

句子对的个数： 23027384

句子长度统计:

均值： 34.554794283189096

最大值： 200

最小值： 5

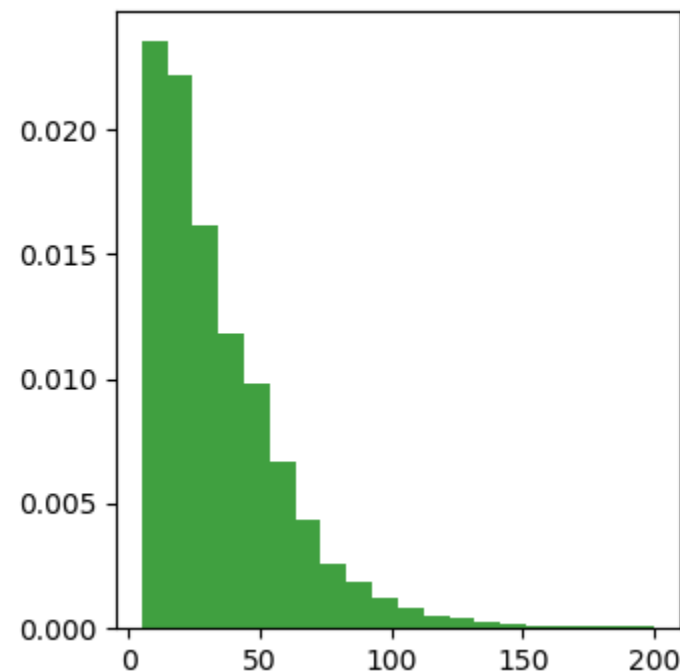
中位数： 28.0

要构造的错误类型:

- 1.同音字错误
- 2.近音字错误
- 3.形近字错误

一句话中的错误个数： 根据句子长短而定

句子长度分布图形



通过统计语料库中句子的长度分布情况,

过于长的句子中设定： 40%的可能性出现两处错误， 50%的可能性出现一处错误， 10%的可能性不出现错误 字数设定： >100

中等长度的句子设定： 10%的可能性出现两处错误， 80%的可能性出现一处错误， 10%的可能性不出现错误 字数设定： 100--10

过于短的句子设定： 2%的可能性出现两处错误， 68%的可能性出现一处错误,30%的可能性不出现错误 字数设定： <10