# 分词：

使用正则按顺序获取符合以下任意一项的字符组合

1. 大写字母和.组合的缩写

e.g. U.S.A.

1. 网址

开头包括[http://|ftp://|https://|www]结尾包括[com|net|cn|gov|edu|org]

1. 人称缩写

包括Mr. Miss. Mrs. Ms. Dr. Pro. Madam.

1. 带’的缩写

包括's 't 'm 're 'll 've 'd o'clock

1. 大小写英文字母的组合
2. 带符号的数字组合

e.g. $12.40, 82%

1. 时间

E.g. 12:01

1. 带小数点的数字

E.g. 0.01

1. 带，的数字

E.g. 200,000

1. 省略号 ...
2. 各种中英文符号

包括[+\*/?!？！；:/][.,，;"'()-\_`，。：、；‘’“”（）]

注：’t 会与前一个单词拼接，如don’t，其他缩写不会

# 分句：

先对句子进行分词（在上面11种情况中加入第12种：空格），然后将所有词拼接起来，遇到终止符则另起一句。终止符包括['.', '!', '?', '！', '？']。若句子为空则删除该句子。

# 分段：

以分段符为依据进行分段，分段符包括[\n|\r|\t]