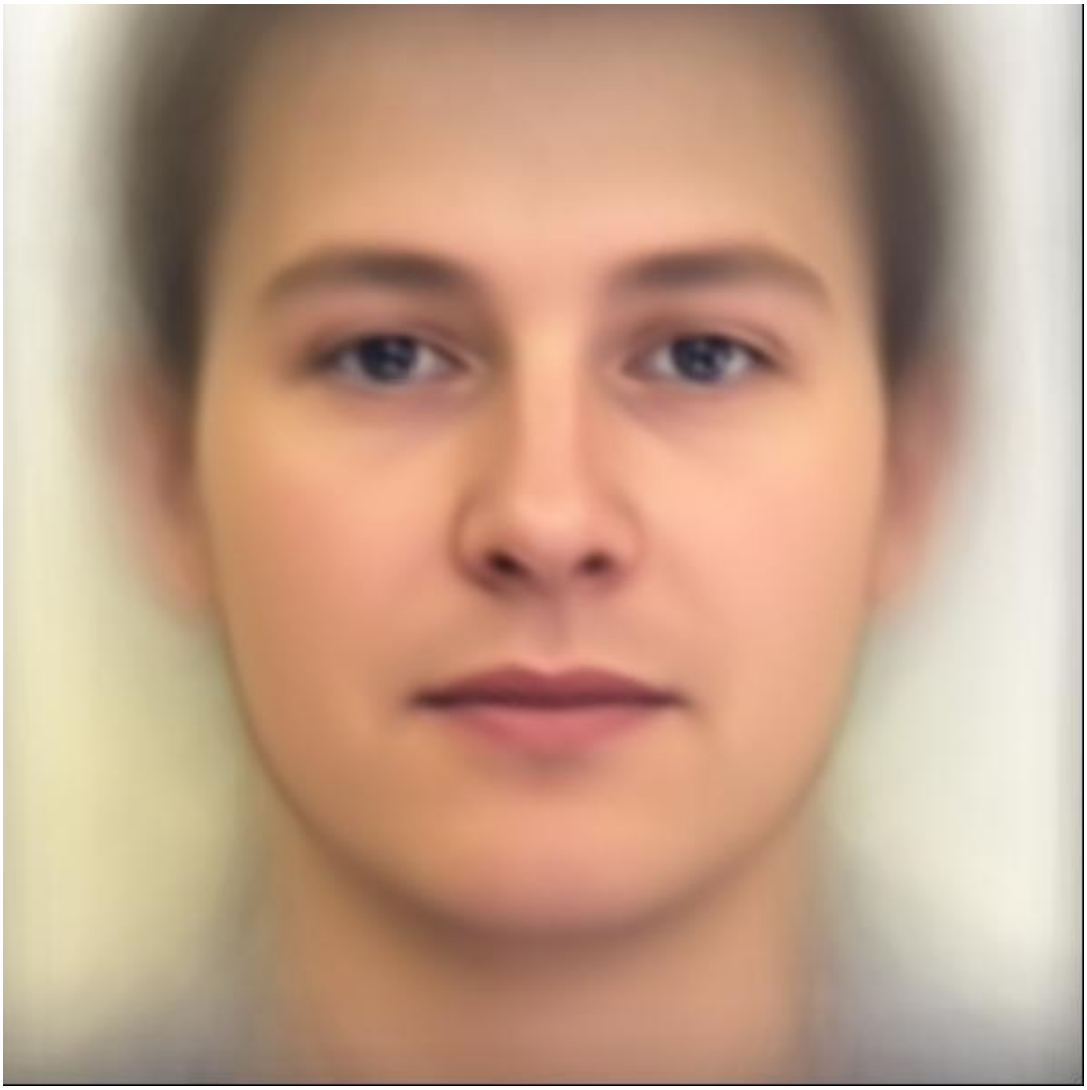
學號：R05943138 系級： 電子所碩二 姓名：賴又誠

A. PCA of colored faces

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。

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。

(a) 原始圖片

(b) Reconstruction 圖(使用前四大 eigenfaces)

