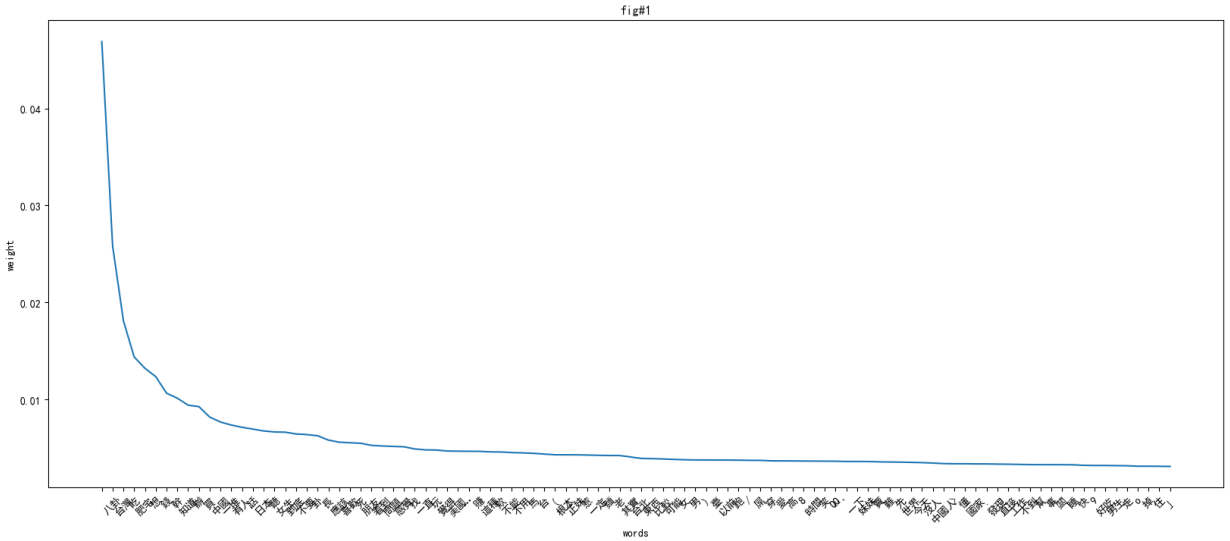
In [1]:
```
# -*- coding: UTF-8 -*-
import pandas as pd
import numpy as np
import jieba
import collections
import math


filename = open("hw1-dataset.txt", mode = 'r')
f = open("stopwords-zh.txt", "r")
stopwords = set()
for word in f.readlines():
    stopwords.add(word[: -1]) #不要\n

doc = {}
for i, line in enumerate(filename.readlines()):
    line = jieba.lcut(line)
    temp = []
    for w in line:
        if (w not in stopwords):
            temp.append(w)
    doc[i] = temp

word_count = collections.defaultdict(int) #字詞 t 在文件 d 出現的次數
all_words = 0
all_article = 0
for line in doc.values():
    for w in line:
        word_count[w] += 1
    all_words += len(line)
    all_article += 1

article_count = collections.defaultdict(int) #包含字詞 t 的文件數
for line in doc.values():
    for w in set(line):
        article_count[w] += 1

tf = {}
for w, n in word_count.items():
    tf[w] = n / all_words

idf = {}
for w, n in article_count.items():
    idf[w] = math.log(all_article / n)

tf_idf = {}
for w, n in tf.items():
    tf_idf[w] = n * idf[w]
```

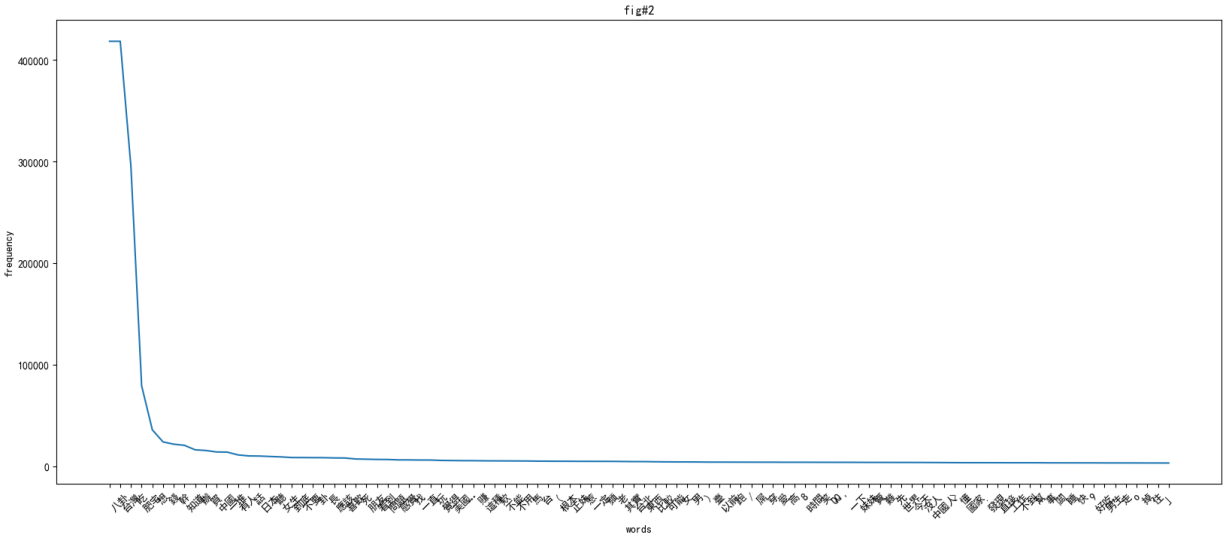
```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/8q/s5rn_lld0qj8zg63cy8t3xtw0000gn/T
/jieba.cache
Loading model cost 0.285 seconds.
Prefix dict has been built successfully.
```

In [2]:
```
import matplotlib
matplotlib.matplotlib_fname()
```

Out[2]:
```
'/Users/hsiu/opt/anaconda3/lib/python3.9/site-packages/matplotlib/mpl-data/matplotlibrc'
```

In [3]:
```
import matplotlib.pyplot as plt

plt.rcParams['font.sans-serif'] = ['SimHei']

data = sorted(tf_idf.items(), key = lambda x: x[1], reverse = True)[: 100
fig1 = plt.figure(figsize=(20,8))
ax1 = plt.axes()
x = [w[0] for w in data]
y = [w[1] for w in data]

ax1.plot(x, y)
plt.xticks(rotation = 45)
plt.title('fig#1')
plt.xlabel('words')
plt.ylabel('weight')
plt.show()

data = sorted(word_count.items(), key = lambda x: x[1], reverse = True)[:
y = [w[1] for w in data]

fig2 = plt.figure(figsize=(20,8))
ax2 = plt.axes()
ax2.plot(x, y)
plt.xticks(rotation = 45)
plt.title('fig#2')
plt.xlabel('words')
plt.ylabel('frequency')
plt.show()
```





In [4]:
```
from wordcloud import WordCloud


freq = {}
for w, n in data[2: 34]:
    freq[w] = n

wordcloud = WordCloud(font_path = "/Users/hsiu/opt/anaconda3/lib/python3.
wordcloud.generate_from_frequencies(frequencies = freq)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.title('fig#3')
plt.show()
```