文本长度分析

以单词数量为计量

合理性：

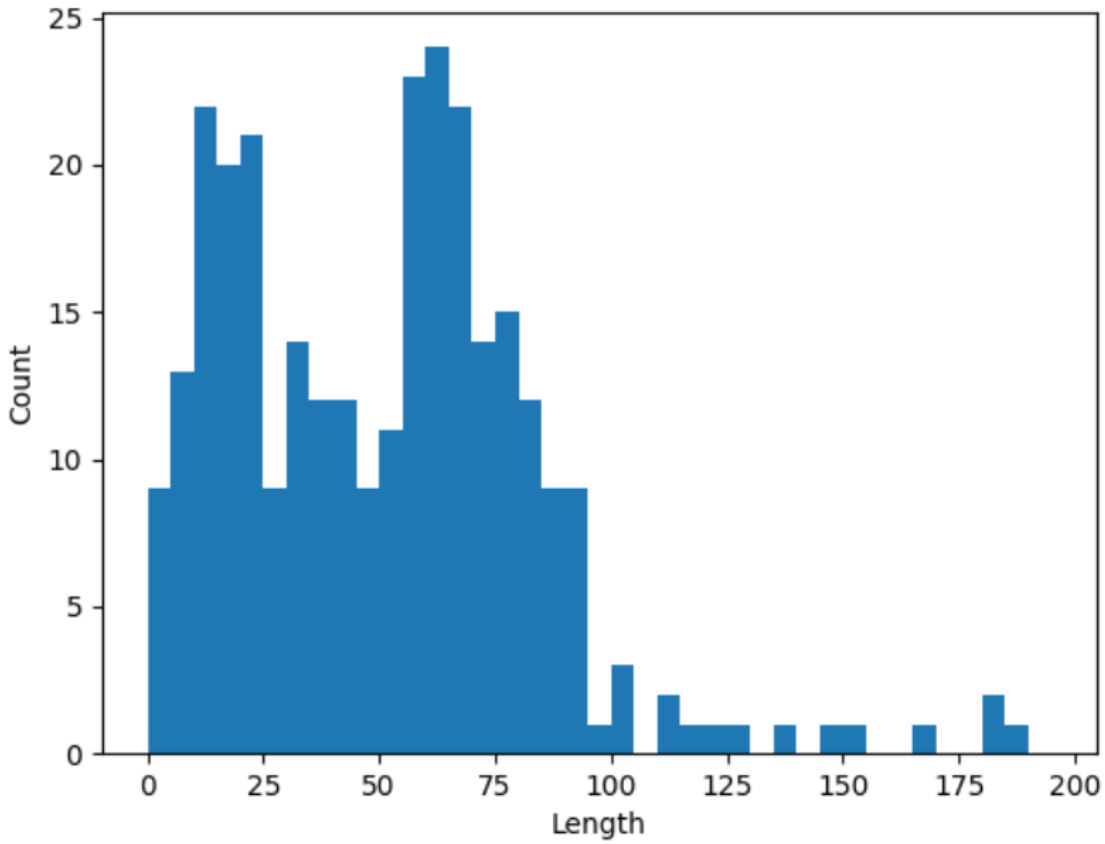
以单词个数作为文本长度能够体现语言无关性，可以适用到英语、中文等多种语言中，使不同语言的句子长度有了一个相对客观的标准。

