

テキスト処理

Linux コマンド 参考サイト

<https://hydrocul.github.io/wiki/commands/>

文字列の置き換え

■ tr

- ・テキストファイル中の特定の文字を置き換える。
 - ・置き換える文字が、一文字ずつ順に対応しているところがポイント
 - ・temp.parsed というファイル内の丸かっこを角かっこに置き換える

```
cat temp.parsed | tr '()' '[]'
```

OpenNLP

<http://opennlp.sourceforge.net/projects.html>

pdf ファイルからテキストを抽出する。pdftotext

```
pdftotext -enc EUC-JP text.pdf
```

これで、text.txt というファイルが生成される。

American National Corpus における n-gram 表現の抽出

1) 本文のみのファイルをひとまとめにする

```
cat *trans.xml > sw20.all
```

2) タグの除去

```
perl -pe 's/<.+?>//g' sw20.all > sw20.txt
```

3) n-gram 抽出プログラムをかける (Ngram.pm を応用)

```
freqNgram sw20.txt > sw20.ngram
```

4) 結果

```
2 you know 980
2 it s 609
2 that s 488
2 don t 392
2 um hum 389
2 i think 270
3 i don t 240
2 i m 234
2 uh huh 224
2 a lot 218
```

- ・まとめてやると

```
cat *trans.xml | perl -pe 's/<.+?>//g' > file.txt
freqNgram file.txt > file.ngram
```

freqNgram は引数にファイル名を要求するので、パイプ処理はできない。

perl で xml のタグの削除

```
perl -pe 's/<[a-zA-Z0-9 ="\.¥?¥!¥/¥-:~]*>//g' file.xml
```

```
perl -pe 's/<[¥w¥s¥d="¥.¥?¥!¥/¥-:~]*>//g' file.xml
```

```
perl -pe 's/<.+?>//g' file.xml
```

繰り返しを表す + の後ろの ? は「最短一致」

これを指定しないと「最長一致」でマッチしてしまう。

emacs で日付の挿入 .emacs.el - sugiura (2006 年 03 月 16 日 13 時 05 分 14 秒)

```
;; insert date
(defun foo ()
  (interactive)
  (insert (concat
    "["(format-time-string "%Y-%m-%d")"]"(" (format-time-string "%A")")"))))
(define-key global-map "%C-cd" 'foo)
```

これで

```
[2006-03-16] ( 木曜日 )
```

とでる。

perl のモジュールのインストール - sugiura (2006 年 01 月 15 日 06 時 37 分 37 秒)

1. <http://search.cpan.org/> からダウンロード
2. 解凍
3. perl Makefile.PL
4. make
5. make test
6. su
7. make install

perl での複数行にわたるコメントアウト - sugiura (2005 年 12 月 08 日 03 時 10 分 00 秒)

=pod

この部分をコメントアウト

=cut

emacs での取り消し - sugiura (2005 年 10 月 04 日 15 時 47 分 48 秒)

C-x u

サクラエディターを使ったテキスト処理 - sugiura (2005 年 09 月 17 日 10 時 44 分 20 秒)

<http://sakura-editor.sourceforge.net/>

- ・ 関連情報
 - ・ <http://www.geocities.jp/maru3128/sakura.html>
- 1. パッケージ版をダウンロードしてインストールする。
 - ・ http://members.at.infoseek.co.jp/sakura_editor/package/sinst1-5-5-0.exe
- 2. 環境を設定する
 - ・ フォント
 - ・ メニューの「設定」から「フォント設定」: MS 明朝の 12 ポイントがよい
 - ・ 行の折り返し
 - ・ メニューの「設定」から「現在のウィンドウ幅で折り返し」
- 3. 検索
 - ・ メニューの「検索」から「検索」
 - ・ メニューの「検索」から「Grep」
 - ・ ファイルが保存していないと「現在編集集中のファイルから検索」が使えない。
 - ・ 正規表現が使える
- 4. 置換
 - ・ メニューの「検索」から「置換」
 - ・ 「正規表現」を使うことでいろいろとテキスト処理ができる。

例: べたうちテキストを、一文一行にする。
「。」があったら改行すればよい。
「。」を「。改行」に置き換える。
「改行」は正規表現で「¥r¥n」(Windows 上の場合)
つまり、「。」を「。¥r¥n」に置き換える、とすればよい。

html ファイルをテキストファイルに整形するには - sugiura (2005 年 08 月 11 日 02 時 13 分 04 秒)

```
lynx -dump URL
```

これで、もとの html ファイルの「見た目」を「一応」保った形で、テキストファイルが出力される。(URL は実際の http.... を入れる)

あとは、これをリダイレクトすれば、様々なテキスト処理ができる。

- ・ 例: 単語数を数える。

```
lynx -dump URL | wc -w
```

- ・ 例: Web ページの変更をチェックする

```
lynx -dump URL > file1.txt  
lynx -dump URL > file2.txt  
diff file1.txt file2.txt
```

perl メモ - sugiura (2005 年 07 月 17 日 02 時 35 分 43 秒)

<http://www.din.or.jp/~ohzaki/perl.htm>

マルチバイトの文字のマッチングの処理方法が参考になった。

これを参考に、webGrep for NESS を修正。

「犬」を入れてもだいじょうぶ。