2023 年 8 月
東京大学情報システム部
前田朗

大学図書館員のための正規表現のススメ

１．はじめに

正規表現という言葉を聞いたことがあるでしょうか。簡単にいうと文字列をパターンで表すための書式です。大学図書館員にとっては文字列のパターンマッチというと、情報検索で使うトランケーション（「*」を使った文字列の部分一致）がまず思いうかぶかもしれません。正規表現はそれに比べパターンの表現力が高く、かつプログラミング言語を含むさまざまなソフトウエアでサポートされています。

２，普段使いするならエディタで

正規表現をお手軽に使えるのはテキストエディタです。あいにく Windows 標準の「メモ帳」では正規表現対応をしていませんが、さくらエディタや EMEditor など高機能と呼ばれるテキストエディタであれば普通はサポートしていると思います。プログラミング言語で使う場合と違って、正規表現とロジックを組み合わせてテキストのフォーマット変換をすることは難しいですが、それでもテキストの「検索」と「置換」を高レベルで行うことができるようになります。

３．基本的な正規表現文法

正規表現自体がひとつのプログラミング言語といってよいほど、さまざまな処理を行えますが、ここではごく単純なテキスト検索に役立つものだけをいくつか紹介します。正規表現の詳しい説明はインターネットで検索すると数多くでてきますので、それを参照してください。なお、ソフトウエアごとに独自の拡張をしていること（方言）があることにご注意ください。

例	意味
.	一文字にマッチします
.+	量指定子「+」を文字のあとにつけることで、指定の 1 文字以上にマッチします。
.*	量指定子「+」を文字のあとにつけることで、指定の 0 文字以上にマッチします。
.{3}	量指定子「{n}」を文字のあとにつけることで、指定の n 文字の繰り返しにマッチします。

.{3,5}	量指定子「{m, n}」を文字のあとにつけることで、指定の文字からのｍからｎの繰り返しにマッチします。
¥d	数値一文字にマッチします。
¥w	アルファベット１文字にマッチします。
¥s	空白文字１文字にマッチします。
¥t	タブ１文字にマッチします。
(a|b)	丸括弧（グループ化）と|を使うと論理和のマッチができます。例ではａもしくはｂにマッチします。
^	文字列の先頭にマッチします。
$	文字列の末尾にマッチします。
¥.	¥(バックスラッシュ)を使うと、正規表現中で特殊な意味をもつ文字（.[]|{}など）を文字として扱えます。例では「.」を一文字のパターンマッチではなくピリオド１文字としてマッチさせます。

４．ISBN のパターンマッチを考える

それでは、具体例として ISBN のパターンマッチを考えてみましょう。たとえば、電子メールの本文中から ISBN だけ抜き出すことができれば、図書館の仕事で使えそうに思えてきませんか。

ISBN にはご承知のとおり、古い 10 桁のものと、新しい 13 桁のものがあります。話を簡単にするために、13 桁の次のような ISBN を具体的な数値を使わずにパターンマッチで表現してみます。まずは、例１の簡単な例から。

例１：　9784326000296

これは、数値が 13 個並んでいるので、次のように表記できます。

¥d{13}

しかしこれだけでは、13 桁以上の数値であればすべてマッチしてしまいます。ISBN は先頭３桁が「978」か「979」と決まっていますので、これを条件に追加してみましょう。

97(8|9)¥d{10}

もうすこし、高度な話として ISBN 中に「-」が入っている例２のケースを考慮してみます。

例２：　978-4-532-32399-8

先頭3桁内に「-」が入ることはない、また「-」の個数が限られるという知識を使うと、「-」のない場合を含めて次のように表記できます。

```
97(8|9)(¥d|-){10,14}
```

ここまで正規表現でISBNと思しきものをマッチするようにしましたが、より厳密にはチェックディジットの確認もしたいところです。ただし、これは正規表現ではなく関数やプログラミングで解決すべき領域となります。

なお、より高度なISBNの正規表現パターンについて、書籍「正規表現クックブック」(ISBN: 978-4-87311-450-7) に記載があります。

５．普段使いにおけるポイント

日常業務で正規表現を使う際に気を付けたいのは、「あまり凝りすぎない」ことです。正規表現は１行で凝った指定ができ、まさに職人芸ともいえるようなものさえあります。正しく指定しようとしたレアケースが実際の処理対象のテキスト中に出現しないなら、労力を抑えて若干簡易につくった正規表現でも同じ結果が返ってくることになります。特に、テキストエディタから正規表現を使う場合は、自身の眼での結果確認が容易ですので、多少の検索ノイズは許容してもよいのではないでしょうか。

６．余談

（１）正規表現チェッカー

正規表現の動作を確かめるための正規表現チェッカーがいくつもあります。

【Webサイト上で使える正規表現チェッカーまとめ】

https://qiita.com/aim2bpg/items/ea0c0d5e5fc0df7e2b6e

（２）単語の活用形までを含めてマッチ

たとえば、テキスト中で「犬が走る」ことに関する情報を検索したいとします。しかし正規表現では次のものをまとめてマッチさせることは難しそうです。

- 「犬が走った」
- 「犬が走る」
- 「犬は走れ」

正規表現の話とは別になりますが、PythonのspaCyというにツールある Mapper 機能を使うと、「犬」+「接置詞」+「『走る』の活用形」といったような、マッチングすらできてしまいます。

(3) ChatGPT は正規表現を教えてくれるか

最近では生成系 AI がプログラムのコードを生成してくれると話題です。そこで、OpenAI が提供する ChatGPT で ISBN の正規表現パターンマッチの方法を聞いてみました。

試した限りでは、聞き方（プロンプト）によって、生成される正規表現が異なるようです。無料の ChatGPT-3.5 ではたしかに ISBN にマッチするようですが、自分が理解できるものや、より厳密なマッチを求める場合は、いまのところ参考程度にしておくのがよさそうです。有料にはなりますが ChatGPT-4 の場合、筆者がみた感じでは十分有用な正規表現を回答してくれました。

ご興味のあるかたは、ぜひお試しください。