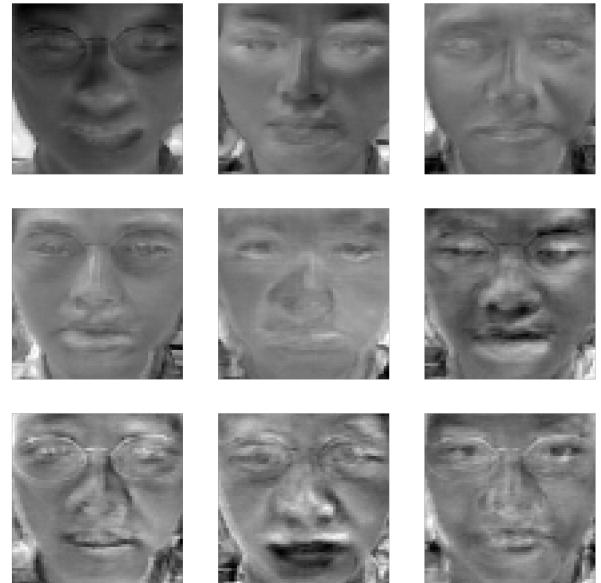
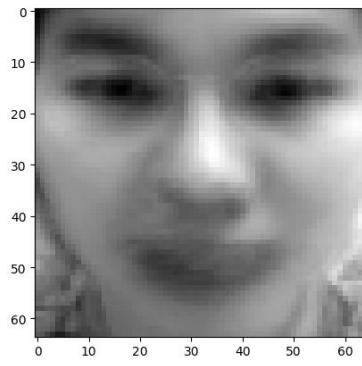


學號: R05944043 系級: 網媒碩一 姓名: 宋焱檳

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:

$k = 59$  的時候，error 達到  $0.99\% < 1\%$

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答: train: 輸入 txt 文件的名稱 (all.txt)

output: 保存輸出的 model (model.bin)

Size: 特徵向量的維度 (250)

Window: 一個句子中當前和預測詞之間的最大距離 (8)

Alpha: 初始 learning rate (0.15)

Min\_count: 忽略總頻率低於此值的所有單詞。 (15)

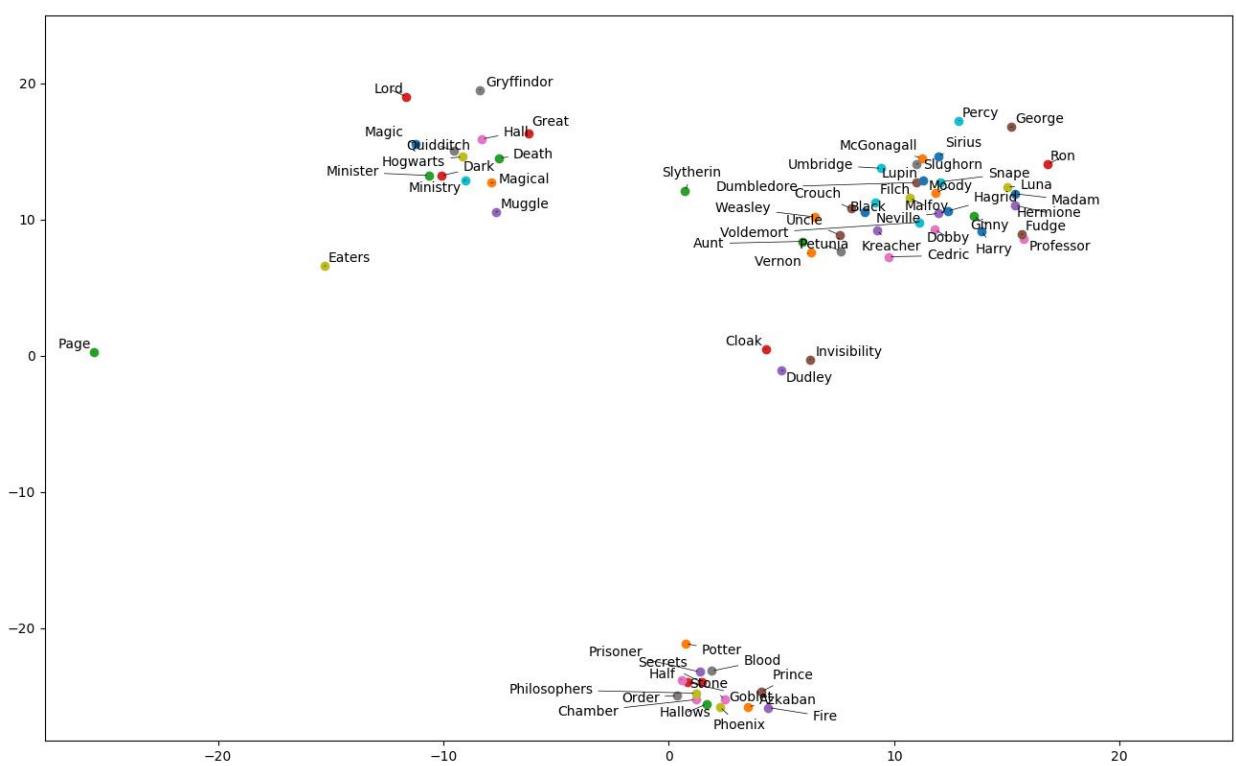
Negative: 應該畫幾個“噪音”詞 (3)

Iter: 語料庫的迭代次數 (200)

Cbow: train 的模型的類型 (1)

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：

可以看出，根據詞與詞之間的聯繫的緊密度，大體上主要分了 3 個區塊。這些詞在出現之後，有更高的概率出現周邊的詞，或者這些詞的含義很接近，或者這些詞的類型很相似。觀察發現，每個區塊中有比較多的名字、名詞、形容詞等。而且，在區塊中靠很近的點互相之間有很多聯繫，比如 Fudge 和 professor。圖中左邊的 Page 和 Eaters 雖然頻率比較高，但是與其他詞的聯繫並不是很大。所以被“孤立”了。

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：

大概想法是通過 KMeans 對每組 data 的 variance 找出一定數量的 cluster，然後觀察 cluster 的中心值，去評估是否離得太近（尋找時 threshold 為本身的 2% 大小），找到 threshold 邊界處的點，記錄下 cluster 分的堆數（因為知道是 100 以內，所以從 50 開始，到 60 就找到了）。然後將分好的 60 組的 cluster 從小到大的排序（1-60），然後依據排序大小（也就是 prediction），給與 200 組 data 屬於不同 cluster 的，不同預測的 dimension。因為，不同維度之間 variance 的差異不同，所以利用這一點，可以很好地猜測 dimension 的大小，所以對 variance 進行劃分 cluster。最後在 public 上的 entries 是 0.11882。通用性一般，因為對於目標的 dimension 的範圍不知道的時候，很難尋找到合適的 scale。或者花費時間太長。

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：

如果在只有 dataset 本身的話，我覺得有一定的可行性，只不過花的時間可能比較多。因為，一開始在評估 prediction 的範圍的時候，就需要對大量的數據進行處理。然後估計出可行的長度，以及可行的範圍（如果發現範圍太大，則很難完成）。之後，就可以用之前的方法利用 KMeans 進行估計。但是得到的結果可能不是很好，因為這個解法有個問題，就是默認了高維度的 variance 一定會比低維度的大，所以如果在維度變化大，而 variance 又相近的情況下，這個方法的結果會很差。