學號：B04902045 系級： 資工二 姓名：孫凡耘

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 = 58, error = 1.02%
k = 59, error = 0.996% ⇒ 59

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

我皆使用default值

cbow <int> Use the **continuous bag of words** model; default is 1 (skip-gram model)

size <int> Set size of word vectors; default is 100

window <int> Set max skip length between words; default is 5

sample <float>

Set threshold for occurrence of words. Those that appear with higher frequency in the training data will be randomly down-sampled; default is 0 (off), useful value is 1e-5

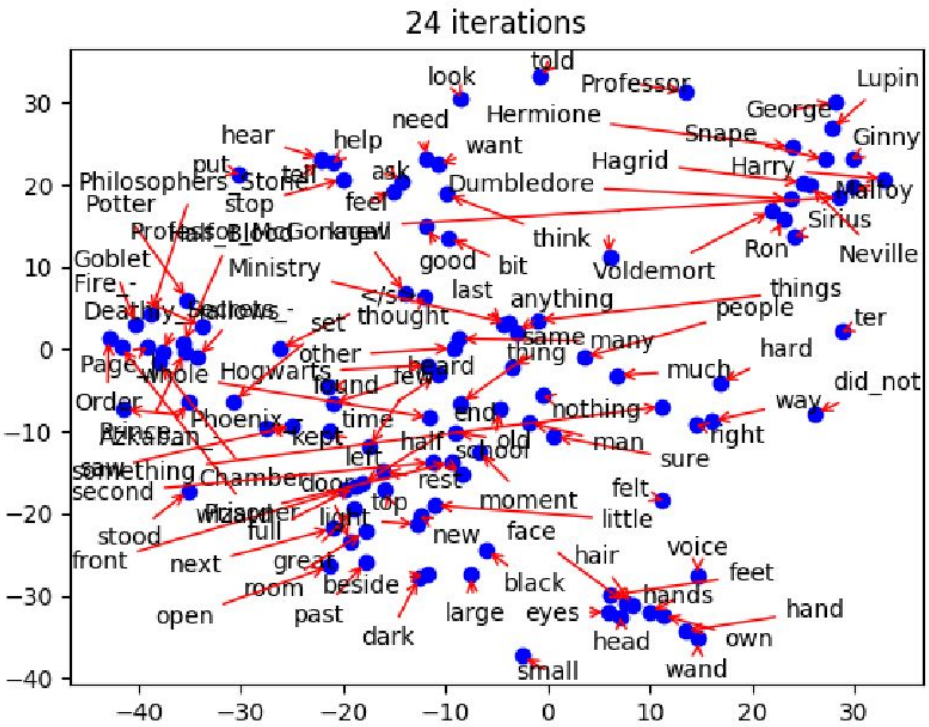
hs <int> Use Hierarchical Softmax; default is 1 (0 = not used)

negative <int> Number of negative examples; default is 0, common values are 5 - 10 (0 = not used)

min_count <int> This will discard words that appear less than <int> times; default is 5

alpha <float> Set the starting learning rate; default is 0.025

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：

可以看到性質相近(包含意思跟詞性的字會比較相近)。大致上可以觀察到人名集中在右上角(雖然Harry Potter的Potter出現在左上角)，人的部位出現在右下角，不定代名詞(ex: anything, nothing)出現在中間，動詞則比較多出現在左下角與左上角。像many, much這兩個幾乎同意的字也非常接近。

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：
我使用4個不同threshold的pca estimator(每個estimator有不同的threshold)，利用助教給的測資產生器生成很多training data(10000筆), 之後用一個decision tree classifier把四維的data(即四個pca estimator預測的結果)map到一個[1, 60]的dimension. 這樣做的原因是發現不同threshold的pca estimator會有各自預測比較準的區間，所以想說用很多個的pca estimator的結果再做預測，而且感覺decision tree很適合做這個task(例如判斷某個estimator預測的結果大於50，那結果就會在什麼區間之類的)。這方法使用助教給的data generator，一定不太通用畢竟真實情況我們是不知道testing data是怎麼從低維transform到高維的，但若能夠用很多不同的data set，或者用很多不同的方法generate training data，這方法應該就能通用許多。這樣說的原因是，下面這個articel(http://www.sciencedirect.com/science/article/pii/S0020025515006179)中提到的很多方法(像是 Camastra–Vinciarelli's algorithm)中，也是需要生測資(但可能不向我這邊的做法是再train一個model，是用測資的結果去fit一個curve之類的)

Best result on public leaderboard: 0.10104

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。
擷取自Conical dimension as an intrisic dimension estimator and its applications *這篇paper
"The hand image dataset is a real video sequence (481 frames) of a hand holding a bowl and rotating along a 1-d curve in space, although the frame size is rather large, 480×512, it actually embedded in a 2 d space."

1. 如何利用前一提的方法來分析
   使用與上一題一樣的測資產生方法與training data set的話，預測出來結果為12。
2. 討論合理性與其原因
   不準確是情有可原，畢竟training data的產生方法跟testing data的方法顯然不一樣(不像作業)，應該使用比較一般性的intrinsic dimension estimating的方法(即在沒有training data下的algorithm，或者是產生training data的方法要多元一點），而不是只使用助教給的測資產生器。上面那篇paper用maximum likelihood esimation當作baseline，baseline預測結果也在2.88，比我的結果好太多。可以說我上述使用的方法如果apply到general 的intrinsic dimension esimation，應該是overfitting到只有在辨識某一種在高維空間中的低維結構表現特別好(或者說只看得出某種transform)。
   另外我覺得結果這麼慘的原因也是因為我用decision tree classifier。如果用dnn作為classifier，結果應該會好一些。不過若要使用這個架構來預測intrinsic dimension的話，最好的解決方法應該還是蒐集不同性質與結構來當作training data.