文本长度分析中使用了第三方jieba库对中文文本进行了分词，以单词的数量作为文本的长度。

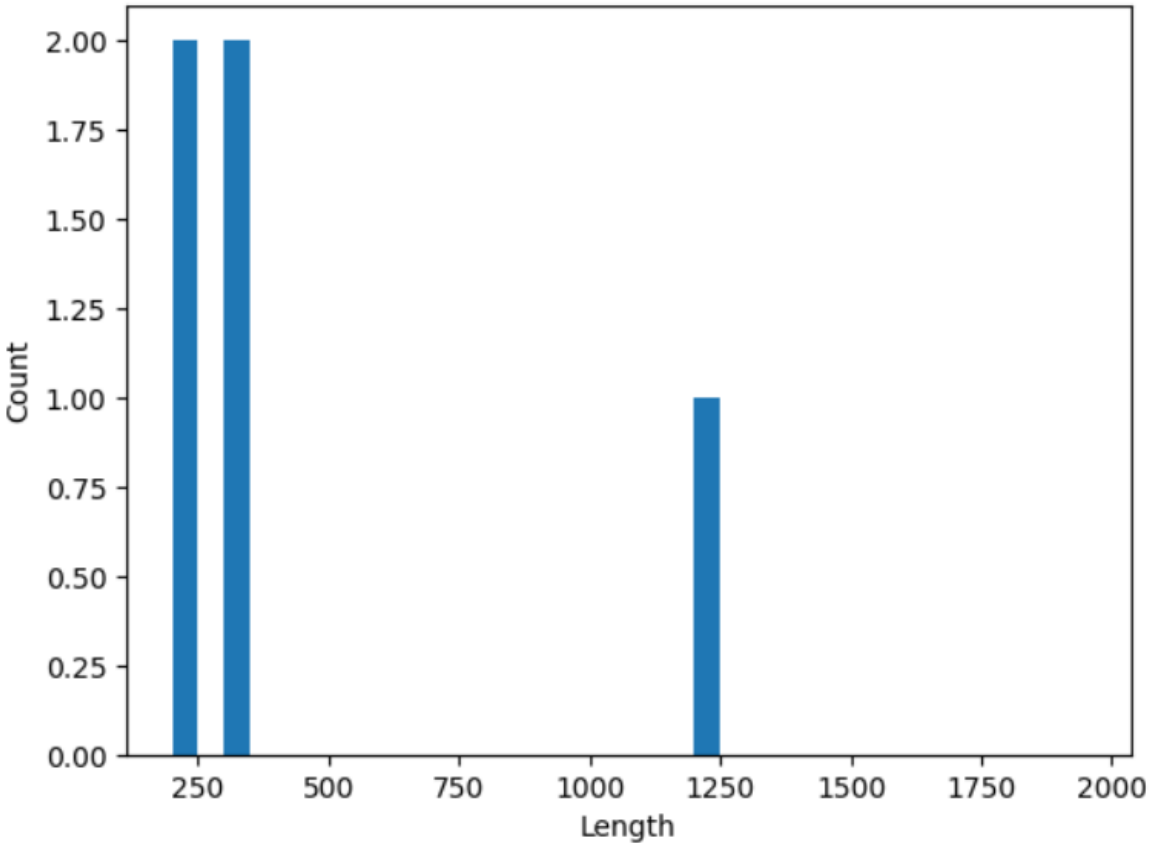
统计结果如下：

度量点	数值
最小长度	1
最大长度	1216
平均长度	57.47176079734219
长度中位数	54
标准差	79.20599745740591

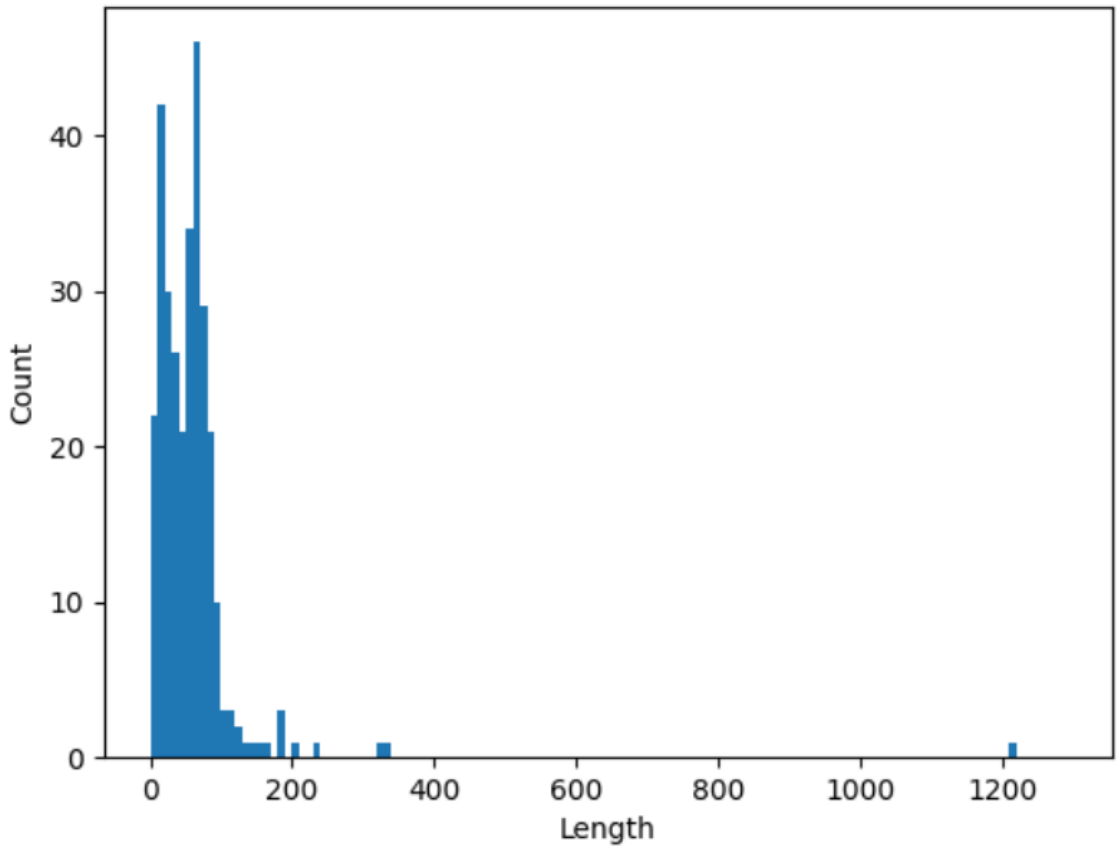
文本长度在0 - 200之间的数量分布，其中横座标以5为单位：



