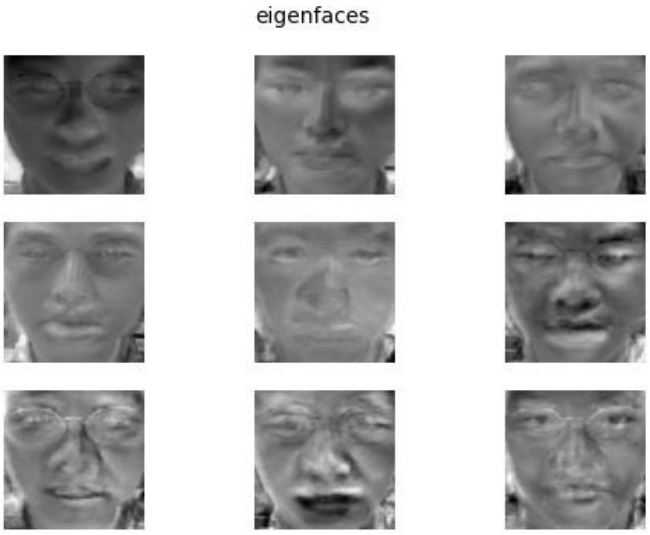
學號：R04945022 系級： 生醫電資碩二 姓名：張君澤

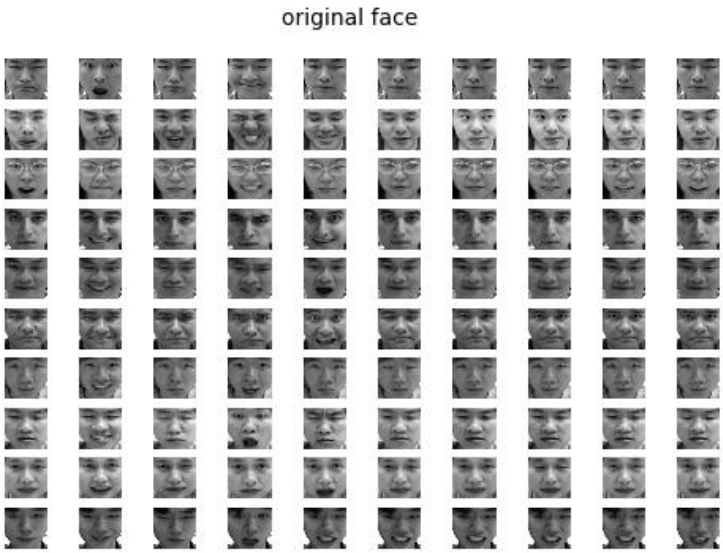
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:
答： (左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)
前十人的前十張照片 average face:

前九個 eigenface 圖:

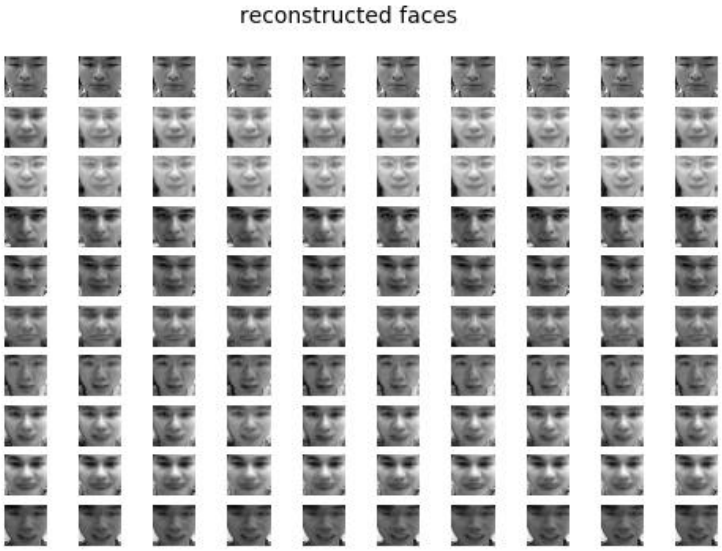


1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):
答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)

原始圖片:



