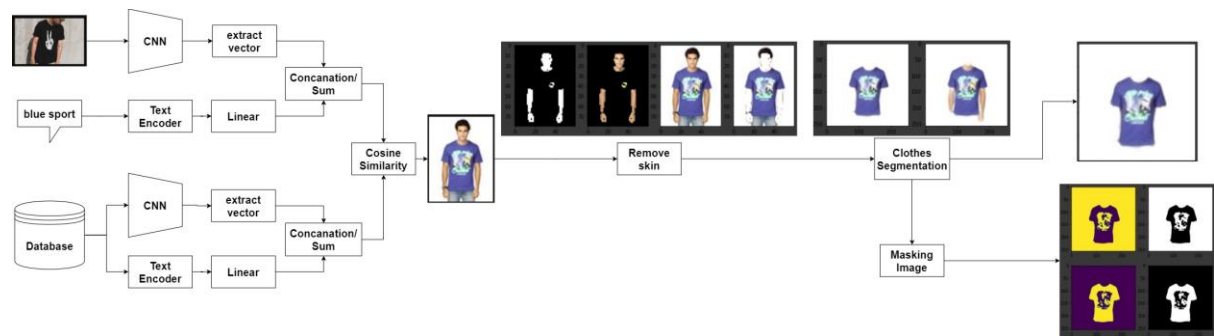


推薦系統



推薦系統流程圖

資料整理

1. 利用 pandas 將資料讀取成 dataframe，並選取類別為上衣的部分

```
[ ] # select topwear
df = df.loc[(df['masterCategory'] == "Apparel") & (df['subCategory'] == "Topwear")]
df.head(10)
```

| | id | gender | masterCategory | subCategory | articleType | baseColour | season | year | usage | productDisplayName | image |
|----|-------|--------|----------------|-------------|-------------|------------|--------|--------|--------|---|------------------|
| 0 | 15970 | Men | Apparel | Topwear | Shirts | Navy Blue | Fall | 2011.0 | Casual | Turtle Check Men Navy Blue Shirt | images/15970.jpg |
| 4 | 53759 | Men | Apparel | Topwear | Tshirts | Grey | Summer | 2012.0 | Casual | Puma Men Grey T-shirt | images/53759.jpg |
| 5 | 1855 | Men | Apparel | Topwear | Tshirts | Grey | Summer | 2011.0 | Casual | Inkfruit Mens Chain Reaction T-shirt | images/1855.jpg |
| 6 | 30805 | Men | Apparel | Topwear | Shirts | Green | Summer | 2012.0 | Ethnic | Fabindia Men Striped Green Shirt | images/30805.jpg |
| 7 | 26960 | Women | Apparel | Topwear | Shirts | Purple | Summer | 2012.0 | Casual | Jealous 21 Women Purple Shirt | images/26960.jpg |
| 15 | 12369 | Men | Apparel | Topwear | Shirts | Purple | Fall | 2011.0 | Formal | Reid & Taylor Men Check Purple Shirts | images/12369.jpg |
| 17 | 42419 | Girls | Apparel | Topwear | Tops | White | Summer | 2012.0 | Casual | Gini and Jony Girls Knit White Top | images/42419.jpg |
| 23 | 13089 | Men | Apparel | Topwear | Sweatshirts | Grey | Fall | 2011.0 | Sports | ADIDAS Men Lfc Auth Hood Grey Sweatshirts | images/13089.jpg |
| 27 | 7990 | Men | Apparel | Topwear | Tshirts | Navy Blue | Fall | 2011.0 | Sports | Fila Men's Round Neck Navy Blue T-shirt | images/7990.jpg |
| 30 | 37812 | Men | Apparel | Topwear | Shirts | Navy Blue | Summer | 2012.0 | Formal | John Players Men Navy Blue Shirt | images/37812.jpg |

2. 將文字敘述進行預處理，將 gender, articleType, baseColour, season, usage, productDisplayName 合併起來成新的 column

| input_text |
|---|
| men shirts navy blue fall casual turtle check ... |
| men tshirts grey summer casual puma men grey t... |
| men tshirts grey summer casual inkfruit mens c... |
| men shirts green summer ethnic fabindia men st... |

向量化

1. 使用預訓練模型 Resnet34 將圖片向量化，並放入新的 column。文字方面，使用 SentenceTransformer 對文字進行 encode，並修改輸出維度為 512，放入新的 column

| image_vec | word_vec |
|--|--|
| [tensor(0.5950), tensor(0.8089), tensor(0.8006...] | [tensor(-0.0311), tensor(-0.0016), tensor(0.31...] |
| [tensor(0.2147), tensor(0.1861), tensor(0.6978...] | [tensor(0.1969), tensor(0.0623), tensor(0.1967...] |
| [tensor(1.0277), tensor(1.1450), tensor(0.3659...] | [tensor(0.1610), tensor(0.1430), tensor(0.1933...] |

2.將圖片與文字的向量進行concatenate與sum，並分別放入新的column

| sum_vec | con_vec |
|--|--|
| [tensor(0.5639), tensor(0.8073), tensor(1.1196...] | [tensor(0.5950), tensor(0.8089), tensor(0.8006...] |
| [tensor(0.4116), tensor(0.2484), tensor(0.8945...] | [tensor(0.2147), tensor(0.1861), tensor(0.6978...] |
| [tensor(1.1887), tensor(1.2880), tensor(0.5591...] | [tensor(1.0277), tensor(1.1450), tensor(0.3659...] |

相似度比對

1.將需要推薦的衣服與文字敘述輸入，並個別向量化

```
# input
image = "drive/MyDrive/test.jpg"
word = "blue sport"
image_v = get_vector(image)
word_v = modelw.encode(word, convert_to_tensor=True)

print(word)
plt.subplot(1, 2, 1)
plt.imshow(cv.imread(image)[: , : , ::-1])
plt.show()
```

blue sport



The image shows a person from the waist up, wearing a black t-shirt with a white graphic that appears to be a peace sign or a similar symbol. The person is standing against a light-colored background. The image is displayed in a subplot with axes ranging from 0 to 4000 on the x-axis and 0 to 3000 on the y-axis.

