```
import jieba
import requests
import re

url = "https://raw.githubusercontent.com/cjwu/cjwu.github.io/master/courses/nlp/hw1-dataset.txt"
data = requests.get(url)
data = data.text

pattern = re.compile(r'[\s+\.\!\/_,$%^*(+\"\']+|[+——！，。？、~@#￥%……&*（）：]+')
data = re.sub(pattern, '', data)

for
words = jieba.cut(data)

print(" ".join(words))
```

```
為 什麼 聖結 石會 被 酸 而 這群 人 不會質 感劇本 成員 都 差 很多 好 嗎 不要 拿 腎 結石來 污辱 這群 人 為 什麼 慶祝 228 會 被 罵
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-21-d06cdce78e11> in <module>
      9 for word, freq in word_freq.items():
     10     tf = freq / len(word_freq)
---> 11     idf = math.log(doc_count / (1 + len([1 for text in [data] if word in
text])))
     12     tfidf[word] = tf * idf
     13

                                  1 frames
/usr/local/lib/python3.9/dist-packages/requests/utils.py in iter_slices(string,
slice_length)
    562         slice_length = len(string)
    563     while pos < len(string):
--> 564         yield string[pos:pos + slice_length]
    565         pos += slice_length
    566

KeyboardInterrupt:
```

## 新增區段