

O recurso de backreference

Transcrição

Nossa regex funciona como o esperado, mas agora ela também deve funcionar para uma tag `h2`. Ok, nada mais fácil do que isso, podemos pensar, basta adicionar a tag no início e fim, junto com a tag `<h1>`:

```
<(h1|h2).+?>([\w\sõãí. ]+)</(h1|h2)>
```

O agrupamento foi necessário para o `|` funcionar, para não interpretar o `|` literalmente. Ao testar, aparentemente a regex funciona, mas ela tem um erro grave. Ela também aceita alvos com `h1` no início e `h2` no final!

A regex pega (`h1` no início, `h2` no final):

```
<h1 id="regex" class="form">Expressões regulares não são tão difíceis.</h2>
```

O que queremos, quando a tag for uma `h1`, no final também deve estar uma `h1`, igualmente para a tag `h2`. Essa dependência podemos definir através de **back-references**, uma **referência** para um texto já encontrado. A sintaxe é simples, usa-se um `\` seguido pelo número do grupo, começando com 1.

Na nossa regex, a tag é definida no primeiro grupo. Ou seja, vamos usar a *back-reference* `\1`:

```
<(h1|h2).+?>([\w\sõãí. ]+)</\1>
```

Agora sim! Se testarmos, a regra só funciona com tags iguais na abertura e fechamento.