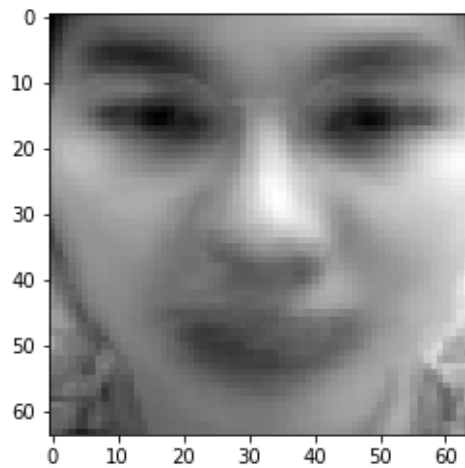
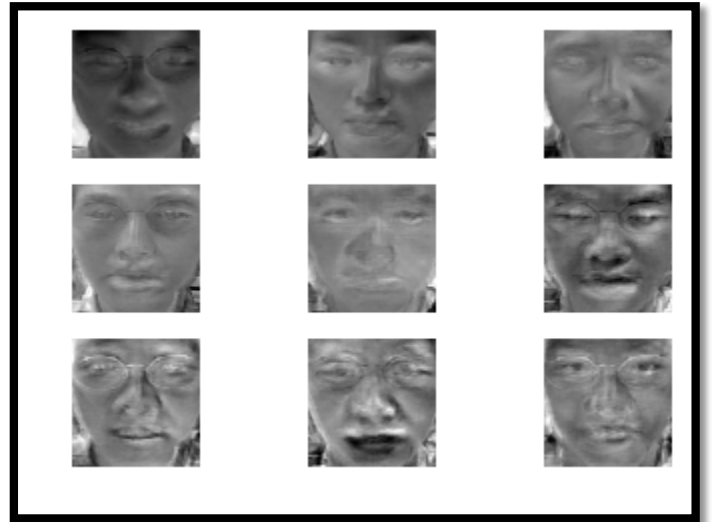


1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

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



圖一：平均臉



圖二：前 9 個 eigenfaces

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)



圖三：原始圖片



圖四：reconstruct 圖（用前 5 個 eigenfaces）

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）

一開始在實作的時候，因為讀進來的圖片自動 mapping 到 0-255，所以算出來的 **k** 很大，後來修正之後，算出來的 **k** 為 59

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

答：

我給的參數數值如下：

cbow=1, size=300, min_count=20, window=5, negative=3, iter_=200, alpha=0.1

cbow：如果是 0 的話，表示用的是 **word vector** 的總和，如果是 1 的話，表示用的是 **word vector** 的平均

size：特徵向量的維度

min_count：當字出現的次數小於這個次數，就忽略這個字不計

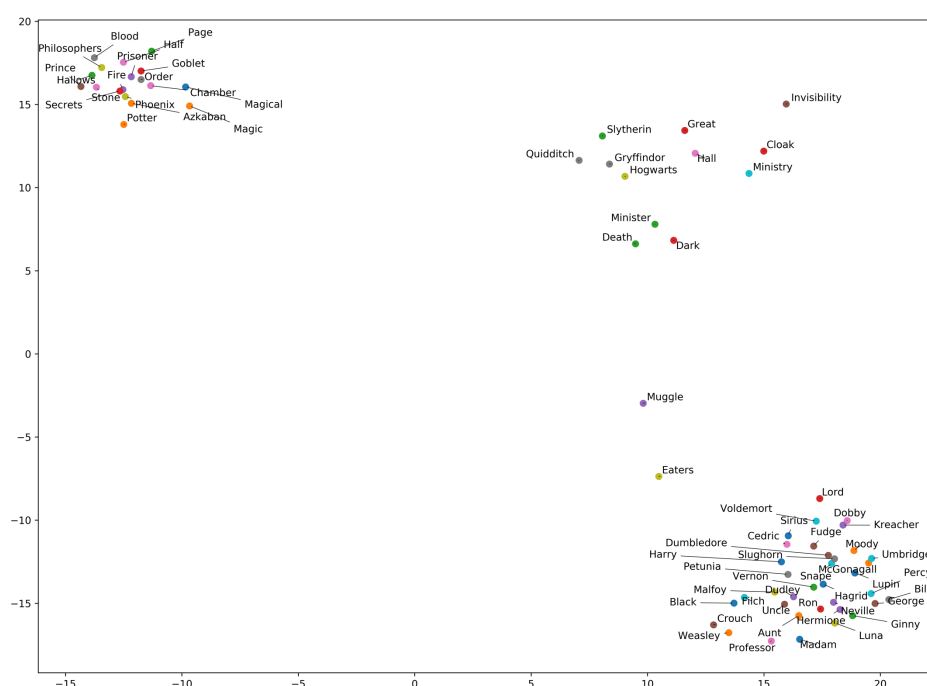
window：要預測的字跟現在的字在句子當中最長距離

negative：會有多少個 **noise word** 會被畫出來

iter_：train 的過程要多少個 **iterations**

alpha：初始的 **learning rate**

2.2. 將 word2vec 的結果投影到 2 維的圖：



圖五：word2vec

2.3. 從上題視覺化的圖中觀察到了什麼？

答：

左上角那群，幾乎都是書的名字；而在中間上方，可以看到魁地奇、葛來分多和史萊哲林靠得很接近，有兩個是學院的名字，有一個是學院之間舉辦的活動，三者之間很有關聯性；右下角則是很多人物的名稱，像是可以看到佛地魔跟他的本名 Lord 很接近。

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：

我的想法是因為這些資料是 normal distribution，而且仔細想想，每增加一維，在空間中各個點的 variance 就會越大，所以我先計算每一筆資料的 variance，然後用 kmeans 將這 200 筆的 variance 分成 60 群，再將這 60 群的 center 做排序，越大的表示維度越大，從 1 維到 60 維。

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：我先將所有的圖片變成一個 vector，然後算出來的維度是 28 維，我覺得不合理，因為用來 train 的 dataset 跟手旋轉的 dataset 是不同的兩個 dataset，用一個 dataset 的標準去比對另外一個 dataset 是很不合理的！