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。

比例	4.1%	2.9%	2.4%	2.2%

B. Visualization of Chinese word embedding

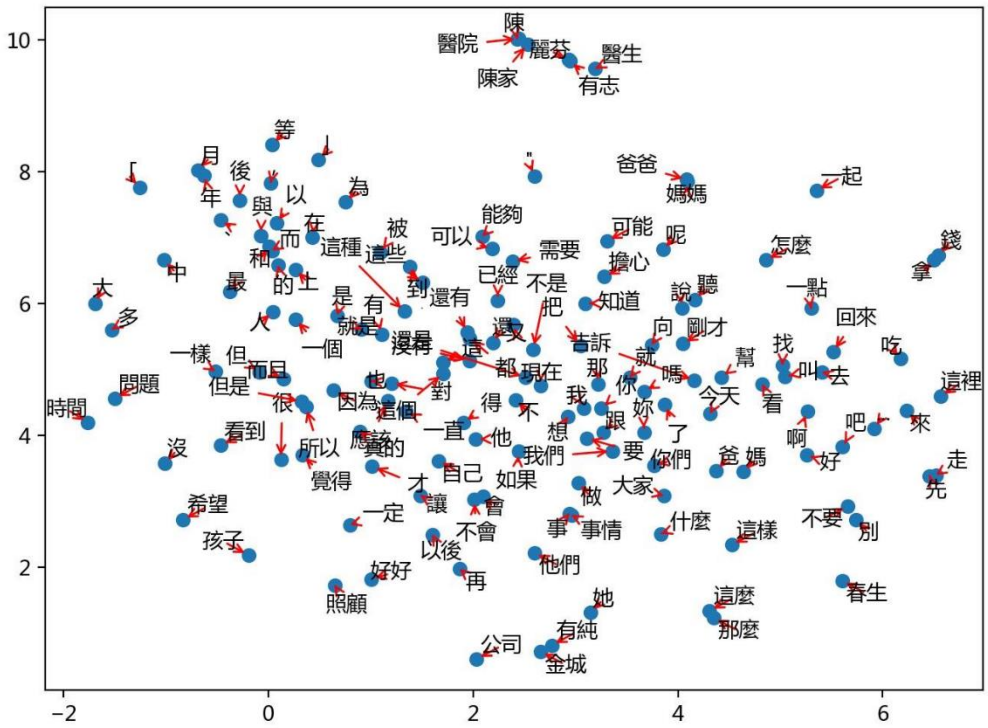
B.1. (.5%) 請說明你用哪一個 word2vec 套件，並針對你有調整的參數說明那個參數的意義。

我使用助教所建議的 jieba 的 dict.text.big 做分詞，之後再使用 gensim 當作我的 word2vec 套件做 word embedding，code 如下:

```
from gensim.models import word2vec
model = word2vec.Word2Vec(sentences, min_count=3100, window=5, size=120)
```