Reconstruct 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：

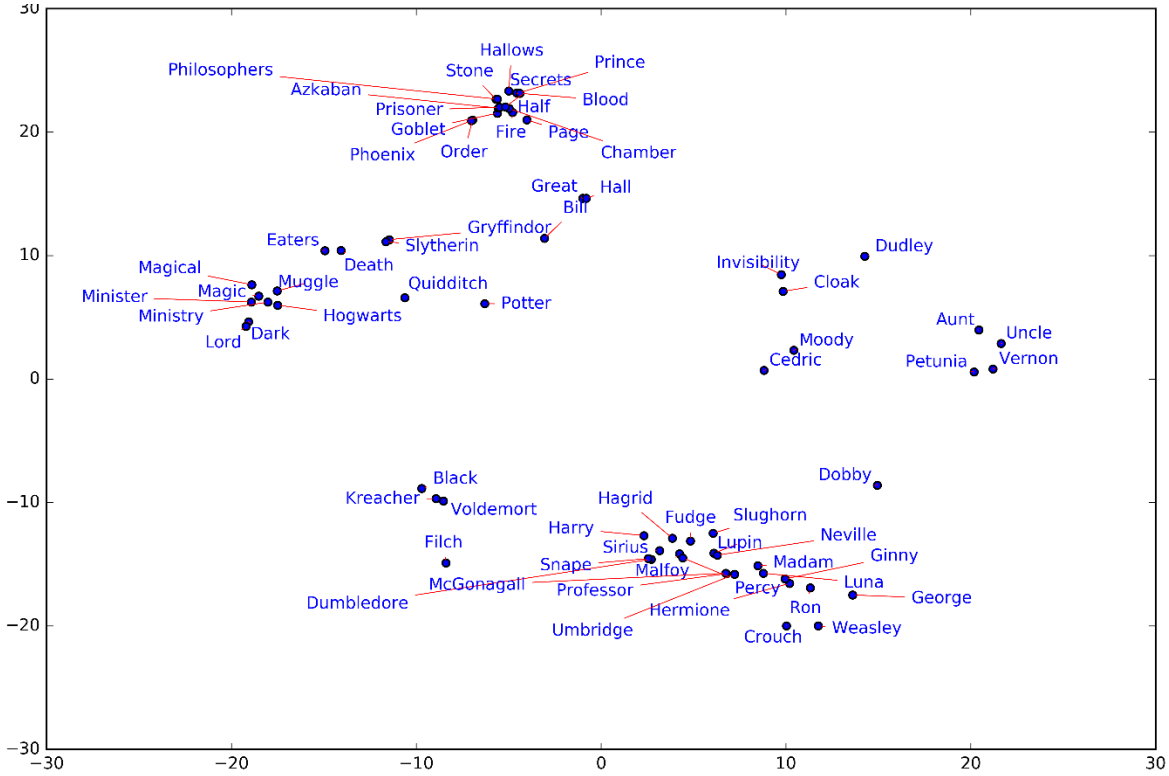
達到 rmse 小於 1%時，K = 60。

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:
答：
在我使用的 word2vec 參數中有 threads=4, window=10, min_count=5, verbose=True

其中 min_count 會忽略整篇文章中出現次數小於 5 次的字，windows 的用意是設定字與字之間有關連性的最大距離，簡單來說有可能是一句的話的平均詞彙數，我將 windows 設為 6 與 10 做比較，發現 windows 為 10 的時候 word2vec 比較有邏輯性。而 verbose 只是記錄詳細的輸出結果，影響不大。
而 model.vocab 裡面放的就是類似 1-of-n encoding 的矩陣，用以儲存每個字的陣列。

2.2. 將 word2vec 的結果投影到 2 維的圖:
答：

圖如下



2.3. 從上題視覺化的圖中觀察到了什麼？
答：
視覺化的圖裡面可以發現到和學校有關的人物會聚集在一起，例如馬份、海格、Weasley, Snape 等校園常出現的好人與壞人。而魔法行政部門則跟舉辦的 Quiddiitch 比較有關，魔法部門跟 Slytherin 的關係也比跟 Gryffindor 的關係更密切。值得一提的是 Harry 這個字出現在學校這個 Cluster，而 Potter 出現在靠近魔法部門那個 Cluster 顯示同儕之間是直接叫名子，而在正式場合或者是面對史萊哲林的壞蛋的時候可能都被稱呼 Mr. Potter。

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：

原本使用 sklearn 的 mds 去進行降維，但 kaggle 上最好的結果在 0.3 左右，因此果斷放棄。後來發現了一個 hub_toolbox.IntrinsicDim 的套件，基本上它的原理應該是類似 nearest neighborhood, 我起始把 k2 的 neighborhood 用預設的 12 效果並不理想，後來把 neighbor 調成 15，效果也不顯著。把 result 列印出來後發現最大的維度只有 25，跟助教所提供的最大維度是 60 的假設相差太遠，推測是 nearest neighbor 在高維度的降維會有更大的誤差。因此我以線性分配的方式將最大的結果放大到 60 維度，大於 16 的維度，以線性的方式對應到更大的維度，小於 16 的維度不調整。

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：

目前在 kaggle 上獲得 0.112 的 error rate。在助教時間後，我也使用了 sklearn 的 nearest neighbor 中的 ball tree, 一樣是取前 5566 比資料來做推測，結果落在 0.135 左右，並沒有比較好(可能是我參數調教不好，或者點取不夠多)，因此可以推測 nearest neighbor 的演算法對於降維準確率的效果比 PCA 還要來得好一些。