Gen\_data.py:

# question: 这里的fact\_in\_annotated\_train代表什么？

真实的（头实体-尾实体-关系）三元组组成的集合

# question: 这里的na\_triple代表什么？

不存在关系的顶点对（实体对）

# question: 这里的bert\_starts\_ends代表什么？

bert\_starts\_ends是一个sen\_tot \* max\_seq\_length \* 2的矩阵，其初始值为max\_seq\_length - 1

# question: token\_start\_idxs的公式中为什么要加1

# question: 为什么要把token\_start\_idxs >= max\_seq\_length-1的都置为max\_seq\_length – 1

Config.py

# question: data\_train\_bert\_mask代表什么？

预训练句子编码时的mask们，这些mask分为两部分，前半部分全是1，后半部分全是零，主要是为了获得使用0补充长度后的token中实际有用的token。（跟划分子网时的掩码概念有点像，都是为了让长度动态变化而设置的）

# question: 这里的neg\_multiple代表什么？

一个权重，它控制sel\_idx的长度，即sel\_ins中元素个数

# question: 为什么要设置一个h\_t\_limit

控制tensor长度，防止炸显存

# question: 这里的self.dis2idx代表什么？

把分类后的距离使用dis2idx映射为20类距离值[0-19]

# question: h\_mapping和t\_mapping有什么区别？

h\_mapping针对的是头实体，t\_mapping针对的是尾实体

# question: max\_h\_t\_cnt有什么作用？

保存不同的实体对的个数

# question: 本代码是如何处理token数量大于512的文档的？

直接忽略超长部分