README

基本框架

本程序使用mongodb作为后端数据库，在text_extractor模块使用nltk,stanford core nlp, stanfore parser,jieba或者ansj进行语句处理，在word2vec进行词向量训练，在predictor使用tensorflow训练模型，最后利用gunicorn和flask进行网页上的模型展示。

本程序有以下接口模块:

- scripts.py 包含用于调试，导入数据，检查的函数
- text_extractor.py 包含分词器，分词同时也会进行词根化，词性还原，专有名词判断等工作。
- plain_predictor.py 基于语法的情感分析工具

mes_holder模块读取data/config/_.yml中的模型信息，可以是LSTM和NOLSTM(分别对应CNNLSTMPL模型和CNNPL模型)。
以下模块的超参数由mes_holder模块统一管理，在使用之前需要确定yml文件中的配置信息正确。

- word2vec.py
  - 用于训练词向量。
  - 需要修改写在**主函数**中的参数，以让程序明确使用哪一份yml配置文件。
- predict_LSTM.py和predict_NOLSTM.py
  - 用于训练，测试模型
  - 使用方法为`python predict_LSTM.py <collection_name> LSTM`或者`python predict_NOLSTM.py <collection_name> NOLSTM`
- demo_service.py
  - 用于模型展示

简明流程

1. 安装python，使用pip安装相关依赖包如tensorflow, flask, gunicorn, yaml, jpype,

matplotlib, nltk, pymongo等
    - 如果出现错误，请检查是否安装了gcc或python本身的c语言依赖包。
2. 安装mongo，运行mongo
3. 导入数据，数据格式为{"text": 评论内容, "tag": 情感正负中极性}
    1. 修改参数并运行scripts.py导入nlpcc2014数据
        - 下载地址: http://tcci.ccf.org.cn/conference/2014/dldoc/evtestdata2.zip 注意一次只能导入一个文件。
    2. 也可以运行ctrip/PageCrawler.py爬取数据。
4. 分词，进行预处理
    1. 修改text_extractor.py中的参数:中英文，collection name。
        - 例如若collection name为nlpcc_en，语言为英文，则将主函数修改为

```
cutter = WordParser()
import utils
nlpcc_en = utils.get_docs("nlpcc_en")
for record in nlpcc_en.find():
    record['words'] = cutter.split(record['text'], "en")
    nlpcc_en.save(record)
print 'completed!'
```

    2. 训练词向量: 修改参数并运行word2vec.py
        - 若collection name为nlpcc_en，要为LSTM模型训练词向量，则

```
mes = mes_holder.Mes("nlpcc_en", "LSTM", "W2V")
```

5. 运行predict_LSTM.py或者predict_NOLSTM.py进行训练。
    - 若collection name为nlpcc_en，要训练LSTM模型，则运行指令 `python predict_LSTM.py nlpcc_en LSTM`，根据提示输入模型名称`<model_name>`。
6. 修改predictors，并运行demo_service.py展示网页。
    - 如仅仅使用nlpcc_en的结果：

```
predictors = {
    "nlpcc_en_NOLSTM": predict_NOLSTM.PredictorNOLSTM('nlpcc_en', <model_name>, trainable=False),
    "nlpcc_en_LSTM": predict_LSTM.PredictorLSTM('nlpcc_en', <mode
```

```
l_name2>, trainable=False)
}
```

yml参数介绍

参数	含义	类型
LANG	语言	en或zh
LABEL_NUM	分类数目	整数
W2V_FILTER_NATURES	过滤稀有词时特殊处理的词性	词性或'all', None组成的字符串数组
W2V_VOC_LIMITS	特殊处理的词性对应的最低词频	正整数数组
W2V_DELETE_RARE_WORD_FFIDS	过滤稀有词的特征编号	正整数数组
W2V_DELETE_RARE_WORD_TFIDS	过滤后的特征编号，对应W2V_DELETE_RARE_WORD_FFIDS	正整数数组
W2V_ONE_HOT_FIDS	word2vec阶段，需要准备转化为ONE_HOT编码的特征编号	正整数数组
W2V_TRAIN_FIDS	word2vec阶段，需要训练词向量的特征编号	正整数数组
W2V_TRAIN_FIDS_EMB_SZ	词向量的维数，对应W2V_TRAIN_FIDS	正整数数组
DG_DIVIDE_FOLD	是否在word2vec时划分fold，用于k-fold，注意若某条数据的is_train字段为false，则会被分到fold_id为0的组中	bool
DG_STEP_BACK	每步数据生成时回退多少个词,LSTM等RNN专用	正整数

参数	含义	类型
DG_STEP_NUM	一次生成多少步数据，LSTM等RNN专用	正整数，建议不要太大
DG_FIDS	生成数据的特征编号	正整数数组
DG_BATCH_SZ	每个训练batch包含的句数	正整数
DG_TEST_BATCH_SZ	每个测试或验证batch包含的句数	正整数
DG_RNUM	重复生成数据的次数	正整数
DG_SENTENCE_SZ	每次处理的单词数	正整数
DG_FOLD_NUM	k-fold划分的集合数，即k	正整数
DG_FOLD_TEST_ID	测试集的fold_id	正整数或-1
DG_FOLD_VALID_ID	验证集的fold_id	正整数或-1
PRE_ONE_HOT_FIDS	正式训练或预测中，需要被转化为ONE_HOT编码的特征编号	正整数数组
PRE_ONE_HOT_DEPTHS	ONE_HOT编码的维数，对应于PRE_ONE_HOT_FIDS	正整数数组
PRE_C_FIDS	正式训练或预测中，需要被转化为离散编码的特征编号	正整数数组
PRE_EMB_FIDS	正式训练或预测中，需要被转化为word2vec编码的特征编号	正整数数组
PRE_CONVS_LEVEL_NUMS	卷积层层数	正整数
PRE_CONVS_KERNEL_NUMS	卷积层的核，每行代表一层	正整数或-1矩阵，-1代表全长
PRE_CONVS_STRIDES	卷积层的步长，每行代表一层	正整数或-1矩阵，-1代表全长

参数	含义	类型
PRE_CONVS_FILTER_NUMS	卷积层的维数，每个元素代表一层	正整数数组
PRE_POOLS_SIZES	池化层的大小，每行代表一层	正整数或-1矩阵，-1代表全长
PRE_POOLS_STRIDES	池化层的步长，每行代表一层	正整数或-1矩阵，-1代表全长
PRE_GOOD_RATE	当验证准确率大于该值时，保存结果	小于等于1的浮点数
PRE_LINEAR1_SZ	线性层1的维数	正整数
PRE_LSTM_SZ	LSTM细胞的维数	正整数
PRE_LINEAR2_SZ	线性层1的维数	正整数
PRE_E_LEARNING_RATE	学习率初始值	浮点数，默认为0.001
PRE_STEP_NUM	学习次数	正整数
PRE_DROPOUT_KEEP_PROB	dropout保留比率	浮点数
PRE_VALID_TIME	每隔PRE_VALID_TIME次，计算准确率并决定是否保存	正整数