

Aquisição de Arquivos e Datasets

Expressões Regulares

Delermendo Branquinho Filho

Expressões regulares

- Expressões regulares podem ser pensadas como uma combinação de literais e *metacharacters*
 - Para desenhar uma analogia com a linguagem natural, pense no texto literal que forma as palavras desta linguagem, e os metacaracteres que definem sua gramática
 - Expressões regulares têm um rico conjunto de metacaracteres
-

Literais

O padrão mais simples consiste apenas em literais. O literal “nuclear” corresponderia às seguintes linhas:

```
Ooh. I just learned that to keep myself alive after a
nuclear blast! All I have to do is milk some rats
then drink the milk. Aweosme. :}
```

```
Laozi says nuclear weapons are mas macho
```

```
Chaos in a country that has nuclear weapons -- not good.
```

```
my nephew is trying to teach me nuclear physics, or
possibly just trying to show me how smart he is
so I'll be proud of him [which I am].
```

```
lol if you ever say "nuclear" people immediately think
DEATH by radiation LOL
```

Literais

O literal “Obama” corresponderia às seguintes linhas

```
Politics r dum. Not 2 long ago Clinton was sayin Obama
was crap n now she sez vote 4 him n unite? WTF?
Screw em both + McCain. Go Ron Paul!
```

```
Clinton conceeds to Obama but will her followers listen??
```

```
Are we sure Chelsea didn't vote for Obama?
```

```
thinking ... Michelle Obama is terrific!
```

```
jetlag..no sleep...early mornig to starbux..Ms. Obama
was moving
```

Expressões regulares

- O padrão mais simples consiste apenas em literais; Uma correspondência ocorre se a sequência de literais ocorre em qualquer lugar no texto que está sendo testado
 - E se nós só queremos a palavra “Obama”? Ou frases que terminam na palavra “Clinton”, ou “clinton” ou “clinto”?
-

Expressões regulares

Precisamos de uma maneira de expressar Limites de palavras em branco - conjuntos de literais - início e fim de uma linha - alternativas (“guerra” ou “paz”) Metacaracteres para o resgate!

Metacaracteres

Alguns metacaracteres representam o início de uma linha

```
^i think
```

Corresponderá as linhas

```
i think we all rule for participating
i think i have been outed
i think this will be quite fun actually
i think i need to go to work
i think i first saw zombo in 1999.
```

Metacaracteres

\$ Representa o fim de uma linha

```
morning$
```

Corresponderá as linhas

```
well they had something this morning
then had to catch a tram home in the morning
dog obedience school in the morning
and yes happy birthday i forgot to say it earlier this morning
I walked in the rain this morning
good morning
```

Classes de caracteres com []

Podemos listar um conjunto de caracteres que vamos aceitar em um determinado ponto no jogo

```
[Bb] [Uu] [Ss] [Hh]
```

Corresponderá às linhas

```
The democrats are playing, "Name the worst thing about Bush!"  
I smelled the desert creosote bush, brownies, BBQ chicken  
BBQ and bushwalking at Molonglo Gorge  
Bush TOLD you that North Korea is part of the Axis of Evil  
I'm listening to Bush - Hurricane (Album Version)
```

Classes de caracteres com []

```
^[Ii] am
```

Vai corresponder

```
i am so angry at my boyfriend i can't even bear to  
look at him  
  
i am boycotting the apple store  
  
I am twittering from iPhone  
  
I am a very vengeful person when you ruin my sweetheart.  
  
I am so over this. I need food. Mmmm bacon..
```

Classes de caracteres com []

Da mesma forma, você pode especificar um intervalo de letras [a-z] ou [a-zA-Z]; Aviso que a ordem não importa

```
^[0-9] [a-zA-Z]
```

Corresponderá às linhas

```
7th inning stretch  
2nd half soon to begin. OSU did just win something  
3am - cant sleep - too hot still.. :(  
5ft 7 sent from heaven  
1st sign of starvagtion
```

Classes de caracteres com []

Quando usado no início de uma classe de caracteres, o “^” também é um metacaracter e indica caracteres correspondentes NÃO na classe indicada

```
[^?..]$
```

Corresponderá às linhas

```
i like basketballs
```

```
6 and 9
```

```
dont worry... we all die anyway!
```

```
Not in Baghdad
```

```
helicopter under water? hmmm
```