dataProcess使用说明：

简介：

dataProcess是本萌新对前段时间数据处理方面所作工作的一次汇总结果，其主要使用了pandas工具包完成一些本萌新在项目中遇到的一些需求

依赖项：

Pandas

nltk及其的停用词库，情感词库

功能介绍：

首先将工具包放入您所需项目文件夹下，用import将其导入  
import data\_process.data\_process

1. 创建一个dataProcess对象

dp = dataProcess(file=“”)

参数：

file:可选参数，是一个路径名（不要求是绝对路径），如果输入不是绝对路径则会在当前工作路径（调用的脚本所在的路径）寻找data文件夹，如果不在data中就会寻找工作路径所在的文件夹，如果都不存在，则不会初始化成功，但不会导致无法使用

2.将数据集文件转化为pandas中的数据框（dataFrame）结构

df = dp.to\_data\_frame(file="", save\_to\_file ="")

参数：

file:可选参数，是路径名（不要求是绝对路径），效果同初始化dp时，注意：若初始化dp时使用不同的路径，则以函数的参数优先

save\_to\_file：可选参数，是路径名（不要求是绝对路径，不是绝对路径时将保存在工作目录下的data文件夹中，没有会创建一个），将转化的dataframe保存到文件中，格式为csv，在其值不为空时有效

返回一个dataframe对象

3.将dataframe保存到文件

dp.save\_to\_file(path, df=pd.DataFrame(), csv\_type =True)

参数：

path：必选参数，保存文件的路径名，不是绝对路径时会在当前路径中寻找data文件夹放入，没有data时会自己创建

df：可选参数，一个dataframe对象，若没有传入则使用对象初始化时的file创建对象

csv\_type：可选参数，默认为真，为真是表示保存为csv格式，否则保存为json格式

无返回值

4. 过滤不必要的列属性

df = dp.filter\_attrs（attrs\_keeped, df=pd.DataFrame()）

参数：  
attrs\_keeped：必选参数，一个列表格式，内容为要保留的列属性

df=pd.DataFrame()：可选参数，内容为一个dataframe，如果没有传入则会使用初始化的file获得的dataframe，都没有则失败返回

返回值：一个dataframe对象

5. 对列名的规范化操作，可以重命名，重排顺序以及合并某些列（默认使用‘-’连接）

df = dp.reform\_attrs(reform\_dict, df=pd.DataFrame())

参数：

reform\_dict：必选参数，重新规范的格式，具体写法如右： {"new\_name": ["old\_name1", "old\_name2"], "new\_name2": "old\_name3"}，字典的key为新 名字，如果值为列表则将列表中的值合并，否则只是用新名字替换旧名字，同时会按照字典中的出现顺序对列名进行重排

df=pd.DataFrame()：同以上方法

返回值：一个dataframe对象

6. 按属性分组，抽取某属性的某个/些值

groups = dp.group\_data\_by\_attr\_value(attr, df=pd.DataFrame(), value="", values=[], gather=False, grouped\_attrs=[], save\_to\_dir="", save\_as\_pickle =False, read\_only=True)

参数：

attr:必选参数，分组所依赖的属性

df:同以上

value：可选参数，字符串，保留的具体值

values：可选参数，列表，保留的值列表

gather：此选项决定是否将得到的一系列分组文件进行合并，若合并则返回一个dataframe，否则返回一个字典，key为属性名，value为dataframe对象

grouped\_attrs：可选参数，列表，选择的保留属性

save\_to\_dir：可选参数，保存至文件或文件夹内，取决于gather的值，若不是绝对路径则在data文件夹中建立文件夹或文件

save\_as\_pickle：该值在save\_to\_dir不为空时生效，若为真则以pickle的形式存储

read\_only:该值在save\_to\_dir不为空时生效，若为真则设置文件为只读

返回值：返回一个字典或dataframe

7.对数据进行简单的统计工作

dp.do\_statics(attrs, df=pd.DataFrame(), show\_axis =1, show\_details=False)

参数：

attrs：必选参数，进行统计的列，要求为列表格式，进行一个大略的统计，展示数目，不重复数目等信息

df：同以上

show\_axis：0或1，为1时将多列的结果作为列展示，否则作为行

show\_details：展示这些列的每个属性值的频数，和这些频数的大概分布

暂无返回值

功能还待完善，尚缺少可视化和保存

8.分割数据集

df1, df2 = dp.splite\_data(extract\_dep\_attr, df=pd.DataFrame(), extract\_persent=0.6)

参数：

extract\_dep\_attr:必选参数，字符串，抓取数据所依赖的属性

df:同上

extract\_persent：在每个依赖属性值中的抽取比例

返回值，两个dataframe对象，前一个是拥有extract\_persent比例的

9.从数据集中抽取评论文本的aspect和opinion以及opinion的sentiment

dp.get\_reviews\_aspect(review\_attr\_name, corpus\_type, keeped\_attrs, group\_attr="", df=pd.DataFrame(), reviews\_path="", save\_to\_file="")

参数：

review\_attr\_name：必选参数，字符串，评论文本的属性名

corpus\_type：必选参数，字符串，数据集种类，例如：“Cell\_Phones”

keeped\_attrs:必选参数，字典，key为要改为的名字，value为要进行保留的属性在数据集中的名字

group\_attr：可选参数，字典，即按改属性将数据划分为若干组后，按组进行aspect抽取，拥有更好的精确度，建议选择传入

df:同上

reviews\_path：可选文件，若已有系列的分组文件，则作为参数传入

save\_to\_file:将结果保存到的文件，格式是json

返回值是一个字典，key为自动找的id拼接而成，值为一个json对象