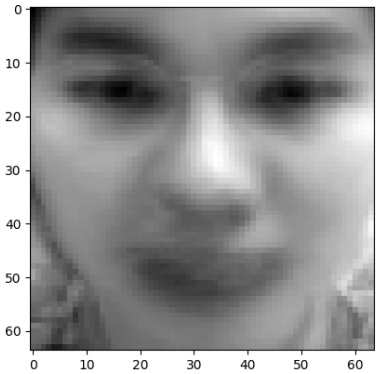
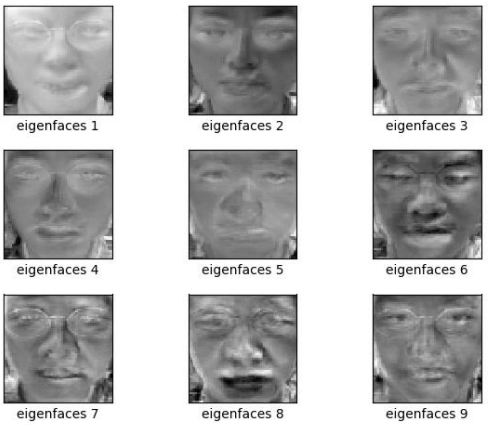
Machine Learning HW4

學號：R05943005 系級： 電子所碩一 姓名：呂丞勛

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：計算出來的結果(除以 255)，得到 k=60

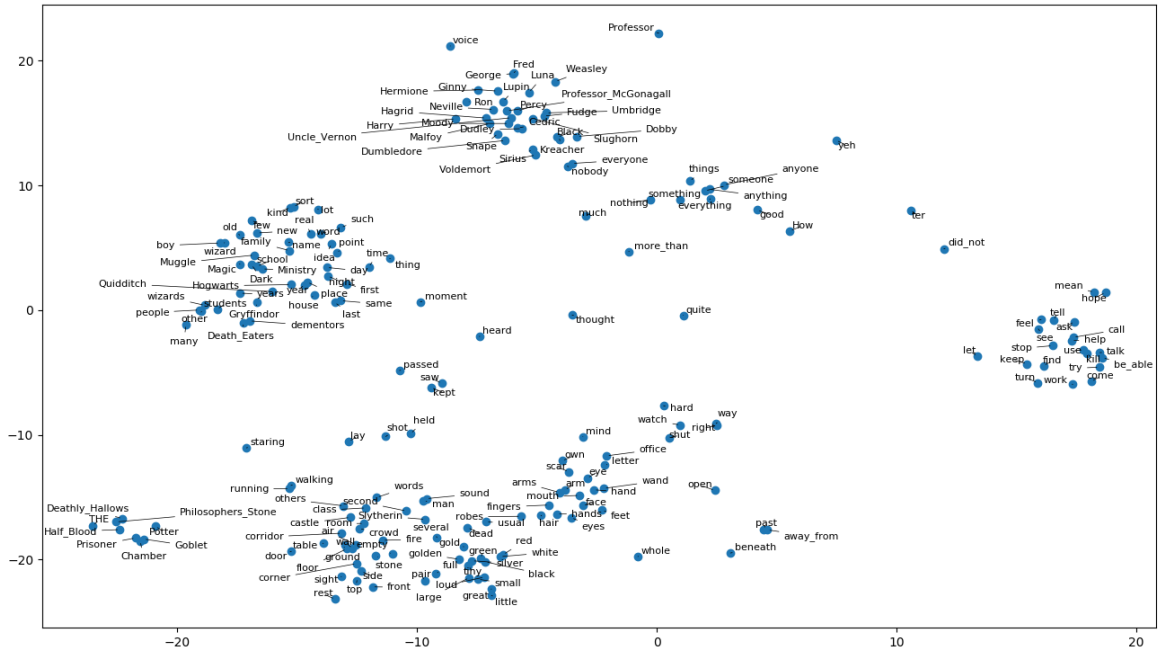
2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

word2vec: 第一個分別 input text，第二個為 output vector，size=100 表示 output 出來的 vector 為 100 維的向量，default 的模型為 skip-gram(模型有兩種 skip-gram & Cbow)，window 大小則用預設值(window 大小表示決定該單詞的時候需要往前往後看多少詞)。

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：

關注某些 clusters，可以發現一些在特定 clusters 中有共同的特性，像是最右邊的 cluster 中，大部分都是原形動詞，而中間下面的那一小群中，則是形容詞，最上面那一些則是有大寫字母或是名字，因此在 transformed 的過程中，還是可以保留一些特性，來組成 clusters。

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：

先利用 SVD 將所得到的資料做分解，可以得到 100 個 eigenvalues(此例子中所得 data 為 100 維)，又知道原始資料的維度在[1, 60]之中，因此將 data 用 eigenvalue 重建回來(選擇的 dimension 從 1 慢慢往上增加，最高到 60)，重建回來的訊號與原始訊號做 RMSE 的估計，並設定一個 threshold，當 dimension 增加到某個$\hat{d}$後，RMSE 變會低於所設定的

RMSE，此時該$\hat{d}$則為預估的 intrinsic dimension。利用 PCA 轉換，選出幾個重要的 eigenvalue 的確為 dimension reduction 的方式之一，但利用這種方式為線性轉換，如果需要非線性轉換可能會有些問題，另外，該方式的 RMSE 的 threshold 該如何選擇也會隨著 Data 的不同而有不一樣的大小，因此通用性可能不高。

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：若使用與 3.1 相同的 RMSE threshold，則會得到預估出來的 intrinsic dimension 為 30，然而有 paper 有拿相同的 dataset 來做估計，估計出來的結果大概是 3 維左右，可以知道在 3.1 使用的方法，非常的 data dependent，因此可能需要利用局部的維度來估計整體的維度，才是一個比較好的方式。