

# 一、训练集、验证集、测试集

请先阅读参考阅读资料文件夹内：[一文看懂 AI 训练集、验证集、测试集（附：分割方法+交叉验证）](#)

形象上来说**训练集**就像是学生的课本，学生 根据课本里的内容来掌握知识，**验证集**就像是作业，通过作业可以知道 不同学生学习情况、进步的速度快慢，而最终的**测试集**就像是考试，考的题是平常都没有见过，考察学生举一反三的能力；

类比到机器学习及深度学习中，为了得到最佳的模型效果，需将处理后的语料分为训练集（train）、验证集（dev）、测试集（test），为了**防止模型过拟合**（可以理解为学生背题，没有理解），一般采用**交叉验证法**来对模型进行评估，得到最佳模型；最常用的如**k-fold交叉验证**，以此来降低数据划分带来的影响

## 二、任务说明

### 2.1 总述

将人工标记好的带标签的语料转化为序列标注语料，并按训练集、测试集进行9:1切分（此次任务略去验证集）

处理前：

昆曲又称<ICH-TERM>昆腔<ICH-TERM>、<ICH-TERM>昆山腔<ICH-TERM>，是元末明初南戏发展到昆山一带，与当地的音乐、歌舞、语言结合而生成的一个新的声腔剧种。

处理后：

昆	O
曲	O
又	O
称	O
昆	B
腔	E
、	O
昆	B
山	I
腔	E
、	O
，	O
是	O
元	O

### 2.2 详细说明

## ①语料切分

本次任务为训练集、测试集、验证集的切分，此次任务文本都在data文件夹下，为191个txt文档，同学们需对这191个文档进行采用10折交叉验证的方式切分；

说的简单点，10折交叉就是对所有文本进行打乱切分，然后得到10份比例为9：1不同的训练集、测试集文件夹

处理前：

es (F:) > A文档 > python学习 > 模型学习 > 非遗论文 > data > txt > 传统戏剧					搜索"传统戏剧"
名称	修改日期	类型	大小		
0.txt	2019-11-22 14:45	文本文档	4 KB		
1.txt	2021-05-30 16:08	文本文档	4 KB		
2.txt	2019-11-22 14:51	文本文档	4 KB		
3.txt	2019-11-22 14:51	文本文档	4 KB		
4.txt	2019-11-22 14:51	文本文档	4 KB		
5.txt	2019-11-22 14:52	文本文档	4 KB		
6.txt	2019-11-22 14:53	文本文档	4 KB		
7.txt	2019-11-22 14:55	文本文档	3 KB		
8.txt	2019-11-22 14:55	文本文档	2 KB		
9.txt	2019-11-22 14:56	文本文档	3 KB		
10.txt	2019-11-22 14:56	文本文档	3 KB		
11.txt	2019-10-25 12:39	文本文档	2 KB		
12.txt	2019-10-25 12:39	文本文档	2 KB		
13.txt	2019-11-22 14:58	文本文档	3 KB		
14.txt	2019-11-22 14:59	文本文档	3 KB		
15.txt	2019-11-22 15:00	文本文档	3 KB		
16.txt	2019-11-22 15:00	文本文档	2 KB		

















处理后：

1	2019-10-24 12:07	文件夹
2	2019-10-25 15:06	文件夹
3	2019-10-24 12:07	文件夹
4	2019-10-24 12:07	文件夹
5	2019-10-24 12:07	文件夹
6	2019-10-24 12:07	文件夹
7	2019-10-24 12:07	文件夹
8	2019-10-24 12:07	文件夹
9	2019-10-24 12:07	文件夹
10	2019-10-24 12:07	文件夹

