代码说明文档

# 【下载数据集至data\_ori文件夹】

# 【word2vec】

1. **训练词向量**

执行[word2vec]文件夹中 train\_word2vec.py 文件

# 【data】

1. **对数据进行统计分析**

|  |  |
| --- | --- |
| 样本数 | 训练集（10.2277w）;测试集（10.2277w） |
| 列名 | article / word\_seg / classify |
| 类别数 | 20类 |
| 有无空缺值 | 无 |
| ‘word\_seg’的句子长度(词级表示) | 2560/mean；24/min；14w/max |
| ‘article’列的句子长度（字级表示） | 5810/mean；111/min；28w/max |

注：

1° 测试集无’classify’列

1. **对数据进行预处理**
2. **代码文件**：

执行**【**data\_process.py】

1. **主要操作**：

删除’article’列；

classify – 1；

对数据进行截断和补零；

word 2 index;

数据转换成numpy数组形式；

划分训练集和验证集 1：0.07；

1. **生成的结果**：

【data\_train.pkl】：(mat\_train, mat\_vali)

【data\_test.pkl】：x

注：

1° x中的样本的顺序在使用过程中不能被改变，必须按id递增的顺序使用。

2° mat\_train/mat\_vali二维矩阵的最后一列为class

# 【models】

存放网络结构模型