

只取评价内容，保存成新的 txt 文件（分隔符保持一致）

```
1 说的32寸可以用，太小了根本没法上
2 东西很好有点挡视线，还行.....这个物流跟乌龟一样快
3 配件方便好用，店家人很好，
4 好好好好好
5 一般般.....
6 先试试，看起来很方便
7 质量不错，使用方便，高清耐脏
8 非常不错
9 非常不错
10 还不错，没用呢
11 质量不错，手感很舒服
12 还可以吧！
13 很好，是正品
14 手感很不错，而且一点味道都没有，很满意，谢谢卖家
15 好好好好好好好好
16 手机壳很漂亮，质地通透，很柔软，而且套上非常贴合，很喜欢透明的手机壳。
17 亮光
18 不错
19 价格很便宜，东西不错
20 挺好
21 感觉一般吧
22 ！()】
23 可以可以.....
24 货太差
25 很不错，很满意
26 好
27 很满意，钢化膜的包装很好看，里面很贴心的有棉布很满意的二次购物
28 不怎么样太差了
29 音质不错，下次继续买。
30 实物不同，发错货
31 声音很大，很好用，老人用很合适
32 123456789
33 物流慢手机钢化膜不好
34
```

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

```
content = ""

try:

    fo = open(filename)

    print("读取文件名: ", filename)

    for line in fo.readlines():

        content += line.strip()

    print("字数: ", len(content))
```

- (4) 使用结巴分词组件做中文分词

```
rawContent = readFile(rawFileName)

r = '[0-9\s+\.\!/\_,$%^*()?;,:-【】+\\"'+|+——! , ;: 。 ? 、
~@#¥%……&* ( ) ]+'

rawContent = re.sub(r, " ", rawContent)

seg_list = jieba.cut(rawContent, cut_all=False)

writeFile(dataFileName, " ".join(seg_list))
```

- (5) 分词结果进行词频统计

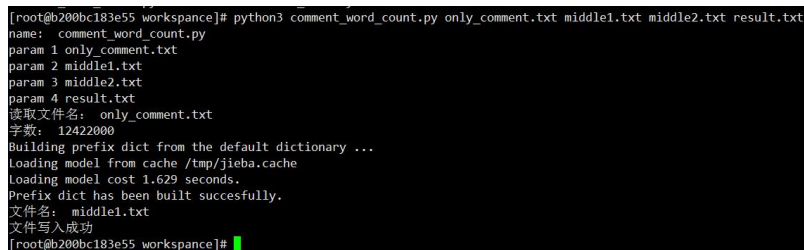
```
with open(dataFileName) as wf, open(sortFileName, 'w') as wf2,
open(tmpFileName, 'w') as wf3:
```

```
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

```
word_items.sort(reverse = True)
for item in word_items:
    wf2.write(item.label+' '+str(item.times) + '\n')
```

3. 运行编写好的 python 脚本，得到词频统计结果文本文件



```
[root@b200bc183e55 workspace]# python3 comment_word_count.py only_comment.txt middle1.txt middle2.txt result.txt
name: comment_word_count.py
param 1 only_comment.txt
param 2 middle1.txt
param 3 middle2.txt
param 4 result.txt
读取文件名: only_comment.txt
字数: 12422000
Building prefix dict from the default dictionary ...
loading model from cache /tmp/jieba.cache
loading model cost 1.629 seconds.
Prefix dict has been built successfully.
文件名: middle1.txt
文件写入成功
[root@b200bc183e55 workspace]#
```

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

四、实验成果

本次实验完成后，需要得到以下结果：

- 实现用户评价信息中的中文分词；
- 实现中文分词后的词频统计；
- 在 hive 中新建词频统计表加载分词数据；