初めての正規表現

プログラム初心者向けの
正規表現

後半は難しめ

作成：Ituki Kirihara/NI

はじめに

正規表現は何に使えるのか

文字列の「検索」と「置換」が主な使い方

- 「文字列を探す」のと、「文字列を置き換える」のに使います
- しばらくは「探す」方の話をします

使いどころ

特定のパターンを探すのに便利

- 例えばこんなパターン

  時間っぽい文字列を探したいとき

  例えば……　10:30　23:59 などなど

  ⟹ 普通に探すと…… 00:00 00:01 00:02 ... 99:99

  １万パターンの文字を用意しないと駄目

  ⇩

  「２桁の数字」、「:」、「２桁の数字」で探したい

これを可能にするのが「正規表現」です

これだけ覚えておけばとりあえずは書ける、基本中の基本

文字について

半角記号以外はそのまま書きます
半角記号は￥を前につけて書きます

- 次の文字は、その文字そのものとして扱われます
  - 半角アルファベット
  - 「_」（半角下線）
  - 「 」（半角スペース）
  - 全角文字
- それ以外の文字は、別の意味になります
  - 半角記号は、文字そのものとは違う意味になります
  - 文字として扱いたい場合は 記号の前に ￥ をつけます
    - 例
      - 「-」（半角マイナス）文字を表したい場合は「¥-」と書きます

これが正規表現の基本形

(文字) { 整数A , 整数B } が基本形

- 世界の半分ぐらいの正規表現は、次の形式です

(文字) { 整数A , 整数B }

- Point
  - 赤で書いた部分は、記号のまま書く
  - 黒で書いた部分は、以下のルールで書く
    - 文字　……　１文字
    - 整数A、整数B …… 整数値。ただし、0 ≦ 整数A ≦ 整数B
- 例えば……　(A){1,1}　　や　(あ){2,4} など
- これと、この省略形が正規表現の基本です

基本形解説

「文字」を「整数A」から「整数B」回繰り返した
全てのパターン

(文字){ 整数A , 整数B }

- 「文字」を「整数A」から「整数B」回繰り返した、すべてのパターンを表します

- これが理解できれば、正規表現の基本はマスター

基本形の例

「文字」を「整数A」から「整数B」回繰り返した
全てのパターン

(文字){ 整数A , 整数B }

- 「文字」を「整数A」から「整数B」回繰り返した、すべてのパターンを表します

- (A){1,1}　とは
  - 「A」のことです
- (あ){2,4}　とは
  - 「ああ」か「あああ」か「ああああ」のことです

複数文字

基本形を連続して書くと、複数の文字列を表せます

(文字){ 整数A , 整数B }

- 「文字」を「整数A」から「整数B」回繰り返した、すべてのパターンを表します

- (A){1,1}(B){1,1}
  - 「AB」のことです
- (あ){2,4} (い){1,1} (う){1,1}
  - 「ああいう」か「あああいう」か「ああああいう」のことです

基本形では長いです

省略形が用意されています

(文字){ 整数A , 整数B }

- 「abcdefg」を正規表現で書くと
  - (a){1,1} (b){1,1} (c){1,1} (d){1,1} (e){1,1} (f){1,1} (g){1,1}
- 長すぎるので、省略記法が用意されています

基本的な省略形

{1,1}は省略可

( )は、他の記号が続かなければ省略可

- (a){1,1}
  - 整数A = 整数B = 1、つまり、{1,1}の場合は{1,1}を省略して良いことになっています
  - つまり、(a){1,1}は、(a)と書けます
- (a)
  - ()の後に他の記号が続かない場合は、(と)を省略して良いことになっています
  - つまり、(a)は、aと書けます
- 「abcdef」を、正規表現で「abcdef」と書けるようになりました

基本形を少しいじる

「文字」には他の正規表現を入れて良い

(文字){ 整数A , 整数B }

- 「文字」の部分に他の正規表現を入れても、正規表現になります
- 正規表現の「abcdef」を「文字」部分に入れてみる
  - (abcdef){整数A,整数B} と書けます
  - (abcdef){1,1} と書くと、「abcdef」そのものです
  - (abcdef){1,2}はどういう文字列とマッチするかというと
    - 「abcdef」もしくは「abcdefabcdef」

少し例でイメージ

- 正規表現の　(あいうえお){1,2}
  - 「あいうえお」か「あいうえおあいうえお」です
- 正規表現の　(かきく){2,4}
  - 「かきくかきく」か「かきくかきくかきく」か「かきくかきくかきくかきく」です
- 正規表現の　(XYZ){1,1}
  - 「XYZ」です

どちらかにマッチする (1)

|で繋ぐと、どちらにもマッチする正規表現が書けます

- 正規表現の「文字」部分を　A|B　と書く
  - 正規表現Aか正規表現Bのどちらかにマッチする正規表現
- 正規表現 (X|Y){1,1}
  - 「X」か「Y」にマッチします
- 正規表現 (X|Y){2,3}
  - （「X」か「Y」）が（２文字か３文字）にマッチします
  - 「XX」「XY」「YX」「YY」、「XXX」「XXY」「XYX」「XYY」「YXX」「YXY」「YYX」「YYY」のどれかにマッチ

