使用分詞工具 斷詞

In [1]:

```
import os
import requests
```

In [2]:

```
import jieba.analyse
import urllib

url = "https://raw.githubusercontent.com/cjwu/cjwu.github.io/master/courses/nlp/hw1-c
text = urllib.request.urlopen(url).read().decode("utf-8")
result = jieba.analyse.extract_tags(text, topK=100, withWeight=True)
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\iwin4\AppData\Local\Temp\jieba.cache
Loading model cost 0.845 seconds.
Prefix dict has been built successfully.
```

In [3]:

```
#輸出
print("Top 100 High TF-IDF Words:")
for i in result:
    print('word:', i[0], 'TF-IDF:', i[1])
```

```
Top 100 High TF-IDF Words:
word: 什麼 TF-IDF: 0.19464509600157406
word: 八卦 TF-IDF: 0.19420501140555646
word: 台灣 TF-IDF: 0.12081408131496592
word: 怎麼 TF-IDF: 0.11186701189176337
word: 肥宅 TF-IDF: 0.07336596927026089
word: 現在 TF-IDF: 0.03800903805058438
word: 不會 TF-IDF: 0.036229733848907944
word: 還是 TF-IDF: 0.03568381096884812
word: 是不是 TF-IDF: 0.0355764318510935
word: 一個 TF-IDF: 0.03550183667549485
word: 中國 TF-IDF: 0.034474018907481
word: 這樣 TF-IDF: 0.0325902479818425
word: 怎樣 TF-IDF: 0.029860633581543415
word: 時候 TF-IDF: 0.02967528939386879
word: 一樣 TF-IDF: 0.028691280252032575
word: 真的 TF-IDF: 0.026822655692091746
word: 沒有 TF-IDF: 0.026099831518909124
word: ... TF-IDF: 0.02543259244328046
word: 應該 TF-IDF: 0.02359936993246231
word: 可以 TF-IDF: 0.022958278161416924
word: 喜歡 TF-IDF: 0.02285462328744244
word: 因為 TF-IDF: 0.022416537025666042
word: 一堆 TF-IDF: 0.021680457381807062
word: 問題 TF-IDF: 0.020815837223021518
word: 感覺 TF-IDF: 0.020714740393380813
word: 哪個 TF-IDF: 0.020643972612632316
word: 女生 TF-IDF: 0.02055531365074145
word: 這麼 TF-IDF: 0.019963253959718225
word: 覺得 TF-IDF: 0.01862540591413954
word: 這種 TF-IDF: 0.01824460785582621
word: 美國 TF-IDF: 0.018055893773830226
word: 正妹 TF-IDF: 0.016286699255117856
```

```
word: 知道 TF-IDF: 0.015811205555870015
word: 其實 TF-IDF: 0.01557902144763291
word: 為何 TF-IDF: 0.015043208250537165
word: 還有 TF-IDF: 0.014736547867293687
word: 東西 TF-IDF: 0.014719698395686903
word: 比較 TF-IDF: 0.01451750473640549
word: 那麼 TF-IDF: 0.014369229386265786
word: 到底 TF-IDF: 0.013855929767928526
word: 有人 TF-IDF: 0.013674560834915146
word: 自己 TF-IDF: 0.013466423468451874
word: QQ TF-IDF: 0.013445878342213998
word: 時間 TF-IDF: 0.013361630984180075
word: 開始 TF-IDF: 0.013206615845397658
word: 這個 TF-IDF: 0.012974093137224033
word: 哪裡 TF-IDF: 0.01282581778708433
word: 不是 TF-IDF: 0.012758688475732879
word: 沒人 TF-IDF: 0.012552856347054421
word: 甚麼 TF-IDF: 0.012468608989020498
word: 出來 TF-IDF: 0.012320333638880795
word: 日本 TF-IDF: 0.012127572493474573
word: 那個 TF-IDF: 0.012054111987493602
word: 發現 TF-IDF: 0.011993453889709177
word: 中國人 TF-IDF: 0.011976604418102393
word: 國家 TF-IDF: 0.011976604418102393
word: 如果 TF-IDF: 0.01195142064483163
word: 不要 TF-IDF: 0.011371977189677437
word: 就是 TF-IDF: 0.011178381502120506
word: 他們 TF-IDF: 0.010851059714769191
word: 大家 TF-IDF: 0.010829155333451544
word: 朋友 TF-IDF: 0.010388757432612226
word: 很多 TF-IDF: 0.010254718969681581
word: 台北 TF-IDF: 0.010090151731776114
word: 已經 TF-IDF: 0.009971517296895042
word: 老師 TF-IDF: 0.009917598987753332
word: 大學 TF-IDF: 0.00989737962182519
word: 變成 TF-IDF: 0.009695185962543778
word: 邊緣 TF-IDF: 0.009459293360048795
word: 我們 TF-IDF: 0.0093885255793003
word: 結果 TF-IDF: 0.0093885255793003
word: 遊戲 TF-IDF: 0.009065015724450039
word: 不用 TF-IDF: 0.00885884470039901
word: 手機 TF-IDF: 0.008744875763921134
word: 一點 TF-IDF: 0.008717916609350278
word: 看到 TF-IDF: 0.008670900198239207
word: 多少 TF-IDF: 0.008670105043160773
word: 男生 TF-IDF: 0.00855216230787021
word: 別人 TF-IDF: 0.008535942315997006
word: 當然 TF-IDF: 0.008482024006855295
word: 10 TF-IDF: 0.008434845486356299
word: 女友 TF-IDF: 0.008252324115040545
word: 如何 TF-IDF: 0.008088109356327164
word: 還好 TF-IDF: 0.008067527005328399
word: 電影 TF-IDF: 0.008013608696186688
word: 新聞 TF-IDF: 0.008013608696186688
word: 還要 TF-IDF: 0.008003499013222618
word: 韓國 TF-IDF: 0.007986649541615832
word: 鄉民 TF-IDF: 0.007952950598402264
word: 起來 TF-IDF: 0.007770976305048992
word: 根本 TF-IDF: 0.007719534763027918
```

```
word: XD TF-IDF: 0.007700208524300497
word: 好吃 TF-IDF: 0.007579158623013259
word: 妹妹 TF-IDF: 0.007550327706300244
word: 的掛 TF-IDF: 0.0075384535968753665
word: 不過 TF-IDF: 0.0073497395148793805
word: 一直 TF-IDF: 0.007303212591178013
word: .. TF-IDF: 0.0072486426852386735
word: ptt TF-IDF: 0.007110477018063041
word: 最強 TF-IDF: 0.006979051139530123
```

