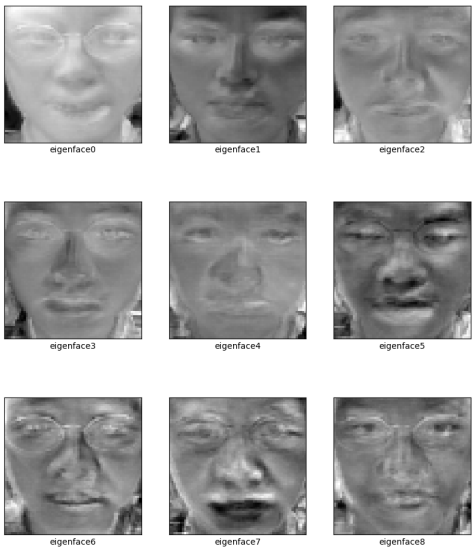
學號：B02902044 系級：資工四 姓名：陳映紅

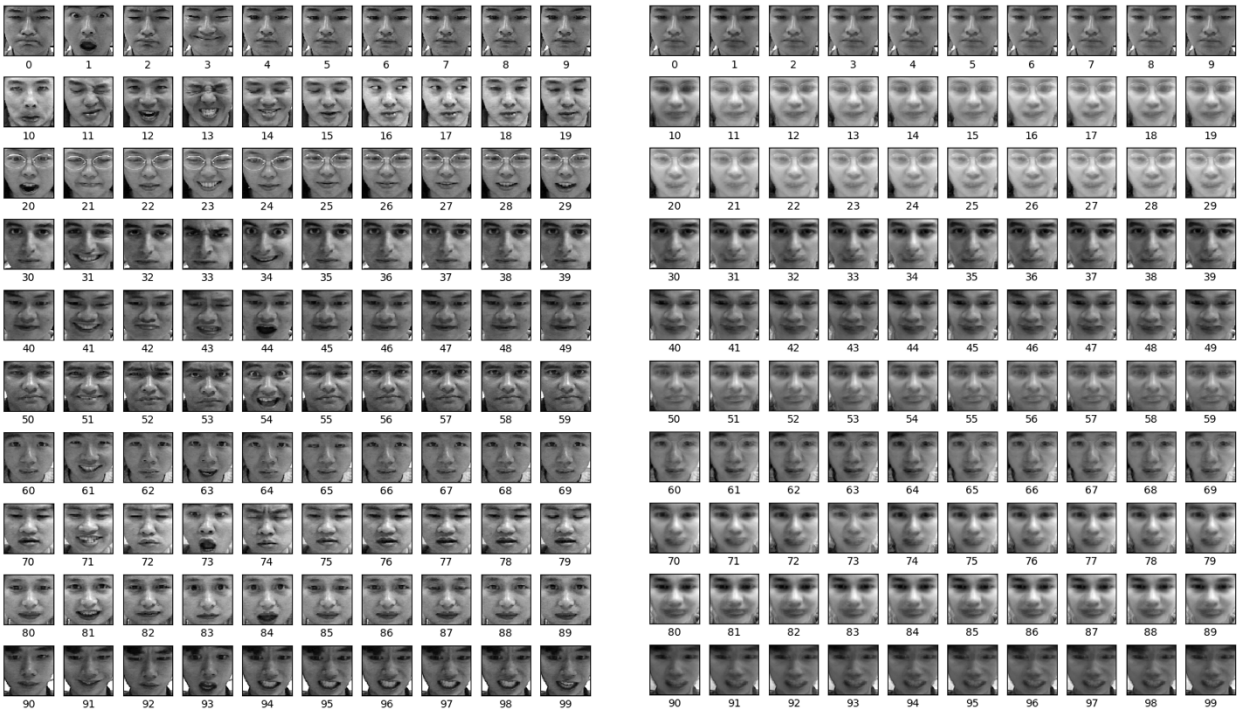
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

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)

