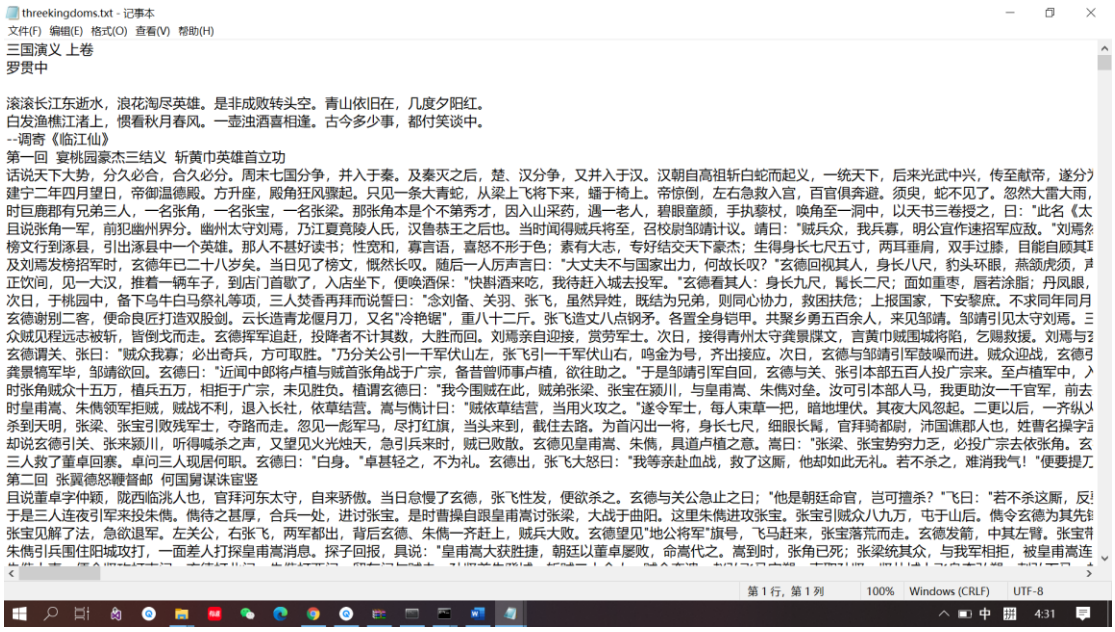
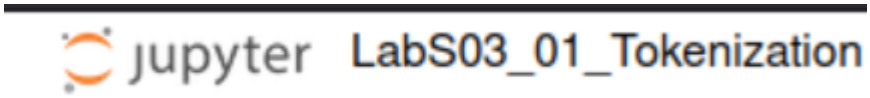
Bonus



The content of the Chinese text file.



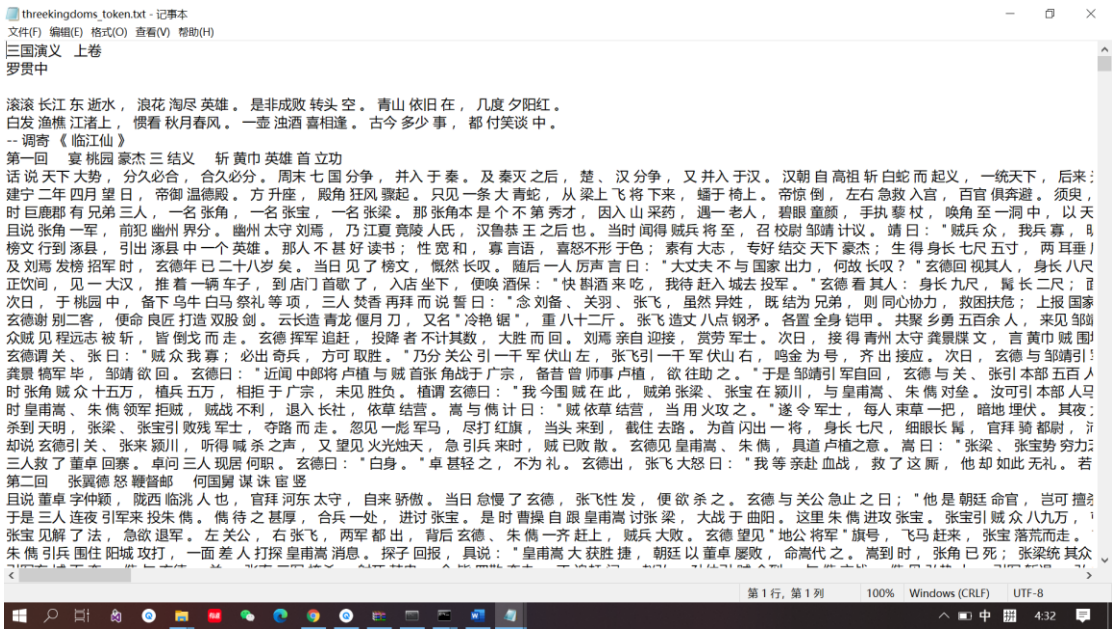
First do tokenization to the text.

