

學號：B04902097 系級：資工二 姓名：陳家棋

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 (左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)

答：



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, 左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)

答：左邊是原始照片，右邊是 前 5 個 eigenfaces reconstruct



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. (回答 k 是多少)

答：算出來的 K = 60



2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

size : 這是 train 完後每個字的維度，大概在 200~400 間可以有較佳的 accuracy ，太小則 accuracy 較低，高於 400 的話 accuracy 沒有太大的差別，反而 train 速度慢 (我取300)

window : training 時每個字前後看的距離，若設的小則 train 出來會偏向該字詞本身，若設的大 train 出來會偏向個字詞間的關係 (我取7)

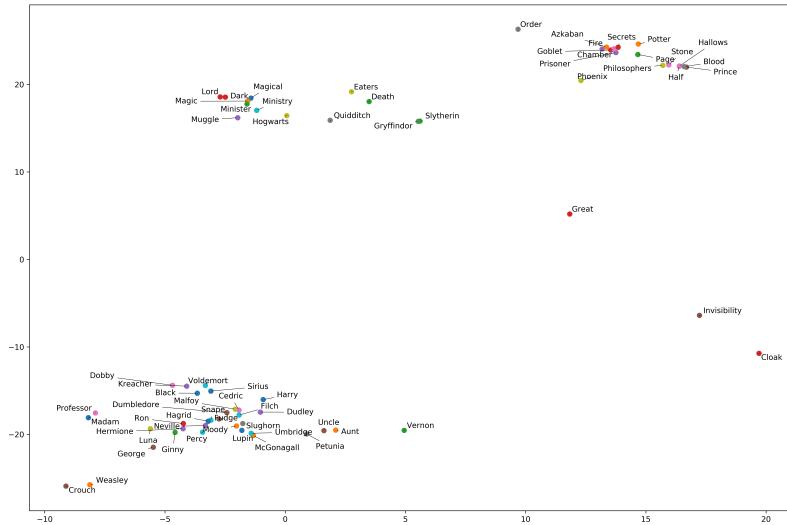
min_count : 把出現次數太少的字捨棄 (我取10)

alpha : 起始的 learning rate (我取預設值0.025)

negative : 讓 training 過程中，不會每次都 update 所有的 weight ，而是部分比例 (我取預設值5)

2.2. 將 word2vec 的結果投影到 2 維的圖:

答：(圖)



2.3. 從上題視覺化的圖中觀察到了什麼？

答：

我不太確定x軸、y軸所代表的含義，但可以明顯看出圖形分為三大群。左下角為各角色的名稱(人名)，上方中間偏左則是有關學校的部分，而右上角則是跟書名、主題有關的詞彙。

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：

1. 利用 PCA 取一個 threshold，當前 k 個 eigenvalue 比例總和大於該 threshold，則以 k 作為他的維度 (效果不好，kaggle 0.33)
2. 利用助教提供的 code 去生測資，設定多個 threshold，若超過該 threshold 且得到的 k 跟對應的 threshold 一致，則以 k 作為他的維度 (效果還可，kaggle 0.12)
3. 利用助教提供的 code 去生測資，把生出來 data 的 eigenvalue 拿去做 DNN (效果不錯，kaggle 0.053)

我覺得我的方法若是在有足夠資訊、且能產生大量 training data 的情形下是可行的，但若是換成估計未知 data 的維度，效果必定很差

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：我用第三個方法將前 60 的 eigenvalue 拿去 predict，得到的維度是 9。我認為不太合理，這筆 data 應該頂多到 5、6 維而已，推斷如同上題所述，我是在知道資料產生的各種資訊去 train 我的 model，實在是無法推廣到一般狀況。