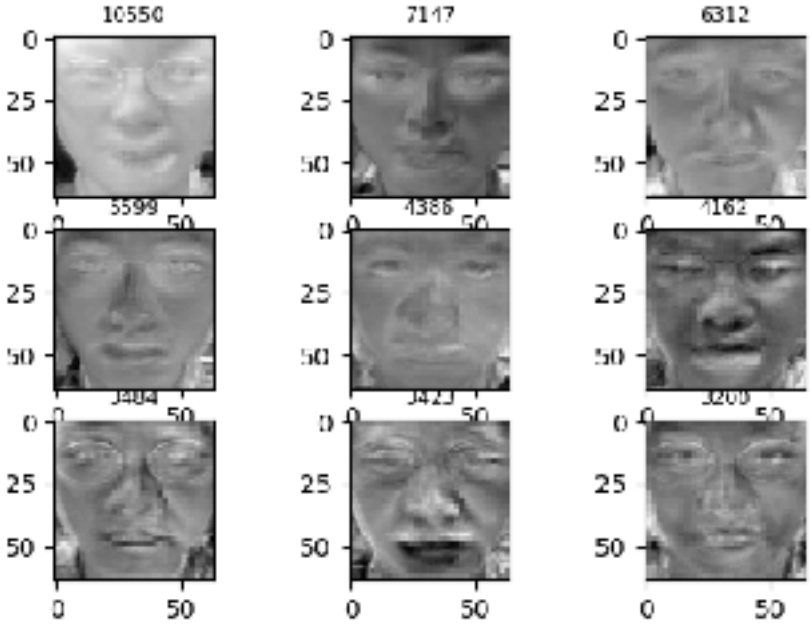
學號：B02901093 系級： 電機四 姓名：吳岳

1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

答：(左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)

註：右圖上方的值是對應的eigenvalue



1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):

答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)

左邊是原圖，右邊是Reconstruct後的圖

1.3. Dataset 中前 10 個人的前 10 張照片投影到 top k eigenfaces 時就可以達到 < 1% 的 reconstruction error.

答：k = 59

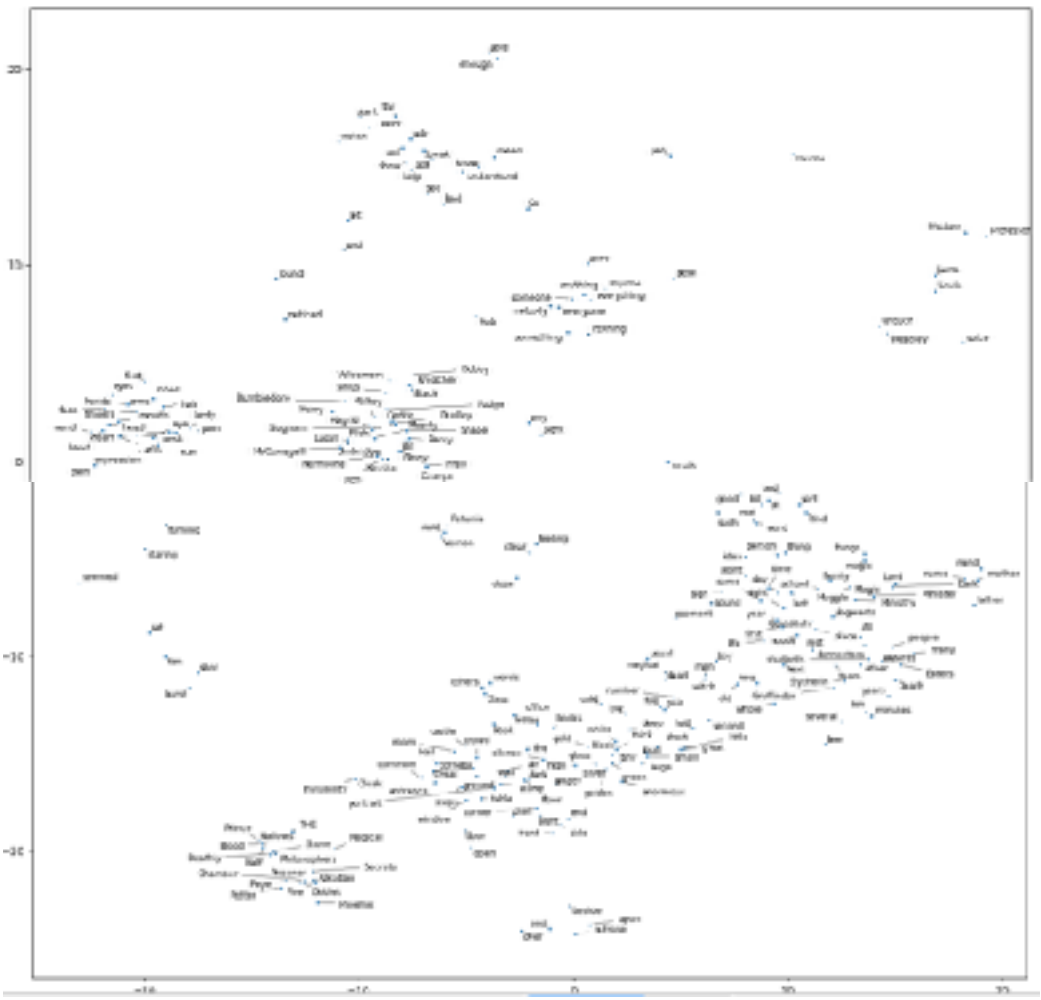
2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：size=100, alpha = 0.025, window = 5, min_count = 5, sample = 0.001

