テキスト道
正規表現篇

市川せうぞー
www.seuzo.jp

セミナーの目標

- 正規表現とはなにかを理解。
- 正規表現でどんなことができるかを知る。
- 正規表現の基本的な考え方がわかる。
- 正規表現を使いたくなる。

正規表現ってなあに？

- Regular Expressionの日本語訳で、regexと略される場合もある。
- 文字列をモデル化し、柔軟なパターンマッチを行う検索方法である。
- ツールというより、一般的なスキルである。
- あまりに強力なので、一度味をしめると、これなしではいられない。諸刃の刃。

なぜ正規表現を使うのか？

- 通常の検索が１：１の検索であるのに対し、正規表現は**１：多の検索**が可能。
- 大量のテキストを自由に加工し、効率の向上。
- 多くのテキストを扱うアプリケーションで使用できる。
- 要するに、**ラクをしたい**から。

どんな場面で使えるか？

- **原稿整理**

  用字用語統一、テキスト洗浄

- **タグづけ**

  HTML、InDesign タグ、XML

- **並べ替えや抽出**

  Excel を起動するまでもない。

- **およそ、テキストを使う上の<u>すべての場面</u>で使える。**

  後述「さまざまなテクニック」も参照。

どんなツールで使えるか？

- **テキストエディタ**

  mi、JEdit、Emacs、vi、MYFES、秀丸など

- **コマンドツール群**

  egrep、sed、など

- **スクリプト**

  Perl、Ruby、awk、JavaScript、PHP など

- **検索エンジン**

  Namazu、Rast など

- **その他アプリケーション**

  GoLive、Becky!、、、など

正規表現の道具箱

- メタ文字表現
- 文字クラス
- 繰り返し
- アンカー（行頭と行末）
- グループと前方（後方）参照
- 選択（論理和）
- 実際の使用時にはヘルプを見るべし

文字列のモデル化とは？

- 包丁を選ぶ前に、**まず材料を見よ！**
- パターンを見つける。
- 箱をつくる。
- どうなるべきかを想像するのが最短の道。
- 例外を考える。

注意事項

- **最左最長一致**
  最短一致をさせる方法（2種類）
- **方言**
  対応レベルの違い、メタ文字の方言
- **ワイルドカード表現**
  似て非なるもの
- **文字コード**
  文字クラス問題
  否定文字クラスには改行も含まれる
- **DFAとNFA**
  エンジンの違い
- **バックトラックの爆発**
  むやみな繰り返し

miの正規表現一覧

文字に一致する正規表現	一致する文字
.（ピリオド）	改行コード以外の任意の一文字
¥w	英数字、アンダーバー
¥W	英数字、アンダーバー以外
¥d	数字
¥D	数字以外
¥s	スペース、改行、タブ
¥S	スペース、改行、タブ以外
¥x	この後に続く２文字を１６進数アスキーコードとした文字
¥c	この後に続く１文字についてaをアスキーコード1、zをアスキーコード26とした文字
¥t	タブコード
¥r	改行コード
[]	文字範囲指定 １　[]内の文字のうちどれか ２　a-zのようにハイフォンでつないだ場合は、aからzのうちのどれか ３　^で始まる場合は、否定を表し、[]内の文字以外に一致 ４　¥w，¥d，¥sの使用が可能（[]外にあるときと同じ意味） ＜例＞　[^0-9a-zA-Z]　英数字以外に一致
それ以外の普通の文字（漢字を含む）	その文字に一致します
繰り返し正規表現＝直前の文字、グループを指定回数繰り返したものに一致	繰り返し数
*	０回以上（最長一致）
*?	０回以上（最短一致）
+	１回以上（最長一致）
+?	１回以上（最短一致）
?	０回、または、１回（最長一致）
??	０回、または、１回（最短一致）
{n,m}	n回～m回（最長一致）※ただしn,mの最大値は65534です
{n,m}?	n回～m回（最短一致）※ただしn,mの最大値は65534です

選択	
|	|より前に記述した正規表現と|以降に記述した正規表現の両方に一致
グループ	
()	()で囲んだ部分は、グループとして記憶され、後方参照や、置換時にそのとき一致した文字列を使うことができる。最初の(から順に、１からの番号が割りふられる。（くり返しの指定範囲を区切るだけのためにも使用する。） ＜例＞　<H1>(.*)</H1>　グループ１は<H1></H1>に囲まれた部分 また、置換文字列中の、$n（n:グループ番号）は、対応するグループの文字列に置き換えられる。 ＜例＞　検索文字列：<H1>(.*)</H1>／置換文字列：<H2>$1</H2> の場合、 <H1>～</H1>が<H2>～</H2>に置換される。
(?:)	繰り返しの指定範囲を区切るためだけにグループを使用したい場合に使用する。 グループ番号が割り振られない。
¥n	後方参照 グループnの文字列と一致する。 ＜例＞　(..)-¥1　ab-ab，cc-ccなど、ハイフォンで区切って、同じ２文字が繰り返される文字列に一致
特定の位置に一致する正規表現	一致するもの
^	行頭
$	行末
¥b	単語と単語の境界
¥B	単語と単語の境界以外
その他	
(?=文字列)	後続指定（一致） A(?=B)のように記述し、正規表現Aに一致し、かつ、正規表現Bに一致する文字列がそのあとにくる場合のみ一致する。また、一致したとみなされる部分は、Aのみになる。
(?!文字列)	後続指定（不一致） A(?!B)のように記述し、正規表現Aに一致し、かつ、正規表現Bに一致する文字列がそのあとにこない場合のみ一致する。また、一致したとみなされる部分は、Aのみになる。
(?#文字列)	コメント