1. **整体说明**

　　predeal\_data.py文件主要实现将训练文本和测试文本转化成具有一定格

　　式，用于后期情感分析任务的处理：

　　训练文本转化的格式为: id\tview\pol\ttext

　　测试文本转化为：id\tview\ttext

　　(id表示记录编号，view表示视角,pol表示情感，text表示相应的句子，\t表示分隔符)

1. **函数介绍**

* **get\_new\_path (train\_or\_test,isSecond,isSeg):**

根据不同的处理方式设置不同的存储路径，

Train\_or\_test:为0表示是训练集

isSecond:1表示是复赛训练集

isSeg:是否拆分，拆分的意思为对于一个句子中出现多个视角，按语义将文本拆分给不同的视角或者直接把整个句子作为视角的文本，此框架处理采用不需要划分方式。

* **predeal\_data(t\_path,t\_view\_path,t\_data\_result\_path,train\_or\_test,isSecond,isSeg):**

t\_path:表示文本train 或者 test，

t\_view\_path：表示抽取的视角文件

t\_data\_result\_path:结果文件

train\_or\_test:表示处理的是训练集还是测试集，0表示训练集，1表示测试集

isSecond：表示处理的是初赛的数据还是复赛的数据，1表示复赛

isSeg:表示要不要按视角对整个句子进行划分，1表示要划分

* **trans\_space(raw\_path,del\_path,train\_or\_test,isSecond,isSeg):**

采取jieba分词，如果视角中出现空格，要做特殊处理，转化成^号，后面相应的句子中出现的此视角中的空格也为^。

Raw\_path:表示待处理文件

Del\_path:表示结果文件

Train\_or\_path：处理的是训练集还是测试集，0表示训练集，1为测试集

isSecond：表示是否是复赛数据集，0表示初赛数据集

isSeg：表示是否需要切分，0表示不需要，1表示需要

* **filter\_view(text\_path,result\_path,train\_or\_test,isSecond,isSeg)**

根据统计得到的规则，对文本进行视角过滤，过滤不符合规则的视角。

1. **运行说明**

对于训练集，只需运行：predeal\_data以及trans\_space函数，得到的结果即为训练集结果(采取初赛和复赛训练集结果拼接而成)

* 运行初赛训练数据

predeal\_data('Train.csv', 'Label.csv', 'train\_id\_view\_pol\_t.txt', 0,0,False)  
trans\_space('train\_id\_view\_pol\_t.txt', 'train\_id\_view\_pol\_trans\_t.txt', 0,0,False)

* 运行复赛训练数据

predeal\_data('TrainSecond.csv','LabelSecond.csv','train\_second\_id\_view\_pol\_t.txt', 0,1,False)

trans\_space('train\_second\_id\_view\_pol\_t.txt','train\_second\_id\_view\_pol\_trans\_t.txt', 0,1,False)

* 手动拼接train\_id\_view\_pol\_trans\_t.txt和

train\_second\_id\_view\_pol\_trans\_t.txt文件作为下步算法的输入

* 运行测试集数据

输入的generate\_view.txt为视角抽取结果

predeal\_data('Test.csv','generate\_view.txt','test\_second\_id\_view\_text\_t.txt', 1,0,False)

trans\_space('test\_second\_id\_view\_text\_t.txt','test\_second\_id\_view\_trans\_t.txt', 1,0,False)

filter\_view('test\_second\_id\_view\_trans\_t.txt','test\_second\_id\_view\_t.csv',1,0,False)

得到的‘test\_second\_id\_view\_t.csv’用于下步情感分析处理