cnc_Regex 使用指南

李亮 撰写

2014 年 6 月 8 日

【主要功能】提取某些特征的单词、生成完整词表（typelist\tokenlist）；处理对象可以是手工复制的一些文本，也可以是针对一个文件夹（包括子文件夹）的所有文本文件；

【安装方法】本软件是 C#语言开发，是基于.Net 2.0 框架，所以，如果你是 Windows Vista 或 Windows 7 或 Windows 8 都能直接运行而不必安装什么。但是，你如果是 Windows XP，你就需要安装.Net 2.0 这个软件，你可以访问下面网址而下载……
http://www.crsky.com/soft/4818.html
http://www.cr173.com/soft/25219.html

【使用方法】
（1）复制或输入一些文本到主界面的文本框，然后输入正则表达式来检索；
（2）点“选择”按钮进行“欲检索的文件夹”的选择，然后再输入正则表达式来检索；

【文字编码】本软件提供了“gb2312、gbk、utf-8、unicode、unicode BE”这 5 种编码，如果你文本文件本身没乱码，却检索结果呈现为乱码，就请你重新选择这 5 种文字编码中的其他一种编码，重新检索，很可能就不会乱码了。

【检索结果】将直接呈现在主界面的文本框；检索之前你可以选择“结果自动去重”，这样就剔除了结果中的重复单词；检索之前你也可以选择“结果自动排序”，这样就让结果中的单词便于观察；而你如果同时选择了“结果自动去重”和“结果自动排序”而且输入正则表达式“\w{1,}”，就让检索结果变为了“有序而不重复的单词表”了!

【模糊检索/正则表达】
（1）如果你希望提取所有的数字串，你就输入
\d{1,}
（2）如果你希望提取所有的单词（包括纯字母串、数字串、字母数字混合串），你就输入
\w{1,}
（3）如果你希望提取所有的纯字母的单词，你就输入
[a-zA-Z]{1,}
（4）如果你希望提取至少由 3 个字母组成的单词串，你就输入
[a-zA-Z]{3,}
（5）如果你希望提取 ful 结尾的单词，你就输入
[a-zA-Z]{1,}ful
（6）如果你希望提取 dis 为前缀的单词，你就输入
dis[a-zA-Z]{1,}
（7）如果你希望搜索以 un 为前缀且以 ful 为后缀的单词，你就输入
un[a-zA-Z]{1,}ful
（8）如果你希望检索以 g 开头或 h 开头的单词，你就输入
[ghGH][a-z]{1,}
（9）如果你希望检索“$100”或“$75”这样的若干美元的数字串，你就输入
\$[0-9]{1,}
（10）如果你希望检索“$100,000”或“$56,234”这样的中间有逗号的美元表达，你就输入
\$[0-9,]{1,}
（11）如果你希望检索“我/代词 喜欢/动词 唱歌/名词”这样的分了词且标注了词类的中文中的

两个汉字所组成的动词，你就输入

../动词

（12）如果你希望检索“我/代词 喜欢/动词 唱歌/名词”这样的分了词且标注了词类的中文中的 1 个或 2 个或 3 个汉字所组成的动词，你就输入

.{1,3}/动词

（13）如果你希望检索“We_pron need_vt money_n”这样的标注了词类的英文中的被标注为 vt 的所有单词，你就输入

\w{1,}_vt

（14）如果你希望检索“We_pron need_vt money_n”这样的标注了词类的英文中的被标注为 pron 与 n 的所有单词，你就输入

\w{1,}_pron|\w{1,}_n

（15）如果你希望检索 big 这个词的所有的左侧第 1 个搭配词，你就输入

\w{1,} (?=big)

（16）如果你希望检索 big 这个词的所有的左侧第 2 个搭配词，你就输入

\w{1,} (?=\w{1,} big)

（17）如果你希望检索 big 这个词的所有的左侧第 3 个搭配词，你就输入

\w{1,} (?=\w{1,} \w{1,} big)

（18）如果你希望检索 big 这个词的所有的右侧第 1 个搭配词，你就输入

(?<=big) \w{1,}

（19）如果你希望检索 big 这个词的所有的右侧第 2 个搭配词，你就输入

(?<=big \w{1,}) \w{1,}

（20）如果你希望检索 big 这个词的所有的右侧第 3 个搭配词，你就输入

(?<=big \w{1,} \w{1,}) \w{1,}

（21）如果你希望提取出“a big bird”或“a small boy”之类的 3 个词组成的语块且第一个词是 a，你就输入

a \w{1,} \w{1,}

（22）如果你希望提取出“many good boys”或“very good advice”之类的 3 个词组成的语块且第 2 个词是 good，你就输入

\w{1,} good \w{1,}

（23）如果你希望提取出“good”或“foot”这样的中间含了 oo 的单词，并且 oo 两侧至少含有 1 个字母，你就输入

\w{1,}oo\w{1,}

（24）如果你希望提取出至少 10 个字母组成的较长单词，例如“international”，你就输入

\w{10,}

（25）如果你希望检索“It_pron kills_vt pains_n .”这样的以下划线方式标注了词类的英文中的“及物动词与名词”的组合，你就输入

\w{1,}_vt \w{1,}_n