第 6 回：正規表現による検索･置換 (実践編)

1. 正規表現の利用例 (1) EmEditor で置換して保存

青空文庫所収の夏目漱石『こころ』のテキスト (kokoro.txt) を使い，正規表現を使って (1) 文末に改行を追加する (2) ルビを削除する (3) 漢字部分をルビで置き換える作業をしてみよう。

(1) ルビを削除する

	正規表現	結果
1. ファイル kokoro_org.txt を開く。		…筆を執《と》っても心持は同じ事である。よそよそしい頭文字《かしらもじ》などはとても使う気にならない。↲ 　私が先生と知り合いになったのは鎌倉《かまくら》…
2. 。の後に改行を追加する。 **補足：**厳密には，「。」の後に改行を入れただけでは全ての文を区切れたとはいえず，また間違った区切りが入ることもある。問題が起こる例を各自考えてみよう。	(置換) 。 → 。¥n (正規表現で行の加工を含む置換をすると時間がかかるので,「正規表現を利用する」はチェックせず,「エスケープシーケンスを使用する」にチェックを入れる。これでも¥t (タブ)，¥n (改行)といった基本的なメタ文字を利用できる。)	…筆を執《と》っても心持は同じ事である。↲ よそよそしい頭文字《かしらもじ》などはとても使う気にならない。↲ ↲ 　私が先生と知り合いになったのは鎌倉《かまくら》…
3. 段落末の。の後に改行が 2 重に追加されているので，段落末のみ元に戻す。	(置換) 。¥n¥n → 。¥n	…筆を執《と》っても心持は同じ事である。↲ よそよそしい頭文字《かしらもじ》などはとても使う気にならない。↲ 　私が先生と知り合いになったのは鎌倉《かまくら》…
4. 検索結果を保存する。ファイル名：kokoro.txt		

(2) ルビを削除する

	正規表現	結果
1. ファイル kokoro.txt を開く。		…広い寺の境内《けいだい》にある… …同じ言葉を二｜遍《へん》繰り返した。
2. ルビに当たる箇所 (《から》まで) を削除する。	(置換) 《[^》]+》→ (何も入力しない) (「正規表現を利用する」をチェックする)	…広い寺の境内にある… …同じ言葉を二｜遍繰り返した。
3. ルビの付く文字列の始まりを特定する記号を削除する。	(置換) ｜→ (何も入力しない)	…同じ言葉を二遍繰り返した。
4. 検索結果を保存する。ファイル名：NoRuby.txt		

(3) 漢字部分をルビで置き換える ((1), (2) に比べ複雑なので，しっかり確認すること！使用す

る正規表現を regex_sample.txt に記載しておいたのでコピー＆ペーストするとよい。)

	正規表現	**結果**
1. ファイル kokoro.txt を開く。		…広い寺の境内《けいだい》にある… …同じ言葉を二｜遍《へん》繰り返した。
2. ルビのつく文字列の始まりが｜でマークされている漢字をルビで置き換える	(置換) ｜([^《ぁ-んァ-ヶ、。｜]+)《([^》]+)》→ ¥2	…同じ言葉を二へん繰り返した。
3. ルビのつく文字列をルビで置き換える	(置換) ([ぁ-んァ-ヶ□、。])([^《ぁ-んァ-ヶ、。｜]+)《([^》]+)》→ ¥1¥3 □は全角スペース (段落先頭のインデントに対応)	…広い寺のけいだいにある…
4. 検索結果を保存する。ファイル名：Ruby.txt		