2.將圖片與文字的向量合併，並利用Cosine Similarity的方法與資料集的每筆向量比較相似度

```
df['similarity_sum'] = df['sum_vec'].apply(lambda x: sim(image_v, word_v, x, 0))
df['similarity_con'] = df['con_vec'].apply(lambda x: sim(image_v, word_v, x, 1))
```

| similarity_sum | similarity_con |
|------------------|------------------|
| [tensor(0.6038)] | [tensor(0.6051)] |
| [tensor(0.6178)] | [tensor(0.6289)] |
| [tensor(0.6156)] | [tensor(0.6348)] |


3.將相似度由大到小排列

```
rec_df = df.copy()
rec_df.sort_values('similarity_sum', inplace=True, ascending=False)
rec_df = rec_df.reset_index(drop=True)
```

| similarity_sum | similarity_con |
|------------------|------------------|
| [tensor(0.7339)] | [tensor(0.7356)] |
| [tensor(0.7319)] | [tensor(0.7371)] |
| [tensor(0.7316)] | [tensor(0.7375)] |

4.顯示相似度最高的圖片

```
id 3365
gender Men
masterCategory Apparel
subCategory Topwear
articleType Tshirts
baseColour Black
season Summer
year 2011
usage Casual
productDisplayName Myntra Men's Brain Black T-shirt
image images/3365.jpg
input_text men tshirts black summer casual myntra mens br...
image_vec [1.51568115, 1.04728103, 0.0459099486, 0.13806...
word_vec [0.161217868, 0.26004976, 0.414260566, -0.4608...
sum_vec [tensor(1.6769), tensor(1.3073), tensor(0.4602)...
con_vec [tensor(1.5157), tensor(1.0473), tensor(0.0459)...
similarity_sum [tensor(0.7356)]
similarity_con [tensor(0.7427)]
Name: 0, dtype: object
```





5.由於圖片的特徵值過於強烈，所以利用文字敘述來篩選出更相近的排列，方法為篩選出input的word都有同時出現在衣服的描述中

```
[53] # Reference https://stackoverflow.com/questions/37011734/pandas-dataframe-str-contains-and-operation
base = r'^{'
expr = '(?=.*){'
base = base.format('').join(expr.format(w) for w in mylist))
base

'^(?=.*blue)(?=.*sport)'
```

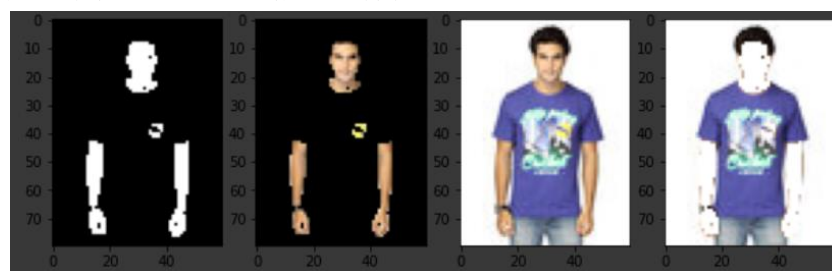
```
[54] for i, c in rec_df.loc[(rec_df['input_text'].str.contains(base))].iloc[:5].iterrows():
    rec = cv.imread(c.image)[: , :, ::-1] # reverse the channel order, because OpenCV reads image in BGR order
    plt.figure()
    print(c)
    plt.imshow(rec)
    plt.axis("off")
    plt.show()
```

```
id 7666
gender Men
masterCategory Apparel
subCategory Topwear
articleType Tshirts
baseColour Blue
season Fall
year 2011
usage Sports
productDisplayName Puma Men's Silly Point Blue T-shirt
image images/7666.jpg
input_text men tshirts blue fall sports puma mens silly p...
image_vec [1.20639336, 1.16069853, 0.544063151, 0.015081...
word_vec [-0.2754606, 0.3669902, 0.1913286, -0.54464316...
sum_vec [tensor(0.9309), tensor(1.5277), tensor(0.7354...
con_vec [tensor(1.2064), tensor(1.1607), tensor(0.5441...
similarity_sum [tensor(0.7063)]
similarity_con [tensor(0.7128)]
Name: 86, dtype: object
```




去除人物

1.將圖片中的人物去除，方法為設定人體顏色在HSV的最小與最大範圍，在利用圖片與我們設置的threshold比較，會回傳binary mask（N-Dimensional array），在顏色範圍內的話會回傳1，範圍外回傳0。最後再利用原圖減去皮膚範圍



2.利用預訓練模型去除人，先去除皮膚的效果比直接用原圖去除的好



3.將圖片進行masking。先將圖片轉成grayscale，再二值化，將大於門檻值的灰階值設為最大灰階值，小於門檻值的值設為0。最後將圖片的值資料進行二進制的not operation

