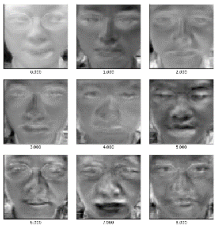
學號：B03901096 系級： 電機三　姓名：周晁德

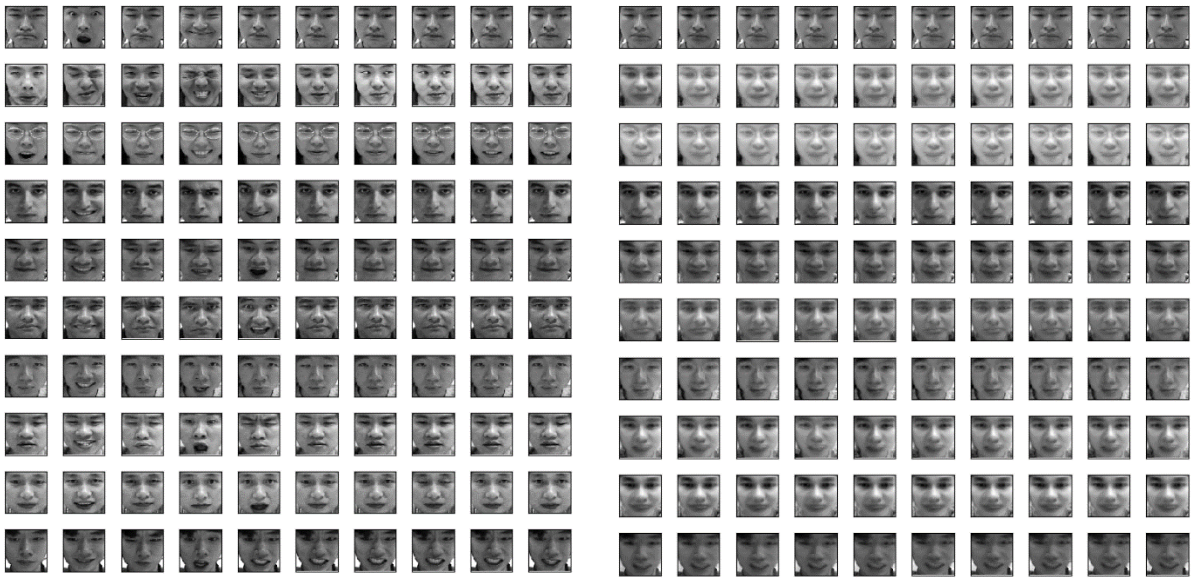
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

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＝59(我是除 256!)

56	57	58	59	60	61	62
1.101997	1.06537	1.029989	0.996823	0.967823	0.938073	0.909149

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

Size=200 : size of word vectors，每個維度可能代表某種性質!

min_count =5 : discard words that appear less than <int> times，基本上是一個減少不重要 training data 的方法!

