第 3 章

テキストファイルとその扱い方

前章まで扱ってきたファイルはすべて「テキストファイル」であった。ファイルの実態は単なるデータの塊であるから，コンピュータにとってはどんなファイルも区別がない。しかし，人間が読んで理解できる文字 (に割り当てられたデータ）だけで構成されたファイルは「文字の集合」として解釈できる。このようなファイルを「テキストファイル」，それ以外のファイル (画像，動画ファイル等）を「バイナリファイル」(binary file) と呼ぶ。本章ではこのテキストファイルをあれこれ操作して UNIX の CUI に慣れるための訓練を行う。

3.1　文字コード = 文字集合 + 符号化方式

文字コードとは，文字集合（使用可能な文字の集まり）に対してどのような整数（2 進数を bit 列で表現) の対応付けを行うか（符号化方法）を定めた方式のことを言う。文字コードが異なると「文字化け」が生じる (図 3.1) が，これは文字コード同士で同じもでも異なる整数を割り当ててているために起きる。異なる文字コード体系では，bit 列との対応付けが異なり，同じ bit 列でも違う文字が割り当てられるからである。

現在使われている主な文字コードは下記のものがある。1 Byte=8 bit を単位として文字を割り当てる方式が主流である。

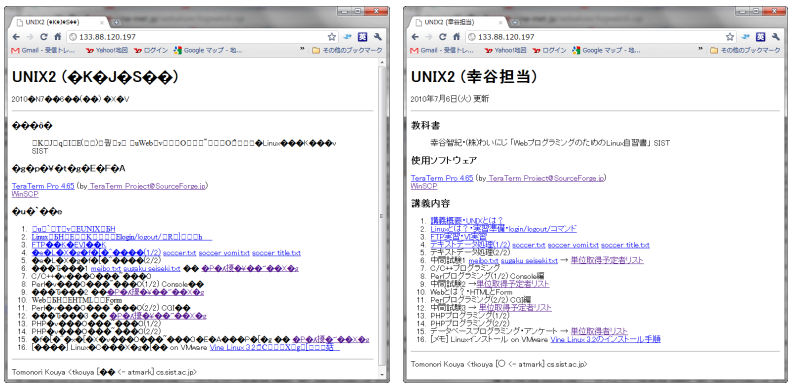

図 3.1　文字化けの例 (左図)：Shift JIS 指定で正しく表示 (右図)

■1 Byte コード・・・ASCII, JIS X 0201　ASCII(American Standard Code for Information Interchange) コードは 7bit で英語で使用されるアルファベット，数字，その他の記号（スペース，タブ，カンマ，セミコロン (;)，コロン (:)，シングルコーテーション (')，ダブルコーテーション (") 等) といった文字集合に 7bit の符号化方式を定めたもの。日本ではかつて「半角英数字」とも呼ばれていたが，「半角」「全角」というのはフォント (文字の形態) の横幅を指す言葉であるので，現在のように様々なフォントを混ぜ込んで使うようになると，必ずしも「半角英数字」が ASCII コードの文字表現とは限らない。

この ASCII 文字集合に日本語で用いられるカタカナと記号 (¥ マークなど) を付加して 8bit = 1 Byte で表現したものが JIS(Japan Industrial Standard) X 0201 である (X は情報関係の規格を意味する)。

■2 Byte コード・・・JIS コード，ISO-2022-JP, Shift JIS, 日本語 EUC　日本人が日常的に使用する感じをすべて表現しようとすると，どうしても 1 Byte では足りず，2 Byte 分の文字集合が必要となる。これを JIS X 0208 と呼び，区と点という 2 次元座標に展開して文字を並べた表が用いられていることから，区点コードとも呼ぶ。古語や特殊な異体字を除けば，現在の日本語表記を表現するにはこれで十分であることが多い。この区点コードをそのまま符号化方式として採用したものを JIS コードと呼ぶ。

この JIS X 0208 という文字集合に対して，現在使用されている符号化方式は 3 つある。

一つは E-mail に日本語表記を用いる際に使用される ISO[4] 2022-JP。古いメーラー (MTA, Message Transfer Agent) を使っている人とメールのやり取りをすると，これ以外の符号化方式に対応しておらず，下手に Unicode なぞを使うと「文字化けして読めない！」というお叱りを受けたりする。

