**实验一 用python处理文本数据实验报告**

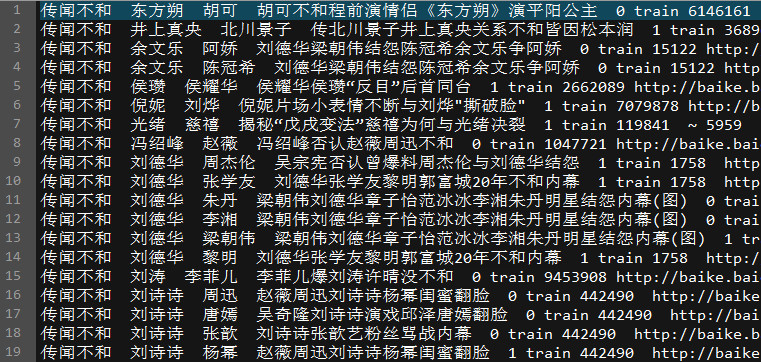
**实验目的：**熟悉python的基本数据结构，以及文件的输入与输出。

**实验环境：**anaconda3

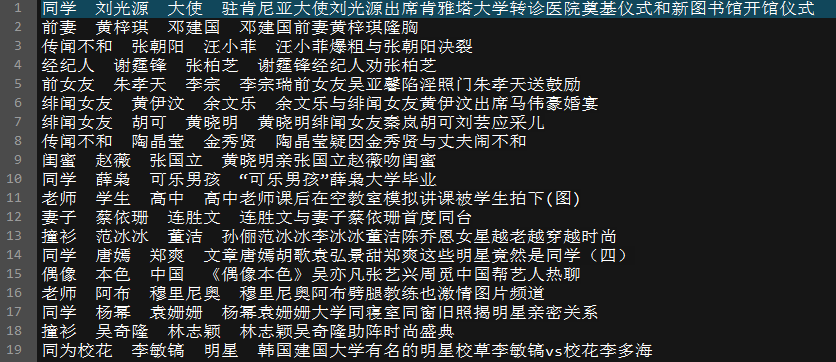
**实验数据：**

利用2015年中国机器学习会议的评测数据和评测任务，数据包括训练集(7813条)和测试集(2610条)，评测任务为通过给定的训练数据，预测测试集中的关系是正例还是负例，在每个样本最后给出1或者0。数据描述如下。

训练集如下图所示，第一列为关系类型，第二列和第三列为人名，第四列是标题，第五列是关系为正例还是负例，1为正例，0为负例；第六列表示训练集还是训练集。



测试集如下图所示，格式基本与训练集类似，唯一不同的是第五列没有关系是正例还是负例的标记。



**实验内容：**

1对训练集数据进行处理，只留下前面五列，输出文本命名为exp1\_1.txt。

2 在第一步得到的数据的基础上对19类关系进行分类，生成的文本存放在exp1\_train文件夹下，按照关系类别出现的顺序，第一个关系类别的数据存放在1.txt中，第二个关系类别存放在2.txt中，直到19.txt。

3 测试集按照训练集的19个类别的顺序将各个样本按照关系类别归类，即相同关系类型的数据放到一个文本文件中，同样生成19个类别的测试文件，格式仍旧和测试文件保持一致。存放在exp1\_test文件夹下，每个类别的文件仍旧命名为1\_test.txt，2\_test.txt…

同时对每个样本在原测试集中出现的位置进行记录，和19个测试文件一一对应起来。比如第一类“传闻不和”的每个样本在原文中处于第几行，在索引文件中进行记录，保存在文件index1.txt，index2.txt….

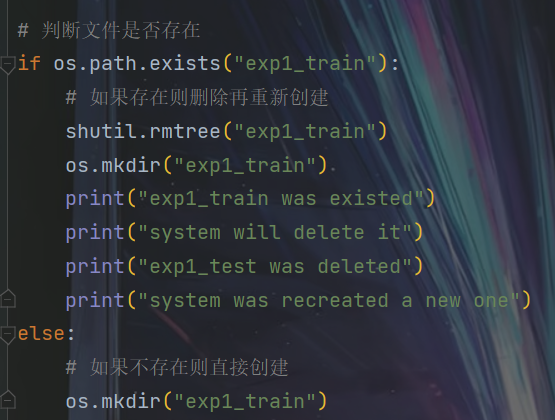
**各功能实现思路：**

**功能1：**对训练集数据进行处理，只留下前面五列，输出文本命名为exp1\_1.txt。

首先打开相关文件，用for循环遍历文件。把文件内容按行进行逐一读取并存入列表。默认的换行分隔符需要改为制表符，测试数据中的每个行的行属性是用制表符分割的。因为文件中的每一行都被存入了列表并且一个制表符一次分割。所以列表的每一个元素就是每一行的相应属性。题目要求留下前五行也就是每一行的前五个属性。因而我们也只需要提取列表的前五个元素用制表符进行衔接字符串，就能会获得每一行的前五个属性。然后把衔接好的字符串写入文件中。就这样逐一遍历测试数据的每一行进行处理再进行写入新文件。即可得到相关功能。

