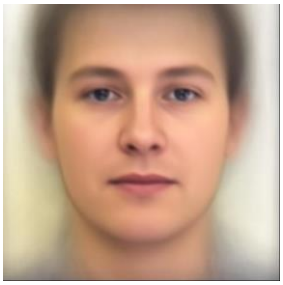
學號：B04901060 系級： 電機三 姓名：黃文璁

A. PCA of colored faces (No collaborators)

A.1. (.5%) 請畫出所有臉的平均。



A.2. (.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

A.3. (.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

A.4. (.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%

B. Visualization of Chinese word embedding (No collaborators)

B.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。

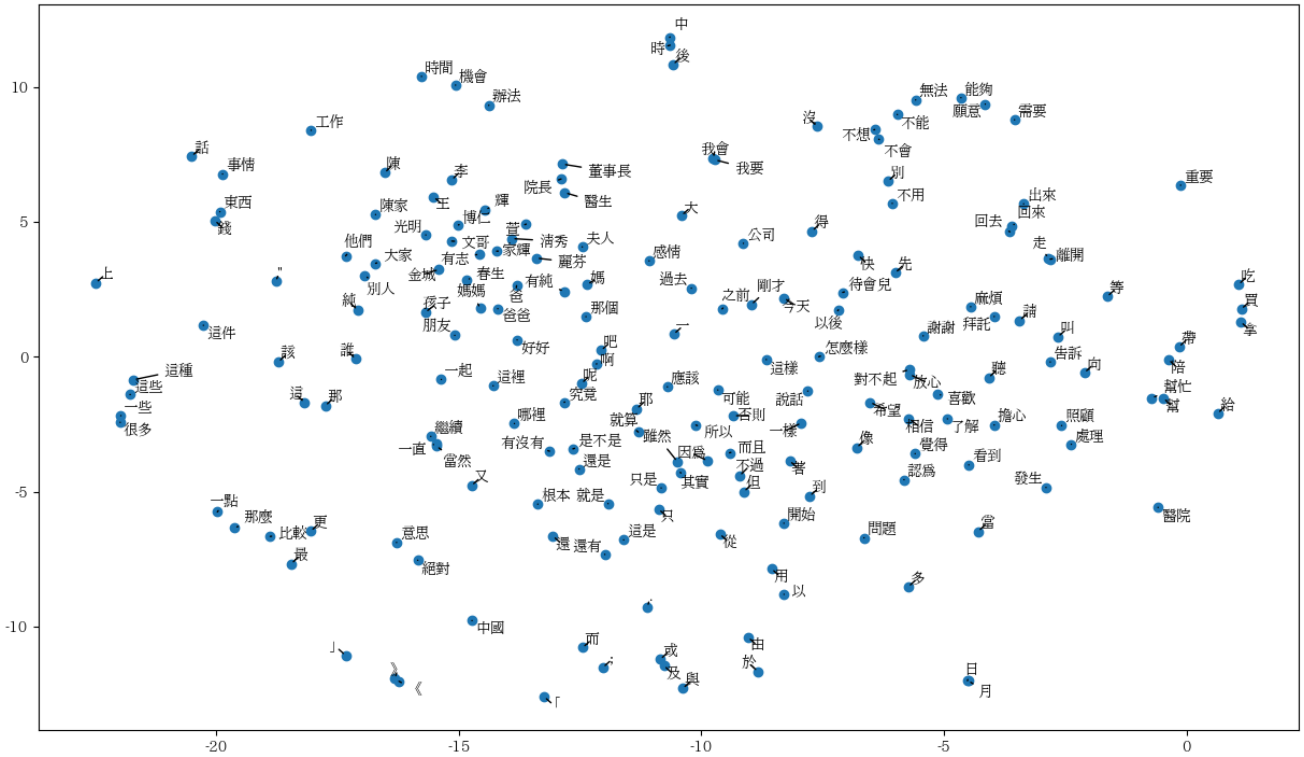
使用 gensim.models.word2vec。

```
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）

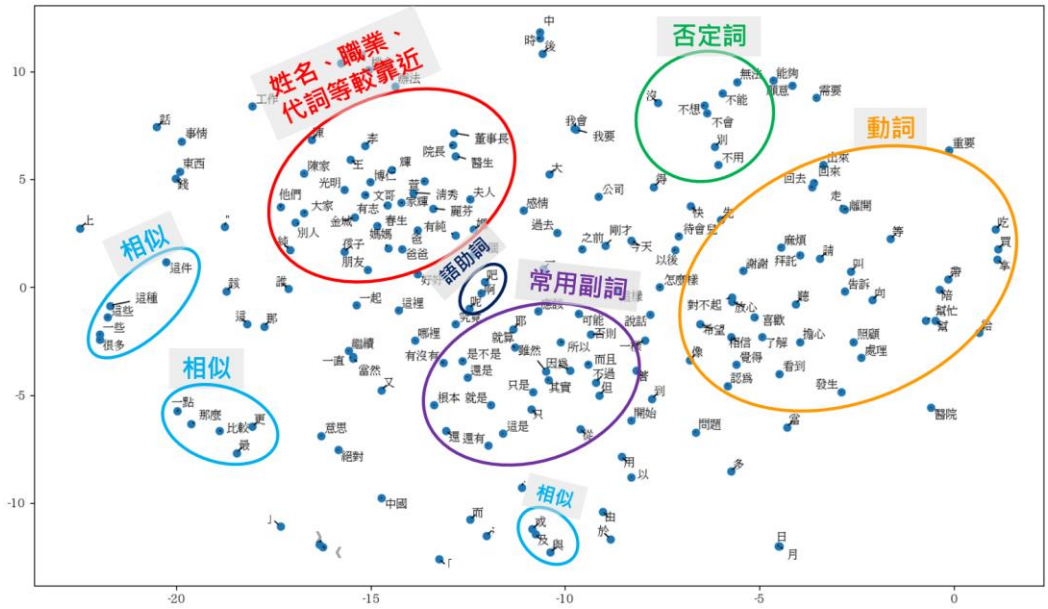