統計前一百個高頻和TF-IDF權重高

In [4]:

```
import jieba.analyse
tags = jieba.analyse.extract_tags(text, topK=5, withWeight=True)

for tag in tags:
    print('word:', tag[0], 'tf-idf:', tag[1])
```

```
word: 什麼 tf-idf: 0.19464509600157406
word: 八卦 tf-idf: 0.19420501140555646
word: 台灣 tf-idf: 0.12081408131496592
word: 怎麼 tf-idf: 0.11186701189176337
word: 肥宅 tf-idf: 0.07336596927026089
```

計算並畫出其統計圖型

fig #1

In [7]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import TfidfVectorizer


docs = text.split('\n') # 一行一個文章

# TF-IDF
vectorizer = TfidfVectorizer()
X_axes = vectorizer.fit_transform(docs)

weights = np.asarray(X_axes.mean(axis=0)).ravel().tolist() # cauculate the weight


#x軸:字詞編號;y軸:權重
# build DataFrame
df = pd.DataFrame({'term': vectorizer.get_feature_names_out(), 'weight': weights})

# sort
df = df.sort_values(by='weight', ascending=False)

# index
df = df.reset_index(drop=True)
df['word_index'] = df.index + 1

# 輸出
top_100_1 = df.head(100)
print(top_100_1)

# 繪製折線圖
weights = top_100_1['weight'].values
print(len(weights))
word_index = top_100_1['word_index'].values

# 繪製摺線圖

# plt.figure(figsize=(10, 6))
plt.plot(word_index, weights)
plt.xlabel('word num')
plt.ylabel('weight')
plt.title('top 100 weighted')
plt.grid(True)
plt.show()
```

```
       term    weight  word_index
0       的八卦  0.001379           1
1       有沒有  0.001273           2
2      沒有資料  0.000736           3
3       認真回  0.000561           4
4        vs  0.000547           5
..      ...       ...         ...
95      為什麼  0.000112          96
96       對了  0.000111          97
97  pokemon  0.000111          98
98       無聊  0.000110          99
99      嗆三小  0.000109         100
```