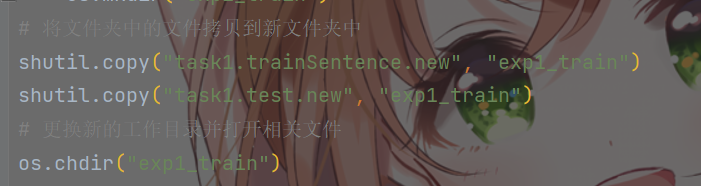
**功能2：**在第一步得到的数据的基础上对19类关系进行分类，生成的文本存放在exp1\_train文件夹下，按照关系类别出现的顺序，第一个关系类别的数据存放在1.txt中，第二个关系类别存放在2.txt中，直到19.txt。

* 首先得判断是否有旧文件夹或者数据。如果有先进行清理操作。

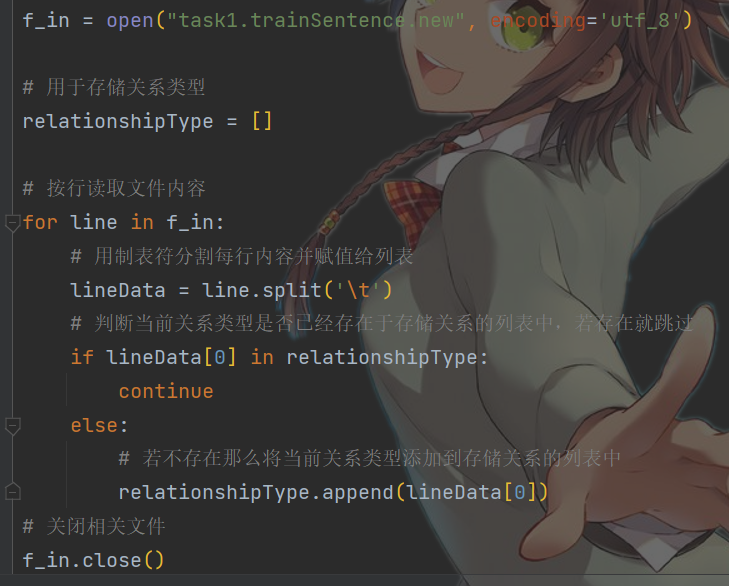


判段是否有旧文件夹，有的话进行全部删除后重新创建。如果没有的话直接创建相关文件夹，不需要删除操作。

* 拷贝一下测试数据，拷贝进目标工作目录，并更换工作目录

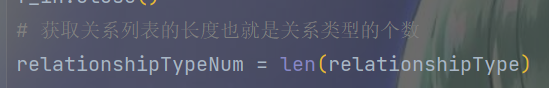


* 也是逐行读取测试数据，观察可以发现，测试数据的第一行都是记录着关系。所有找个列表把测试数据中的关系种类提取出来。



列表的第一个元素就是关系类型，判断关系类型是否已经被记录。若已经被记录那么跳过，若未被记录那么进行记录。这样就能获得一个全数据的关系类型的记录列表。

* 获取关系列表的长度我们就可以获得关系类型的数量。



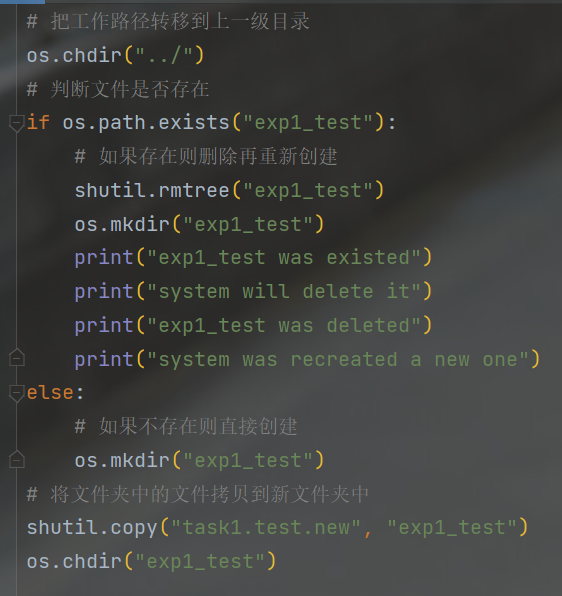
* 按照关系类型列表进行关系分类查找

****

关系类型列表内的顺序就是各个关系出现的顺序。外一层循环遍历关系类型列表。内层循环从测试数据中挑选出符合条件的关系数据进行写入相关文件。文件命名顺序和命名规则就是关系列表元素的顺序和下标+1。直到遍历完关系列表的所有元素即可获得相关分类文件。

**功能3：**测试集按照训练集的19个类别的顺序将各个样本按照关系类别归类，即相同关系类型的数据放到一个文本文件中，同样生成19个类别的测试文件，格式仍旧和测试文件保持一致。存放在exp1\_test文件夹下，每个类别的文件仍旧命名为1\_test.txt，2\_test.txt…

同时对每个样本在原测试集中出现的位置进行记录，和19个测试文件一一对应起来。比如第一类“传闻不和”的每个样本在原文中处于第几行，在索引文件中进行记录，保存在文件index1.txt，index2.txt….

* 首先也是确定好工作目录和处理好旧文件夹和文件，再进行相关文件的创建工作。再把测试数据拷贝到工作目录
* 把测试数据内的关系进行分类存储到单个文件中和建立索引文件



也是利用功能2的方式借助关系类型列表进行关系类型的分类并创建相关文件。但要加入一个索引文件的概念。索引文件命名规则也是用关系类型列表的下标+1来命名。内循环遍历测试文件进行匹配寻找关系类型的时候用一个变量进行记录当前行数。如果在当前行发现了符合遍历条件的函数的话把当前行数写入索引文件内。遍历完关系列表那么就能得到相关文件了。