```
#pip install jieba
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: jieba in /usr/local/lib/python3.9/dist-packages (0.42.1)
```

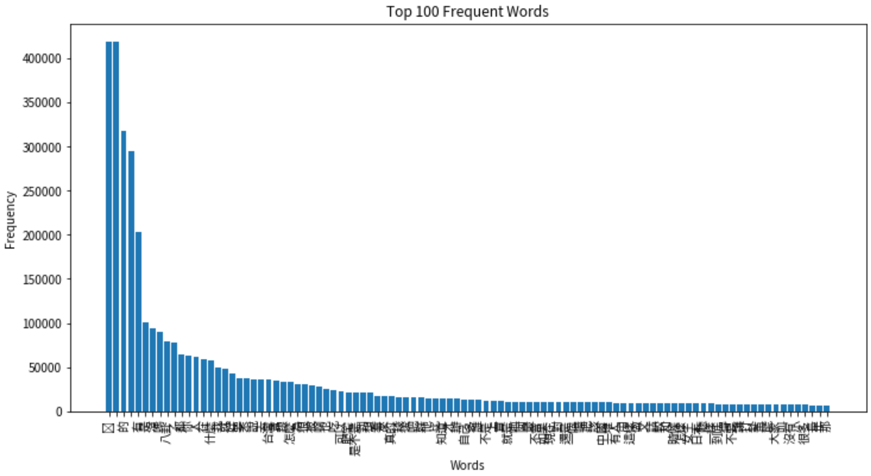
```
import jieba
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer
import re
import requests
# 設定字體為中文字體
#plt.rcParams['font.family'] = ['Microsoft JhengHei']

# 讀取文本資料
#with open('input.txt', 'r', encoding='utf-8') as f:
#  text = f.read()
response = requests.get('https://raw.githubusercontent.com/cjwu/cjwu.github.io/master/courses/nlp/hw1-dataset.txt')
text = response.text
text = re.sub('[^\w\s]', '', text)

# 使用jieba進行分詞
words = list(jieba.cut(text))
word_count = Counter(words)
top100_freq = word_count.most_common(100)#取前100


#計算TF-IDF權重
tfidf = TfidfVectorizer()
tfidf_matrix = tfidf.fit_transform([text])
tfidf_scores = zip(tfidf.get_feature_names_out(), tfidf_matrix.toarray()[0])
tfidf_scores = sorted(tfidf_scores, key=lambda x: x[1], reverse=True)

# 取前100個
top100_tfidf = tfidf_scores[:100]
print(top100_tfidf)
```

```
[('的八卦', 0.3433089395621334), ('有沒有', 0.3387618012897872), ('vs', 0.3105045848830648), ('沒有資料', 0.2111171340732135), ('o_o', 0.1727912
```

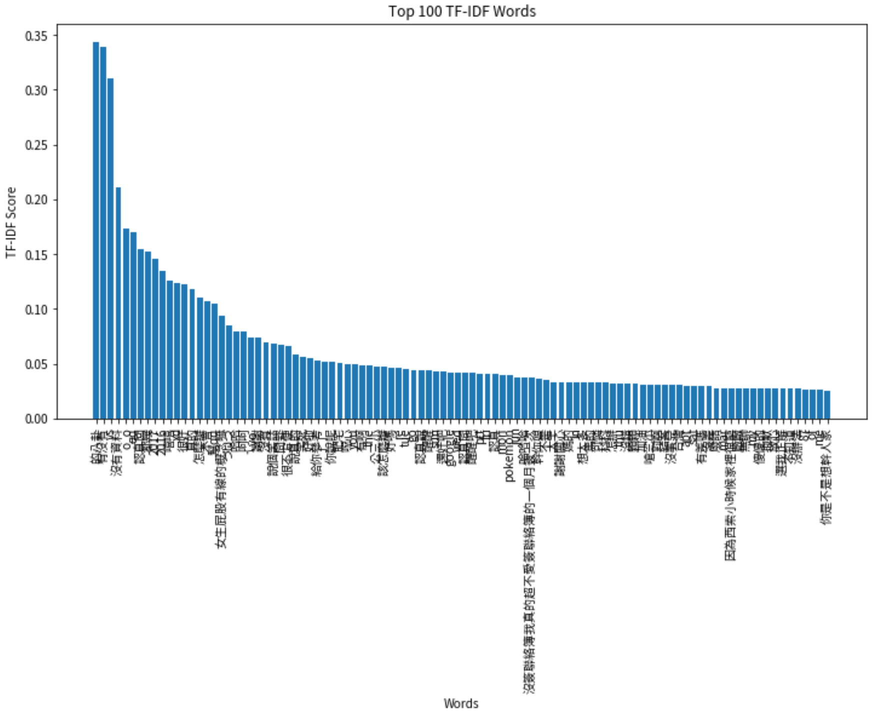
```
#!wget -O TaipeiSansTCBeta-Regular.ttf https://drive.google.com/uc?id=1eGAsTN1HBpJAkeVM57_C7ccp7hbgSz3_&export=download
```

```
--2023-03-18 15:46:41--  https://drive.google.com/uc?id=1eGAsTN1HBpJAkeVM57_C7ccp7hbgSz3_
Resolving drive.google.com (drive.google.com)... 142.250.99.101, 142.250.99.138, 142.250.99.139, ...
Connecting to drive.google.com (drive.google.com)|142.250.99.101|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-0k-9o-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/h4v1407sp4idrfam3igavt3jjhlnrhk9/167915437
Warning: wildcards not supported in HTTP.
--2023-03-18 15:46:44--  https://doc-0k-9o-docs.googleusercontent.com/docs/securesc/ha0ro937gcuc7l7deffksulhg5h7mbp1/h4v1407sp4idrfam3igavt3jjhl
Resolving doc-0k-9o-docs.googleusercontent.com (doc-0k-9o-docs.googleusercontent.com)... 172.253.117.132, 2607:f8b0:400e:c0a::84
Connecting to doc-0k-9o-docs.googleusercontent.com (doc-0k-9o-docs.googleusercontent.com)|172.253.117.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 20659344 (20M) [application/x-font-ttf]
Saving to: 'TaipeiSansTCBeta-Regular.ttf'

TaipeiSansTCBeta-Re 100%[===================>]  19.70M  --.-KB/s    in 0.1s

2023-03-18 15:46:45 (175 MB/s) - 'TaipeiSansTCBeta-Regular.ttf' saved [20659344/20659344]
```