2. **正規表現の利用例** (2) EmEditor **で検索→マークアップして保存**

『こころ』のテキストから「〜してしまう」を含む文を抽出し，検索結果に **<てしま>** のように三角括弧のマークをつけて保存しよう。

	正規表現	**結果**
1. ファイル NoRuby.txt を開く。		
2. 試しに通常の検索をする。	しま	…乗ってしまった。 …聞いていたが、しまいに…
3. NoRuby.txt に対し GREP 検索 (ファイルから検索) をする。「しまう」のさまざまな変化形のパターンを指定する。	(ファイルから検索) 検索対象としてファイル NoRuby.txt を指定しておこなう： **(て|で)しま(っ|う|い|わ|え|お)**	以下のパターンが検索できる： **(し)て** **(ん)で** **(っ)て** } **しま** { **わ(ない)** **い(ます)** **う(とき)** **え(ば)** **お(う)** **っ(た)**
4. 変化形が網羅的に検索できたら，検索結果をマークアップする。	(置換) 検索結果に対しておこなう： **(て|で)(しま)(っ|う|い|わ|え|お)** → <¥1¥2>¥3	…乗っ**<てしま>**った。 (3. の検索文字列にマッチを記憶する括弧が追加されていることに注意)
5. 検索結果を保存する。ファイル名：simau.txt		

3. **正規表現の利用例** (3) EmEditor **を使った本格的なテキスト加工**

EmEditor の正規表現による置換機能を使ったより本格的な加工を体験しよう。『こころ』のテキストに形態素解析システム ChaSen [1] で品詞情報を追加したデータを利用して，助詞「に」を含む文を検索し，さらに余計な品詞情報を削除し読みやすいデータに加工してみよう。

[1] ChaSen (茶筌) は奈良先端科学技術大学松本研究室が開発しているフリーの日本語形態素解析システム。URL: http://chasen.aist-nara.ac.jp/hiki/ChaSen/

担当：千葉庄寿 (CHIBA Shoju) e-mail: schiba@reitaku-u.ac.jp

	正規表現	結果
1. EmEditor でファイル (keitai.txt) を開く		
2. GREP 検索 (ファイルから検索) をする。	(ファイル検索) に@助詞-格助詞-一般	国家@名詞-一般 に@助詞-格助詞-一般 つかえる@動詞-自立
3. 変化形が網羅的に検索できたら，検索結果をマークアップする。同時に，品詞情報を削除してしまおう。	(置換) 検索結果に対しておこなう： (に)(@助詞-格助詞-一般)→ <¥1>	国家@名詞-一般 **<**に**>** つかえる@動詞-自立 (2. の検索文字列に括弧が追加されていることに注意)
4. 検索結果から，不必要な品詞情報を除く	(置換) @[^]+ → (何も指定しない)	国家 **<**に**>** つかえる
5. 検索結果を保存する。ファイル名：kokoro_ni.txt		

- NoRuby_chasen_original500.txt は，ChaSen のオリジナル出力である (最初の 500 行のみ)。オリジナル出力には読み (カタカナ) や見出し語，活用形の詳細などが含まれる。NoRuby_chasen_2cols500.txt は，出現形と品詞情報のみをタブで区切って出力したもの (最初の 500 行のみ) [2] 。keitai.txt は後者のフルテキストを Perl スクリプト seikei.pl で加工したもの。
- 第 3 回授業で実習したように, NoRuby_chasen_2cols500.txt のようなタブ区切りのテキストは，Excel で簡単に利用できる (Excel の活用方法は次回以降実習する)。
- 3. で <¥1> とする代わりに，¥t<¥>1¥t と両側にタブを入れ，データをコピーし，Excel に貼り付けることで, KWIC (KeyWord In Context) 風の表示をおこなうこともできる。(ただし，キーワードが複数含まれている場合には，このやり方では網羅的な KWIC 表示ができない。解決策については「プログラミング言語 Perl によるテキスト処理」で紹介する。)

4. エディタによるテキスト加工からより複雑なテキスト処理へ

エディタの検索・置換機能を使ったテキスト加工作業には，限界がある：

- 手間がかかる (検索・置換を何度も繰り返す必要がある)
- 正規表現の処理に時間がかかる (kokoro.txt のような大きなファイルの改行記号を追加･削除するといった作業を EmEditor の正規表現でおこなおうとすると，めっぽう遅い)
- 複雑な加工が不得意 (たとえば,「単語の出現回数を数える」というような単純な作業でも，EmEditor の検索・置換機能だけでは実現が難しい)

より柔軟で複雑なテキスト処理をおこなうには, テキスト処理用のスクリプト言語による自動処理をおこなうのが効果的である。スクリプト言語は sed, awk, perl が有名で，どれも正規表現をサポートする (もちろん，実装内容は言語によって，またバージョンによって多少なりとも異なる)。

