表4.1 语料库检索中的常用正则表达式（改编自梁茂成等 2010：17-18）

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **符号** | **含义** | **举例或说明** |
| 1 | \b | 字符串的边界 | \ban\b匹配an，不匹配a, and和sand |
| 2 | \w | 任何字母或数字 | \w匹配字母a-z, A-Z, 数字0-9，以及下划线\_ |
| 3 | . | 任何字符或非字符 | 2.4匹配204, 214, 2t4, 2 4, 2.4, 2-4等 |
| 4 | + | 出现1次或更多 | A+匹配A, AA, AAA, AAAA等 |
| 5 | \* | 出现0次或更多 | BA\*匹配B, BA, BAA, BAAA等 |
| 6 | ? | 有或者无 | BA?匹配B和BA |
| 7 | () | 组合 | act(ing)?可以匹配act和acting；B(AA)?匹配B和BAA；B(AA)\*匹配B, BAA, BAAAA等 |
| 8 | [] | []中的任意字符 | [abc]匹配a或b或c；[abc]+匹配a, b, c, ab, abc, ba, cba, bc, bca等 |
| 9 | \s | 一个空格 | \w+\s\w+匹配that is等连续的2个词 |
| 10 | \1 | 表达式中的第1个组合 | (\w+)\s\1可以匹配very very等连续重复使用的词 |
| 11 | | | 或者 | (analyze|analyse)匹配analyze或analyse |
| 12 | \S | 非空格 | 任何除空格之外的字符及非字符 |
| 13 | \w+ | 任意数目的字母、数字及两者的组合 | \w+匹配a, the, and, 21th, bookshelf等，但不匹配book-shelf；compu\w+匹配computer, computers, computation等 |
| 14 | \w\* | 0个或多个任意数目的字母、数字及两者的组合 | act\w\*匹配act, acted, acts, acting等；book\w\*匹配book, books, bookshelf； |
| 15 | \S+ | 任何连续的非空格 | 可以匹配文本中任何可见的符号 |
| 16 | \S+\_\S+ | 分隔符为“\_”的标注文本 | 匹配被分隔符隔开的任何可见字符，是检索词性标注文本中任何词或标点的常用方法 |
| 17 | [\r\n]+ | 换行符 | [\r\n]+可以匹配ASCII格式文本的换行符，也可以匹配Unicode格式文本的换行符。删除[\r\n]+会使多行文本连成一行文本 |
| 18 | .\*? | 任何字符串 | ^.\*?[\r\n]+匹配任何格式文本中的任何一行 |