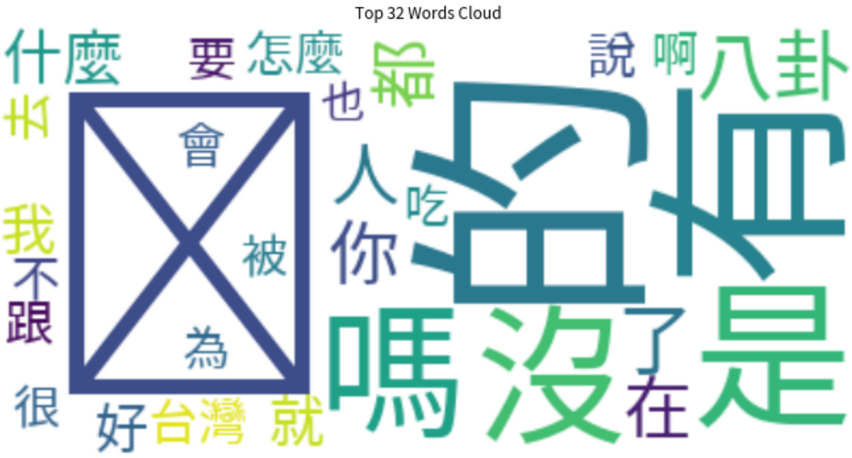
```
import matplotlib as mpl
from matplotlib.font_manager import fontManager
#中文字體
fontManager.addfont('TaipeiSansTCBeta-Regular.ttf')
mpl.rc('font', family='Taipei Sans TC Beta')


# 繪製高頻詞統計圖
x1 = range(len(top100_freq))
y1 = [f[1] for f in top100_freq]
plt.figure(figsize=(12,6))
plt.bar(x1, y1)
plt.xticks(x1, [f[0] for f in top100_freq], rotation=90)
plt.title('Top 100 Frequent Words')
plt.xlabel('Words')
plt.ylabel('Frequency')
#plt.tight_layout()
plt.show()
```



```
# 繪製TF-IDF權重詞統計圖
x2 = range(len(top100_tfidf))
y2 = [f[1] for f in top100_tfidf]
plt.figure(figsize=(12,6))
plt.bar(x2, y2)
plt.xticks(x2, [f[0] for f in top100_tfidf], rotation=90)
plt.title('Top 100 TF-IDF Words')
plt.xlabel('Words')
plt.ylabel('TF-IDF Score')
#plt.tight_layout()
plt.show()
```



```
# 製作前32個文字雲
wc = WordCloud(background_color="white", contour_width=3, contour_color='steelblue', font_path= 'TaipeiSansTCBeta-Regular.ttf')
wc.generate_from_frequencies(dict(word_count.most_common(32)))
plt.figure(figsize=(12,6))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.title('Top 32 Words Cloud')
#plt.tight_layout()
plt.show()
```


Top 32 Words Cloud
什麼
要
怎麼
都
說
啊
八卦
去
也
會
人
有
吃
我
被
你
不
跟
為
嗎
沒
了
是
很
好
台灣
就
在