5. スクリプト言語 Perl を使ったテキスト処理の例

- Perl (ぱーる) は正規表現を使ってテキストを加工できる強力なスクリプト言語で，テキスト処理でもっともよく用いられる。
- 大学の PC にはバージョン 5.8 が入っており, さまざまなエンコードのテキストを処理できる。
- 単語の頻度情報を調べる **count.pl** と 単語の出現場所 (行数) を調べるスクリプト

[2] chasen の具体的なコマンドは `chasen -F "%m¥t%P-¥n" NoRuby.txt`

reference.pl で，その威力を調べてみよう。

- [スタート] → [プログラム] → [アクセサリ] → [コマンドプロンプト] を開き，以下のコマンド (網掛け部分, すべて半角で入力) を打ちこんで出力結果を EmEditor で確認しよう。

MS-DOS コマンドの解説：

dir: ファイルの一覧を表示
cd: フォルダを移動
<: 入力ファイル指定
>: 出力ファイル指定

```
>F: [Enter]
>cd  kenkyuu2005 [Enter]
>dir [Enter]
>cd  No6 [Enter]
>dir [Enter]
>perl count.pl < keitai.txt > keitai_count.txt [Enter]
>perl reference.pl < keitai.txt > keitai_refs.txt [Enter]
>dir [Enter]
>exit [Enter]
```

スクリプトによるテキスト処理には，グラフィカルなインターフェース (GUI, graphical user interface) ではなく，上記 Windows の「コマンドプロンプト」のような，文字テキストで処理を指定する旧来のインターフェース (CUI, character-based user interface) を使えることが望ましい。形態素解析システム ChaSen のように，有用なテキスト処理ツールの多くも CUI での利用を前提としている。

6. 参考文献：

6.1. 正規表現について (Perl の入門書にも正規表現の詳しい解説がある)

- IDEA・C (2001) 『正規表現の達人』 ソフトバンク.
- 大名力「正規表現によるテキスト検索」オンライン資料 URL: http://infosys.gsid.nagoya-u.ac.jp/~ohna/re/index.html
- 中尾浩 他 (2002) 『コーパス言語学の技法Ⅰ：テキスト処理入門』夏目書房
  同 (2004) 『コーパス言語学の技法Ⅱ：言語データの収集とコーパスの構築』夏目書房
- Friedl, Jeffrey E. F. (歌代和正監訳, 2003) 『詳説正規表現』第 2 版 オライリー・ジャパン

6.2. スクリプト言語 Perl について (Perl の入門書は他にもたくさんある)

- クロス, デイビット (宮川訳, 2003)『Perl データマンジング: データ加工のテクニック集』ピアソン・エデュケーション
- 平田 豊 (2004) 『Perl トレーニングブック―これから始める人の Perl プログラミング練習帳―』ソーテック社
- 武藤健志・トップスタジオ (2004 年) 『独習 Perl』第 2 版. 翔泳社
- 目黒編集室 (2004) 『これだけで身につく Perl 入門 例題 80』 日経 BP.
- Wall, Larry 他 (2002) 『プログラミング Perl』第 3 版 (全 2 巻.) オライリー・ジャパン.
- Christiansen, Tom 他 (Shibuya Perl Mongers 監訳, 2004) 『Perl クックブック』第 2 版. (全 2 巻) オライリー・ジャパン
- Hammond, Michael (2003) *Programming for Linguists: Perl for Language Researchers*. Oxford: Blackwell.

6.3. Perl 以外のスクリプト言語について

- 美吉明浩 (1998) 『Grep, Sed, Awk Manual & Reference』秀和システム. (絶版)
- Dougherty, Dale 他 (1997) 『sed & awk プログラミング』改訂版. オライリー・ジャパン.
- Aho, Alfred V (2001) 『プログラミング言語 AWK』シイエム・シイ.
- 上田博人 (1998) 『パソコンによる外国語研究 (II) 文字データの処理』くろしお出版. (awk を使った文字列処理)
- Barnbrook, Geoff (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press. (awk を使った言語分析用スクリプトを紹介)

担当：千葉庄寿 (CHIBA Shoju) e-mail: schiba@reitaku-u.ac.jp