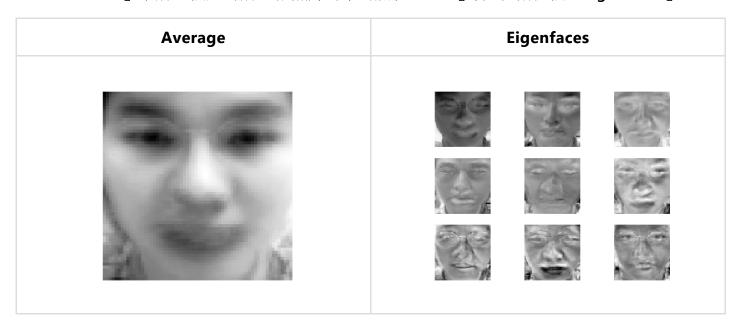
機器學習作業四 報告

學號	系級	姓名
B03902015	資工三	簡瑋德

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」



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」)

Origin	Reconstruction

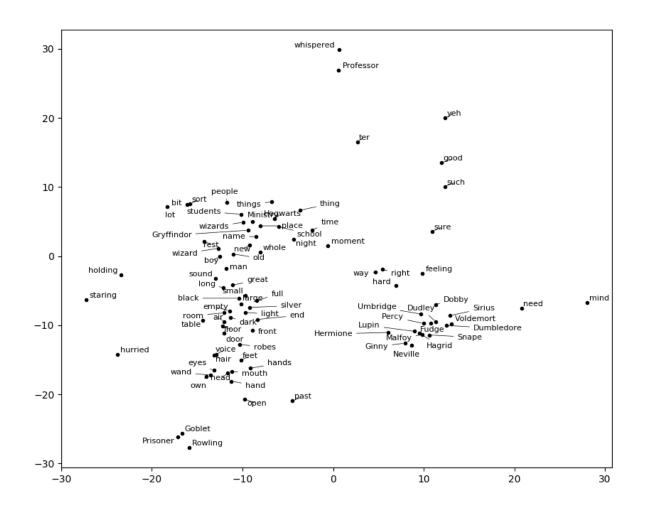
1.3. 「Dataset」中前10個人的前10張照片投影到「top k eigenfaces」時就可以達到<1%的「reconstruction error」

k = 59, RMSE= 2.55186750057

- 2.1. 使用「word2vec toolkit」的各個參數的值與其意義
 - size=50 <int> 詞向量的維度
 - window=5 <int> 考慮附近多少個詞語
 - hs=1 <int> 是否使用「Hierarchical Softmax」來優化
 - negative=0 <int> 是否使用「negative sampling」來優化
 - threads=1 <int> 訓練時使用的執行緒數量
 - min_count=5 <int> 出現多少次以下(不含)的詞彙就省略
 - alpha=0.025 <float> 就是「learning rate」

• cbow=1 <int> - 使用「cbow(0)」還是「skip-gram(1)」模型

2.2. 將「word2vec」的結果投影到2維的圖



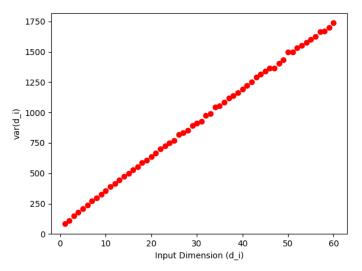
2.3. 從上題視覺化的圖中觀察到了什麼

- 書中角色名字的「embedding」都滿相近的(偏右下角的那一群)
- 魔法學院相關詞,如「students」、「wizards」、「school」等詞彙也有相近的「representation」
- 頭、臉部、五官則是另一群,像是「hair」、「eyes」、「mouth」和「head」
- 上面的「wispered」和「professor」感覺一點關係都沒有,座標卻很接近,猜測要從其他維度才能看出它們的差別

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

- 原理
 - 1. 對每個原始維度 d_i ,用 gen.py 分別產生200次「N=400、 $d_h=\mathrm{randint}(60,80)$ 」的「sample dataset」
 - 2. 此時,每個「sample dataset (d_i,t) 」都有400個 \mathbb{R}^{100} 的向量,40000個數值,定義 $var(d_i,t)$ 為這四萬個數值的變異數,其中 $d_i\in\{1,2,\ldots,60\}$ 且 $t\in\{1,2,\ldots,200\}$

3. 替每個 d_i ,計算 $var(d_i) = rac{\sum_{t=1}^{200} var(d_i,t)}{200}$,可以得到下面的關係圖



- 4. 我們可以發現,隨著原始維度提高,「dataset」中所有數值之變異數的期望值會跟著上升
- 5. 之後,給定一個未知原始維度的「dataset」,我們只要替這個「dataset」中的所有數值 計算變異數,並找出最接近這個數值的 $var(d_i)$,就猜測 d_i 是這個資料集的原始維度
- 6. 在「kaggle public」的表現 MAE=0.11561
- 合理性
 - o 簡單來說,就是使用 gen.py 產生許多的資料集,並記錄原始維度上升對資料集變異數的 影響
 - o 原始維度越高,輸出向量的數值們本來就越容易分歧,與觀察相符
- 通用性
 - 基本上不是很實用
 - o 除非知道如何從原始維度產生資料以及資料在原始向量空間的分布,不然沒辦法大量產生「sample dataset」來計算某個維度的變異數期望值

3.2. 將你的方法做在「hand rotation sequence datatset」上得到什麼結果?合理嗎?

- 得到的結果
 - o 「hand rotation sequence dataset」中共有481張 480×512 的相片,即481個245760維的向量,共118210560個數值
 - \circ 這些數值的變異數約為944.391,最靠近這個數值的 $var(d_i)$ 是990.0857 (當 $d_i=33$)
 - o 因此,猜測這個資料集的原始維度為33
- 我認為不合理,因為這些圖片明顯不是由前一題的方式產生的,觀察到的維度也不同,所以資料集「原始維度與變異數」的關係應該會改變,不能直接拿來使用