每一个数字文件夹内部为：

test	2019-10-24 12:07	文件夹
train	2019-10-24 12:07	文件夹

继续打开每个train及test文件夹：里面都是原始的txt文档

 txt_4.txt	2019-10-24 11:31	文本文档	2 KB
 txt_11.txt	2019-10-24 11:31	文本文档	2 KB
 txt_17.txt	2019-10-24 11:31	文本文档	3 KB
 txt_24.txt	2019-10-24 11:31	文本文档	3 KB
 txt_28.txt	2019-10-24 11:31	文本文档	2 KB
 txt_43.txt	2019-10-24 11:31	文本文档	3 KB
 txt_62.txt	2019-10-24 11:31	文本文档	2 KB
 txt_66.txt	2019-10-24 11:31	文本文档	2 KB
 txt_94.txt	2019-10-24 11:31	文本文档	1 KB
 txt_95.txt	2019-10-24 11:31	文本文档	2 KB
 txt_97.txt	2019-10-24 11:31	文本文档	2 KB
 txt_106.txt	2019-10-24 11:31	文本文档	2 KB
 txt_110.txt	2019-10-24 11:31	文本文档	2 KB
 txt_151.txt	2019-10-24 11:31	文本文档	3 KB
 txt_156.txt	2019-10-24 11:31	文本文档	3 KB
 txt_159.txt	2019-10-24 11:31	文本文档	2 KB

这几个图应该解释的差不多了，说白了就是打乱原始文件夹，把原始所有txt文档处理成包含不同txt文档的10份，同时每一份都要进行9:1切分为训练集及测试集

**一点小提示：这个任务的相关代码文件网上比较容易找到，难度不是很大，大家重点在于理解切分的意义**

## ②序列标注

详细的使用场景、背后的意义不展开说明了，大家可以自己去找相关资料进行学习；这里只讲一下任务的安排

在上述完成10折交叉切分后，对每一折（处理后的一个文件夹）的训练集及测试集转为**BIEOS序列标注**方式；

实体表示标签包括的内容，如本次需处理的文本都是“<ICH-TERM>实体<ICH-TERM>”，两个标记包含的部分即为实体；

- B 表示标签实体的开头
- I 表示标签实体中间部分
- E 表示标签实体的结尾
- O 表示该字不是实体
- S 表示该字为单标签实体

其中都是以字符为单位进行标注，  
两个字的实体如<ICH-TERM>昆腔<ICH-TERM>，转化后为

昆    B  
腔    E

三个字及以上的实体如<ICH-TERM>昆山腔<ICH-TERM>，转化后为

昆    B  
山    I  
腔    E

单个字的实体如<ICH-TERM>净<ICH-TERM>，转化后为  
净    S

非标签的内容，则都标记为O

### ③格式说明

此次需读取该文件夹下的所有txt文档，对每一个文档进行序列标注转换，每一个非空白字符都需要进行处理

处理前：

昆曲又称<ICH-TERM>昆腔<ICH-TERM>

处理后：

昆	O
曲	O
又	O
称	O
昆	B
腔	E

处理后的格式要求为：字tab标记

也即为两列的csv文本，第一列为字，第二列为标记（BIEOS），中间用tab换行隔开，每一个字符单独成行；

同时若遇见句号、问号、感叹号等表示一句话结束的标点符号需再次换行，如下所示：

处理前：

包括<ICH-TERM>老生<ICH-TERM>、<ICH-TERM>小生<ICH-TERM>、<ICH-TERM>丑<ICH-TERM>等。各行脚色在表演中

处理后：

包	O
括	O
老	B
生	E
、	O
小	B
生	E
、	O
丑	S
等	O
。	O
各	O
行	O

最终处理所有语料后存入一个txt文件中；train文件夹下的所有txt文档处理后存入train.txt，test文件夹下的存入test.txt

因为本次采用了十折交叉验证，所以要处理十次；

### ④代码提示

语料切分部分搜索关键词：k-fold交叉验证

序列标注部分：python正则表达式（提取标签内容），os包（文件夹批处理），文本的读取及写入，序列标签转换算法的编写（这个就是自己想了，主要涉及if、else等逻辑判断）

## ⑤最终提交内容

- 任务中涉及的python文件，最好带注释
- 10折交叉切分后的10个文件夹，每个文件夹下包含一个训练集一个测试集，且训练集内txt文档数量和测试集txt文档数量约为9:1
- 序列标注处理后汇总的train.txt、test.txt，按照上述每一折放在不同的文件夹下