Student ID: B0928002

Name: 林力

```
!pip install zhon
```

```
    Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
    Requirement already satisfied: zhon in /usr/local/lib/python3.9/dist-packages (1.1.5)
```

```
import re
import jieba
from zhon.hanzi import punctuation

# 把用到的東西載入
articles = [line.strip().replace(' ', '') for line in open('/content/drive/MyDrive/Colab Notebooks/nlp/HW1/hw1-dataset.
# print(articles)
punctuations = [line.strip().replace(' ', '') for line in open('/content/drive/MyDrive/Colab Notebooks/nlp/HW1/punctuat
# print(punctuations)
punctuation = list(punctuation)
stopwords = [line.strip().replace(' ', '') for line in open('/content/drive/MyDrive/Colab Notebooks/nlp/HW1/stopword.tx
print (len(stopwords))
print (stopwords)

word_counts = []

for i in range(len(articles)):
  # 中文斷詞
  articles[i] = articles[i].replace('\t', '').replace('\u3000', '')
  articles[i] = re.sub('[^\u4e00-\u9fa5]', '', articles[i])
  articles[i] = jieba.lcut(articles[i])
  count = {}
  for word in articles[i]:
    # if word not in punctuations and word != " " and word not in punctuation and word not in stopwords:
    if word not in punctuations and word not in punctuation and word != " " :
      if word in count:
        count[word] += 1
      else:
        count[word] = 1
  word_counts.append(count)
```

```
    1611
    ['壹', '壹.', '壹壹', '壹下', '壹個', '壹些', '壹何', '壹切', '壹則', '壹則通過', '壹天', '壹定', '壹方面', '壹旦', '壹時',
```

```
print (word_counts[1:100])
```

```
    [{'為': 1, '什麼': 1, '慶祝會': 1, '被': 1, '罵': 1, '可是': 1, '慶': 1, '端午': 1, '不會': 1, '因為': 1, '屈原': 1, '不
```

```
# 計算字詞出現次數
word_frequency = []
word_total = {}

for word_count in word_counts:
  all_count = sum(word_count.values())
  freq = {}

  for word, count in word_count.items():
    freq[word] = round(count/all_count, 5)

    if word in word_total:
      word_total[word] += count
    else:
      word_total[word] = count

  word_frequency.append(freq)
word_total = (sorted(dict(word_total).items(), key = lambda x:x[1], reverse = True))
print (word_total[1: 100])
print (word_frequency[1: 100])
```

```
    [('有', 203509), ('是', 100215), ('沒', 96145), ('嗎', 89377), ('八卦', 79064), ('了', 77296), ('你', 64577), ('都',
    [{'為': 0.0625, '什麼': 0.0625, '慶祝會': 0.0625, '被': 0.0625, '罵': 0.0625, '可是': 0.0625, '慶': 0.0625, '端午': 0.
```

```
all_words = []
for word in word_counts:
  all_words.extend(list(word.keys()))
```

```
word_appear = {}

for word in all_words:
  if word in word_appear:

    word_appear[word] += 1

  else:
    word_appear[word] = 1


print (word_counts[1:100])
```

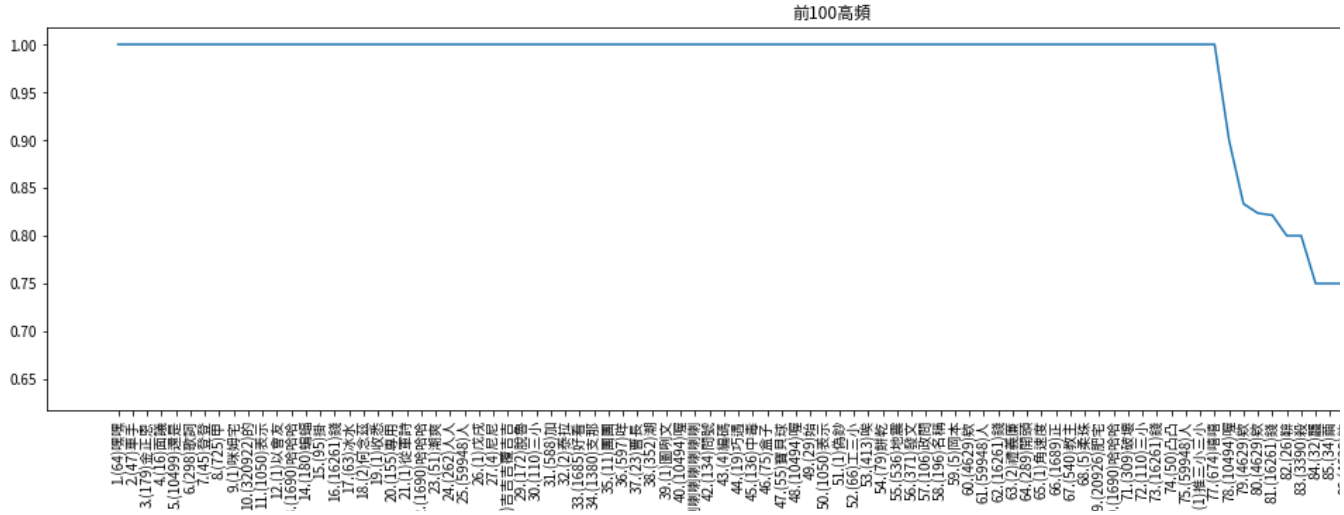
```
[{'為': 1, '什麼': 1, '慶祝會': 1, '被': 1, '罵': 1, '可是': 1, '慶': 1, '端午': 1, '不會': 1, '因為': 1, '屈原': 1, '不
```

```
# 計算 idf值
import math

idf = []

for word_count in word_counts:

  invertFreq = {}

  for word in word_count.keys():
    appear = word_appear[word]
    invertFreq[word] = math.log(round(len(word_counts)/appear), 4)
    # print(appear, word)
  idf.append(invertFreq)


print(idf[1:100])
```

