实验 4-3 中文分词实现用户评价分析

建议课时： 60 分钟

一、 实验目的  
λ 了解中文分词的原理；  
λ 掌握 python 结巴分词的代码实现；  
λ 掌握分词基础上的词频统计方法；  
二、 实验环境  
Dsight 智慧实验室的 hadoop 环境、 python3 环境、 第三方包有 jieba  
Pycharm 、 NotePad++、 Sublime Text 等代码编辑工具  
三、 实验步骤  
本节实验所做内容如下红色标注：  
本节实验主要是通过 python 自然语言分析算法对用户评论信息进行分词统  
计， 将统计结果保存到 hive 中。  
具体实验步骤如下：  
1. 规范用户评价信息  
只取评价内容， 保存成新的 txt 文件（分隔符保持一致）  
2. 编写代码实现中文分词  
（1） 开发语言： python  
（2） 运行环境： hadoop 环境中的 python 环境  
（3） 读取源文本文件内容

代码：

import csv

import jieba

content=""

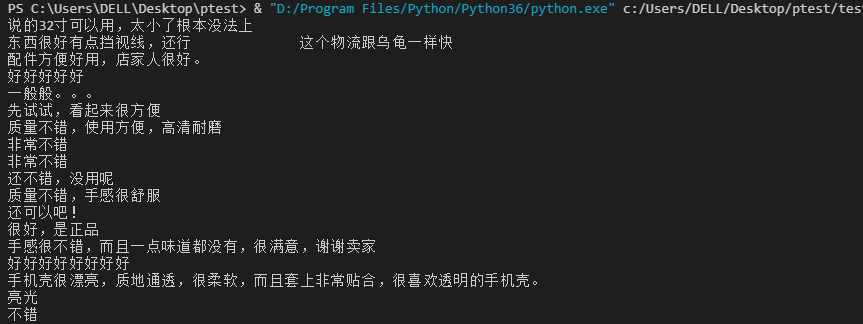
with open("comment.csv","r",encoding='utf-8') as csvfile:

    reader=csv.reader(csvfile)

    for line in reader:

        print(line[1])

        content+=line[1]

  
（4） 使用结巴分词组件做中文分词

代码：

wordlist=jieba.cut(content,cut\_all=True)

stopwords=[line.strip() for line in open('stopwords.txt', 'r', encoding='utf-8').readlines()]

outstr = ''

for word in wordlist:

    if word not in stopwords:

        if word != '\t':

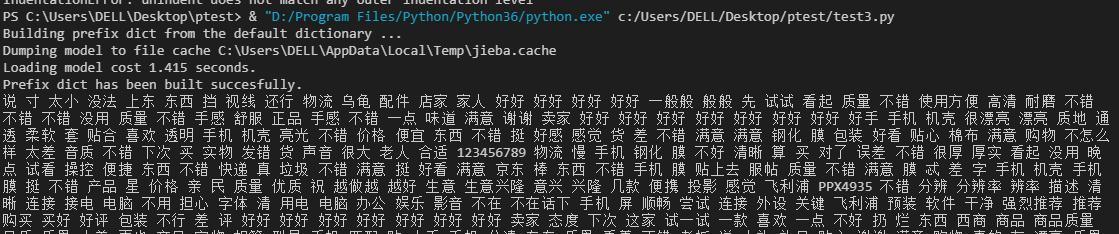
            outstr += word

            outstr += " "

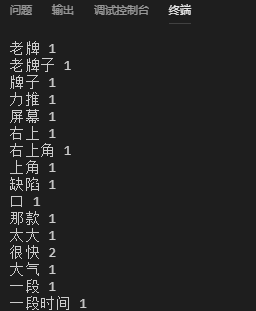
print(outstr)

worddict=dict()

wordlist1=outstr.split(' ')



（5） 分词结果进行词频统计  
with open(dataFileName) as wf, open(sortFileName,'w') as wf2,  
open(tmpFileName, 'w') as wf3:  
for word in wf:  
word\_lst.append(word.split(' '))  
for item in word\_lst:  
for item2 in item:  
if item2 not in word\_dict:  
word\_dict[item2] = 1  
else:  
word\_dict[item2] += 1



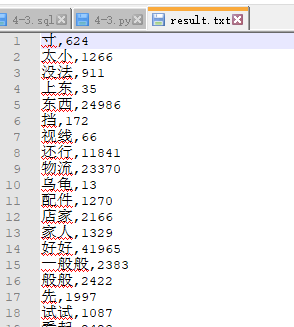
（6） 词频统计结果写入新的 txt 文件

with open("result.csv","w",encoding='utf-8') as f:

    for i in worddict:

        f.writelines(str(i)+","+str(worddict[i])+"\n")

3. 运行编写好的 python 脚本， 得到词频统计结果文本文件



4. 新建 hive 表加载词频统计结果数据  
drop table if exists comment\_word\_count\_tb;  
create table comment\_word\_count\_tb(  
word string,  
count int  
)

row format delimited fields terminated by ' ';

load data local inpath '/xxx/result.txt' into table comment\_word\_count\_tb;  
5. 查看数据是否保存成功