Negative=5 : number of negative samples，這個參數主要是為了加速 training 的步驟

window=10 :max skip length between words，決定多遠要算是有關係的 context

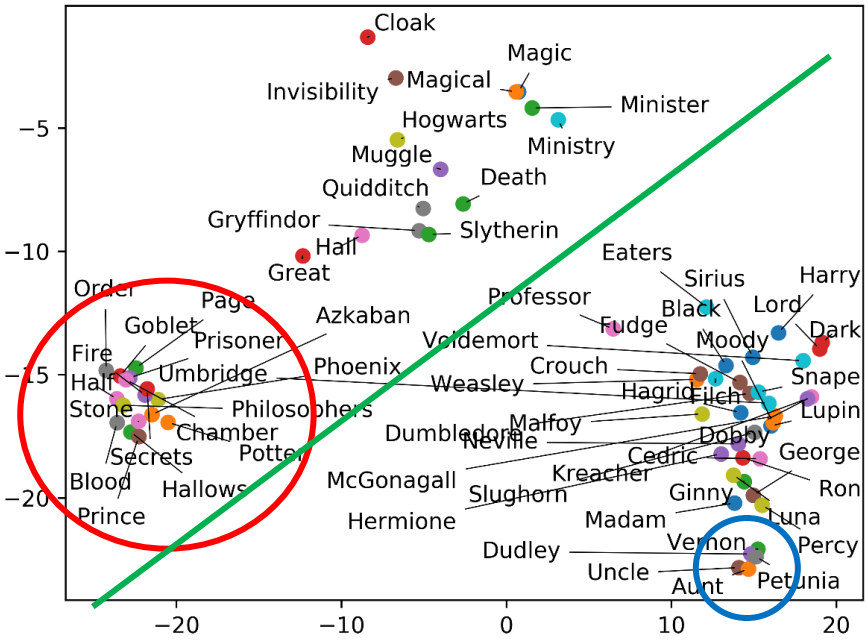
iter=50: number of iterations

alpha=0.025: learning rate

選擇 800 -most-frequent words

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：

(1)大致可以由綠線切開，線的右方主要是人物，線的左方主要則非人物

(2)右下角藍色圓圈中的人物是哈利唯一活在世上的親人，他們的分布也非常接近

(3)藍圈上方那些主要是魔法界的人物，從中發現衛斯李一家(Ron,GinnyPercy,George)位置很接近，Ginny 和 Luna 是好友因此分布接近。

(4)紅圈中幾乎都是每集的書名中的關鍵字!(例如:Half blood prince)

(5)magic 和 magical 幾乎疊在一起（關係為形容詞和名詞）

(6)影行斗篷 invisibility 和 cloak 很接近(在其它圖中它們會疊在一起)

(7)同為家庭小精靈的 Dobby(多比)和 Kreacher(怪角)分布也很接近

(8)同為學院名稱的 Gryffindor(葛來分多)和 Slytherin(史萊哲林)分布接近

(9)Dark lord 幾乎黏在一起是因為那是食死人對佛地魔的稱呼

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：但後來改使用助教的方法，用助教提供的 code 自己產生 data set，然後在”從 data set 中選取 k 筆 data，將每個被選中的 data 與其周圍 200 個 data 做 PCA，因此會得到 k 筆 200 維的 eigenvalues array，每個 array 中的值都已經除以 array 中的最大值，再對這 k 個 eigenvalues array 取平均，而這就是一筆 data”，產出很多筆 data 之後再用這些 data 以及對應的維度，train 一個 svr(我本來有想過用 DNN 但因為 data 少容易 overfitting!)，在預測維度的時候，就把 200 個 set 中的 data 做雙引號("")中的步驟，再用剛剛 train 好的 svr 做預測!
這次作業難是因為會人想用 DNN，但(1)要用 unsupvervise(2)就算用助教給的 code 產生很多 data set，可是每筆 data set 的 data 比數不同如何給 DNN 當 inout 也是個問題。
這個訪法是藉由解析每個 data set 中小範圍的 data 的維度，並多 sample 幾個點，避免一些極端值得引影響，最後藉由 eigenvalues array 曲線形狀來判斷維度，如果 intrinsic dimension 是低維的，則曲線一開始會降很快，且會有很突出的轉折點，若是高維資料則曲線平滑。因此對於不同 input 是不同維度的 data set，我只要重新產生 data set，用上述方法在 train 一個 svr model 就可以合理的推測數據的 intrinsic dimension 了!

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：
我採用上面的方法，但因為要產生維度為 512*480 的 data 會有 Memory Error，且一張圖中重要資訊其實每有這麼多，所以我用 PIL 將影像縮小為 32*30 的大小，並且產生 intrinsic dimension∈[1,100]的 data set，每個 intrinsic dimension 各 4 個 data set，每個 data set 有 481 筆 data(=hand rotation 影像的數目)，用這些 data 來 train SVR，接著直接讀圖片來做預測，因為 get eigenvalies 中有隨機的成分，所以預測結果為[-5.56723,-2.65769]之間，但不管是多少這個結果都不合理，我懷疑是因為每個維度的 data 的數值大小問題，因此我把全部數值做了 normalize，也就是把這筆 data set 中數值最大(max)和數值最小(min)找出來，把每個數值減去 min 除以(max-min)，得到預測結果為[-4.066742,-1.65893]之間，結果仍然不合理。所以我推測可能原因如下:
(1)使用助教提供的 gen.py 所產生的高維度 data set，data 分布並不會有特別扭曲的狀況，也就是產生的 data set 比較平滑，因此比較好做預測，而 SVR 用這些 data set train，因此在預測真實 data set 的時候就預測不準了!
(2)這個 data set 感覺維度應該蠻低的，因此 eigenvalues 所形成的曲線應該下降非常快(比自行生成之 intrinsic dimension=1 的 data set 還快)，使得 SVR 誤判為負的