|  |  |  |
| --- | --- | --- |
| **模块** | **函数命令** | **说明** |
| from \_future\_ import division | len(set(myfiles))/len(myfiles) | 浮点数计算 |
| from nltk import FreqDist | FreqDist() | 词频排序 |
| from nltk import ngrams | ngrams() | 提取N连词 |
| from nltk import word\_tokenize | word\_tokenize() | 分词处理 |
| from nltk.book import \* |  | 加载NLTK固有语料库 |
| from nltk.corpus import brown | brown.tagged\_sents() | 读取经词性标注的Brown语料库 |
| from nltk.corpus import brown | brown.categories() | 显示Brown语料库的文本类型 |
| from nltk.corpus import inaugural | inaugural.fileids() | 读取inaugural语料库中的相应文本 |
| from nltk.corpus import inaugural | inaugural.words() | 对inaugural语料库做分词处理 |
| from nltk.corpus import PlaintextCorpusReader | PlaintextCorpusReader() | 加载纯文本语料库阅读器 |
| from nltk.corpus import stopwords | set(stopwords.words(‘english’)) | 加载停用词模块 |
| from nltk.corpus import stopwords | stop\_words.update() | 添加自定义停用词 |
| from nltk.corpus import treebank | treebank\_chunk.chunked\_sents() | 读取宾州树库标注结构 |
| from nltk.corpus import wordnet | synet.definition() | 显示WordNet词典 |
| from nltk.stem import WordNetLemmatizer | lemmatize() | 还原词形 |
| from nltk.tokenize import TreebankWordTokenizer | TreebankWordTokenizer() | 分词处理 |
| from nltk.tokenize import | word\_tokenizer() | 分词处理 |
| from pattern.en import tag | tag() | 词形标注 |
| from PIL import Image | Image.open() | 打开自定义图形文件 |
| from scipy.stats import chisquare | stats.chi2\_contingency() | 用于卡方检验 |
| from string import punctuation | if word not in punctuation | 清除文本中的标点符号 |
| from tabulate import tabulate | tabulate() | 绘制表格 |
| from wordcloud import ImageColorGenerator | ImageColorGenerator() | 生成自定义图形 |
| from wordcloud import WordCloud | WordCloud().generate() | 生成词云图 |
| import codecs | codecs.open() | 创建一个符合编码要求的文件 |
| import docx | docx.Document() | 打开相应的docx文件 |
| import docx | doc.paragraphs | 读取docx文本 |
| import jieba | jieba.lcut() | 分词后可直接返回链表 |
| import jieba.analyse | jieba.analyse.extract\_tags() | 中文分词后提取关键词 |
| import jieba.analyse | textrank() | 提取关键词 |
| import matplotlib.pyplot | plt.imshow() | 显示二维图形 |
| import nltk | nltk.Text() | 把链表转换成NLTK文件 |
| import nltk | nltk.pos\_tag() | 词性标注 |
| import nltk | nltk.UnigramTagger() | 训练一元标注器 |
| import nltk | nltk.ConditionalFreqDist() | 创建条件频率分布 |
| import numpy | numpy.array() | 图形矩阵运算 |
| import numpy | numpy.arrange() | 图形矩阵运算 |
| import os | os.listdir() | 列出当前目录下的所有文件 |
| import xlrd | xlrd.open\_workbook() | 打开相应的xlsx文件 |
| import xlsxwriter | xlsxwriter.Workbook() | 创建一个用于输出的xlsx文件 |