文本长度在200-2000之间的数量分布，其中横座标以50为单位：



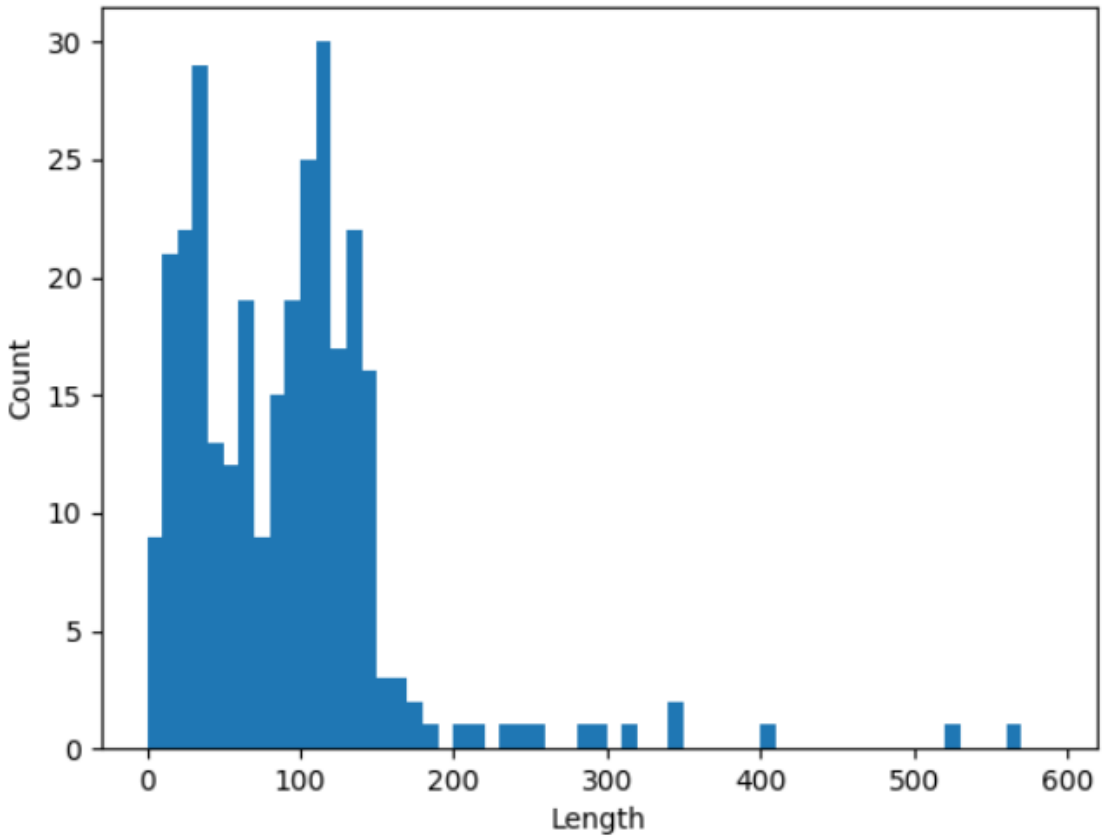
总的分布直方图，横座标以10为单位：



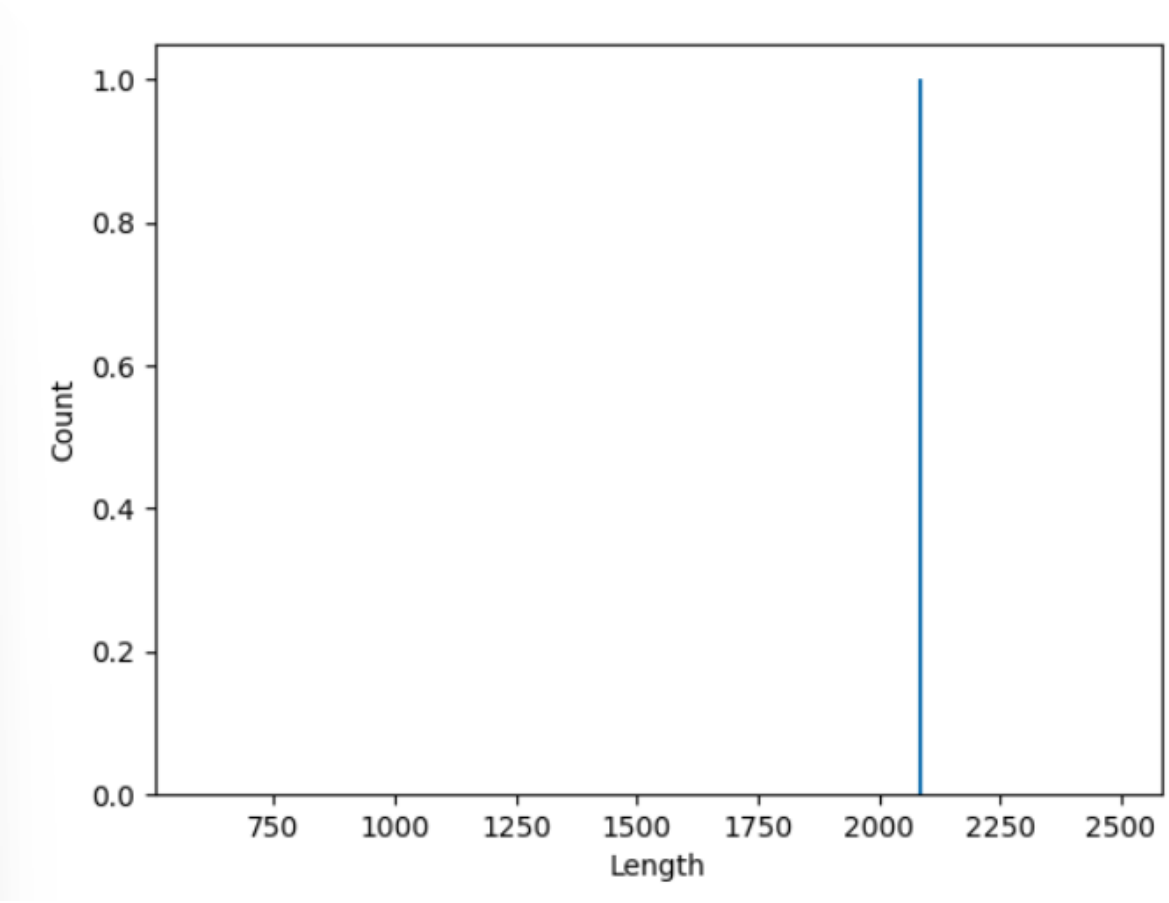
以文字数量为计量

度量点	数值
最小长度	2
最大长度	2085
平均长度	96.85714285714286
长度中位数	91
标准差	134.78235365634055

