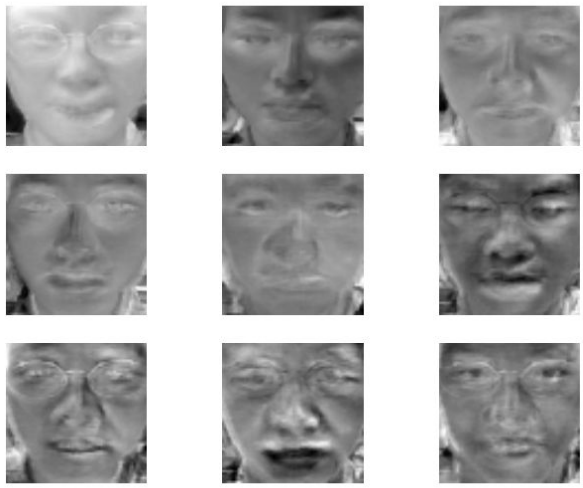
學號：B03b02014 系級： 生技三 姓名：張皓鈞

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)

平均臉

Top-9 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)

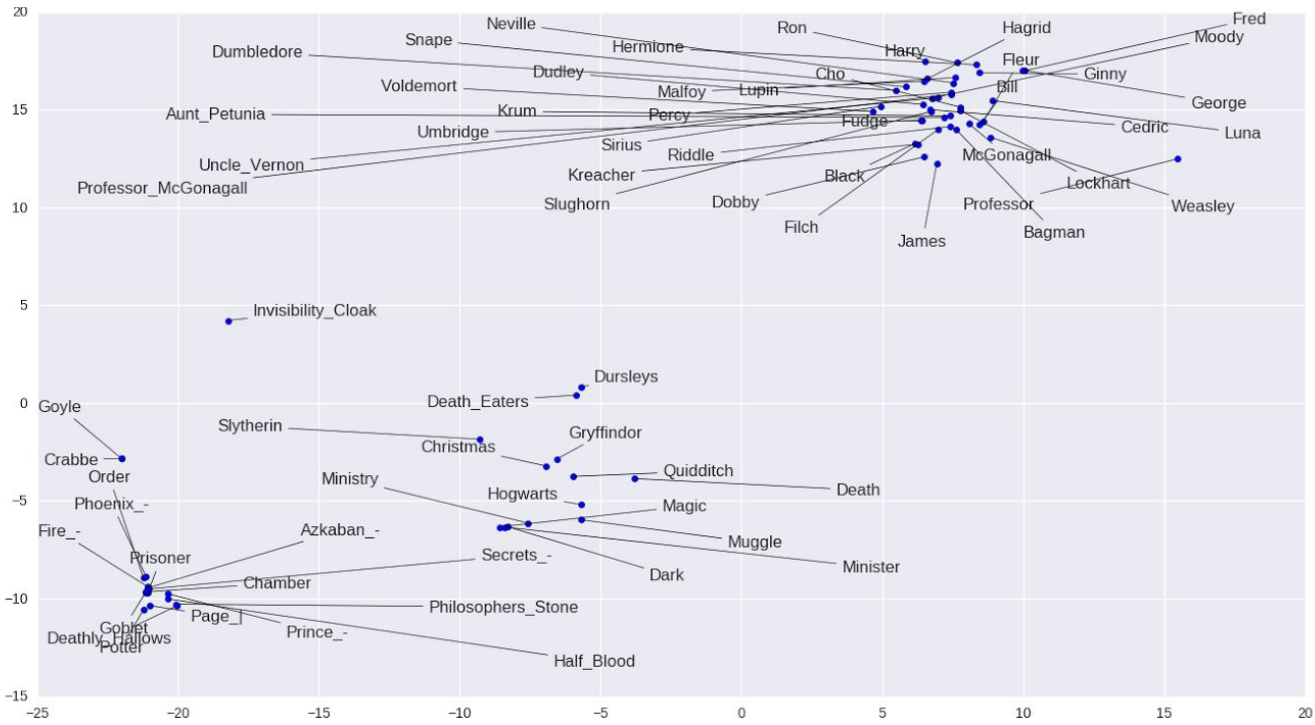
原始圖片

Reconstruct圖

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 = 59

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:
答：訓練參數：size=110, alpha=0.025, window=5, min_count=5, negative=5, cbow=1, min_count=5, sample=0, hs=1
參數說明：
size = 將文字轉換的維度。
window = 同一句子裏用來預測下一個字的最大數目。
alpha = 初始learning rate(之後會線性下降到min_alpha=0.0001)
min_count = 單字出現頻率的最低門檻，低於此值就忽略此單字。
negative = 隨機取樣Negative words的數目（通常在5~20之間）。
cbow=1代表使用的是skip-gram模型
min_count=出現次數小於min_count的單字會被捨棄掉。
sample=高頻率出現的單字會比較不會被sample到，而此參數設定成為"高頻率"的門檻值
（0表示off，0<sample<1表示單字佔總數的比例，sample>=1表示單字出現次數比sample高的就視為"高頻率"）
hs=1代表使用Hierarchical Softmax。

2.2. 將 word2vec 的結果投影到 2 維的圖:
答： (使用TSNE將前1000多的單字投影到2維)



2.3. 從上題視覺化的圖中觀察到了什麼？
答： 從圖中可以看到關於哈利波特標題的單字會形成一群，而人物的名稱（例如：hermione, ginny, weasley等等）也形成一群比較分散的cluster在圖中的右上角。
中間散落著其他沒有顯著cluster的單字，但是，

有趣的是，Dursleys和Death Eaters非常的接近，推測在文中Dursleys和Death Eaters都代表著主角不幸的感覺。

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：

