网络不良信息动态监测技术+半监督深度学 习文本分类

© Created @Mar 8, 2021 7:36 PM

开发环境

Python + Tensorflow + Selenium

- 1. Python 3.8,需要安装下列库,一般直接 `pip install` + 包名即可
 - numpy
 - tensorflow
 - tensorflow-hub
 - beautifulsoup4, 4.9.3
 - lxml
 - 其他如果有报错说缺什么,就装什么
- 2. PyCharm Professional 版本
- 3. Selenium 需要一个浏览器的驱动程序,建议安装好 chrome 浏览器,项目中把 chrome 用到的 driver 放 讲去了

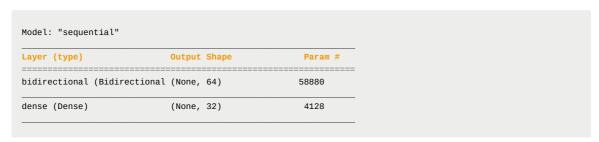
项目结构

数据模块、文件

- 1. raw_data/data.txt: 用来训练模型的原始数据文件,通过 data_util.py 脚本进行格式化处理,写入到 data 文件夹下的 training、test、unlabel 三个文件中,作为训练、测试数据
- 2. nnlm-zh-dim50:来自 tensorflow-hub 的中文文本词嵌入模型,可以将中文文本转化成向量表示。参考 链接: https://tfhub.dev/google/tf2-preview/nnlm-zh-dim50/1
- 3. data_util.py 完成数据相关的处理逻辑,还提供了从中文文本到向量化表示的接口,在其他模块中被引用

深度学习模块

- 1. Tensorflow 实现的双向循环神经网络模型,进行文本分类: Istm.py 中 SemiLSTM 类的 build_lstm 方法 构建了一个深度双向循环神经网络模型(bidirectional),通过一层循环神经网络得到的文本特征,加上两 个全连接层(Dense),输出一个数值,根据最终向量运算结果转化成分类的结果:概率大于 0.5 即为"不 良文本"
 - 模型配置(样例)



模型参数

```
lstm = SemiLSTM(lr=1e-4, epochs=20, batch_size=50)
lstm.build_lstm(lstm_dims=[32], dense_dim=32)
```

SemiLSTM 的构造函数中,传入学习率 Ir,训练迭代轮数 epochs 和批数据大小 batch_size build_lstm 则定义了这个深度网络中,LSTM 层的数量和规模 lstm_dims,数组结构,数组长度即为层数,数值为对应层的特征维数;以及全连接层的输出维度 dense_dim

2. 半监督学习: SemiLSTM 中的 train_semi 即为半监督学习的过程。

```
train_semi(self, train_data, train_label, test_data, test_label, unlabeled_data, round, saved_model='my_lstm'):
```

主要策略:由 round 决定共进行多少轮的半监督学习。每一轮迭代都会选取"未标记数据"中最有可能的几个正负样本,加入到训练集合中去,在下一轮一起更新模型参数,使模型得到增强

爬虫获取微博动态分析

- 1. weibo_spider.py,通过 Selenium 驱动启动一个浏览器,模拟人的行为,获取 HTTP 请求中的数据,获取当前能看到的热门微博,获取微博的文本、作者、时间等信息
- 2. 根据获取到的文本,调用训练好的深度神经网络模型,进行文本分类,连同结果一起写入到本地文件中

运行方式

- 1. 数据清洗部分,运行 data_util.py 程序
- 2. 深度学习模型运行 main.py
- 3. 微博爬虫+动态分析: 运行 weibo_spider.py,根据 weibo_spider.parse(pages=3) 中的 pages 决定爬取 多少数量的微博