分词与情感分析

Chinese Word Segmentation and Sentiment Analysis

PART 1.分词

- 将一个汉字序列切分成一个一个单独的词
- This is a cat. → ["This", "is", "a", "cat", "."]
- 这是一只猫。 → [("这", "pronoun"), ("是 ", "verb"), ("一", "numeral"), ("只", "classifier"), ("猫", "noun"), ("。 ", "punctuation mark")]

为什么要分词?

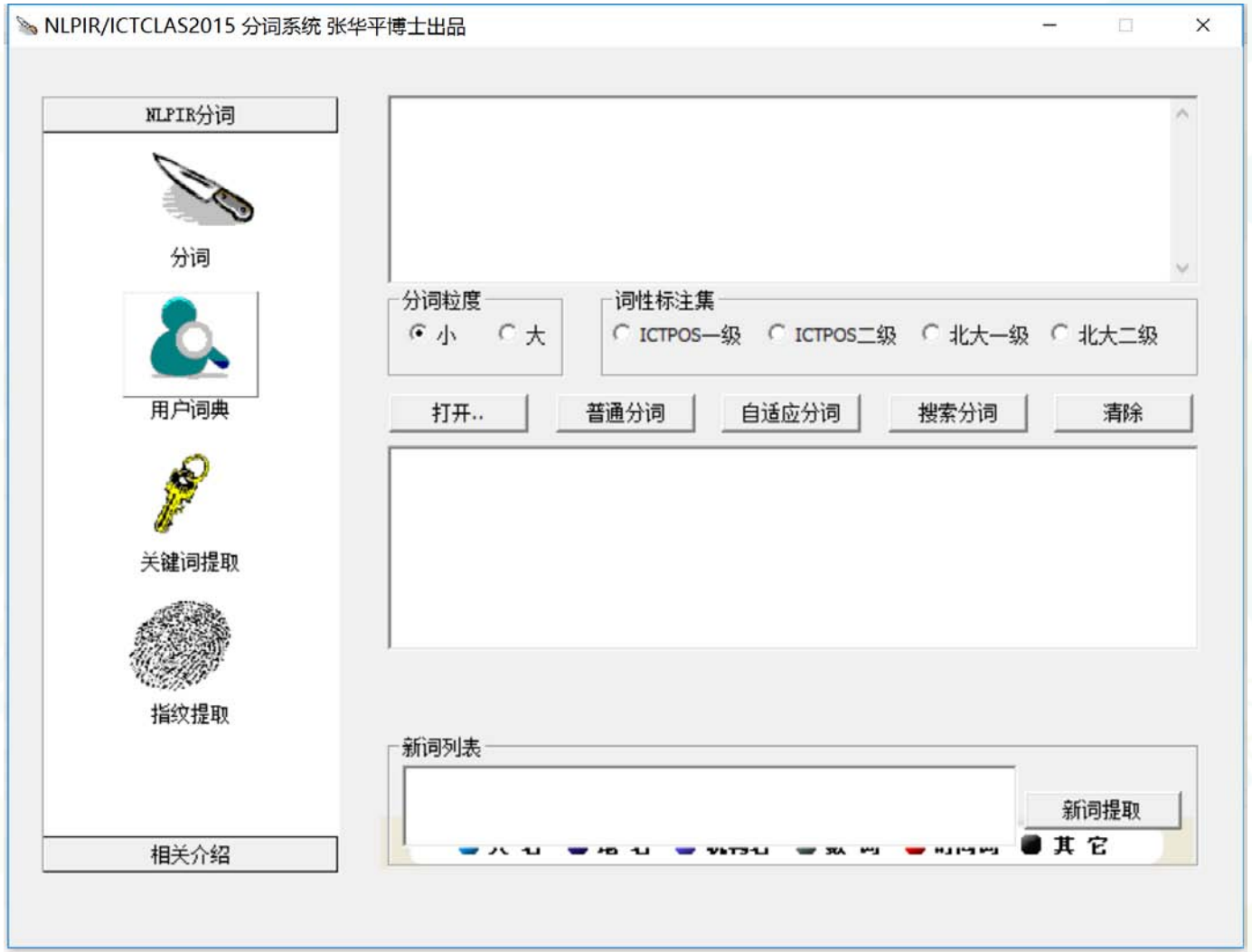
- 词是最小的能够独立运用的语言单位。
- 国际上常用的NLP算法，深层次的语法语义分析通常都是以词作为基本单位
- 中文文本是由连续的字序列构成，词与词之间是没有天然的分隔符
- 武汉市长江大桥 → 武汉市 长江 大桥

  武汉 市长 江大桥

分词算法

分词工具

- ustrwordcount()与ustrword()
- ICTCLAS/NLPIR
- jieba
- BosonNLP


NLPIR/ICTCLAS2015 分词系统 张华平博士出品
NLPIR分词
分词
用户词典
关键词提取
指纹提取
相关介绍
分词粒度
小
大
词性标注集
ICTPOS一级
ICTPOS二级
北大一级
北大二级
打开..
普通分词
自适应分词
搜索分词
清除
新词列表
新词提取
其 它

Stata与Python交互

Stata与python交互

- pynlpir: 提供了NLPIR/ICTCLAS汉语分词的Python接口。

- jieba: 结巴(jieba)是国人出的一个精品插件，可以对一段中文进行分词，有三种分词模式，可以目前已有Python、JAVA、C++和Nodejs版本。适应不同需求。

curl(CommandLine Uniform Resource Locator)

- curl是利用URL语法在命令行方式下工作的开源文件传输工具。部分功能包括：
  - -H：自定义头信息传递给服务器
  - -d：HTTP POST方式传送数据
  - -G：以get的方式来发送数据
  - -c：操作结束后把cookie写入到这个文件中
  - -b：读取cookie
  - -o：把输出写到文件中
  - -x：使用HTTP代理
  - ……

Unicode编码

- 以“\uXXXX”格式表示的字符(其中X为16进制数字）在JS中被称为Unicode转义字符。通过ustrtohex()和ustrunescape()函数可以实现这一字符的编码与解码。

BosonNLP API调用

- http://docs.bosonnlp.com/tag.html
- curl -X POST \
  -H "Content-Type: application/json" \
  -H "Accept: application/json" \
  -H "X-Token: YOUR_API_TOKEN" \
  --data "\"\u6b66\u6c49\u5e02\u957f\u6c5f\u5927\u6865\"" \
  "http://api.bosonnlp.com/tag/analysis?space_mode=0&oov_level=3&t2s=0"

- [{"tag":["ns","ns","n"],"word":["武汉市","长江","大桥"]}]

Part 2. 情感分析

- 情感分析又称倾向性分析、意见挖掘。目的是为了找出说话者/作者在某些话题上或者针对一个文本两极的观点的态度。

情感分析主要方法

情感分析主要方法

- 基于词典的方法主要通过制定一系列的情感词典和规则，对文本进行段落拆借、句法分析，计算情感值，最后通过情感值来作为文本的情感倾向依据。

- 文本分词→检索情感词→检索情感词前的程度词→检索情感词前的否定词→计算情感得分

- 他高兴。→ 他非常高兴。→他非常不高兴。

情感分析主要方法

- 基于机器学习的方法大多将这个问题转化为一个分类问题来看待，对于情感极性的判断，将目标情感分类2类：正、负。对训练文本进行人工标标注，然后进行有监督的机器学习过程。例如想在较为常见的基于大规模语料库的机器学习等。