

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)

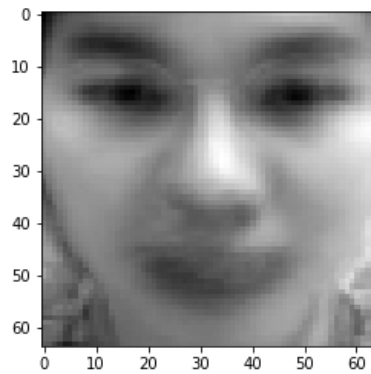


Fig1. Mean Face

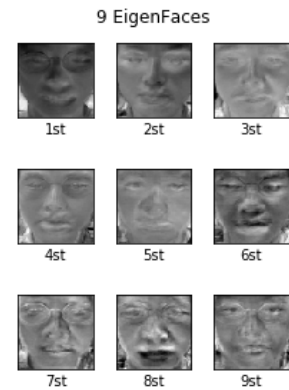


Fig2. 9 Eigen Faces

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)



Fig3. 100 Original Faces



Fig4. 100 Reconstructed Faces (Top 5)

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)

經過計算發現在 K=59 的時候其 RMSE 會小於 1%。

2.1. 使用 word2vec toolkit 的各個參數的值與其意義：

答：

train: training data 路徑

output: model 寫檔案路徑

cbow: 0 為用所有 vector 的 sum 來計算, 1 為用所有 vector 的 mean 來計算, 這邊設定為 0

min_count: 出現頻率低於 N 次就不顯示, 這邊設定為 10

Size: 代表每個 word 要有幾個維度, 這邊設定為 500

window: 字詞關聯的最大距離 (以字為單位), 這邊設定為 5

negative: 顯示 N 個被認為是 noise 的字詞, 這邊設定為 5

iter_: epoch 次數, 這邊設定為 200

alpha: 初始 learning rate, 這邊設定為 0.001

2.2. 將 word2vec 的結果投影到 2 維的圖：

答：(圖)

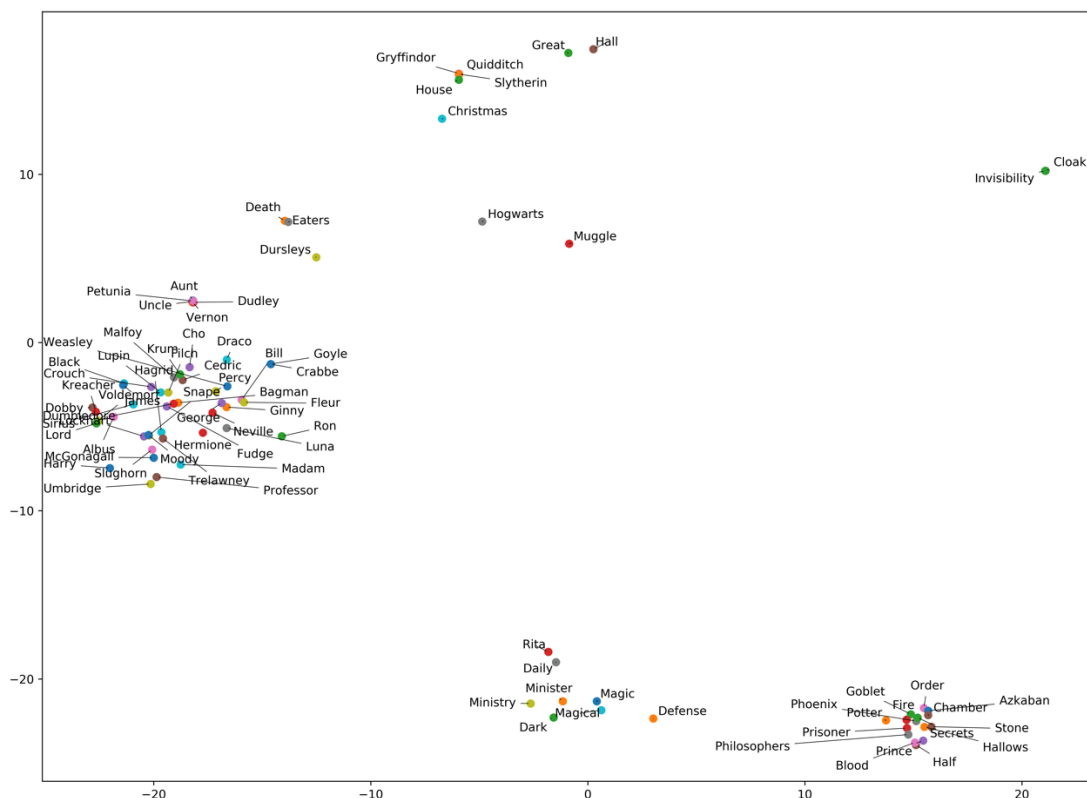


Fig5. Word2vec visualization

2.3. 從上題視覺化的圖中觀察到了什麼？

答：

我覺得最明顯的是德思禮一家族他們被歸類在同一區（包含 Uncle, Aunt, 各個名字還有那個死屍孩兒子），然後葛來分多以及史萊哲林也被分類在同一群，比較驚訝的是魁地奇也被分類在同一群。接著看圖的中間左邊那塊，基本上裡面出現的都是人名，可以看到 Lord 以及 Voldemort 距離很近，不過很奇怪的是 Harry 與 Potter 竟然分超開，Potter 被分類到右下角那一塊感覺不太正常，不管我怎麼調整參數都無法改善。還可已看到看到下面的 Magic 竟然跟 Dark 很近，可見裡面的魔法都很黑暗的！接著看到右上角的 invisibility 以及 cloak，這也算是比較準確的距離，皆為隱形斗篷。最後要提到 Page 這個字也會被分類在最右下角，但這個字對於分類是不重要的，因此我把它設為過濾掉的文字內容。

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：

首先我將每個 set 的所有 value 都展開至一維，針對每個 set 的所有資料計算其變異數，因此會有 200 個變異數。由於 TA 有公布資料的 intrinsic dimension 維度在 1-60 之間，因此接著將這 200 個變異數做 KMEANS (K= 60) 分群演算法得到 60 個 center。緊接著，由於 TA 有說明所有係數皆為 normal distribution 中萃取且 variance 較小的資料其 distribution 較集中，因此可視為同一個維度，當 variance 越大代表會有越多群同個維度的資料，因此這邊將剛剛 KMEANS 的 60 個 center 由小到大編號，越小的 center value 其 dimension 數目越小。然而這樣的方法我認為僅限於助教出的題目，當每個數值不再是從 normal distribution 中萃取，其利用 variance 所算出的分佈將不再像這題所算出來的這麼漂亮。

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：

利用上述的方法算 variance 則先計算每張圖片間的 variance 後計算每個 pixel 的 variance 再比對上述的 KMEANS 結果查表得到其維度為 60。上述的方法是建立在已經知道資料的 intrinsic dimension 落在 1~60 之間，因此可以算 variance 後用 KMEANS 來求出資料大概的維度，但是用在真實的 dataset，如此題的 hand rotation sequence dataset 上則完全無法預測，因此我認為這個方法用在這個 dataset 是非常不合理的，若想要找出這些圖的 intrinsic dimension 我認為需要像第一題的方法一樣使用 PCA 找出 eigenvector，接著計算 RMSE。