**Machine Learning 2017 HW4 Report**

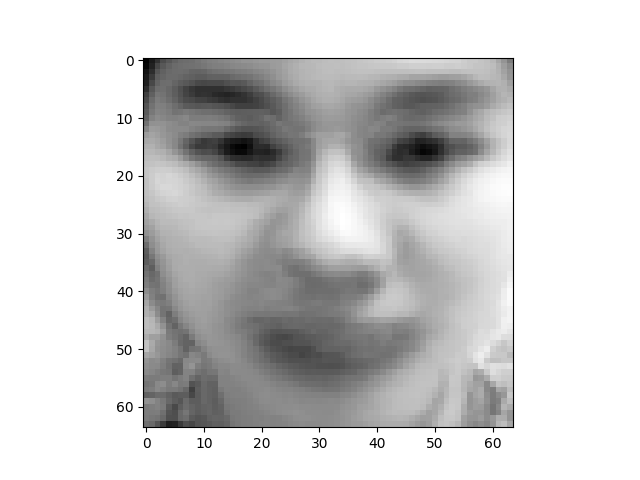
*Predict PM 2.5*

B03901156 Yu Xuan Huang

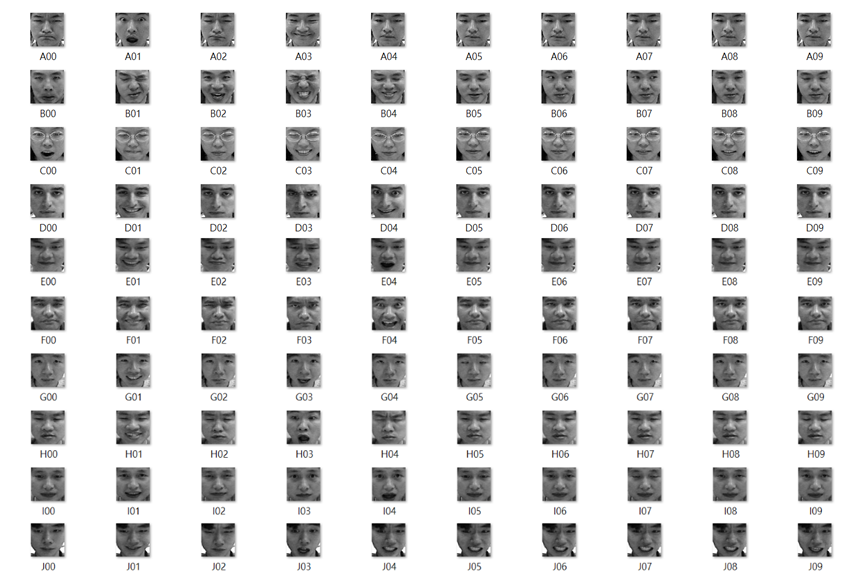
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

ANS.

Average Face 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):



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.

Ans. From my experiment, even when k = 100, the reconstruction error is still >1%; when k =100, the smallest reconstruction error = 0.017834625142 can be obtained.

