**隐式篇章关系识别**

**一、任务介绍**

隐式篇章关系识别旨在没有连接词的情况下，判断两个文本片段（称作论元，argument）之间的语义关系，如下面的例子：

**Argument-1:** *We have seen a big advance of the project.*

**Argument-2:** *(But) the others are still very crude.*

两个论元之间存在“对比”关系，*But*为两个论元之间的连接词，我们需要在没有连接词的情况下，通过提取论元的语义特征，将它们之间的关系识别出来。

通常将该问题视作一个分类问题，即给定两个论元，判断它们属于哪种预定的语义关系。

**二、数据集介绍**

实验所用数据集是从PDTB 3.0采样得到，共包含8000个样本，其中6000个为训练数据，2000个为测试数据，分别在train.tsv和test.tsv文件中。如下图所示，数据文件中每行为一个样本，共包含4个元素，分别是样本序号、语义关系标签、论元1、论元2，它们之间用制表符( \t )分隔。（测试数据的标签未给出，文件中对应位置为Unknown）



数据集中共有四种语义关系，分别是对比（Comparison）、因果（Contingency）、扩展（Expansion）、时序（Temporal），每个样本对应其中一种语义关系。

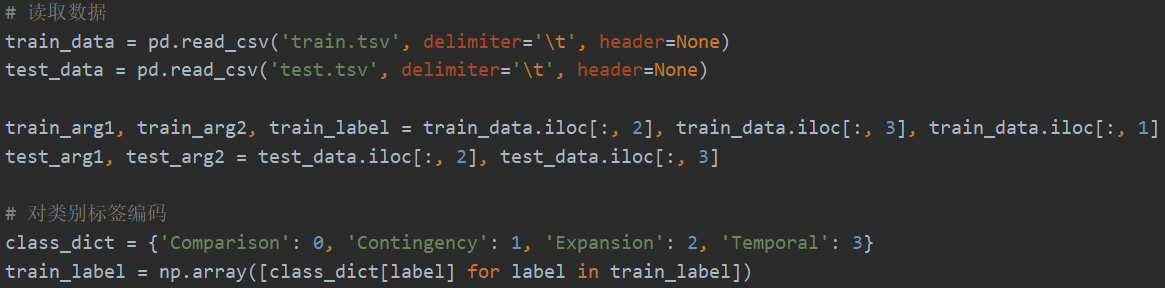
**三、评价指标**

采用准确率（Accuracy）和macro-F1作为分类器的性能评价指标，二者的计算方式如下：

**四、实验思路**

1、加载数据

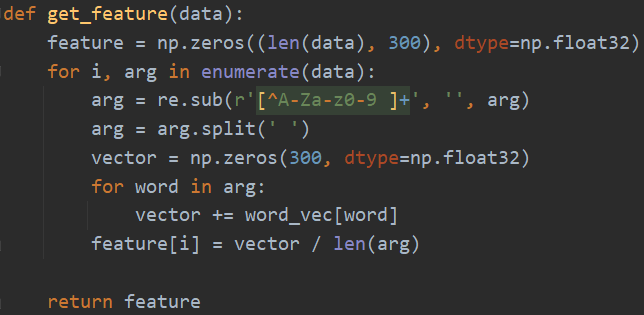
使用pandas库读取数据，并对类别标签进行编码：



2、提取论元特征

由于论元为文本形式，因此需要提取语义特征，将其表示为向量。这里给出一种基于词向量的特征提取方式。

文件glove\_300.pickle中存储了数据集中所有词的向量表示，这些词向量是在大规模语料库中预训练得到的，维度为300维，蕴含了每个词的语义特征。因此，可以把一个论元中所有词的词向量求和再取平均，作为这个论元的特征向量表示，然后把一个样本中两个论元的向量表示拼接起来，输入分类器，得到该论元对的语义关系分类结果。提取特征的具体实现如下：



（**注：**上述论元特征提取方式可作为参考，大家可以基于所给的词向量，自行设计其他的特征提取方法）

3、分类算法

在完成每个论元对的特征提取后，就可以构建分类器进行训练，具体的分类算法，大家可以自行尝试。

文件demo.py中给出了使用SVM进行分类的示例，**仅作参考**。

**五、结果提交**

将测试集的预测结果保存在一个txt文件中，**格式：**每行一个标签，该标签为映射后的数字标签，映射关系为{'Comparison': 0, 'Contingency': 1, 'Expansion': 2, 'Temporal': 3}。然后把预测结果txt文件上传至系统。文件命名为“IDRR实验结果\_班级\_<学号>\_<姓名>.txt”