1% * 255 = 2.55

使用 60 個 eigenfaces 時，RMSE = 2.4776269501197086

使用 59 個 eigenfaces 時，RMSE = 2.5518673845707722

因此 k 應為 60。

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

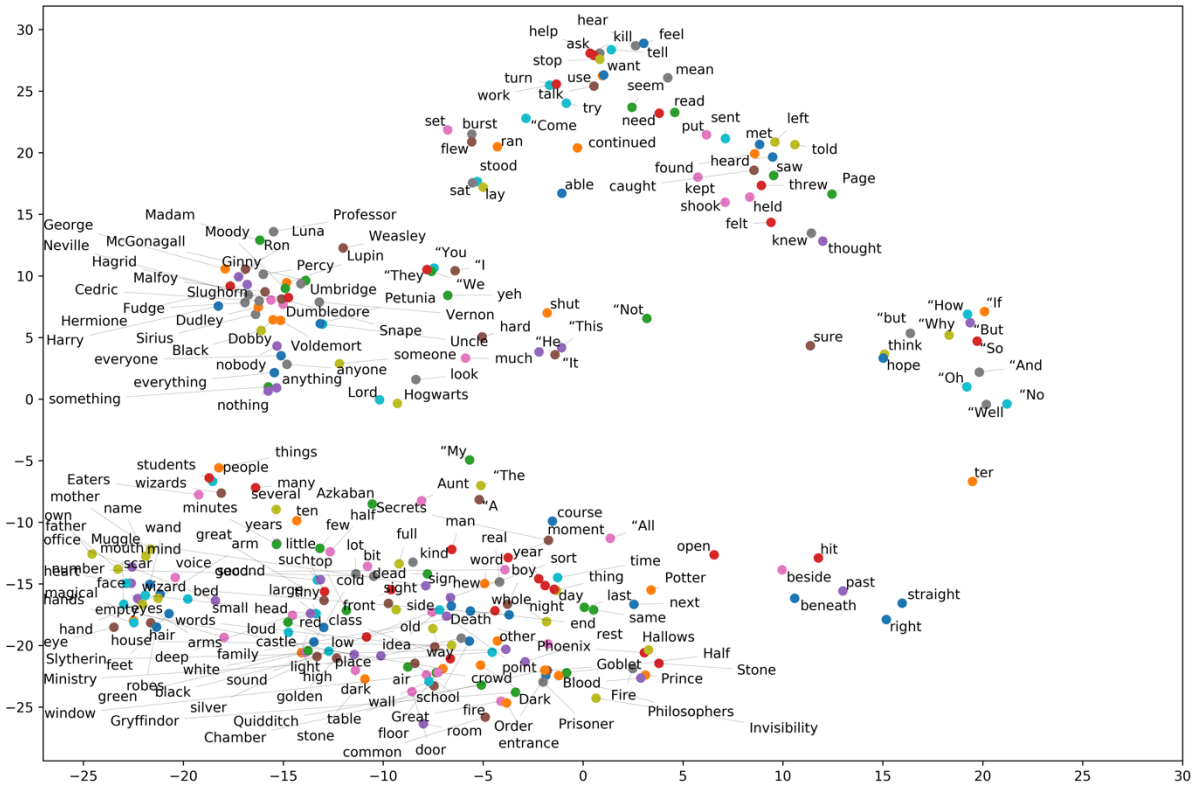
答：word2vec.word2vec 的預設使用方式及我的參數使用方式如下

```
def word2vec(train, output, size=100, window=5, sample='1e-3', hs=0,
             negative=5, threads=12, iter_=5, min_count=5, alpha=0.025,
             debug=2, binary=1, cbow=1, save_vocab=None, read_vocab=None,
             verbose=False):
```

參數	使用之數值	參數功能解釋
train	all.txt	訓練模型的文字庫路徑（不影響結果）
output	all.bin	模型路徑（不影響結果）
size	150	訓練產出向量的維度大小，過小無法正確分隔語意，過大則可能使 TSNE 降維效果不佳
window	2	skip gram 中的最大 skip 長度，將此值由預設值 5 降為 2 後發現動詞現在式及過去式的 cluster 較接近，故使用 2
sample	0	將出現頻率高於此值的字任意降低頻率。使用建議值 1e-5 後發現 cluster 較不明顯，故使用 0
hs	1	使用 Hierarchical Softmax 加快運算速度
negative	5	使 window 內與外的字分開，使用建議值 5
min_count	50	忽略出現頻率小於此值的字，使用 50 是不想讓圖表太複雜
cbow	1	是否使用 skip gram

註：未使用 threads（是否使用多條執行續）、debug（除錯模式）、binary（將模型存為 binary 格式）、save_vocab（訓練後之字彙的儲存路徑）、read_vocab（訓練時使用此文字庫而不使用 train）、verbose（印出訓練過程）、iter_（訓練次數）、alpha（learning rate）

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：有幾個明顯的 cluster：[-3:5, 23:29]的原形動詞，原形動詞附近（[-7:12, 13, 22]）有動詞的過去式，人稱代名詞（I、You、They 等）[-8:-6, 10:11]，人名[-18:-13, -5:14]，人名附近（[-17:-13, 0:4]）有不定代名詞，疑問詞、助詞或開頭的連接詞[15:22, 0:8]，另外在人名附近有一大區[-25:-5, -25:-5]為名詞。

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：

- 方法：我先將每個 dataset 裡的資料 flatten 成一維，計算每個 dataset 各自的 variance 後，利用 kmeans 將每個 dataset 依照其 variance 分成 60 個 cluster，再利用 60 個 cluster 的 center 由小至大排序，依序將 cluster 內的每個 dataset 對應至維度 1 至 60。
- 原理：對一個 dataset 來說，若原始維度很小，增至一百維後，大部分的資料會有較低的亂度，所以 variance 較低。而若原始維度大，增維後亂度仍大，因此 variance 會較大。
- 合理性及通用性：此方法必須知道原始維度的範圍，才能依這個範圍做 clustering，且若 dataset 數量不夠，則 kmeans 可能不精確（甚至不能使用），因此無法找到單一 dataset 的原始維度。

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：

- 方法及結果：我先將手轉杯裡所有資料 imresize 成 16*15 維，再用 PCA 降至 100 維。將這個 dataset 作為第 201 個 dataset，利用 3.1 的方法，求得原始為度為 23。（見 rotate.py）
- 合理性：手轉杯 dataset 降維後的維度應該三、四維，23 維偏離合理值很多。因所有 dataset 的差異性很大，所以在沒有此 dataset 其他資訊的情況下，直接套用 3.1 的方法時將不會有好結果。