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

ANS. train : the file path of the training data

output: the file path of the output saved model

size: 向量的維數，此為 100

window: 上下文窗口大小，此為default值=5

negative: 負例的數目 (用於negative sampling), 此設為0

min\_count: 被捨棄/截斷的低頻詞閥值, 此設為5

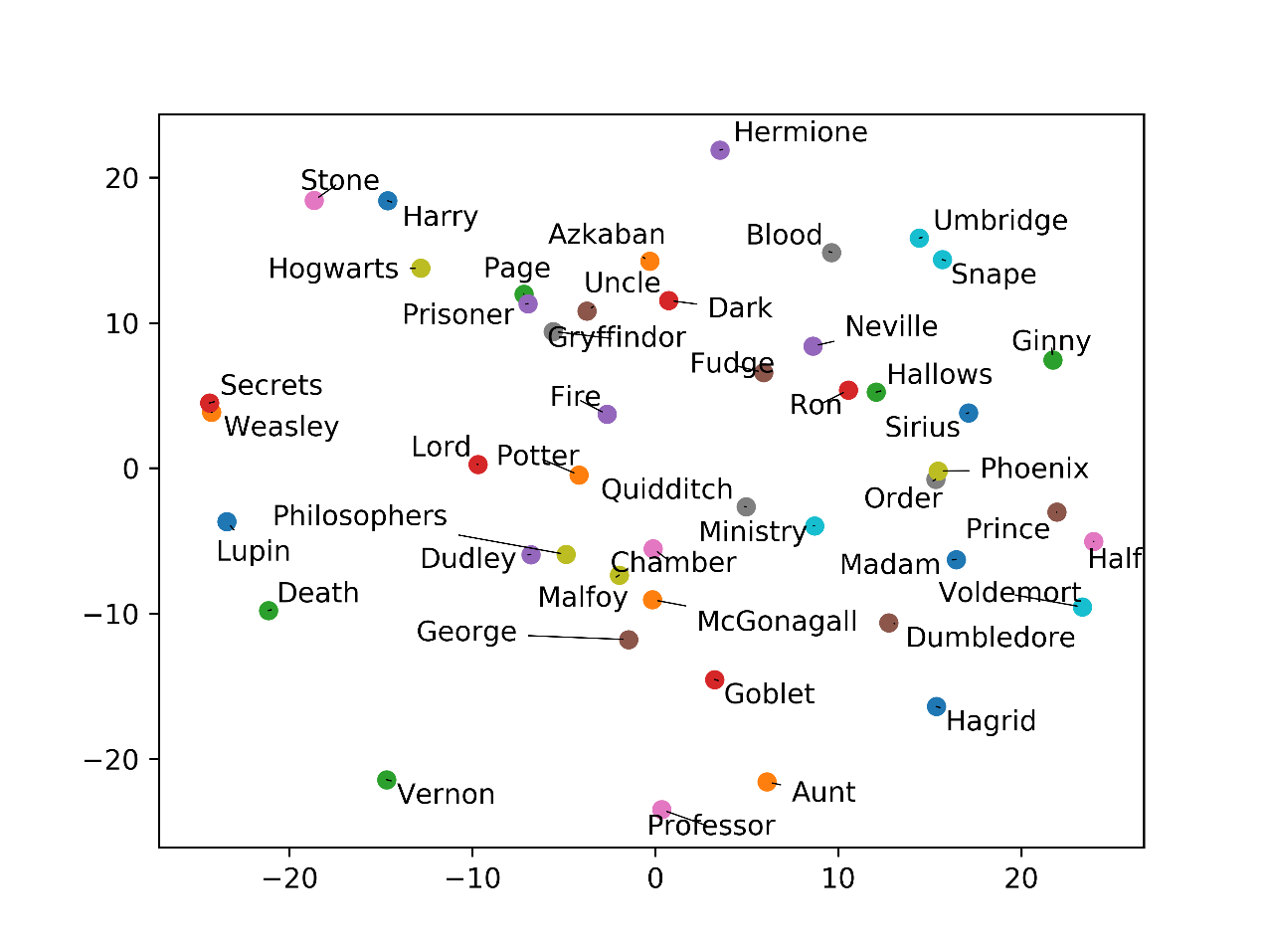
alpha: 起始的learning rate, 此設為0.025

iter: number of training iteration, 此設為10000

cbow: 是否使用CBOW算法(0為不使用), 此設為1

2.2. 將 word2vec 的結果投影到 2 維的圖:

ANS.