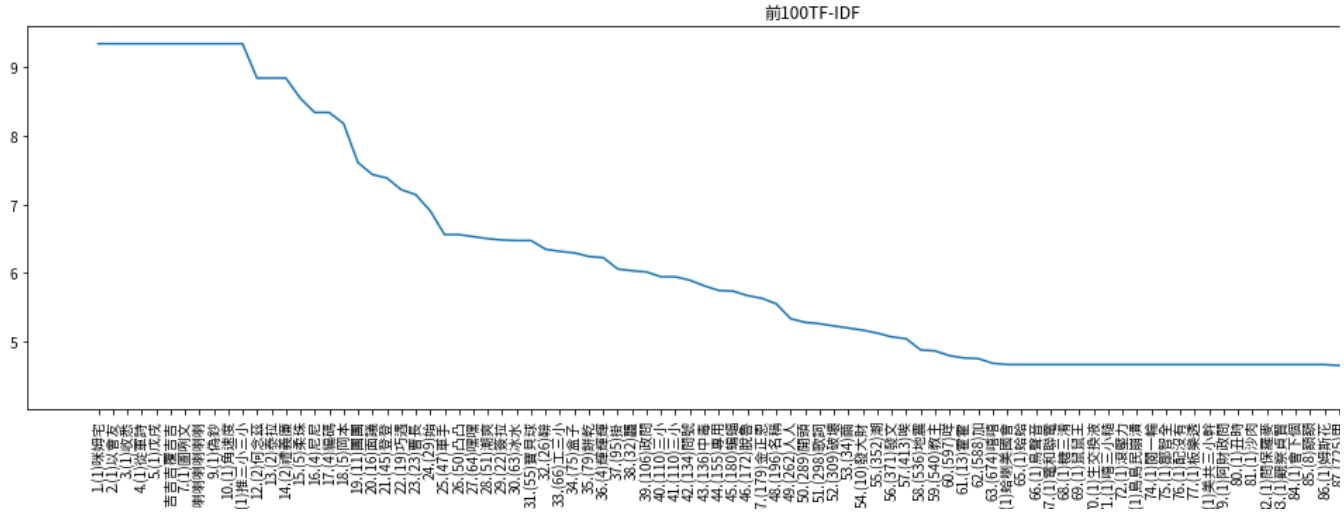
```
[{'為': 1.9036774610288019, '什麼': 1.5, '慶祝會': 8.836920217946837, '被': 1.9534452978042594, '罵': 3.825525845589
```

```
# tf-idf值
tf_idf_all = []
for i, word in enumerate(word_frequency):
  tf_idf = {}

  for word, freq in word.items():
    tf_idf[word] = freq*idf[i][word]
  tf_idf_all.append(tf_idf)


# 高頻前100
frequency_100 = []
for freq in word_frequency:
  if len(freq)>0:
    frequency_100.append((max(freq.items())))
frequency_100.sort(key = lambda x : x[1],reverse = True)


# tf-idf前100
tf_idf_100 = []
for tf_idf in all_idf:
  if len(tf_idf)>0:
    tf_idf_100.append((max(tf_idf.items())))
tf_idf_100.sort(key = lambda x : x[1],reverse = True)


import matplotlib as mpl
import matplotlib.pyplot as plt
from matplotlib.font_manager import fontManager
fontManager.addfont('/content/drive/MyDrive/Colab Notebooks/nlp/HW1/TaipeiSansTCBeta-Regular.ttf')
mpl.rc('font', family='Taipei Sans TC Beta')

x = []
y = []
i = 0
for word in frequency_100[: 100]:
  i+=1
  x.append(str(i)+"."+"("+str(dict(word_total)[str(word[0])])+")"+str(word[0]))
  y.append(word[1])

plt.figure(figsize = (20,5))
plt.plot(x,y)
```

```
plt.title("前100高頻")
plt.xticks(rotation = 90)
plt.show()
```




```
x = []
y = []
i = 0
for word in tf_idf_100[:100]:
  i+=1
  x.append(str(i)+"."+"("+str(dict(word_total)[str(word[0])])+")"+str(word[0]))
  y.append(word[1])
plt.figure(figsize = (20,5))
plt.plot(x,y)
plt.title("前100TF-IDF")
plt.xticks(rotation = 90)
plt.show()
```




```
from wordcloud import WordCloud
WordCloud(collocations=False,
                font_path='/content/drive/MyDrive/Colab Notebooks/nlp/HW1/TaipeiSansTCBeta-Regular.ttf',
                width=1000,
                height=250,
                background_color = "black" ,
                margin=2
                ).generate_from_frequencies(dict(tf_idf_100[:32])).to_image()
```

✓ 0 秒 完成時間：晚上9:09