**TensorFlow｜熟悉中文情感分析库SnowNLP**

1. 基本环境

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

1. 核心问题：了解中文情感分析库SnowNLP,并利用其进行分词
2. 解决过程：

* 导入SnowNLP包
* 读入数据集
* 对数据进行基本操作，熟悉phthon操作
* 利用SnowNLP处理数据，得到s
* 利用s.words进行分词

1. 代码：

#导入SnowNLP包

from snownlp import SnowNLP

import pandas as pd

#读入数据集

df=pd.read\_excel('trainData.xlsx')

count = 0

#对读入数据进行基本操作

df.head()

df.head(1)

df.dtypes

df.index

df.describe

df.iloc[3:5,1:4]

#利用SnowNLP处理数据

while (count < 3):

print ("The count is:", str(df.iloc[count,1]))

strdata=str(df.iloc[count,1])

s = SnowNLP(strdata)

print("---------------")

print(s.words)

print("---------------")

count = count + 1

print ("Good bye!")

#对得到的s进行操作

s.words

1. 结果：

s.words

Out[68]:

['热水器','加','热','时间','太','长','，','安装','费','太','贵','，','预留', '太阳能','口', '摆设','，','根本','用','不','到','，',...,'斜','着','看','才','能','看到','，']

1. 报错记录：

* 乱码

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xc6 in position 8: inva

解决：subline查看文件乱码，修改后不是乱码

* ParserError: Error tokenizing data. C error: Expected 5 fields in line 17077, saw 7

原因：\r的错

解决：尝试另一种输入方法

df=pd.read\_xlsx('/Users/xuyizhou/Desktop/trainData.xlsx')

//false

df=pd.read\_excel('/Users/xuyizhou/Desktop/trainData.xlsx')

//true