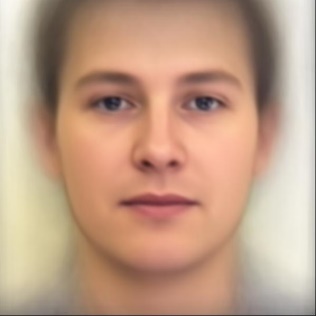
學號：B04901060 系級： 電機三 姓名：黃文璁

1. **PCA of colored faces (No collaborators)**
   1. **(.5%) 請畫出所有臉的平均。**



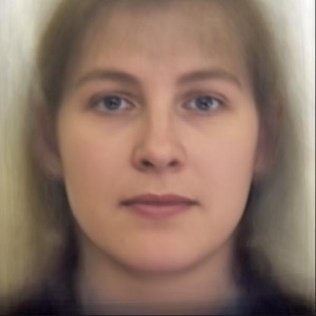
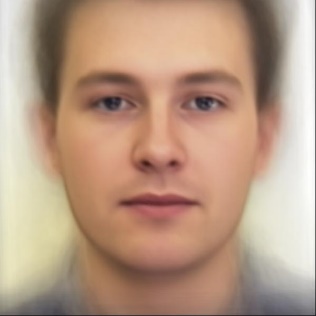
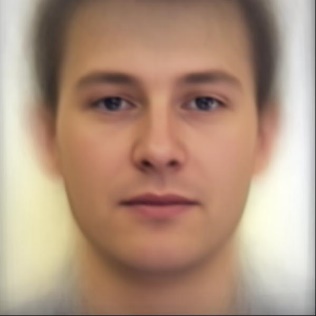
* 1. **(.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。**

Eigenface 0 Eigenface 1 Eigenface 2 Eigenface 3



* 1. **(.5%) 請從數據集中挑出任意四個圖片，並用前四大Eigenfaces進行reconstruction，並畫出結果。**

23.jpg 96.jpg 187.jpg 253.jpg



reconstructed

* 1. **(.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。**

Eigenface 0: **4.1%**

Eigenface 1: **2.9%**

Eigenface 2: **2.4%**

Eigenface 3: **2.2%**

1. **Visualization of Chinese word embedding (No collaborators)**
   1. **(.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。**

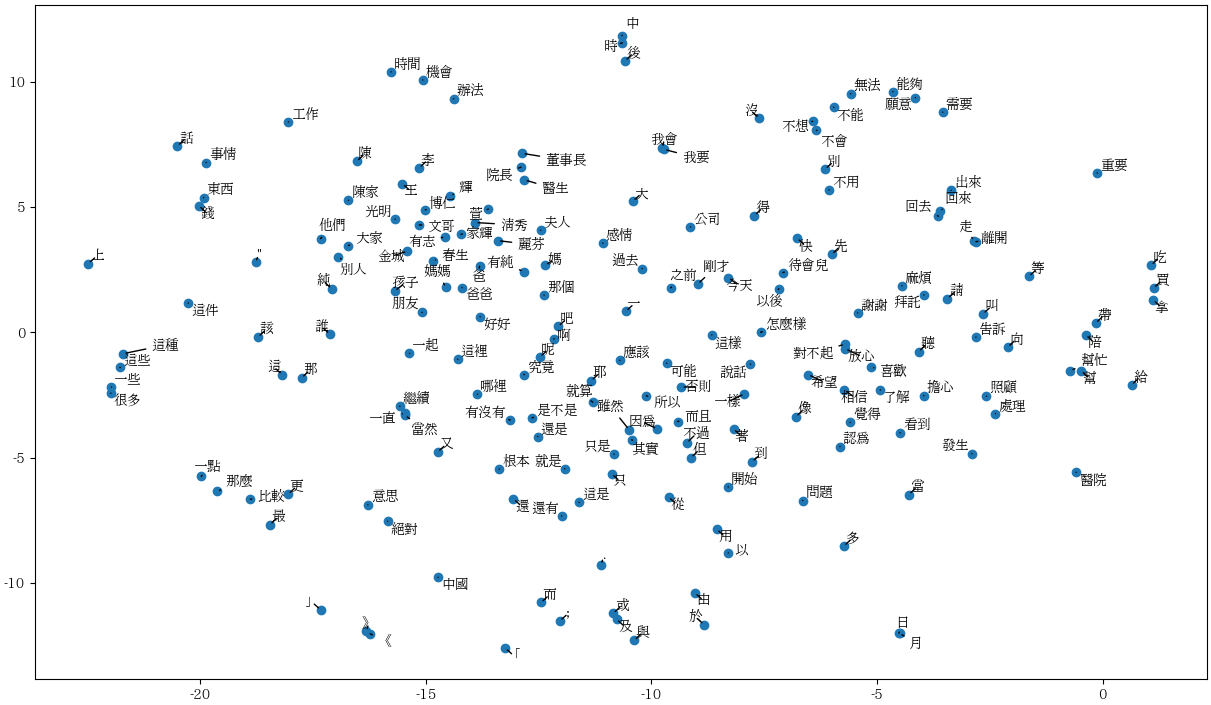
使用gensim.models.word2vec。

model = Word2Vec(lines, size=300, min\_count=16, workers=8, iter=20)