我所調整的是 min_count，其參數意義是某個詞出現的頻率超過 min_count 才把該詞存入字典中，而我設計在超過 3100 才存入字典內。

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



B.3. (.5%) 請討論你從 visualization 的結果觀察到什麼。

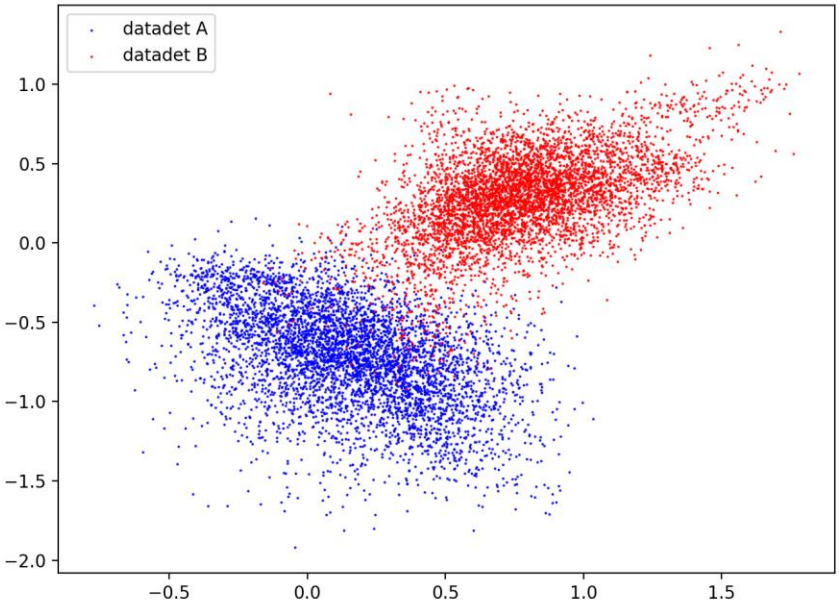
由(x,y)=(3,10)處可發現，主角名子相關的字眼都放在一起，(x,y)=(6.5,6)則是拿的下一個字是:錢，(x,y)=(4.2,3.5)是爸、媽等等意思較為接近在附近，等等的結果顯示 word embedding 做得還算可以。

C. Image clustering

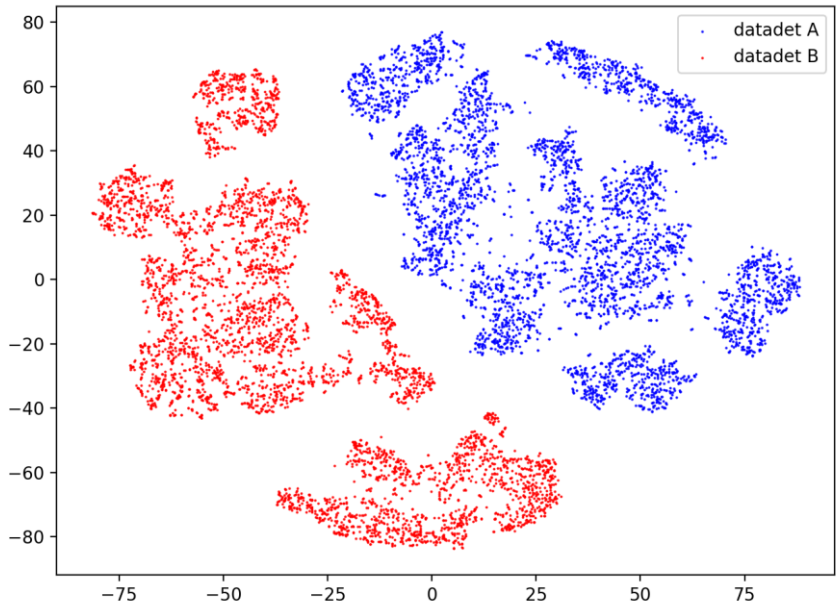
C.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

Model	Auto-encoder	PCA	KMean	Acc(private)
1	✓ (降到 2 維)	✗	✓	0.03894
2	✓ (降到 32 維)	✓ (降到 2 維)	✓	1

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



(a) 取降到 32 維的 feature 前兩個維度作圖



(b) 把降到 32 維後的 feature 再用 t-SNE 投影到二維

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。

我使用的方法是 auto-encoder 降成 32 維，再用 t-SNE 投影到二維，再使用 KMean 做分類，再把分類完的資料的前 5000 張的 label 全部加起來剛好等於 5000，後 5000 張的 label 加起來也剛好是 0，可知道完全分類正確，再由(b)圖可看出 datasetA 和 datasetB 分的很開，結果和真正的 label 一模一樣，完全正確把 dataset 分為 A，B 兩類。