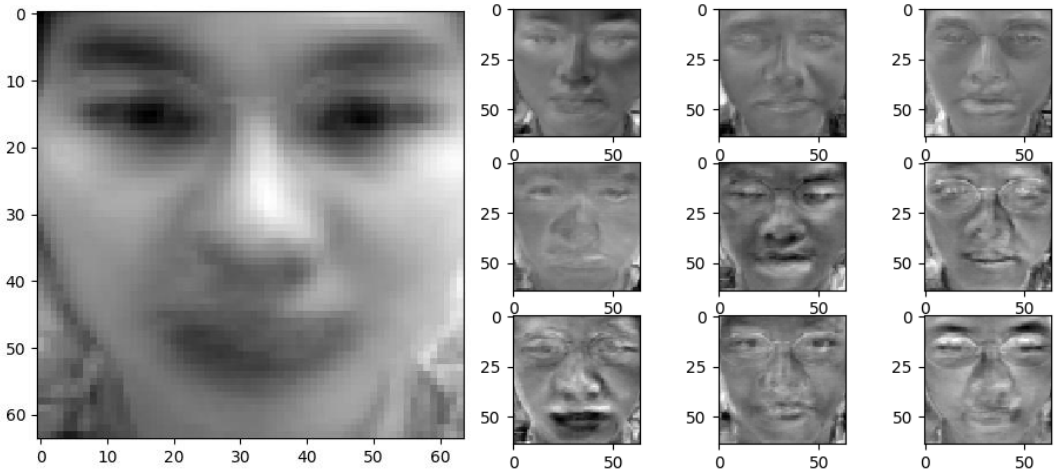
ML2017 HW4 Report

學號：B02901124 系級： 電機四 姓名：黃柏翔

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):

Original: **Reocnstruct:**

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： K = 59

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

min_count = 100:

表示在整份文章裡，一個單詞必須至少出現100次才會被列入考慮

size = 300:

表示要將一個詞彙用一個300維的向量來表示

window = 5:

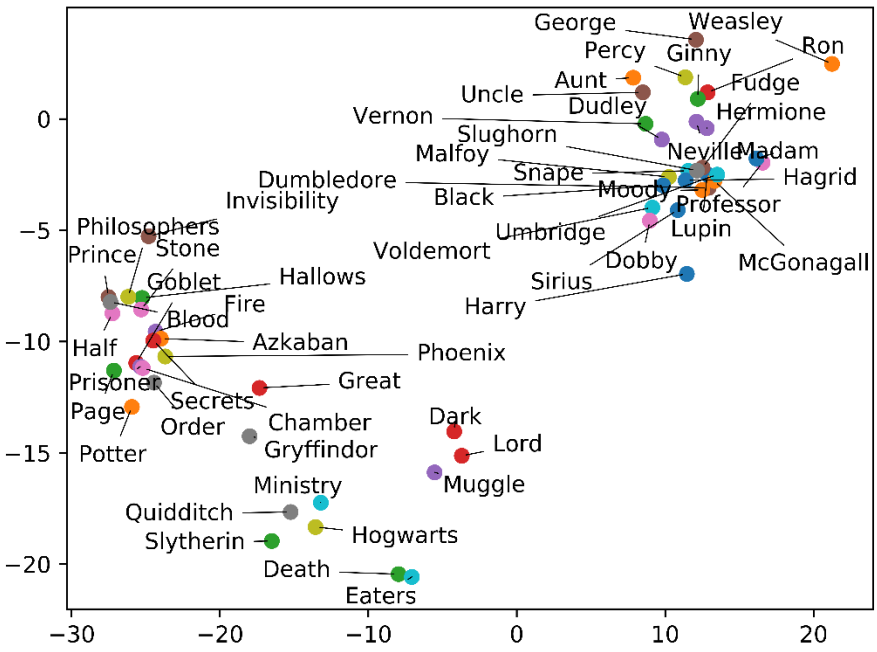
要去計算這些字詞的相關性時，會考慮到單詞附近出現單字，window的大小表示考慮到前後單字的數量。設window為5表示會考慮到這個單詞前後五個單字。

cbow = 1:

使用 Continuous Bag of words model(設為 0)或是 skip-gram model (設為 1)。前者大概是在給定了 input $W_{i-2}W_{i-1}W_{i+1}W_{i+2}$ ，可以找出最高機率的 output W。後者則相反，是給定中間字去預測上下文。

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：

大致上可以看得出詞彙有經過分類。

首先右上角大多是人名的類別，可以看到與哈利比較親近的人(ex.衛斯理家族、妙麗)集中在這個區塊的上半部，而右上角區塊的下半部幾乎是有在霍格華茲任教的一些教授，稍微靠左的地方則有像是恩不理居、佛地魔、馬份這類與哈利算是敵對關係的人。

左邊區塊則有一些類似跟每集主題有關的詞彙(ex.阿茲卡班、混血王子、聖火、神秘石頭、鳳凰會)。

在下方區塊大概有兩類：一個偏向是跟魔法學校相關的詞彙(霍格華茲、葛來分多、史萊哲林、魁地奇)；另一類則是跟佛地魔比較有關的詞彙(黑暗、lord、麻瓜、食死人、死亡)。

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：

我參考 TA 時間所講的方式。先隨機產生出 1~60 個 dimension 的資料，再按照題目所講的方式將這些含有 10K~100K 筆 data 的 set 經由一些 elu 及線性轉換到 100 維度上。利用這些帶有原始 dim 資訊的資料去 train 一個 SVR。並再利用這個 SVR 去預測 “data.npz”的結果。

參數上的調整在實驗後發現取 Sample=5、Neighbor=200、SVR(C=10)時的效果是最好的。

對於這樣的方式我覺得在這個題目上十分的合理，畢竟原本 “data.npz”裡的資料就是用一樣的方式生成的，利用這樣的方式生成的資料去 train SVR 的話，效果必然不會太差。

但對於這個方法的通用性我會抱持著一個懷疑的態度，畢竟很多應用方面的 data 他的資料並不是這麼的隨機，若只是利用隨機產生的資料去 train 出一個 SVR 便要應用到這些 data 上的話感覺並不是這麼合理，因此我認為這個方法的通用性不佳。

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：

為了省時間先將圖片做 down sample 後，利用上題的方式更改 generate data 的維度便開始做 training。最後得到的維度大略等於一維，以結果來說算是十分合理，因為全部都是杯子的圖片作旋轉差異只在轉動的角度上，因此只需要一個維度就可以表示了。