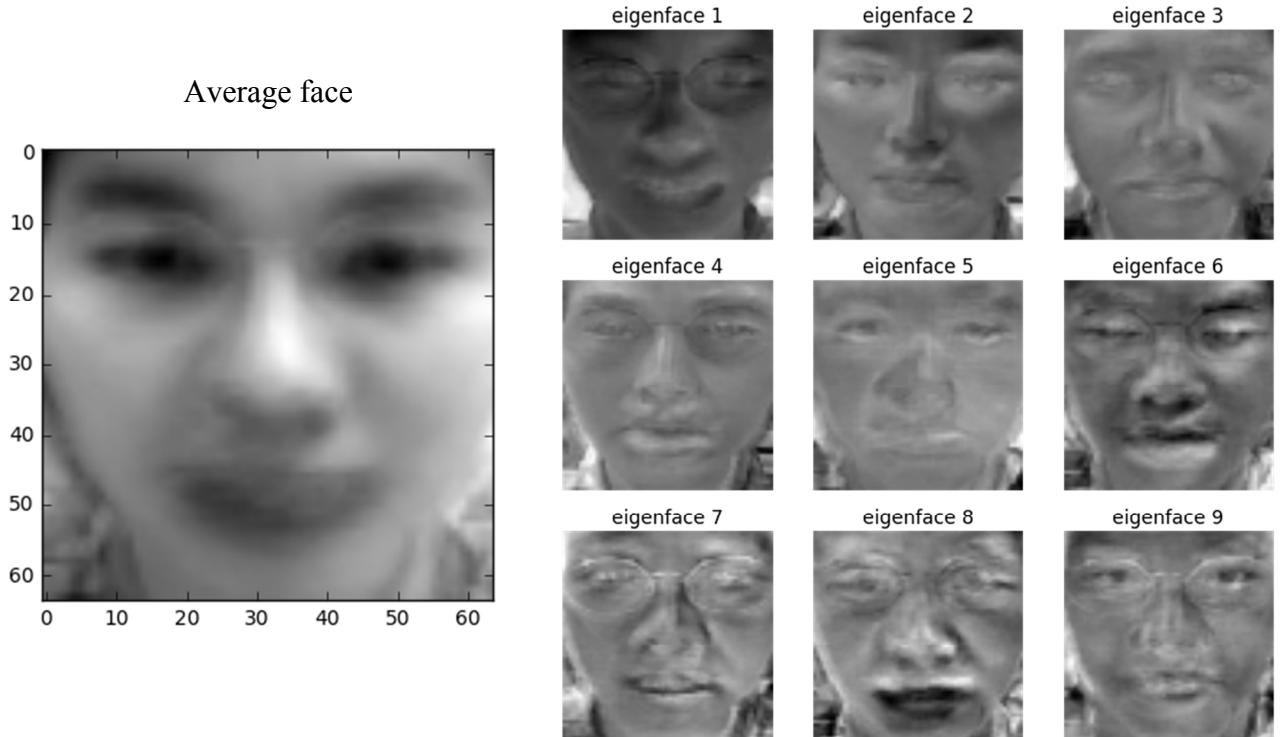


學號：B03901076 系級：電機三 姓名：林尚謙

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)



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)



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：(回答 k 是多少)

