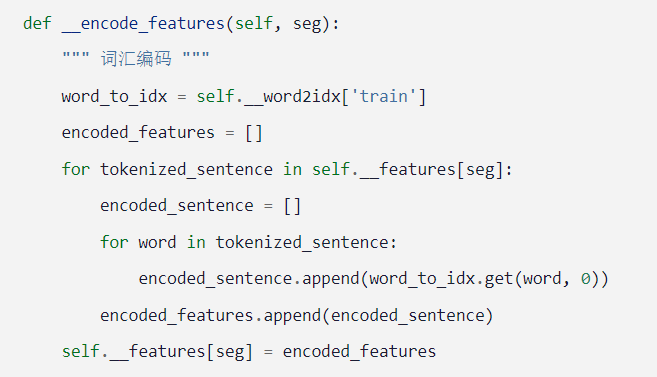
数据预处理

对文本数据集进行词嵌入处理，包括分词、编码、对齐，使之能够适应网络结构。

流程：*sentence->tokenized->encoded->padding->features*

*tokenized：将单词分开，并进行小写处理*

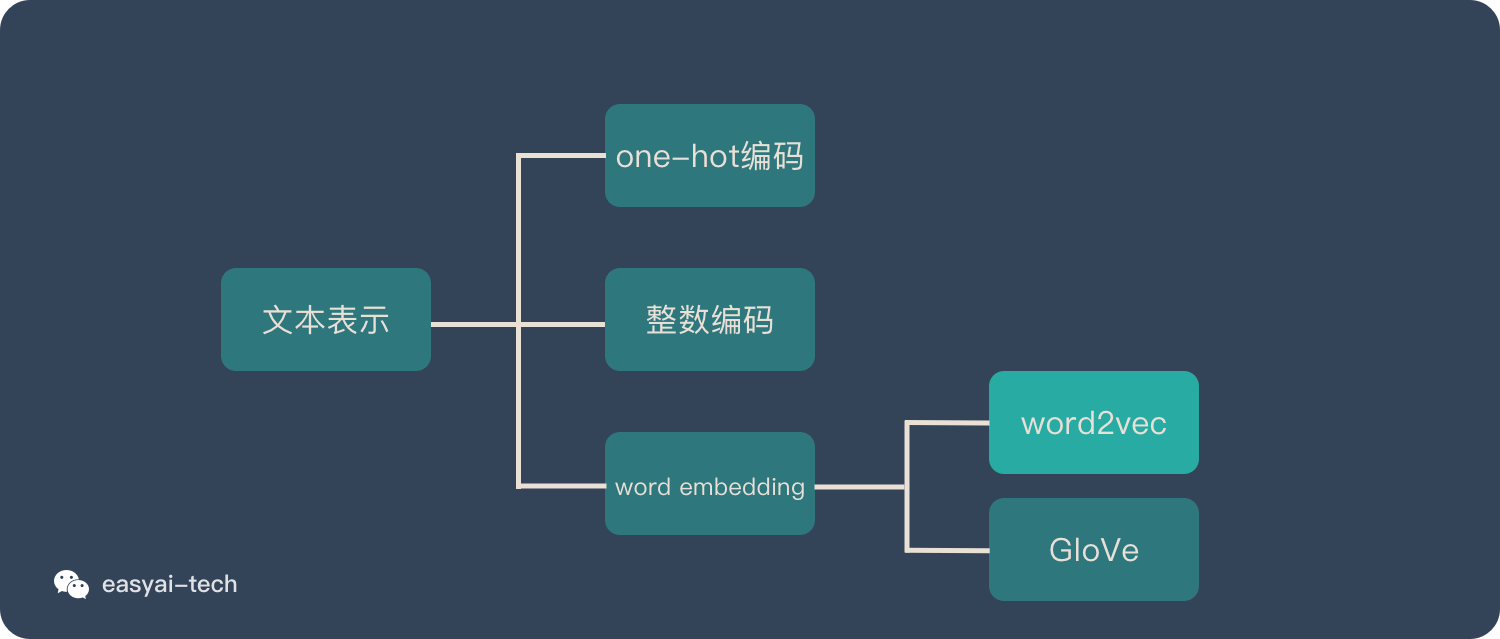
*encoded：先构建词汇表后，再进行句子的编码*



*Padding：将编码的句子填充到最大长度500，方便处理*

数据预处理涉及到的

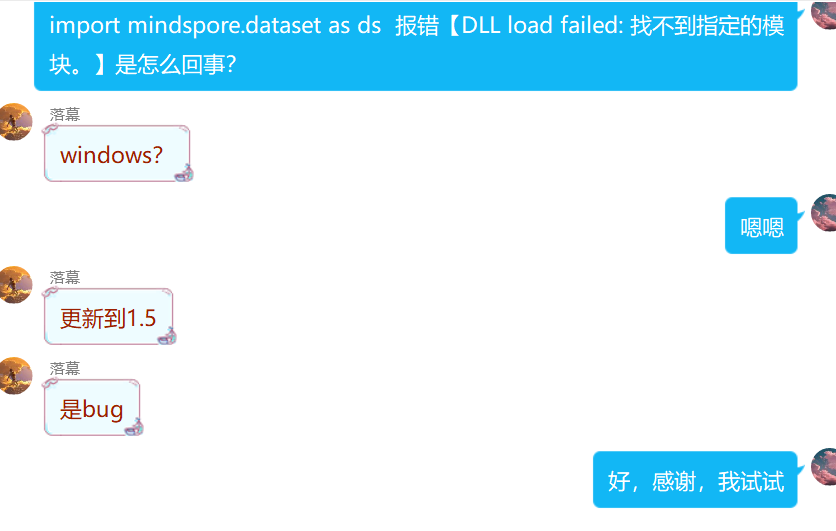
词嵌入算法之Word2vec算法



Word2vec**是 Word Embedding 的方法之一，有两种训练模式CBOW和Skip-gram**

**分别达到通过上下文来预测当前值和用当前词来预测上下文的效果。**

**遇到的问题和总结：**

**在本地运行的时候，在windows系统Mindspore1.3的版本下，导入mindspore.dataset包时会报错找不到模块。最终在询问Mindspore官方交流群后得出了答案，1.3版本存在漏洞，升级成1.5版本即可解决。**