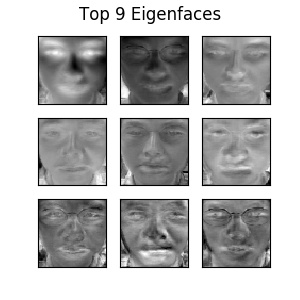
學號：b04902053 系級：資工二 姓名：鄭淵仁

**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)**

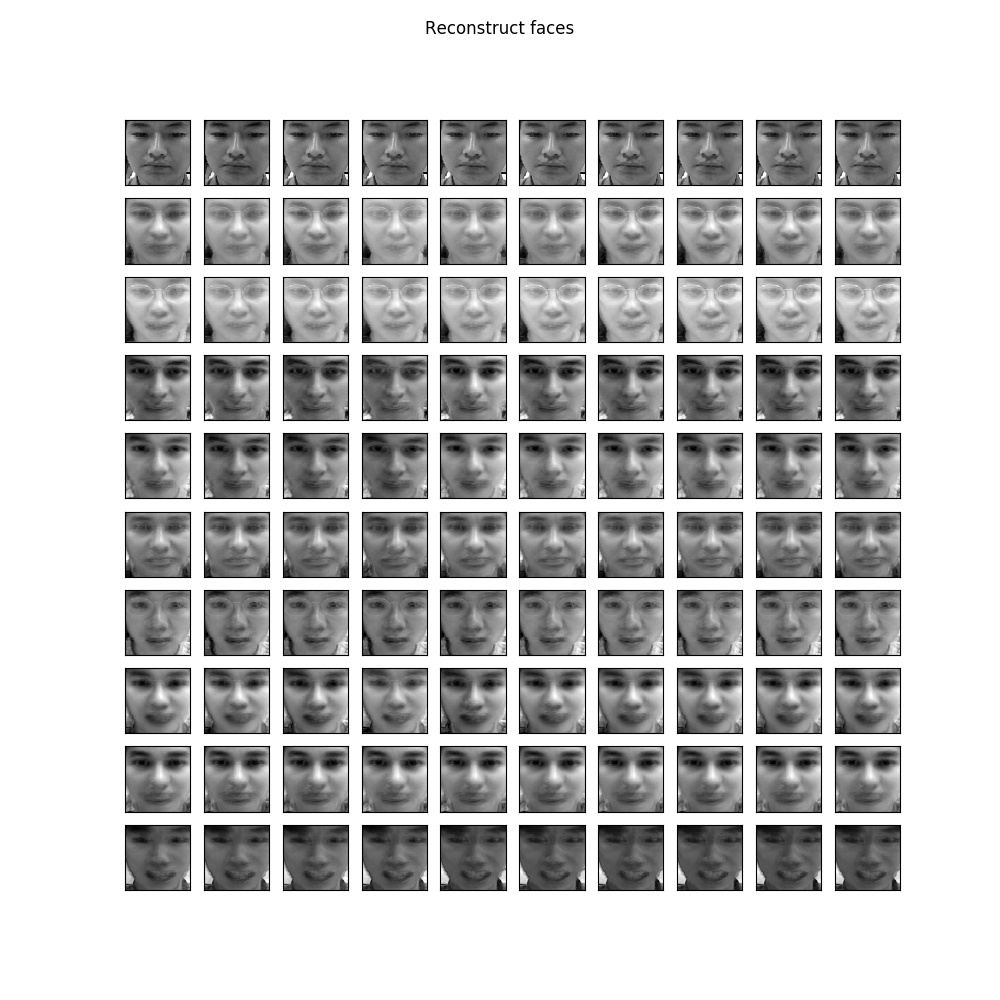




上圖中，左圖是平均臉，右圖是PCA得到的前9個 eigenfaces。

**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)**



上圖中，左圖是原始的圖片，右圖是reconstruct之後的圖片。

**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是60。

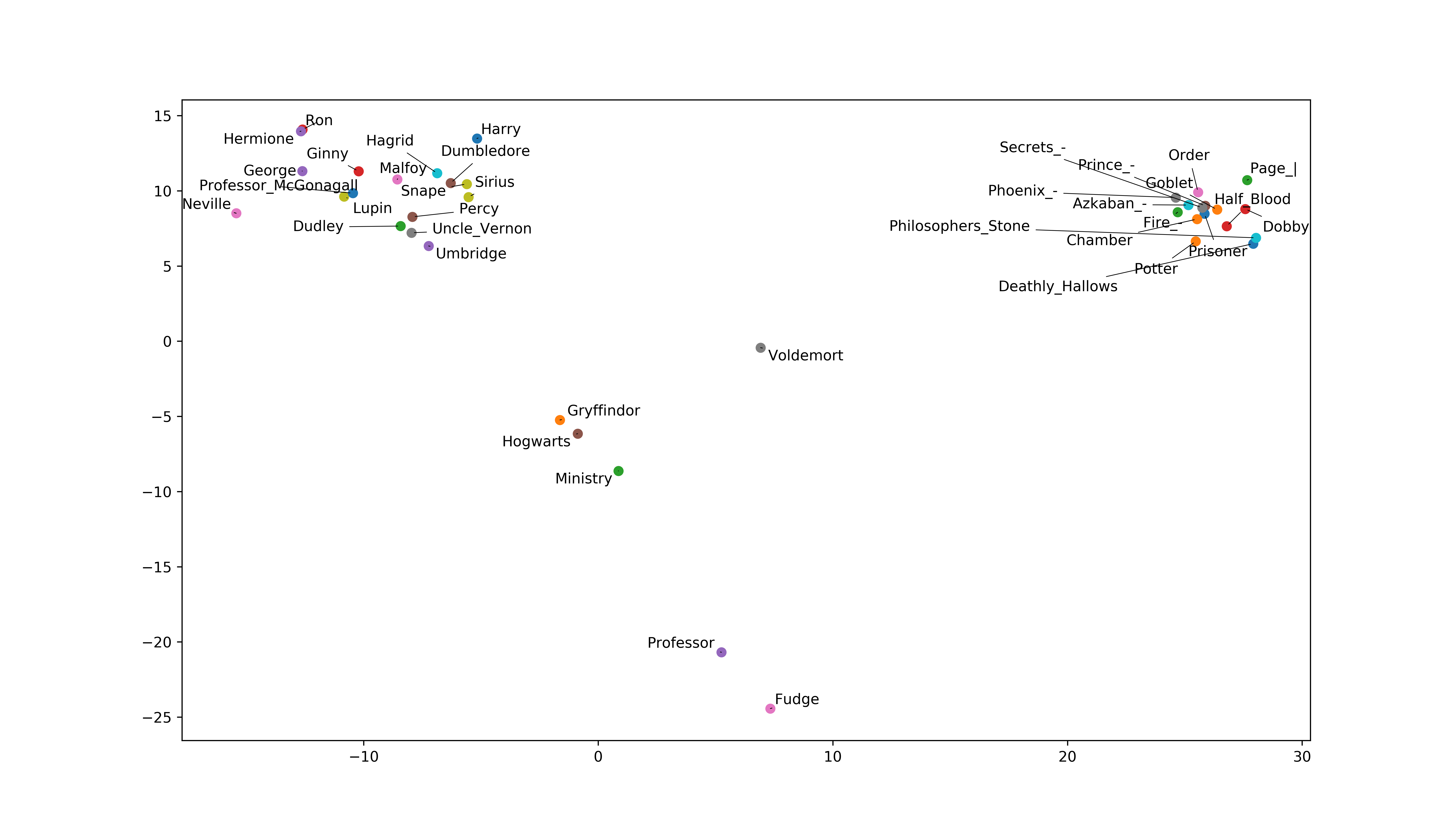
**2.1. 使用 word2vec toolkit 的各個參數的值與其意義：**

**答：**

|  |  |  |
| --- | --- | --- |
| 使用的參數 | 值 | 意義 |
| train | ./all-phrases | train的data的路徑 |
| output | ./all.bin | 要把train好的word vector的結果存到的路徑 |
| size | default(100) | word vector的大小 |
| window | 25 | word跟word之間最大跳躍的長度 |
| sample | default(0) | 設一個詞出現次數的threshold。出現太多次的詞會被隨機的down-sample |
| hs | default(1) | 使用Hierarchical Softmax |
| negative | 5 | negative sample的數量 |
| min\_count | 100 | 把出現次數少於min\_count次的字去掉 |
| alpha | default(0.25) | starting learning rate |
| cbow | default(1) | 使用continuous back of words model |

**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：**

我發現比較相關的詞會出現在一起，例如Gryffindor、Hogwarts、Ministry會是很接近的一群；而其他如Ron和Hermione幾乎疊在一起了，可見他們在文本裡面關係很密切；Sirius、Snape、Percy、Malfoy也是很接近的一群。