其中lines為句子，size為word vector的維度，min\_count為training時只考慮出現次數超過這個數的詞，iter為training過程迭代的次數。（workers為training時用到的thread數）

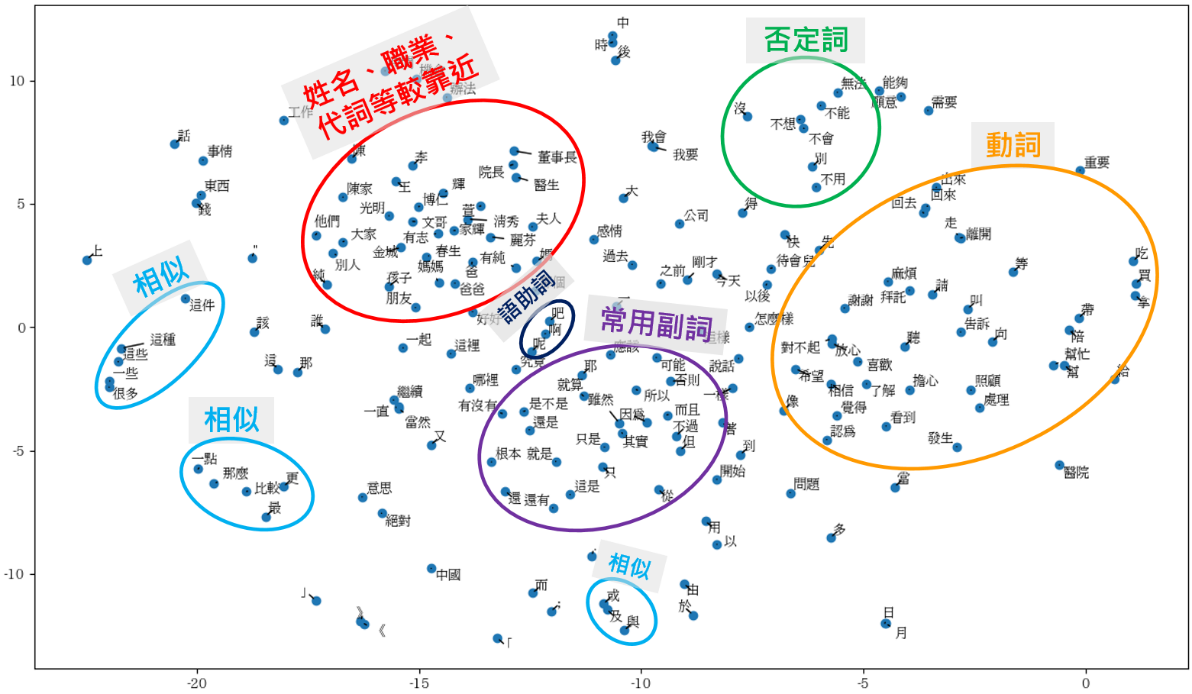
* 1. **(.5%) 請在 Report 上放上你 visualization 的結果。**

取頻率介於2000～8000的詞。

****

* 1. **(.5%) 請討論你從 visualization 的結果觀察到什麼。**

下圖大致將上圖中一些比較明顯的關聯標示出來。整體來說詞性或詞意接近的詞會更靠近。



1. **Image clustering (No collaborators)**
   1. **(.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)**

**- 方法一（Best model）：**

利用AutoEncoder降維至32維，接著利用K-means分成20個cluster，接著再人眼判斷這20個cluster分別屬於哪個dataset，最後再計算答案。