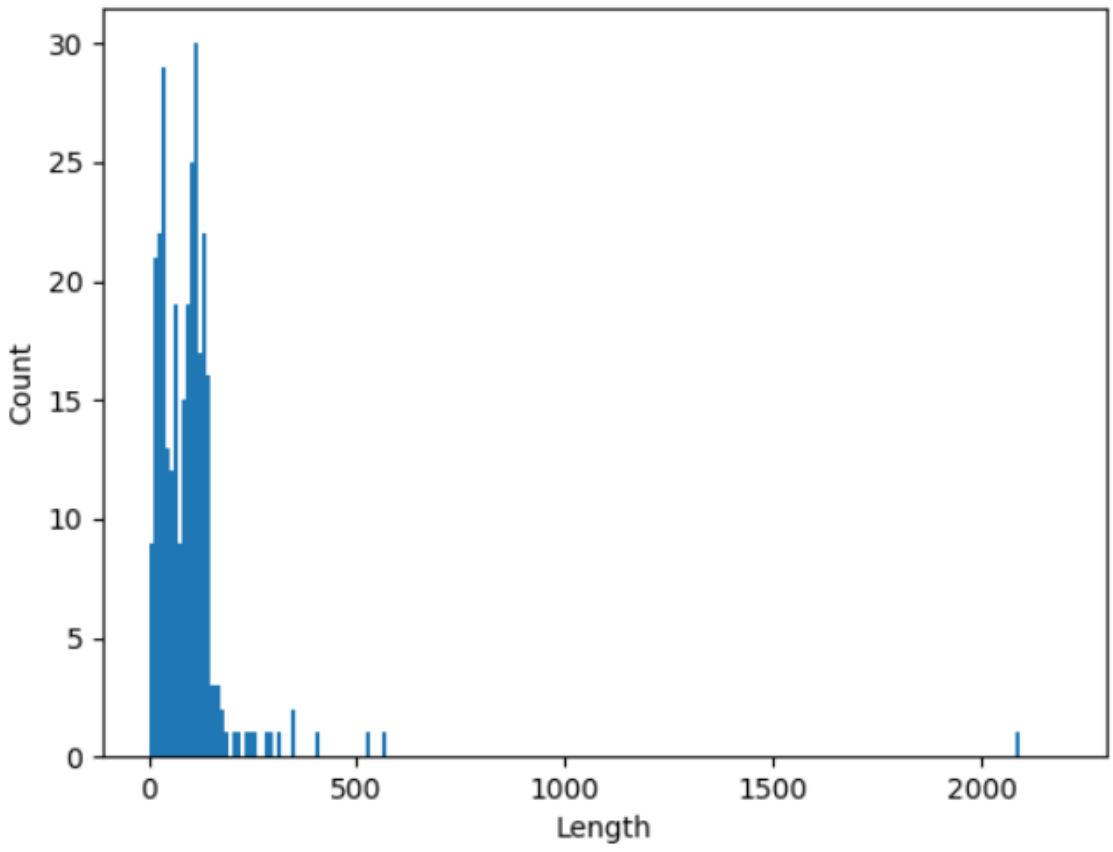
文本长度在0 - 600之间的数量分布，其中横座标以10为单位：



文本长度在600-2500之间的数量分布，其中横座标以10为单位：



总的分布直方图，其中横座标以10为单位：



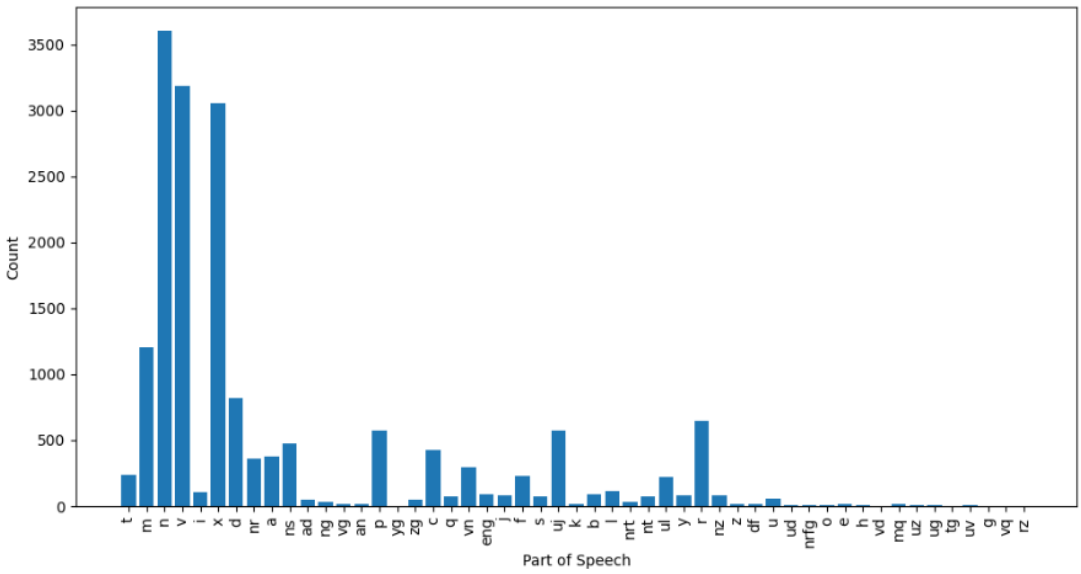
词性分析

分别使用了jieba和THULAC库进行词性分析。

jieba库的词性标记：

- t: 时间词
- m: 数量词
- n: 名词
