

# テキストデータの統計科学入門

## 2章 テキストのクリーニングと関連ツール

茨城大学情報工学科  
國井慎也

# テキストのクリーニング

分析する対象を電子化したりデータ形式をそろえたり、 unnecessary データを削除すること

ここでは、 unnecessary データを削除することを取り上げる

# 正規表現

文字列の集合を1つの文字列で表す方法

- 例 《》とそれに囲まれる文字列

《[ ^ > ] >>

[ ]: 全ての文字列

^: 右に来た表記を否定

# 不必要なデータの削除

データの削除自体は、テキストエディタ上でも可能  
しかし、プログラミング言語を使用したほうが効率が良い

```
use encoding 'shift-jis'  
while(<>){  
s/<<[^\>]>>//g;  
print "$_";  
}
```

ルビとそれに囲まれる文字列を削除するperlのプログラム

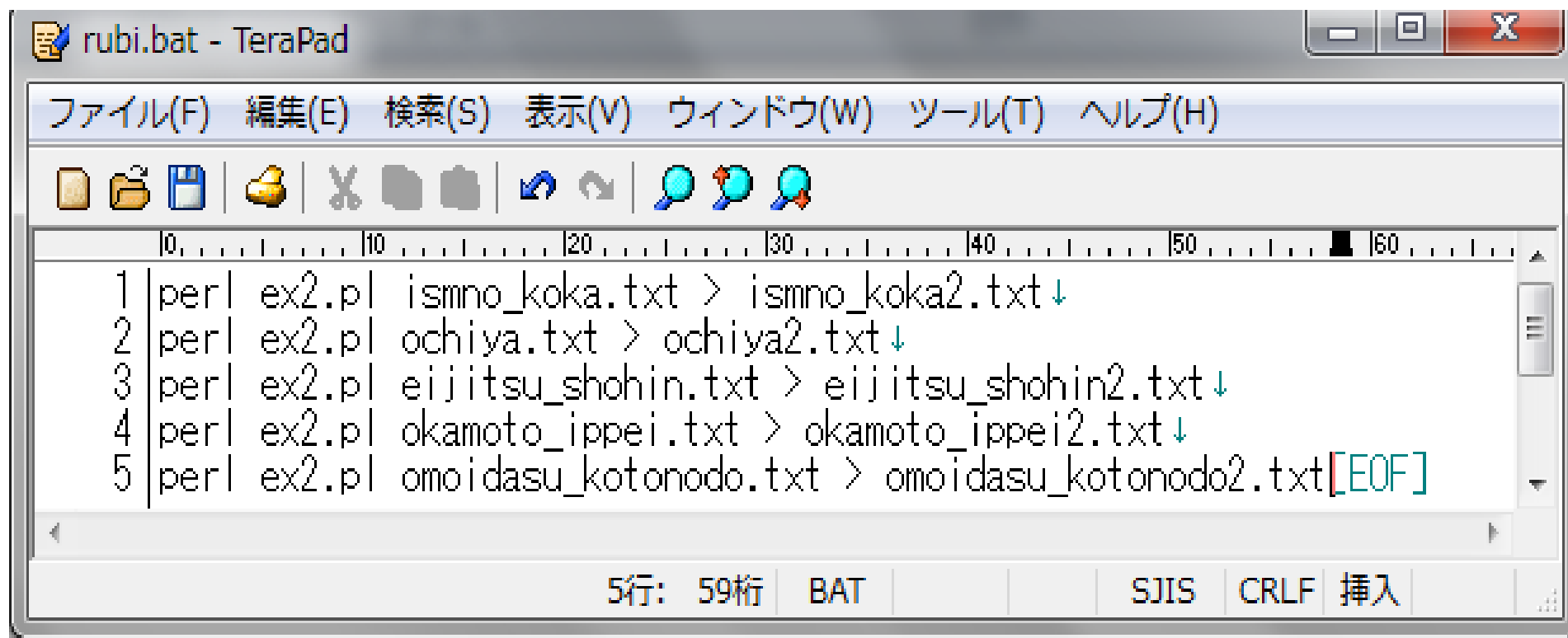
# バッチ処理

コマンドプロンプト上で実行すると、一回の実行で一つのファイルしか実行できない

バッチ処理を使うと、複数のテキストファイルも実行できる

# バッチ処理の例

- 以下のように記述し保存すると、ダブルクリックで処理が実行される



```
rubi.bat - TeraPad
ファイル(F) 編集(E) 検索(S) 表示(V) ウィンドウ(W) ツール(T) ヘルプ(H)
1 | perl | ex2.p | ismno_koka.txt > ismno_koka2.txt ↓
2 | perl | ex2.p | ochiya.txt > ochiya2.txt ↓
3 | perl | ex2.p | eijitsu_shohin.txt > eijitsu_shohin2.txt ↓
4 | perl | ex2.p | okamoto_ippei.txt > okamoto_ippei2.txt ↓
5 | perl | ex2.p | omoidasu_kotonodo.txt > omoidasu_kotonodo2.txt [EOF]
5行: 59桁 | BAT | SJIS | CRLF 挿入
```