我的方法是利用轉換資料的函式皆為連續且可微的性質，假設每一個高維度的資料點和鄰近的資料點可以投影到同一個space(tangent space)。再加上透過觀察發現：如果從d維原始維度sample出來的資料點map到100維上，所取得的eigenvalues會在第d個和第d+1個有明顯的改變，而且每一個原始維度產生出來的eigenvalues分佈皆不太一樣。

因此我的作法如下：

1. 隨機取樣1/100的樣本點，計算每個樣本點和周圍的20-nearest neighbors。
2. 對每個樣本點及其鄰近點作PCA的投影。
3. 將eigenvalues進行normalizationa並由大排到小。
4. 平均各樣本點所得到的eigenvalues，得到100個平均eigenvalues。
5. 利用gen.py產生的資料點訓練svr模型(使用參數C=1.5)。
   （輸入為資料點的平均eigenvalues，預測目標為ln原始維度。）
6. 用svr來預測每筆測資的平均eigenvalues的分佈。

此方法的在kaggle上public的分數為0.09041。

此方法是透過eigenvalues的分佈作為代表資料點的原始維度，以此來估計原始維度。

通用性上的限制是假設產生模擬資料的方式是從N(0, 1)的分佈取樣出來的，若資料產生的方式和假設不符，就得使用其他方法來估計原始維度。

（例如：計算 $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{100} \lambda_i}$，（$\lambda_i$ 為ith eigenvalue），找最小的d使之大於門檻值。）

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：

我的方法估計hand rotation sequence datatset中每一張圖實際可以用4維(平均出來的估計值為4.27750215445)的vector表示，此結果我認為算是合理，因為hand rotation sequence是只有一維地旋轉手持物，透過觀察照片，我推測可以描述這種圖片的維度為”手的旋轉角度”，”手的水平位置”，”手的垂直位置”和”照片的亮度”這四維。

但是，根據其他文獻多數解釋hand rotation sequence可以用水平位置、垂直位置和照片亮度這三維解釋，而我的方法有其誤差的原因可能是因為這方法的假設是產生資料的方式是從N(0,1)的分佈取樣出來的，而可以真正代表hand rotation sequence datatset的資料點產生方式不一定是從這樣的分佈取樣出來的，因此在預期上會有誤差也是合理的。