- v: 动词
- i: 成语
- x: 非语素字
- d: 副词
- nr: 人名
- a: 形容词
- ns: 地名
- ad: 副形词
- ng: 名形词
- vg: 动形词
- an: 名形词
- p: 介词
- yg: 语气词
- zg: 状态词
- c: 连词
- q: 量词
- vn: 名动词
- eng: 英文
- j: 简称
- f: 方位词
- s: 处所词
- uj: 助词
- k: 后缀
- b: 区别词
- l: 习用语
- nrt: 名词性惯用语
- nt: 机构团体名
- ul: 时态助词
- y: 语气词
- r: 代词
- nz: 其他专名
- z: 状态词
- df: 副词
- u: 助词

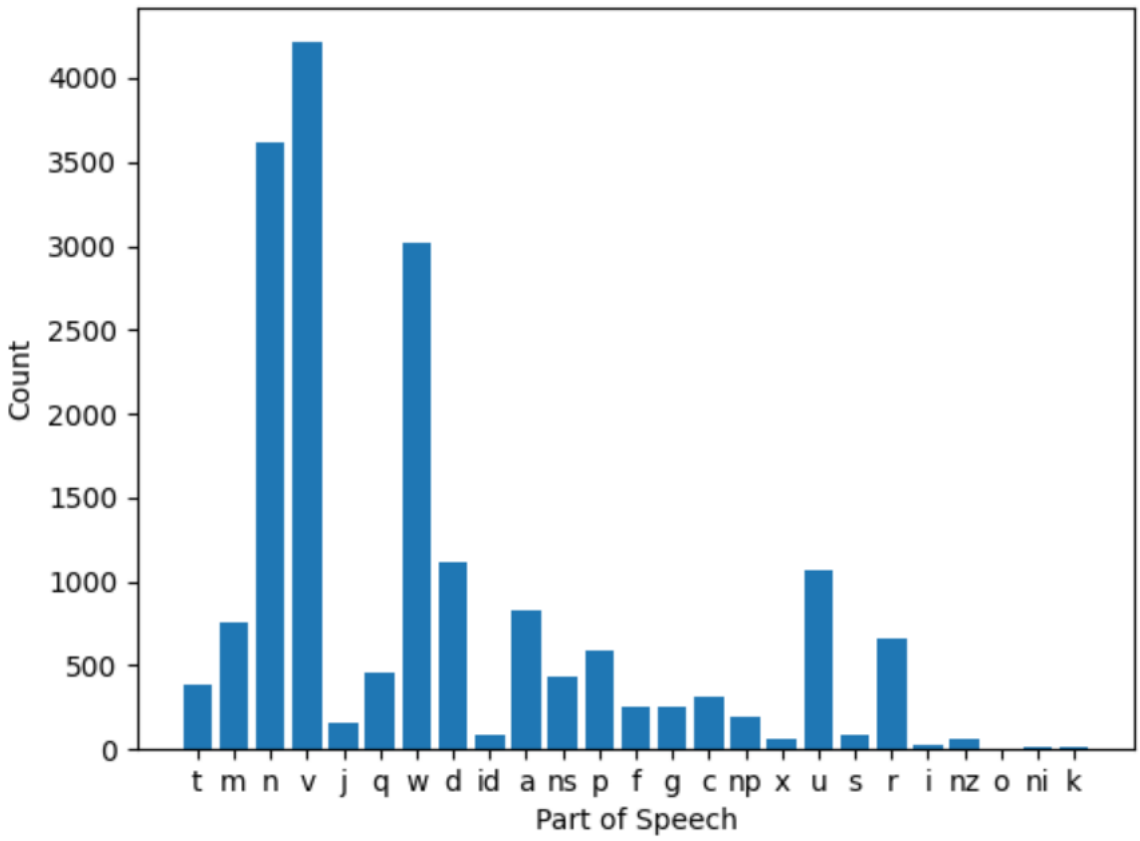
- ud: 助词
- nrfg: 人名
- o: 拟声词
- e: 叹词
- h: 前缀
- vd: 副动词
- mq: 数量词
- uz: 助词
- ug: 助词
- tg: 时态词
- uv: 助词
- g: 语素
- vq: 动词
- rz: 代词