どちらかにマッチする (2)

|の左右に正規表現で|が入っているものも書けます
複数パターンのいずれかにマッチします

- 正規表現の「文字」部分を　A|B　と書く
  - 正規表現Aか正規表現Bのどちらかにマッチする正規表現
- つまり、Bを C|Dと書けば、 A|C|Dとも書ける
- 正規表現 (X|Y|Z){1,1}
  - 「X」か「Y」か「Z」
- 正規表現 (X|Y|Z){2,3}
  - 「X」か「Y」か「Z」が、２文字か３文字
  - 「XX」「XY」「XZ」「YX」（以下略）「ZZZ」

あとはこれの省略記法です

基本はこれでおしまいです
よく使うパターンは省略記法が用意されています

- まとめ

- (文字){整数A,整数B}
- {1,1}の省略
- (文字)の () を省略
- 「文字」には他の正規表現が入れられる
- |で結ぶと、どちらかにマッチする正規表現が作れる

とてもよく使う省略記法

文字の省略記法 (1)

|で沢山の１文字をつなげて書くのは、[]で囲むのと同じ
-でさらにその間の１文字全てを表せる

- 「半角小文字アルファベット１文字のどれか」にマッチする正規表現
  - (a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z){1,1}
- 長すぎるので、次の書き方が用意されています
  - [abcdefghijklmnopqrstuvwxyz]{1,1}
    - aかbかcか……かz　を [ ] で囲うことで略記できます
- これでも長いので、次の書き方が用意されています
  - [a-z]{1,1}
    - [abc……xyz]　を　[a-z]と、-で端と端を繋ぐことで略記できます

文字の省略記法 (2)

任意の１文字 .
含まれない1文字 [^文字]

- 何でも良いから１文字欲しい時
  - . で、何かの１文字にマッチします([]の中では使えません)
  - 正規表現 (a.c){1,1}
    - 「a」、何かの１文字、「c」にマッチします
    - 「aac」「abc」「adc」…「azc」「aあc」「aいc」…（以下略）
- 「これ以外の文字」が欲しい場合
  - [^文字]と書けます(文字は、１文字もしくはa-zなどの表記)
  - 正規表現[^a]{1,1}
    - 「a」意外の何か１文字にマッチします
    - 「b」「c」「d」……「z」「あ」「い」……（以下略）

文字の省略記法 (3)

行頭マッチは^
行末マッチは$

- １行の先頭にマッチして欲しいとき
  - ^で始めると、１行の先頭にマッチします
  - 正規表現　^abc
    - 行がabcで始まっているときにマッチします
- １行の末尾にマッチして欲しいとき
  - $で終わると、１行の末尾にマッチします
  - 正規表現 abc$
    - 行がabcで終わっているときにマッチします
- 正規表現 ^abc$
  - 行がabcだけの時にマッチします

数字の略記法 (1)

{整数A,∞} は{整数A,}

{整数A,整数A}は{整数A}

- 正規表現　(A){1,∞} 的な意味(n文字以上)の表記
  - (A){1,} と表記します
  - , を「文字以上」と日本語に置き換えると良い
- 正規表現 (A){3,3} の略記として (A){3} と書けます

数字の略記法 (2)

{0,}は、*と書き換え可能、{1,}は、+と書き換え可能
{0,1}は、?と書き換え可能

- 正規表現 (A){0,}の略記として (A)* と書けます
  - Aが0回以上出現したらマッチします
- 正規表現 (A){1,} の略記として (A)+と書けます
  - Aが1回以上出現したらマッチします
- 正規表現 (A){0,1} の略記として (A)?と書けます
  - Aが0回か1回出現したらマッチします

正規表現の例

時間っぽい表記

[0-9]{2}¥:[0-9]{2}

- 「２桁の数字」「:」「２桁の数字」を正規表現で
  - (0|1|2|3|4|5|6|7|8|9){2,2}(¥:){1,1}(0|1|2|3|4|5|6|7|8|9){2,2}
  - もしくは、簡単に[0-9]{2}¥:[0-9]{2} など

16進数

0x[0-9A-Fa-f]+

0x([0-9A-Fa-z]{2}){1,4}

- 「0x」で始まって「0～9」か「A～F」か「a～f」が任意の桁
  - 0x[0-9A-Fa-f]+
- 桁数が、2,4,6,8桁のいずれか
  - 0x([0-9A-Fa-z][0-9A-Fa-z]){1,4}
  - 0x([0-9A-Fa-z]{2}){1,4}
  - など

メアドっぽい表記

[a-zA-Z0-9¥-¥_¥.]+¥@([a-zA-Z0-9]+¥.)+[a-zA-Z0-9]+

- メアドっぽい文字列はどんなの？
  - @の前は、英数字、-、_と.の組み合わせ
  - @の後は、「英数字1文字以上と.の組み合わせ」が1回以上、で最後が「英数字1文字以上」
  - [a-zA-Z0-9¥-¥_¥.]+¥@([a-zA-Z0-9]+¥.)+[a-zA-Z0-9]+
  - など

URLっぽい文字列

https?¥:¥/¥/[^]+

- URLっぽい文字列はどんなの？
  - http://かhttps://で始まる
  - 半角空白じゃない文字まで
  - https?¥:¥/¥/[^]+
  - など

あとは、挑戦あるのみ！

END