预训练部分

- [x] 加载数据集
- [x] 将数据集转换成两个txt文件
- [x] 对txt文件分词
    - [x] 分词（数字也要分词）
    - [x] 每个句子加上标记符 并填完
- [x] 加载预训练词向量
    - [x] 中文：腾讯中文词汇/短语向量（Tencent AI Lab Embedding Corpus for Chinese Words and Phrases）V0.2.0 dim=200
    - [x] 英文：谷歌新闻， dim=300
- [x] 将句子列表转换为词向量列表（dim2 to 3）
    - [x] 英文:
        - [x] 添加标记符

> **Tip**
>
> 手动添加的特殊标识符的词向量可能不会包含有效的语义信息，但在实践中，这些特殊标识符主要用于控制序列的边界或填充序列长度，它们的语义信息并不像普通词向量那样重要。

```
special_tokens = {
    '<bos>': np.random.uniform(-0.25, 0.25, vector_dim),
    '<eos>': np.random.uniform(-0.25, 0.25, vector_dim),
    '<pad>': np.zeros(vector_dim),
    '<unk>': np.random.uniform(-0.25, 0.25, vector_dim)
}
```

    - `<bos>` 和 `<eos>` 的向量是随机初始化的。
    - `<pad>` 的向量是全零向量。
    - `<unk>` 的向量也是随机初始化的。
        - [x] 生成输入target序列张量并保存
    - [x] 中文:
        - [x] 添加标记符
        - [x] 生成源语言序列并保存

模型生成部分

- [ ] 准备数据
    - [ ] 编码器输入
    - [ ] 解码器输入
    - [ ] 解码器输出标签
    - [ ] 掩码长度

令$\frac{77}{777}$上下乘10可以得到

$$\frac{77}{777}=\frac{770}{7770}$$

令$f(x)=\frac{a+x}{b+x}$,设$a<b$，$b\neq 0$

则可求得$f(x)$的导数$f'(x)=\frac{b-a}{(b+x)^2}$

当$a<b$时，$f'(x)$恒大于0，所以$f(x)$在$x\in R$上单调递增

令$a=770,b=7770$

$$f(7)=\frac{777}{7777}>f(0)=\frac{770}{7770}=\frac{77}{777}$$

证毕

令