You can also provide a text file for Jieba Tokenization

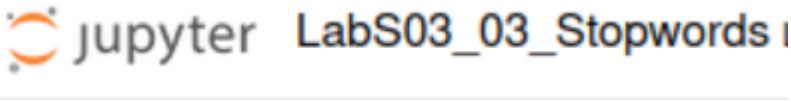
In [2]:
```
import jieba
textFile = 'threekingdoms.txt'
tokenFile = 'threekingdoms_token.txt'

with open(textFile, 'r', encoding = 'utf-8') as sourceFile, open(tokenFile, 'a+', encoding = 'utf-8') as targetFile:
    for line in sourceFile:
        seg = jieba.cut(line.strip(), cut_all = False)
        output = ' '.join(seg)
        targetFile.write(output)
        targetFile.write('\n')
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\conda_temp\jieba.cache
Loading model cost 1.739 seconds.
Prefix dict has been built successfully.
```

Read the text file as txt. And apply jieba.cut to txt. Save the tokenization in another file.



Open the tokenization file. It gives all the tokenization of the original text file.

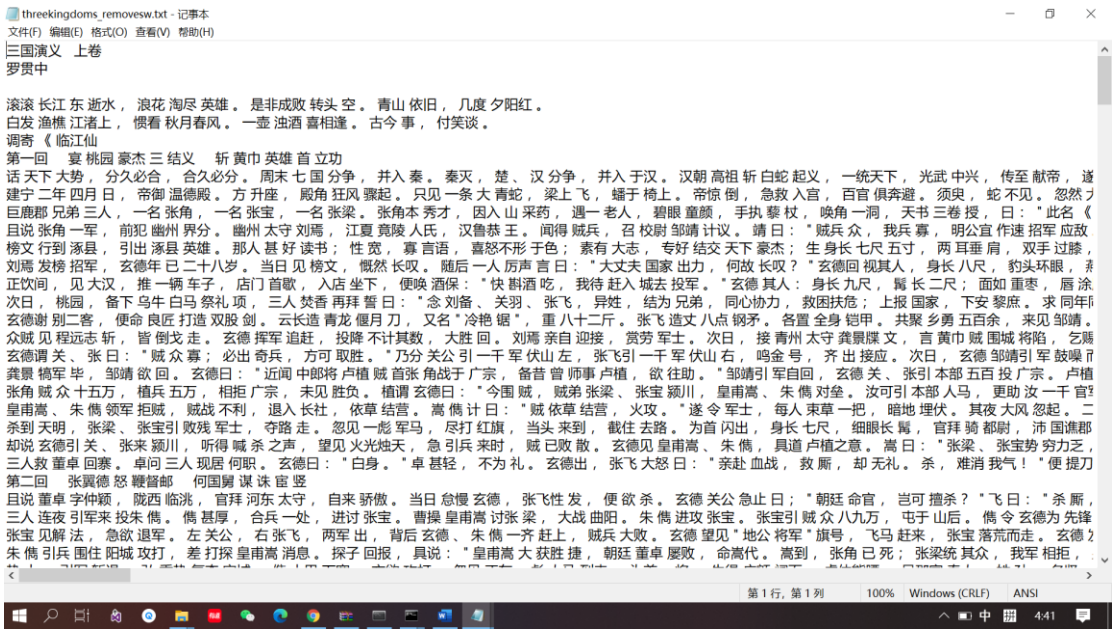


Second, remove stop words in the text file.

```
def seg_sentence(sentence):
    sentence_seged = jieba.cut(sentence.strip())
    stopwords = stopwordslist('baidu_stopwords.txt')
    outstr = ''
    for word in sentence_seged:
        if word not in stopwords:
            if word != '\t':
                outstr += word
                outstr += " "
    return outstr
```

You can also provide a text file for stopwords removal, go through the stopwords dictionary word by word, and remove them

In [12]:
```
processedFile = 'threekingdoms_removesw.txt'

inputs = open(textFile, 'r', encoding='utf-8')
outputs = open(processedFile, 'w')
for line in inputs:
    line_seg = seg_sentence(line)
    outputs.write(line_seg + '\n')
```



Use seg_sentence function to remove stop words in baidu_stopwords.txt, and write the output to a file.