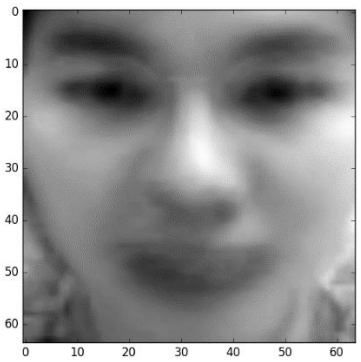
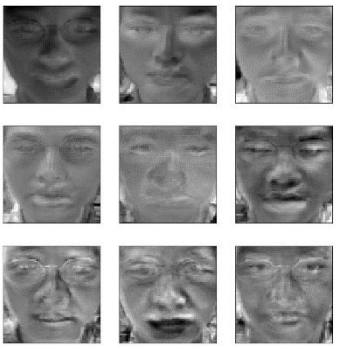
ML2017 HW4 Report

學號：B04611015　系級： 資工二　姓名：陳佳佑

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.

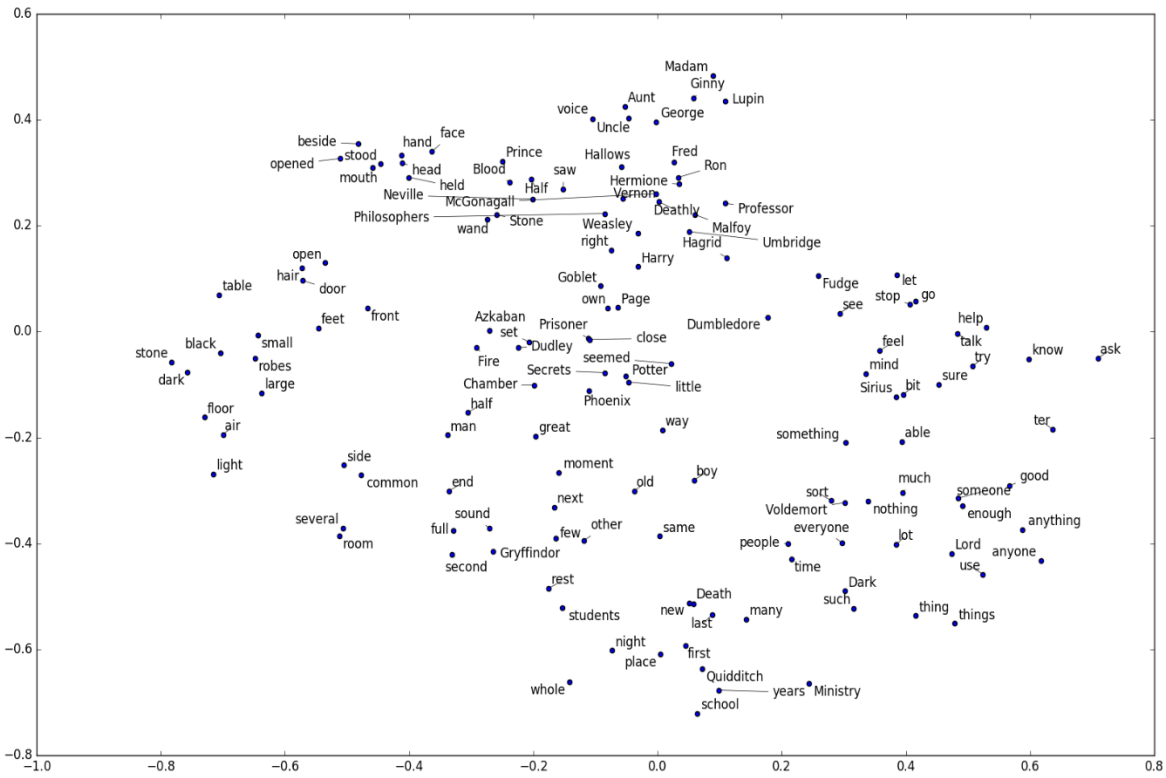
答： (回答 k 是多少)

K = 60 (除以 255 的結果)

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:
Size = 50： 代表 word2vec 後的維度

Min_count =50： 代表頻率的 threshold，超過才會列入運算

2.2. 將 word2vec 的結果投影到 2 維的圖:



2.3. 從上題視覺化的圖中觀察到了什麼？
答：
名字的部分被集中到了上半部，而右下角像是 anything, thing, things, anyone, someone 等詞語都被集中在一起。而左方有像是 small, large, floor, air, dark, light, black。效果並沒有很好，有可能是因為實作上，我是使用 PCA 而不是 TSNE 在最後的降維上，因此造成沒辦法維持太多高維度上的結構。

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？
原理

因為在課堂上，講最細的降維方法就是 PCA，因此一開始我就用 PCA 嘗試，threshold 設在 90%左右，但是丟到 kaggle 上的結果是很差的(0.30)。因此我開始想到，是否是因為每個維度對應到的 threshold 精準度不一樣造成的結果，

換句話說，就是在給定的 threshold 下，是否對某些維度的 predict 結果是較為精準的，而對於某些維度的結果是爛掉的。因此，我利用 generator 產了許多筆的資料，計算出每個維度的資料，預測其自己的維度的 threshold 會是多少，然後多產幾筆平均後，再拿去預估 testing data，丟上去的結果來到(0.117)。

Predict 方法:

透過 generate data 我們會得到一個 dimension_threshold 陣列。
然後對於每筆 testing data，我們會將 eigenvalue (normalized) 一個一個維度相加，直到超過該維度的 dimension_threshold。

合理性

此方法雖然不像 TSNE 能把維持高維的局部結構，獲得好效果。但是以一個線性的方法而言，PCA 的 threshold 的確對於不同 dimension 有不一樣的準確度，並且結果而言，是有顯著增進的效果(0.30 -> 0.117)。

通用性

此方法有一個很大的限制是，dimension_threshold 陣列必須透過 supervised 的方式取得，也就是說，我們必須要有一部分資料的 label，造成其在實際應用上有可能受到阻礙。

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。

Preprocess : 將每張圖都壓成 100*100

將我的方法實作在 hand rotation sequence dataset 獲得的結果是 12 維，這個結果看起來似乎還算是合理，因為手的結構的確有許多彎曲。然而，在這裡我用了一個很強的假設就是 dimension_threshold 陣列會是通用的，但是很明顯，在實際的情況下，這個 claim 並不嚴謹，在這裡，有可能只是手的資料跟上題的資料有某種程度的相似性，造成答案還算是合理。