import nltk

**nltk语料库：**nltk.corpus

**\*Tokenize** 分词效果不佳。对于一篇文章，往往先分句再分词

nltk.sent\_tokenize(字符串) #分句

nltk.word\_tokenize(字符串) #分词

**\*Lemmatization** 还原各种词形、时态变形, 如are→be

from nltk.stem import WordNetLemmatizer

wordnet\_lemmatizer = WordNetLemmatizer()

wordnet\_lemmatizer.lemmatize(‘dogs’,\*POS=词性标签)

#词性标签用于指明处理多词性词的方式



**\*Stemming** 除去英文词各种后缀，提取词干，转小写，保留其原始形式，如walked→walk

from nltk.stem import PorterStemmer

from nltk.tokenize import sent\_tokenize, word\_tokenize

ps = PorterStemmer()

ps.stem(word)

**\*去除停用词**

from nltk.corpus import stopwords

words\_filtered=[word for word in str if word not in stopwords.words('english')]

**\*生成窗口词对**

nltk.bigrams(分词的list)

nltk.trigrams(分词的list)

**\*词性标注**

nltk.pos\_tag(word)

也可以对一个分好词的list进行词性标注，标注其中的每个词

