Metacharacters

```
$ ( ) * + ? { . [ ^ \ |
```

Quantifiers

Quantifiers specify how many times a character or a group of characters occur.

- `?` means 0 or 1.
- `*` means 0 or more.
- `+` means 1 or more.
- `{n}` repeats preceding character exactly `n` times.
- `{n,}` repeats preceding character at least `n` times .
- `{n,m}` repeats preceding character between `n` and `m` times.

Matching One or Several Characters

- `.` match any one character except a newline character (`\n`). We can switch newline character setting an argument in a function (stringi functions): `dotall = TRUE`.
- `[]` match one of many characters out of the list in square brackets. This notation is called *character class*. Note that only: `\`, `^`, `]` and `-` have a special functions. Other metacharacters are just literal text.
- `[^]` caret `^` is a negated class: don't match the list or match all but the list.
- `-` a hyphen creates a range when it's placed between two characters. e.g.: `[A-Z]` includes all capital letters (form ASCII table).

Anchors

Anchors don't match a character but a location in a string:

- `^` - The beginning of a line or string (also: `\A`), e.g. `^Bat`.
- `$` - The end of a line or string (also: `\Z`), e.g. `bat$`. Note: empty rows `^$`.
- `\b` - Word boundaries, e.g. matches `" ton "` but not `"ton"` or `"newtons"`.
- `\B` - Non-word boundaries (negate `\b`). `\Bton\B` doesn't match `" ton "`, but match `"ton"` in `"newtons"`.

Shorthands character classes

Shorthands can be used both inside (e.g. `[\W\d]`) and outside character classes.

- `\d` - a single digit;
- `\D` - any character that is not a digit;
- `\w` - a single word character (letters, digits, and underscores). In English it's identical to `a-zA-Z0-9_`. But in R it isn't, because `\w` includes diacritic marks, e.g.: `ęąłüśşźćż` ;
- `\W` - any character that is not a word character;
- `\s` - a single whitespace character (spaces,tabs, line breaks: `[ \t\n\r]`);
- `\S` - any character not matched by `\s`.

Unicode Categories

- `\p{L}` - Any kind of letter from any language;
- `\p{Lu}` - An uppercase letter that has a lowercase variant;
- `\p{Ll}` - A lowercase letter that has an uppercase variant;
- `\p{N}` - Any kind of numeric character;
- `\p{P}` - Any kind of punctuation character;
- `\p{S}` - symbols (math symbols, currency signs, dingbats, box-drawing characters etc.);
- `\p{Pd}` - any kind hyphen or dash;
- `\p{Ps}` - any kind of opening bracket;
- `\p{Pe}` - any kind of closing bracket.