

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):

答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)



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 = 59

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

我的參數：window=10, sample=1e-3, size=500, hs=0, negative=10, cbow=1

size=500 表示 word vector 的長度是 500

cbow=1 表示我用的模型是 skip-gram

window=10 表示 context 取的範圍是字的前10個字和後10個字

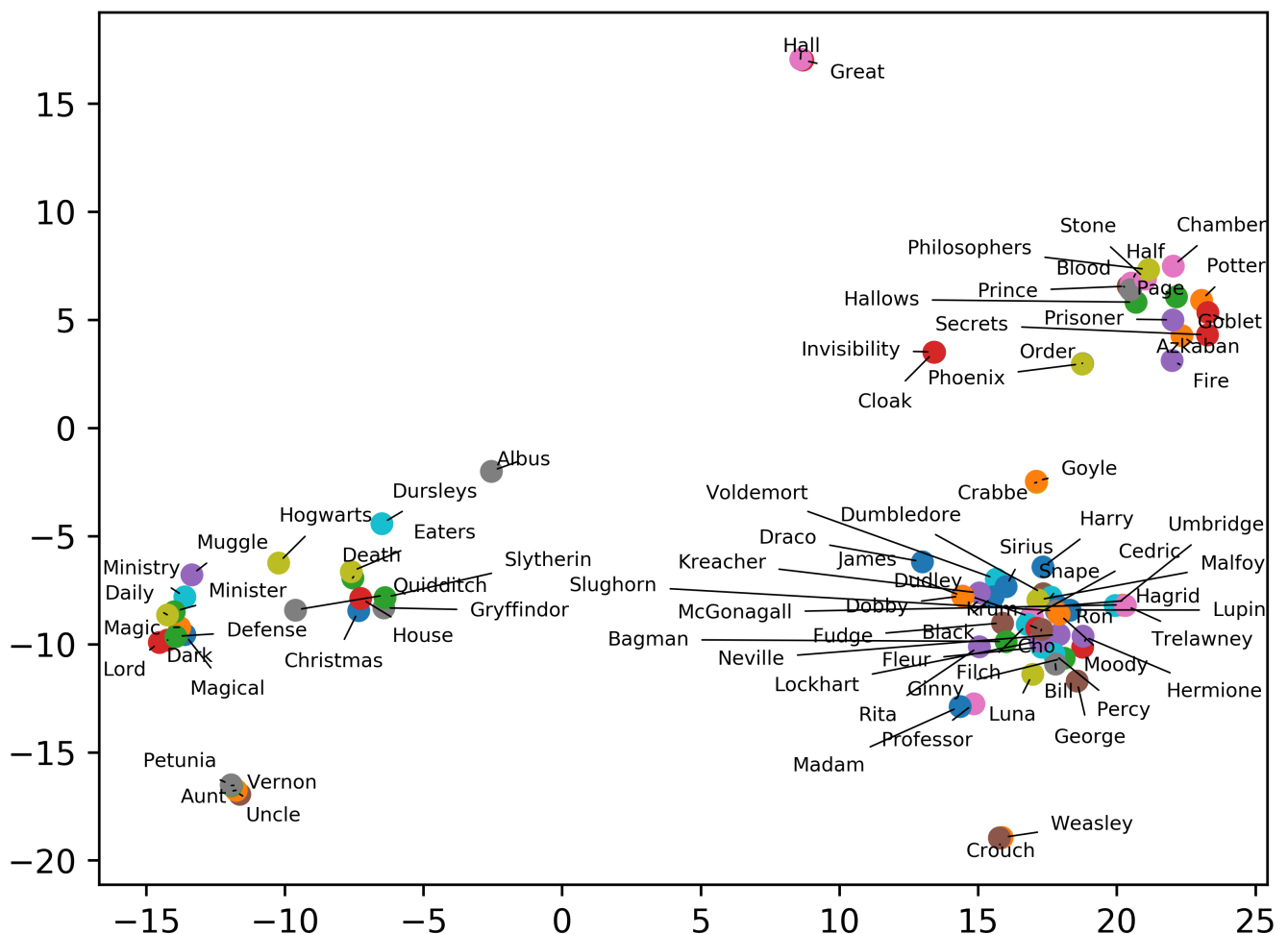
sample=1e-3 表示 subsampling 的 threshold，頻率超過的字會被 subsample

hs=0 表示不用 hierarchical softmax

negative sampling 表示每次在 train 的時候隨機製造 10 個 noise，minimize noises' probability and maximize context's probability.

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：

意思相近的字的點會比較靠近，如 (Aunt, Uncle), (Gryffindor, Slytherin), (Azkaban, Prisoner)

因為這是從小說做出來的結果，許多字的關係會比較偏向於在小說中的意思，如 (Cloak, Invisibility)

小說裡的常出現的那個斗篷有著隱形的能力，不斷的出現，使斗篷幾乎成了隱形能力的代名詞，所以這兩個字的詞向量很接近。

這次的 corpus 比較小，結果不是很好，有些詞向量很接近，但是難以理解他們彼此間的關係，如 (Diary, Ministry)

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：

我用的是助教的方法，隨機取 30 個點，找他鄰近 50 個鄰居，算出 eigenvalue 值後用 linear svr ($C = 75$) train。估計時也用同樣方法算出 eigenvalue 值，丟進 linear svr 估計原始維度。我覺得這樣蠻合理的，用大量資料去訓練可以大概知道資料與原始維度的關係。

這方法的通用性我想是還不錯，kaggle 上的分數是 0.09841

3.2. 將你的方法做在 hand rotation sequence dataset 上得到什麼結果？合理嗎？請討論之。

答：

我算出的原始維度是7維

我認為還算合理，人腦判斷大概是3或4維，因為這只是一隻手的不同角度、擺在不同位置，或許加上些微的不同。考慮到機器比較難從手心、手背而得知是同一隻手，我覺得算出7維還算是合理的。