


人智导·情感分析实验

计94 周葵 2019011301
2021.6.6

一、文件目录

文件目录如下：















<input type="checkbox"/> 名称	修改日期	类型	大小
 isear_v2	2021/6/5 10:54	文件夹	
 model	2021/6/6 1:55	文件夹	
 out	2021/6/5 10:03	文件夹	
 vec	2021/6/5 11:02	文件夹	
 __init__	2021/6/5 22:15	Python 源文件	1 KB
 2019011301_周葵	2021/6/6 14:45	Markdown File	20 KB
 2019011301_周葵	2021/6/6 14:45	Microsoft Edge PD...	1,937 KB
 config	2021/6/5 23:28	配置设置	1 KB
 EarlyStopping	2021/6/5 15:57	Python 源文件	2 KB
 model	2021/6/6 1:25	Python 源文件	17 KB
 pre_w2v	2021/6/5 11:00	Python 源文件	1 KB
 README	2021/6/6 14:35	Markdown File	1 KB
 README	2021/6/6 14:49	Microsoft Edge PD...	124 KB
 test	2021/6/5 11:12	Python 源文件	8 KB
 w2v	2021/6/5 11:02	Python 源文件	1 KB
 wash	2021/6/5 11:01	Python 源文件	2 KB

/isear_v2/

存有数据集、停用词以及经清洗过的数据集 (train.csv, valid.csv 与 test.csv)。

/model/

训练完的不同模型，包含以下几个：

 cnn_32_0.001.pkl	2021/6/5 17:56	PKL 文件	4,278 KB
 cnn_64_0.001.pkl	2021/6/5 17:59	PKL 文件	4,473 KB
 cnn_128_0.001.pkl	2021/6/5 22:01	PKL 文件	4,863 KB
 cnn_128_0.01.pkl	2021/6/5 22:04	PKL 文件	4,863 KB
 cnn_128_0.1.pkl	2021/6/5 21:46	PKL 文件	4,863 KB
 cnn_256_0.001.pkl	2021/6/5 19:24	PKL 文件	5,643 KB
 gru.pkl	2021/6/5 16:16	PKL 文件	4,861 KB
 gru_attention.pkl	2021/6/5 17:00	PKL 文件	4,862 KB
 lstm.pkl	2021/6/6 1:55	PKL 文件	5,119 KB
 lstm_attention_32.pkl	2021/6/5 22:30	PKL 文件	4,245 KB
 lstm_attention_64.pkl	2021/6/5 23:17	PKL 文件	4,473 KB
 lstm_attention_128.pkl	2021/6/6 1:55	PKL 文件	5,121 KB
 lstm_attention_256.pkl	2021/6/5 23:41	PKL 文件	7,185 KB
 mlp.pkl	2021/6/5 11:20	PKL 文件	4,203 KB

/out/

不同模型的输出结果 (label)，其中 std.txt 为标准输出。

/vec/

包含了预训练完的 word2vec，以及构建词向量所用的词库。

config.ini

配置文件，包含了网络所需要的参数，分为以下 7 块：

[ALL], [CNN], [LSTM], [GRU], [LSTM_ATT], [GRU_ATT], [MLP]

分别对应全部模型共同的参数以及每个模型各自的参数。

wash.py

对原始数据进行清洗的 python 源文件。

pre_w2v.py

预训练词向量前先构建词库的 python 源文件。

w2v.py

预训练 word2vec 词向量的 python 源文件。

EarlyStopping.py

用于训练提前终止的模块，参考了一个 github 项目，具体在报告中已经说明。

model.py

模型训练所用的 python 源文件。

test.py

模型测试所用的 python 源文件。

二、运行方法

1. 数据清洗以及预训练

(1) 数据清洗

命令行运行：

```
python wash.py
```

则可以看到在 /isear_v2/ 目录下出现了 train.csv, valid.csv 与 test.csv 三个经过清洗的 csv 文件。

(2) 构建词库

命令行运行：

```
python pre_w2v.py
```

则可以看到 /vec/ 目录下出现 sentence.txt，包含了语料库的所有词语。

(3) 预训练 word2vec

命令行运行：

```
python w2v.py
```

则可以看到 /vec/ 目录下出现了 my_w2v_128.w2v，里面存有预训练的 128 维词向量。

2. 模型训练

在 config.ini 中，可以设置不同模型的参数。

参数设定完毕后，在命令行运行：

```
python model.py MODEL
```

这里的 MODEL 必须是下列模型名中的一种：

cnn, lstm, gru, lstm_attention, gru_attention, mlp

示例：

```
python model.py cnn
```

该 python 源文件会根据命令行参数自动训练对应种类的模型。

训练完毕后，模型的输出结果位于 /model/ 目录下。

3. 模型测试

命令行运行：

```
python test.py MODEL FILE
```

这里的 MODEL 和训练中一样，必须是上面的六个中的一个。而第二个参数 FILE 则表示模型文件名字，是在 /model/ 目录下的文件的名字。

示例：

```
python test.py lstm_attention lstm_attention_128.pkl
```

该 python 文件将在命令行中显示 accuracy 以及 f1-score。预测得到的标签，将输出到 /out/ 目录下对应模型名字的 txt 文件中。