在 $k \geq 59$ 後可使 RMSE 小於 1% (除以 256)。

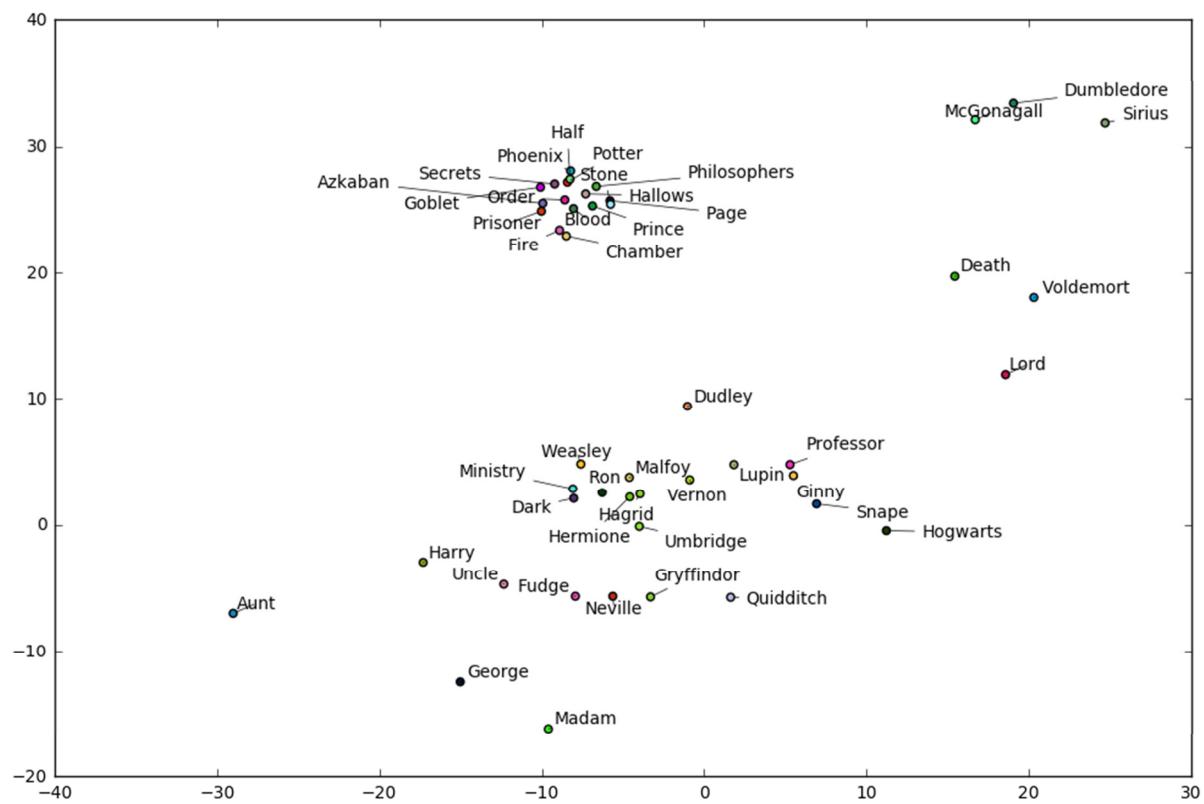
2.1. 使用 word2vec toolkit 的各個參數的值與其意義：

答：(參數=實作中設定的值)

1. cbow = 1 — 是否使用 Cbow(Continuous Bag-of-Words)，若 0 則使用 skip-gram 模型
 2. size = 1500 — 詞向量的維度，等同於一個詞具有的特徵個數
 3. min_count = 5 — 去掉出現次數少於此值的詞
 4. window = 6 — 上下文之間的窗口大小，等於一次看到的詞數量
 5. negative = 5 — Negative samples 的數量，可以減少一次更動的 weights 數量
 6. iter_ = 150000 — iteration 的次數

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：

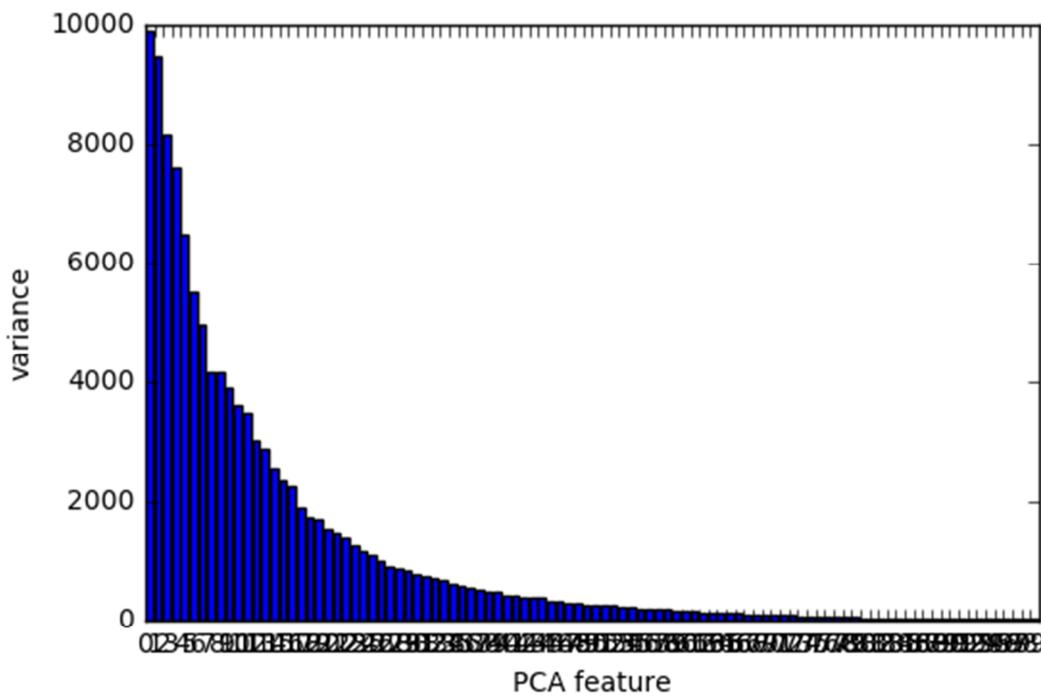
從視覺化的圖可以觀察到，詞之間大致分成較靠近的兩堆，彼此之中較靠近的詞會有某種程度上的關聯(考量原著劇情內容)，例如 Azkaban 和 Prisoner(阿茲卡班監獄)、Lupin、Snape 和 Professor(都是霍格華茲的教授)等等，代表在訓練出來的詞特徵上兩

者之間較為接近，但比較意外的是主角的名字 Harry 和 Potter 距離較遠，可能是剛好訓練出來的詞特徵所致。

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：

我是使用 PCA 來估計原始維度，藉由 scikit-learn 的 PCA tool，可以直接算出一組 data 的 PCA features，並且依照 variance 由大到小排列，而 variance 較大的幾個 feature 就是這筆 data 的原始維度，設定一個閾值並計算 variance 高於這個值的 features 個數，就是估計的原始維度。但這項作法的麻煩點在於對於不同的 data 會有不同的最佳閾值，要慢慢去調整，而且最終跑出來的結果效果也沒到很好(rmse 約 0.13)，好處是適用於很多不同種類的 data，並且執行的速度相當快。



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

答：