**拼写autocorrect：**

|  |  |
| --- | --- |
| **Speller(lang='en')** | **实例化拼写检查器** |
| **SPELL(word)** | **纠正单词拼写** |

**中文jieba：**

|  |  |  |
| --- | --- | --- |
| **分词：** | | |
| **cut(sent, cut\_all=False)** | **返回单词分割结果** | **精确模式** |
| **cut(sent, cut\_all=True)** | **全模式** |
| **cut\_for\_search(sent, HMM)** | **搜索模式** |
| **tokenize(sent, mode, HMM)** | **返回单词分割及起始位置** | **mode：可选'search'** |
| **posseg.cut(sent, HMM)** | **返回单词分割及词性结果** | |
| **POS.flag** | **访问pair单词实例的词性** | |

|  |  |  |
| --- | --- | --- |
| **分析analyse：** | | |
| **extract\_tags(sent, topK, withWeight, allowPOS)** | **TF-IDF** | **语句关键词统计** |
| **textrank(sent, topK, withWeight, allowPOS)** | **TextRank** |

**外语nltk：**

**<https://gitcode.net/mirrors/nltk/nltk_data/-/archive/gh-pages/nltk_data-gh-pages.zip>**

**复制并命名 (内容需解压)：package -> Anaconda\lib\nltk\_data**

**英文语料预处理：分词 -> 纠错 -> 过滤停用词 -> 词性标注 -> 词干提取 -> 词形还原 -> 词义消歧**

|  |  |
| --- | --- |
| **基本：** | |
| **pos\_tag(tokens)** | **返回词性元组 (token, pos)** |
| **ngrams(seq, n)** | **返回单词n元组列表** |
| **ne\_chunk(token\_pos, binary=True)** | **专有名词 -> Tree实例 (len=1)** |

|  |  |
| --- | --- |
| **语料corpus：** | |
| **stopwords.words('english')** | **返回停用词列表** |
| **wordnet.synsets(word)** | **返回同义词列表** |

|  |  |  |
| --- | --- | --- |
| **词干stem：** | | |
| **PorterStemmer()** | **实例化词干提取器** | **Porter** |
| **SnowballStemmer('english')** | **Snowball** |
| **LancasterStemmer()** | **Lancaster** |
| **ST.stem(word)** | **返回词干** | |
| **wordnet.WordNetLemmatizer()** | **实例化词形还原器** | |
| **WN.lemmatize(word, pos)** | **返回原词** | |

|  |  |  |
| --- | --- | --- |
| **分割tokenize：** | | |
| **通用** | **sent\_tokenize(text)** | **句子分割** |
| **word\_tokenize(sent)** | **单词分割** |
| **对象** | **TweetTokenizer(strip\_handles=True)** | **实例化推特分词器** |
| **MWETokenizer()** | **实例化词组分词器** |
| **MWE.add\_mwe(seq)** | **添加词组** |
| **TN.tokenize(sent)** | **返回分词器的单词分割结果** |

|  |  |
| --- | --- |
| **歧义wsd：** | |
| **lesk(tokes, word)** | **返回同义词** |
| **SYN.definition()** | **返回同义词解释** |

**机器sklearn：**

**from sklearn.feature\_extraction import text**

**from sklearn.metrics.pairwise import cosine\_similarity**

**向量相似度：余弦相似度、Jaccard相似度 (交并比)**

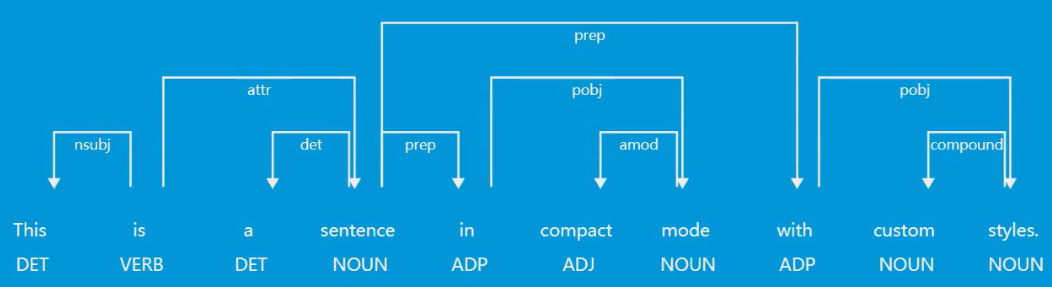
|  |  |  |
| --- | --- | --- |
| **词库：** | | |
| CountVectorizer(tokenizer, stop\_words,  ngram\_range, max\_features, vocabulary) | | **实例化词库模型** |
| **实例方法** | fit\_transform(corpus) | **返回词频矩阵** |
| get\_feature\_names\_out() | **返回高频词** |

|  |  |  |
| --- | --- | --- |
| **TF-IDF：** | | |
| TfidfVectorizer(tokenizer, stop\_words,  ngram\_range, max\_features, vocabulary) | | **实例化TF-IDF模型** |
| **实例方法** | fit\_transform(corpus).todense() | **返回TF-IDF矩阵** |
| get\_feature\_names\_out() | **返回高频词** |
| TfidfTransformer() | | **实例化TF-IDF转换器** |
| **实例方法** | fit(count) | **词频矩阵 -> TF-IDF矩阵** |

**管道spacy：**

**预训练模型下载：https://spacy.io/models/zh**

|  |  |
| --- | --- |
| **分析：** | |
| explain(term) | **返回术语解释** |
| load(model\_name) | **载入并返回预训练模型** |
| NLP.vocab | **返回模型词汇表** |
| NLP.pipe\_names | **返回管道组件名称** |
| NLP.disable\_pipes(\*pipe) | **禁用管道组件** |
| NLP(sent) | **返回各个单词的Doc实例 (Token序列)** |
| displacy.render(DOC, style) | **分析结果可视化** |
| DOC.ents | **返回命名实体Span序列** |
| TOKEN.text | **返回单词文本** |
| TOKEN.pos\_ | **返回单词词性** |
| TOKEN.dep\_ | **返回单词依存关系** |
| SPAN.label\_ | **返回单词实体类别** |



|  |  |
| --- | --- |
| **匹配matcher：** | |
| Matcher(vocab, patterns) | **实例化匹配器** |
| MAT(word) | **返回匹配结果** |

**轻量textblob：**

|  |  |  |  |
| --- | --- | --- | --- |
| **TextBlob(str)** | | **实例化语料库** | |
| **实例属性** | **words** | **返回单词分割结果** | |
| **tags** | **返回单词词性** | |
| **实例方法** | **ngrams(n)** | **返回单词n元组列表** | |
| **Word**  **实例方法** | **singularize()** | **返回单词** | **单数形式** |
| **pluralize()** | **复数形式** |

**翻译translate：**

|  |  |  |
| --- | --- | --- |
| **Translator(to\_lang='Chinese, from\_lang='en')** | | **实例化翻译器** |
| **实例方法** | **translate(sent)** | **返回翻译结果** |

**词云wordcloud：**

|  |  |  |
| --- | --- | --- |
| **WordCloud(width, height, background\_color,**  **max\_words, stopwords, min\_font\_size)** | | **实例化词云提取器** |
| **实例方法** | **generate(text)** | **提取词云** |
| **to\_file(file\_name)** | **输出文件** |
| **to\_image()** | **返回PIL.Image** |
| **to\_array()** | **返回np.array** |