





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。

Eigenface1	Eigenface2
	
Eigenface3	Eigenface4
	

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。

Reconstruction24	Reconstruction92
	
Reconstruction130	Reconstruction206
	

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重 (explained variance ratio)，請四捨五入到小數點後一位。

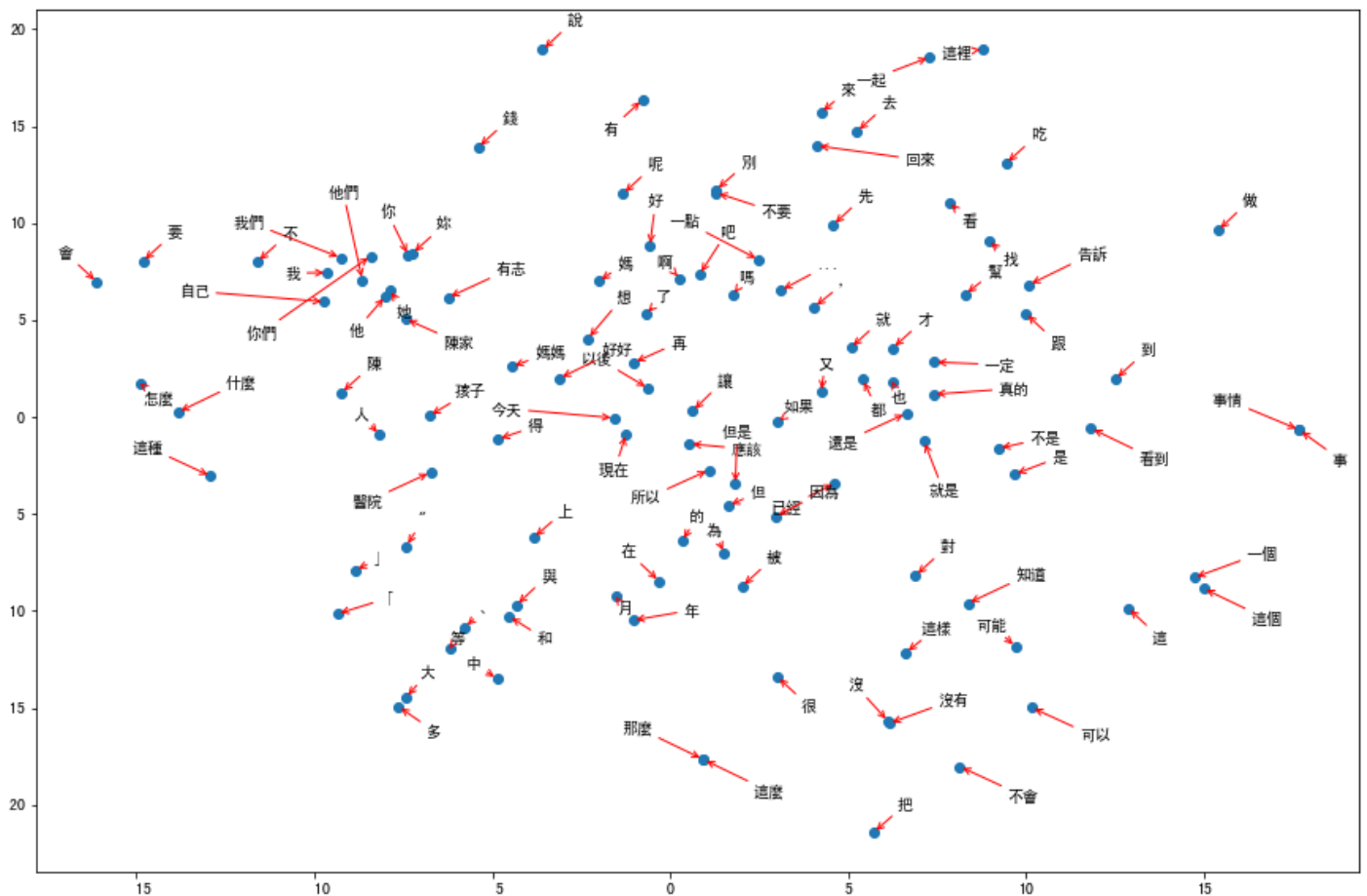
1. 4.2%
2. 3.0%
3. 2.4%
4. 2.2%

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。

使用gensim的word2vec，調整了min_count = 1500, 代表只會對出現1500次以上的詞作計算，size = 128，將每個詞編成128維的向量。

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。

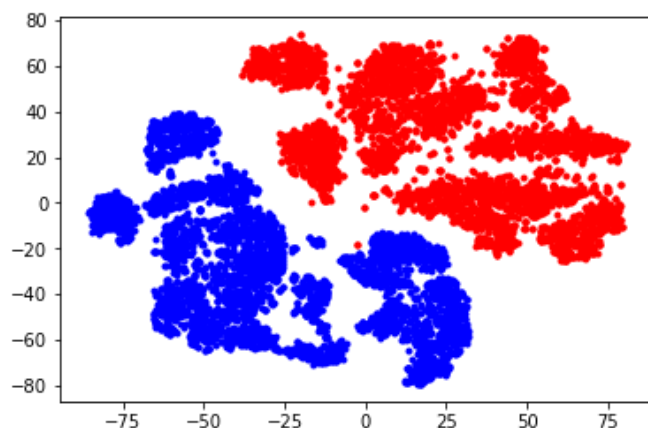
意思或性質比較接近的詞，編成向量後，他們在2維平面上的分布也會比較接近，例如圖中右邊的'事情'跟'事'就幾乎重疊在一起，而左邊的[你,妳,他,他們,你們,我]等等代名詞也都比較接近，可以看出word2vec的向量編法是有考慮詞的意義的。

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)

	Autoencoder + Kmeans	PCA + Kmeans
kaggle public	1.00	0.03
kaggle private	1.00	0.03

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



預測準確率為100%

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。

左圖是我用autoencoder編成384維向量後用TSNE降成兩維視覺化後的正確結果(100%準確率), 但可以看到其實有一小塊紅點的位置怪怪的, 我使用自己384維向量下去做kmeans時, 可以得到100%的正確率, 但如果是在TSNE之後用2維的向量去做kmeans的話, 會有10~15%左右的誤差, 如右圖:

