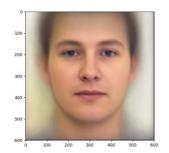
## Part A: PCA of colored faces

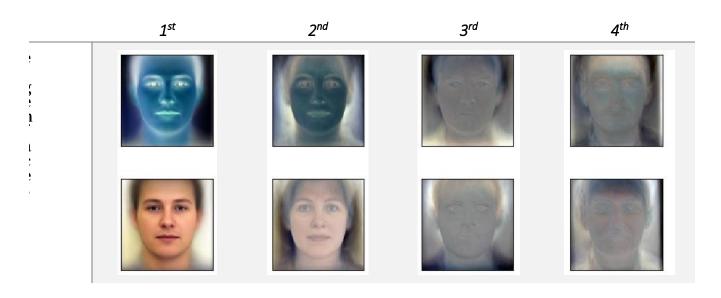
Collaborators: 黃禹程 R06944034、鄭克宣 R06921083、蔡孟庭 R05922078

▶ A.1 (.5%) 請畫出所有臉的平均。

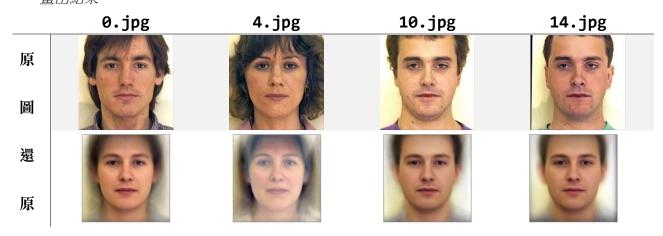


■ 將415 張600\*600\*3 之影像加總並平均之

▶ A.2 (.5%) 請畫出前四個 Eigenfaces,也就是對應到前四大 Eigenvalues 的 Eigenvectors。



- 由左到右依序為最大、次大、第三、第四大之 eigenface
- *礙於計算資源有限,影像已先* resize 成 100\*100\*3
- 為方便助教批改,故列舉正反之兩種可能性
- ➤ 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.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 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>
比重	7.3*10^(-2)	3.7*10^(-2)	2.8*10^(-2)	2.3*10^(-2)

- *礙於計算資源有限,上述數據 resize 成 100\*100\*3 之結果*
- 分別為s[0]/s.sum() \s[1]/s.sum() \s[2]/s.sum() \s[3]/s.sum()

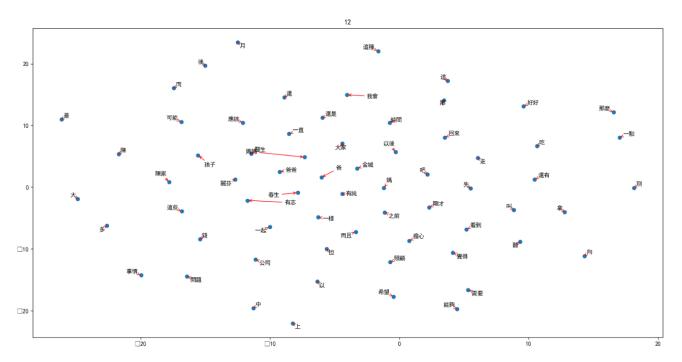
## Part B: Visualization of Chinese word embedding

Collaborators: 黃禹程 R0694403、鄭克宣 R06921083、蔡孟庭 R05922078

- ▶ 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。
  - Word vector model 使用 gensim
  - *斷詞則採 Jieba (dict=dict.txt.big),並用 TSNE 降至二維*

参數	數值	意避
MIN_COUNT	5	一字需至少需出現 5 次才會造其 word vector,藉
		此略除較不重要、罕見之字
WINDOW	5	規範一字與其他字之相鄰範圍之大小
SIZE	100	Word vector之大小,此處則以(1, 100)表示一字

▶ 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。
  - 按助教之建議,僅針對出現次數介於 3000~6000 次之字降維並繪製
  - 「可能」、「應該」、「這些」…等常見詞彙皆出現於圖上
  - 「爸爸」、「爸」、「媽媽」、「媽」等字大抵成平行

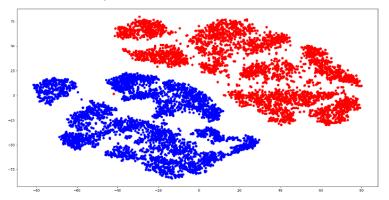
## Part C: Image clustering

Collaborators: 黃禹程 R06944034、丁縉楷 R06922129、葉孟元 R04921094

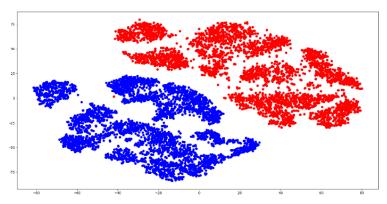
▶ C.1 (.5%) 請比較至少兩種不同的 feature extraction 及其結果。

方法	PRIVATE 分數	PUBLIC 分數	平均
AUTOENCODERS+KMEANS	1	1	1
PCA+KMEANS	0.03040	0.03020	0.03030

- 採TSNE 降維,並用 matplotlib 之 scatter 繪圖
- 紅色來自 dataset A,藍色來自 dataset B,於圖上呈兩大區塊分布
- Autoencoder 參考助教之設計,PCA 之參數則是 n\_components=3
- Autoencoder 約在 train epoch=500 約可達到 0.93,而附上之版本則為 epoch=14900 次
- 與同學討論時,發現PCA亦有潛力提高準確率,礙於時間有限並無成功實作
- ▶ C.2 (.5%) 預測 visualization.npy 中的 Label,在二維平面上視覺化 Label 的分佈。



- 採TSNE 降維,並用 matplotlib 之 scatter 繪圖
- 紅色來自 dataset A,藍色來自 dataset B,於圖上呈兩大區塊
- C.3 (.5%) visualization.npy 中前 5000 個 images 來自 dataset A,後 5000 個 images 來 自 dataset B。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。



- 紅色來自 dataset A,藍色來自 dataset B,於圖上亦呈兩大區塊
- 與自行預測的 Label (上上圖) 相似度極高,應該是 model 於 kaggle 上高準確率之表現有關