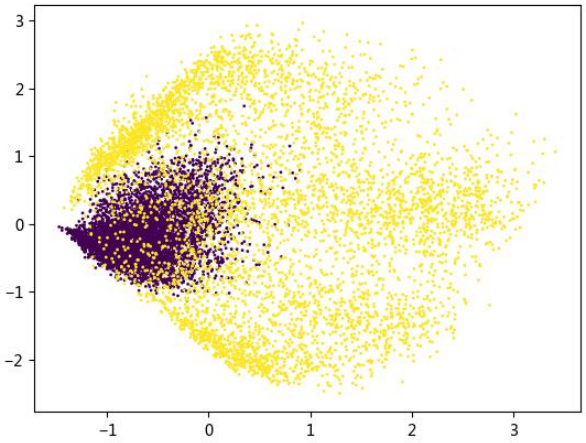
最終的結果中其實可以看到左下角聚集了一些主詞(我們你們爸媽等)在那裡,而上半部那邊多半是動詞(找叫聽等),比較靠近的字通常在詞性上比較接近,雖然經過降維會有一定程度的影響,但大致可以看到 word2vec model 有抓到各詞性間的關係

C. Image clustering

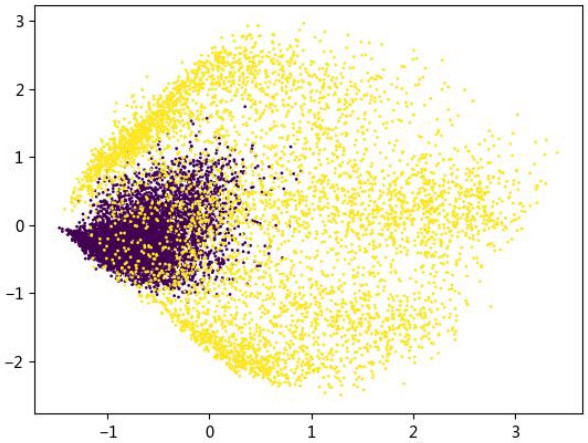
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

我最後的模型是用 pca(有先對資料做 whiten 的前處理) 降至 300 維後再利用 kmeans(cluster = 2)進行預測,最後達到 0.99875 的 f1 score,我另外嘗試了降維之後在計算 cosine similarity 的方式並將大於 0 的歸為同一類,不過在 kaggle 上並沒有很高只來到 0.12 左右,我猜想是因為相似度對給定的門檻會非常敏感,如果沒有經過調校可能會有效能上的損失,但 kmeans 經由距離來計算並強制分為兩類,只要一開始的降維不要讓資料損失的太過誇張都會有還行的成果。

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。



黃色的為一類紫色為一類,直接取其 300 維的前 2 維作為 x,y 軸,預測的與標準答案並沒有差太多是因為本來 model(至少在 kaggle 上面看來)應該就不錯,可以看到很明顯的辨別能力