**結果**：Kaggle上F1 score為：1.00000

**- 方法二：**

　　利用AutoEncoder降維至32維，接著利用K-means直接分成2個cluster。

**結果**：Kaggle上F1 score為：0.44182

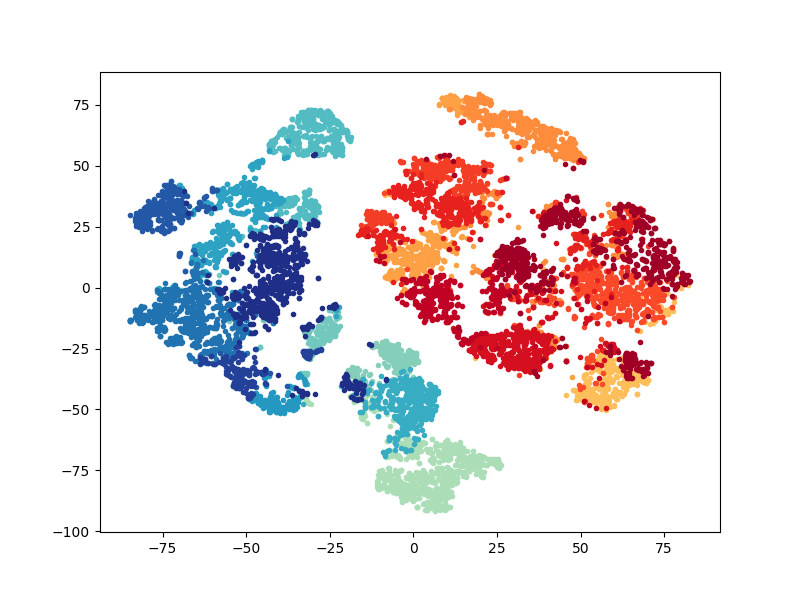
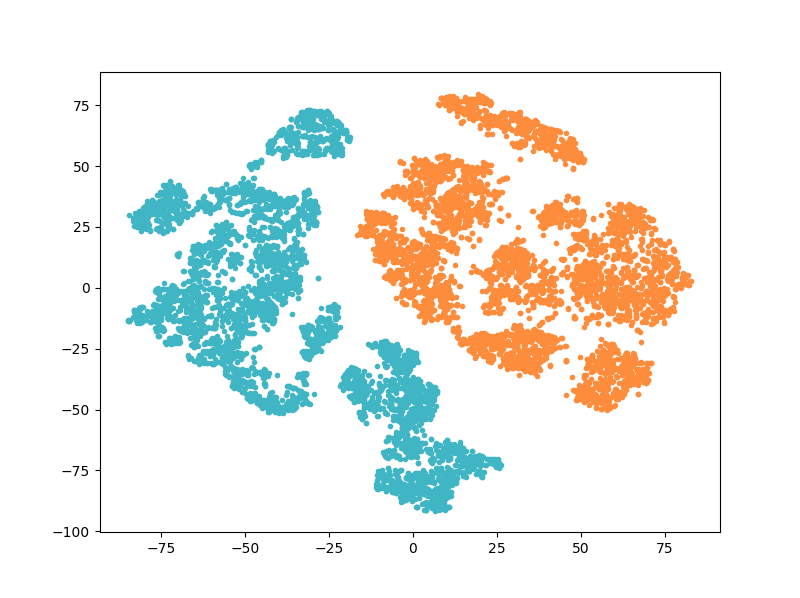
**- 方法三：**

利用Convolutional AutoEncoder降維至32維，接著利用K-means分成20個cluster，再人眼判斷這20個cluster分別屬於哪個dataset。

**結果**：Kaggle上F1 score為：0.42328

**討論：**若先分成更多個cluster而不是直接分成2個cluster的話，可以更準確的將兩個dataset分開，這應該是由於降維後的向量並不完全將兩個dataset分成兩群（例如不同數字可能屬於不同群，但屬於同一個dataset）。此外CAE效果不如AE的原因可能是由於CAE降維後的結果不如AE來的連續，故較難進一步利用K-means分群。

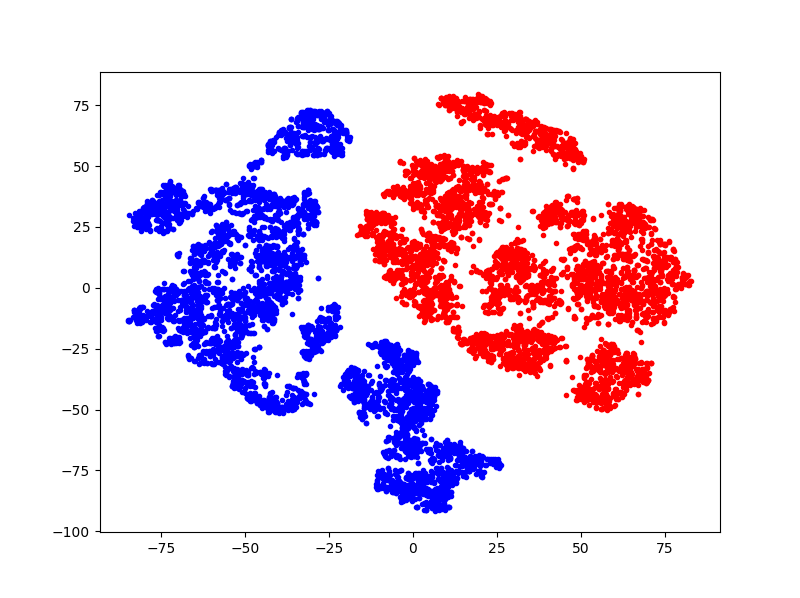
* 1. **(.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。**

左圖為預測出的20個cluster，分別用不同顏色標記。

右圖為將這20個cluster用人眼判斷後再分為兩群後的結果。

* 1. **(.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。**



紅色為前5000張圖，藍色為後5000張。注意到上題第二張圖已完全將兩個dataset分開，和本題ground truth相比是一樣的，而且兩個 dataset之間有一道很明顯的界線。