Size	設定Word Vector的大小	100
Alpha	起始learning rate	0.025
Window	設定word 跟 word之間的max skip length	5
Min-Count	不考慮出現次數小於多少次的單字(default = 5)	5
Sample	設定word的出現門檻，頻率高的字會隨機down-sample	0.001

2.2. 將 word2vec 的結果投影到 2 維的圖:

答：



2.3. 從上題視覺化的圖中觀察到了什麼？

答：

在視覺化的圖裡面，不同分群會有不同的特性。舉例來說，左邊是專有名詞、中間是動詞、右邊則是身體相關的名詞。

另外，透過幾次的測試，我發現每次跑出來高頻群的相對位置都差不多(例如說專有名詞、動詞都差不多分布在同一個位置)，然而，每此跑群組內詞的分佈還是有差異(如右圖)

3.1. 請詳加解釋你估計原始維度的原理、合理性，這方法的通用性如何？

答： 在原始的維度未給定之下，我必須去估計不同維度下的特徵。在做出結果以前，我先「猜測」原始維度跟投影到100維後的變異數應該呈正相關。舉例來說，原始維度1維投影到100維，其他99維應該不會有太大的變化，然而原始維度60維投影到100維，僅有其他40維不會有太大變化。故此，我就先根據這樣的推論算出200個dataset的標準差，再用Kmeans分成60個群，依據標準差的大小設定群的Id。

上傳到Kaggle之後的MSE介於11.5~13之間(因為Kmean種子不同，跑完的結果不會都一樣)，就成果而言是合理的，推論在跟其他助教討論之後也沒有太大的盲點。然而，這樣的方法並沒有太大的通用性，畢竟我一開始就知道原始維度的分佈介於1~60之間，若一開始不知道維度的分佈，Kmeans的方法就沒有太多的可預測性。

3.2. 將你的方法做在 hand rotation sequence datatset 上得到什麼結果？合理嗎？請討論之。

答：首先，我依然是用前面kmeans對映變異數的原理去推測維度，然而，這邊碰到的大問題是**我不知道它維度分布的區間**。助教在投影片裡面說大約3~4維，我就用4群kmeans去做，得到一個1~4維的預測值。

然而，這樣的做法並不合理，假設它有其他的分佈區間呢？這樣我頂多**能比較維度高低的排序，但精確的維度是多少就無法預測**。(例如說給定100個群在200個不知維度區間的dataset裡面，可推論第50群dataset維度比第1群的維度大，但第50群的dataset無法推論是50維)