B.2. (.5%) 請在 Report 上放上你 visualization 的結果。

取頻率介於 2000～8000 的詞。



B.3. (.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。



C. Image clustering (No collaborators)

C.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）：

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

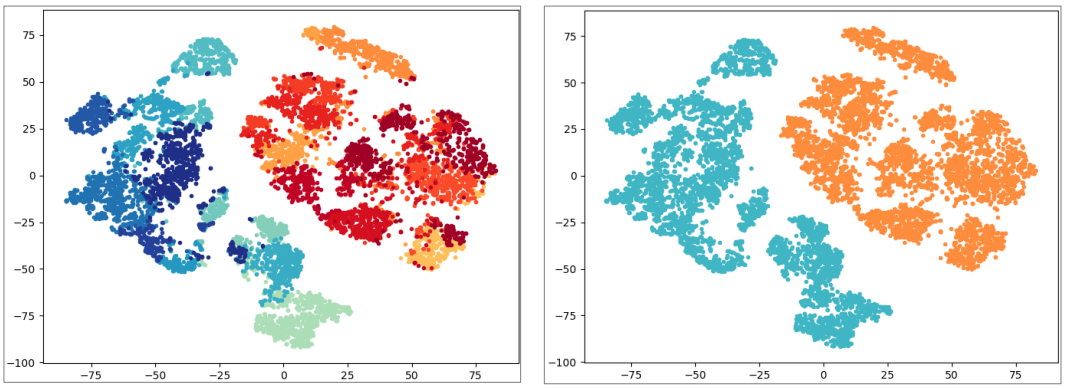
- 方法三：

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。

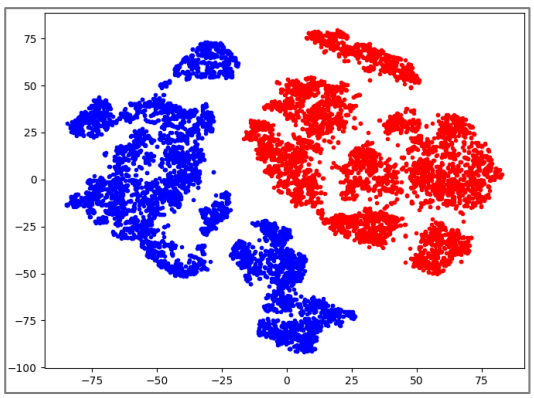
C.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



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。

C.3. (.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。