2.3. 從上題視覺化的圖中觀察到了什麼？

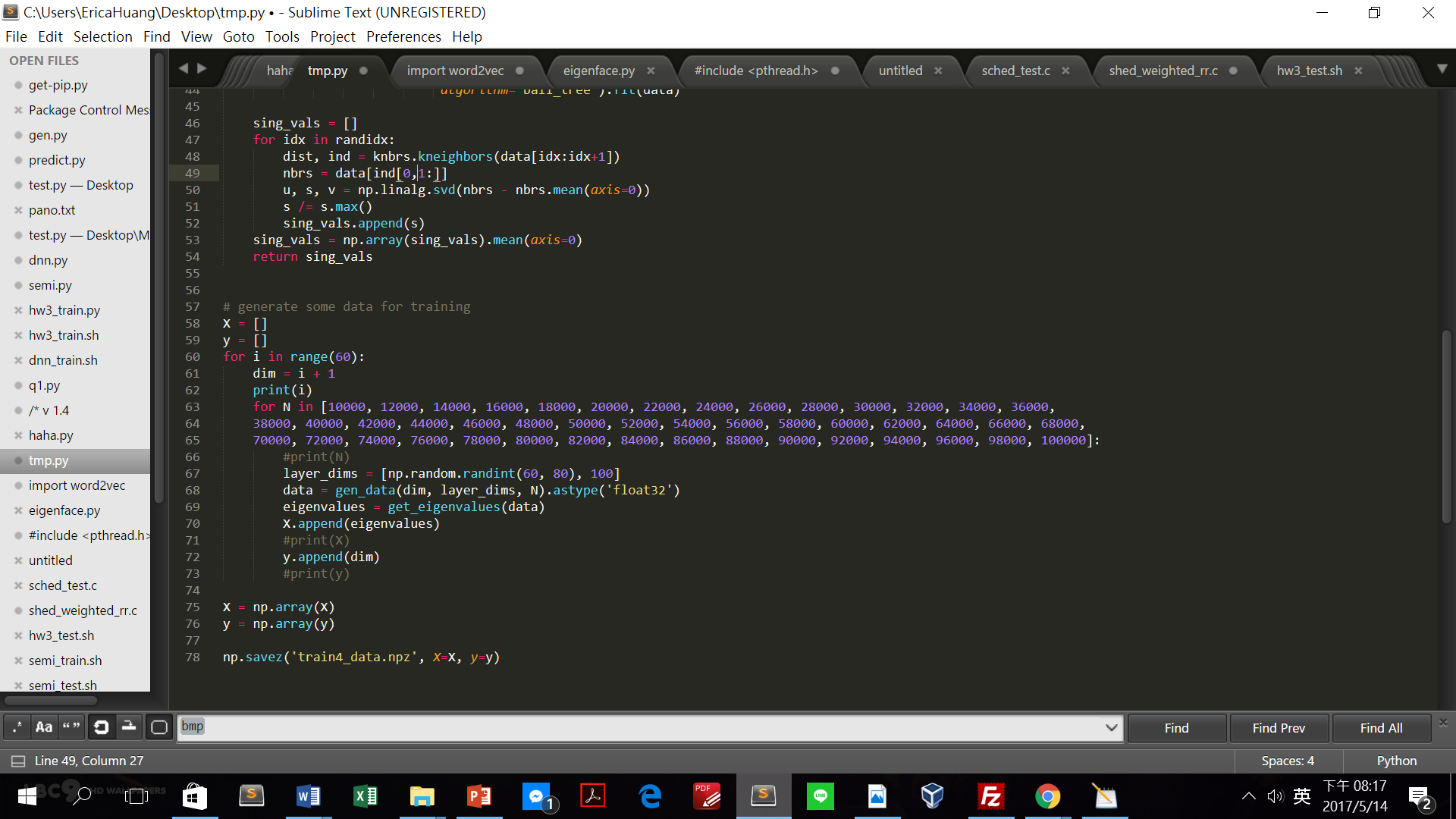
ANS. 詞類較為相近的字詞理應較為接近(如上圖中紅圈處，Page 和 Prisoner, 或Phoenix和Order)，但從此圖觀察可知，其實驗結果的分布點較為分散，較難看出相似詞類的集中程度

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？

ANS.

1. 根據題目所給的dataset資訊，如維度分布、oracle network結構產生一些training data

2. 產生不同維度的training data



3. 從所產生的每個dataset中隨機選取20個sample points，對於每個point找出200個nearest neighbors

4. 從data subset 中計算平均的eigenvalues

5.利用線性的SVR去預測testing dataset 的維度

合理性: 藉由產生類似原始資料性質的data作為training data，模擬其資料特性，再利用SVR預測未知的資料維度，相信這樣的model是能有效反映資料特性

通用性: 由於此model 是根據此dataset的原始資料性質所設計的，若用於估計具相似特性的dataset 效果將會不錯，但若性質差異較大，則預計其估計結果較差

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：

我認為其結果較為不合理，可能是因為此model 是根據此dataset的原始資料性質所設計的，而hand rotation data的資料性質與之差異較大，其估計結果較差