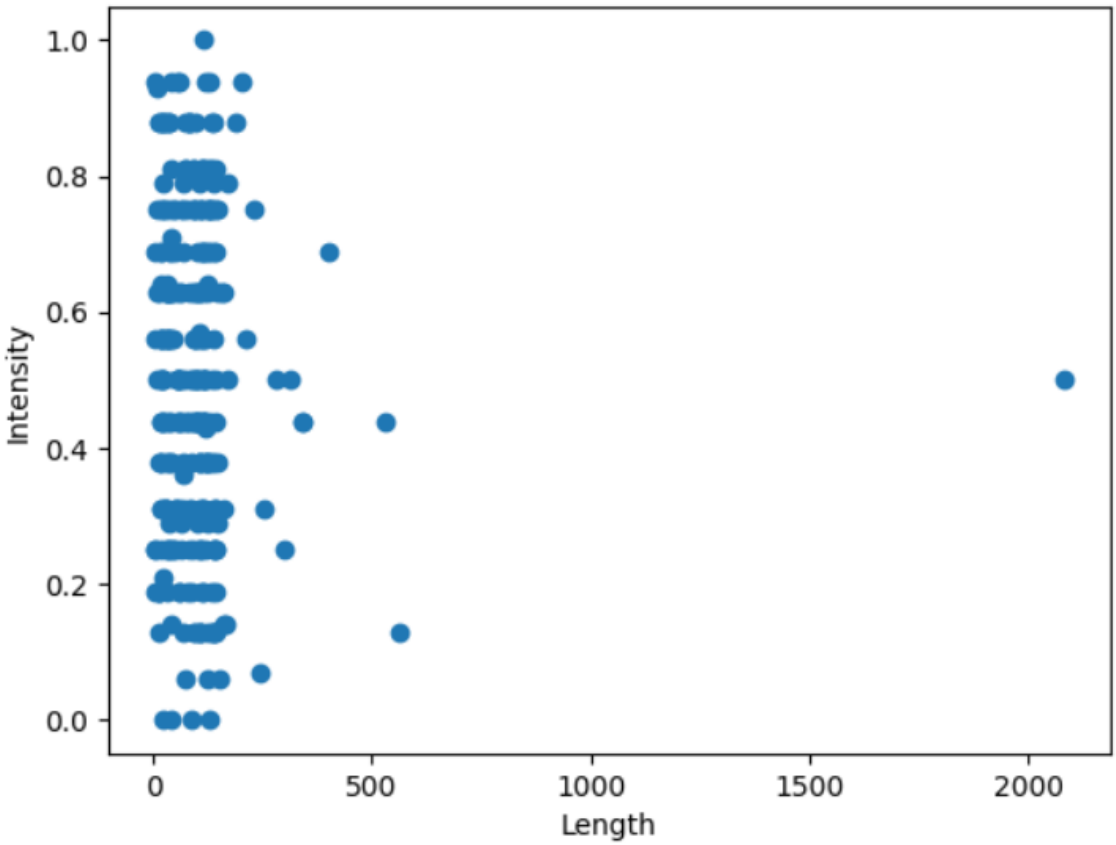
THULAC库的词性标记：

- t: 时间词
- m: 数量词
- n: 名词
- v: 动词
- j: 简称略语
- q: 量词
- w: 标点符号
- d: 副词
- id: 成语
- a: 形容词
- ns: 地名
- p: 介词

