**TensorFlow｜2｜中文分词**

1. 基本环境

|  |  |
| --- | --- |
| 程序名称 | segWord.py |
| 硬件环境 | MacOSX |
| 软件环境 | Python3, Spyder, Anaconda |

1. 核心问题：利用结巴分词进行分词
2. 解决过程：

* 导入jieba包
* 按行读文件
* 利用结巴分词对文件内的文本进行分割
* 测试结果
* 保存分割后的文字

1. 代码：

import jieba

# 确定读、存文件目录

filePath='/Users/xuyizhou/Desktop/corpus.txt'

fileSegWordDonePath ='/Users/xuyizhou/Desktop/corpusSegDone.txt'

# 按行读文件

fileTrainRead = []

with open(filePath) as fileTrainRaw:

for line in fileTrainRaw:

fileTrainRead.append(line)

# 利用结巴分词分割

fileTrainSeg=[]

for i in range(len(fileTrainRead)):

fileTrainSeg.append([' '.join(list(jieba.cut(fileTrainRead[i][:],cut\_all=False)))])

if i % 1000 == 0 :

print(i)

# 测试结果

#PrintListChinese(fileTrainSeg[10])

# 保存结果

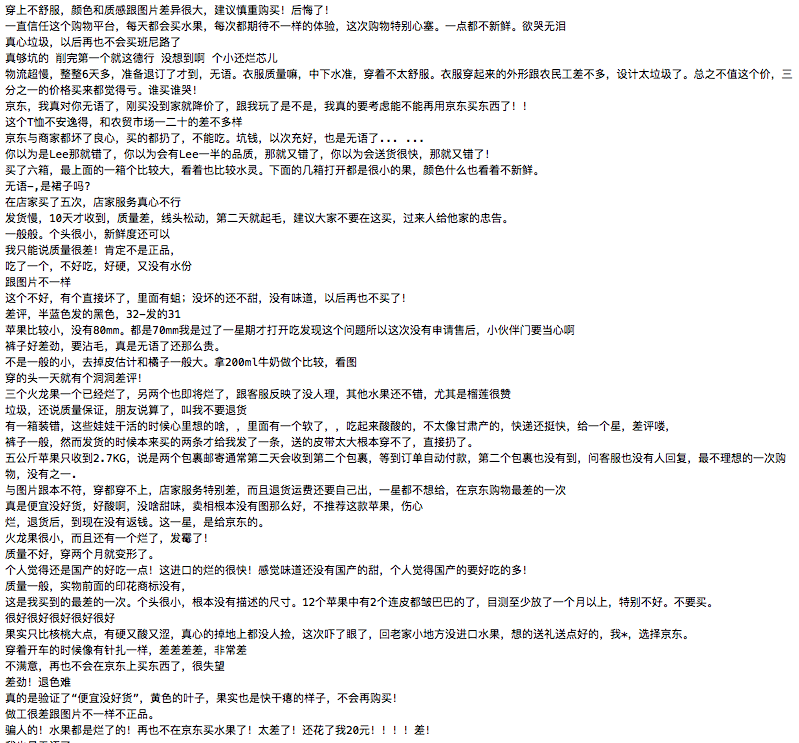
with open(fileSegWordDonePath,'w') as fW:

for i in range(len(fileTrainSeg)):

fW.write(str(fileTrainSeg[i][0]))

1. 结果：

原始文本：corpus.txt



分词文本：corpusSegDone.txt