```
[100 rows x 3 columns]
```

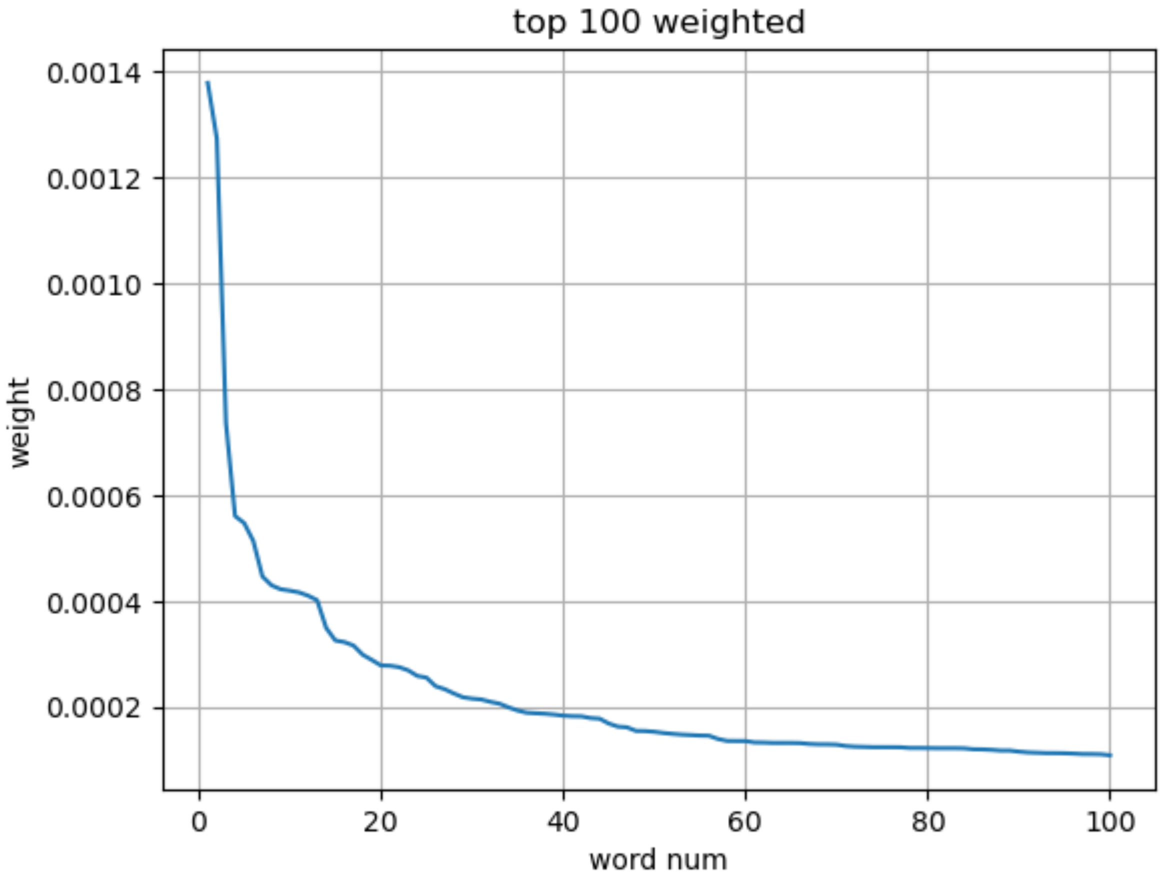
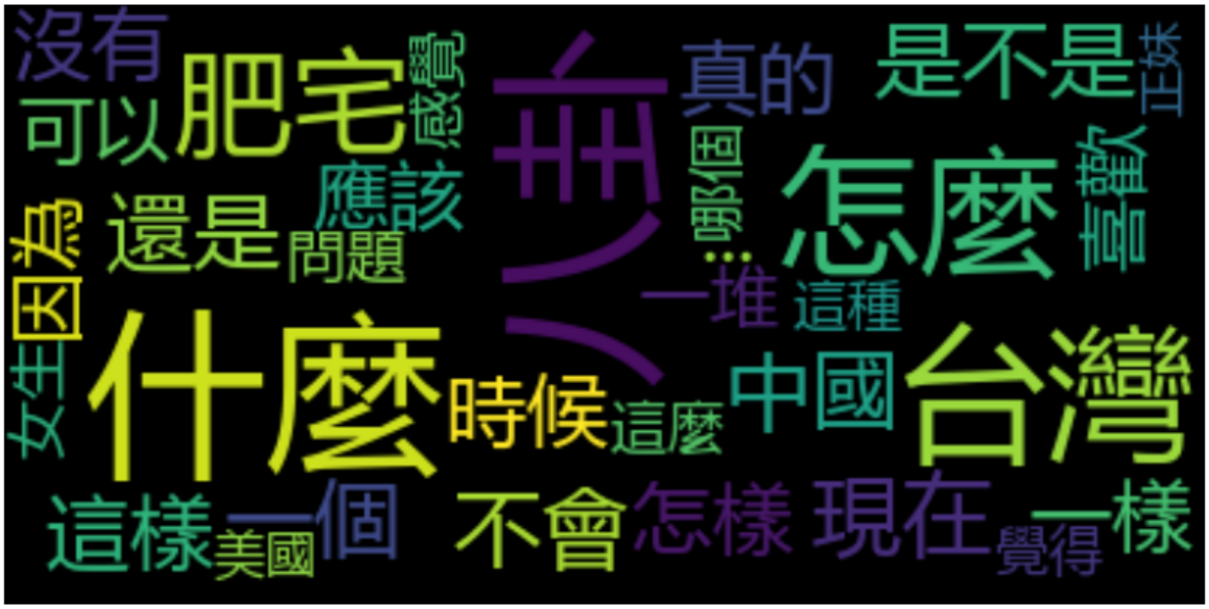


fig #2

fig #3

In [9]:

```
from wordcloud import WordCloud
import jieba.analyse
import matplotlib.pyplot as plt

result = jieba.analyse.extract_tags(text, topK=100, withWeight=True)

# 32 words
result = result[:32]

# transger DataFrame to dict
dict_data = top_100_1.set_index('term')['weight'].to_dict()

# transfer words and weighted dict
words_dict = dict(result)

# WordCloud obj
wc = WordCloud(font_path='msyh.ttc')

# 將 dictionary 中的詞彙及權重傳給 WordCloud 物件
wc.generate_from_frequencies(words_dict)

# 繪製文字雲
plt.figure(figsize=(10, 6))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```


沒有
肥宅
感覺
女
真的
是不是
正妹
可以
應該
喜歡
怎麼
哪個
因為
還是
問題
一堆
這種
什麼
時候
這麼
中國
台灣
女生
這樣
美國
一個
不會
怎樣
現在
覺得
一樣