- f: 方位名词
- g: 学术词汇
- c: 连词
- np: 人名
- x: 非语素字
- u: 助词
- s: 处所名词
- r: 代词
- i: 成语
- nz: 其他专名
- o: 拟声词
- ni: 机构团体名
- k: 后缀

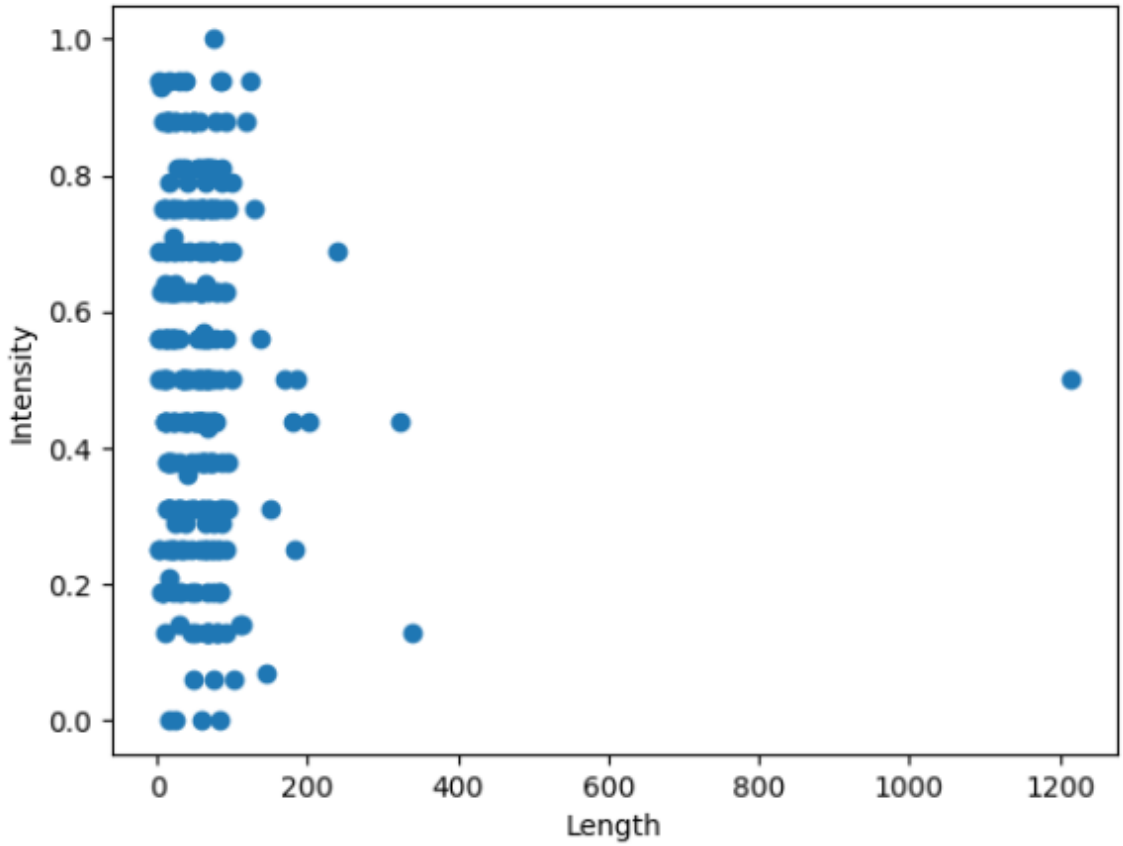


情感强度与文本长度(以文字数量为度量)相关性对比



上图为情感强度和文本长度散点图，计算所得相关性为NaN。

情感强度与文本长度(以单词数量为度量)相关性对比



上图为情感强度和文本长度散点图，计算所得相关性为NaN。