**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：**

* **估計的方法：**

我使用**DNN**來估計原始維度。使用的方法如下：

1. **生成訓練資料：**
   1. 使用題目提供的生成資料的python script生成不同維度的data point，共300,000筆資料。
   2. 使用PCA把每一個data point都降維成100維的矩陣，並且只存下最大的前60維的數字作為training data。
   3. 把各個data point的維度取log後作為training data的label。
2. **生成測試資料：**
   1. 把要測試的200筆資料都使用PCA降維成100維的矩陣，並且只取求60維的資料作為testing data。
3. **Train and Predict：** 
   1. 使用DNN訓練出很多組model。
   2. 把不同model predict出來的值取平均，輸出到csv。

* **原理：**

由於ELU的運算對於資料的原始維度影響不大，所以使用PCA後得出來的值，只要設好threshold，應該都能預測出正確的維度。但是正確的threshold可能會因為原始維度的不同、資料數量、或資料的值的大小差異而有不同的值。所以就使用DNN訓練出一個，可以從經過PCA轉換後的資料算出原始資料維度的function，再使用這個function計算出結果。

* **結果&合理性：**

在validation set上面的RMSE是0.050711，在kaggle上面，public data的RMSE是0.05280，private data的RMSE是0.04020。

預測出來的值沒經過調整就都在1~60之間，以值的大小而言，很合理。

* **通用性：**

由於我的訓練資料是基於題目提供的生成資料的python script生成出來的，所以只會對使用相同方法生成出來的資料有效，在其他情況下的資料會是沒有效果的。

**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：**

計算出來的維度是exp(-468961)。

結果很明顯是**不合理**的，維度並不會有負值，而且負的值還那麼大。

會產生這樣的結果的原因就如同上一小題所述：我的訓練資料是基於題目提供的生成資料的python script生成出來的，所以只會對使用相同方法生成出來的資料有效。但是hand rotation這些圖很明顯不是由題目提供的python script生成出來的，所以結果會差很多而且不合理。