二つ目は Shift JIS という，Microsoft 社が提案して自社の MS-DOS(Windows 以前の CUI OS) で活用された符号化方式である。そのため，MS 漢字という言い方もされる。その名の通り，JIS X 0208 をシフト (shift,「ずらす」の意) して，JIS X 0201 と干渉しないように 2 byte 文字の割り当てを行い，最初の 1 Byte 目をチェックするだけで JIS X 0201 との区別ができるようにしたものである。現在の Windows は後述する Unicode を OS 内部で使用しているが，エディタ等では未だに Shift JIS がデフォルトの文字コードとして使用される。

三つ目は日本語 EUC (Extended UNIX Code) である。その名の通り，UNIX 用の日本語表記のための符号化方式である。中国語でも EUC が定められているが，日本語 EUC との互換性はないようである。Linux でも Vine Linux のように日本人がサポートしているディストリビューションではまだ日本語 EUC が使用されているが，主流は Unicode 系に移りつつある。

■4 Byte コード・・・Unicode　Unicode(ユニコード)[5] は，世界中で使用されている文字集合を可能な限り集めて 21 bit < 4 Byte で表現した文字集合である。中国，台湾，韓国，日本で使用されている漢字は，それぞれ字体が異なっているものが多いが，それを一つにまとめている (CJK, China-Japan-Korea) ため，制定当初はナショナリスティックな小反発が日本では起きた。現在，Windows や UNIX といった主要 OS では Unicode をサポートしているものが増えてきている。これは国際化が容易であること，4 Byte コードを用いても何ら支障がないほどコンピュータの性能が上がった (CPU の処理速度が上がり，

メインメモリも大容量化した）ことで Unicode の負担が気にならなくなった，という事情も寄与しているようである。

この Unicode 文字集合を，1 Byte 単位，2 Byte 単位, 4 Byte 単位で符号化する方式を，それぞれ UTF-8, UTF-16, UTF-32 と呼ぶ。ASCII コードとの互換性が高いせいもあって，多くの OS，アプリケーションでは UTF-8 を使うことが多いようである。本書で使用している Linux の環境では，特に断らない限りこの UTF-8 を使っている。

3.2 リダイレクトとパイプ

ここでは入出力系統を適宜変更する**リダイレクト** (redirect) と，コマンド間の入出力を接続する**パイプ** (pipe) の使い方を簡単に解説する。

3.2.1 リダイレクト

まず本講義用の資料 (P.vi 参照) からつぎの三つのファイルをダウンロードし，自分の UNIX マシンへ cat コマンドを利用して転送してみよう。

prime_minister.txt ・・・日本の首相データ
yomi.txt ・・・首相の氏名（漢字表記）とよみがな
title.txt ・・・首相データのタイトル部分（4 行）

まず，Web ブラウザで Web ページを開き，テキストファイルを表示する。文字化けしていないことを確認したら，TeraTerm 上で cat コマンドを使って

```
$ cat > prime_minister.txt
```

とする。この「cat > **ファイル名**」内の不等号>がリダイレクトを意味する。この場合は，cat コマンドの出力を指定ファイルに流し込む，という意味になる。不等号の向きを逆にすると (<)，ファイルの内容を入力としてコマンドに流し込む，という意味になる。

これを実行すると，cat コマンドは標準入力からの入力待ち状態になるので，Web ブラウザで表示したテキストファイルの内容をコピーし，TeraTerm の上にペーストする。するとペーストしたテキストファイルの内容が標準入力から流し込まれるので，終了したらファイルの終了を意味する [CTRL] + [z] を入力してリダイレクト処理を完了させる。

```
$ cat prime_minister.txt
1       伊藤博文        1885/12/22      1888/04/30      861     44      第一次
伊藤内閣
2       黒田清隆        1888/04/30      1889/10/25      544     47
(略)
98      安倍晋三        2017/11/01      2020/09/16      1051    63      第四次
安倍内閣
99      菅義偉  2020/09/16      2021/10/04      384     71
100     岸田文雄        2021/10/04      2021/11/10      38      64      第一次
岸田内閣
101     岸田文雄        2021/11/10                                      第二次岸田内閣
^Z ←[CTRL]キーを押しながら[z]キーを押す。
[1]+  停止                 cat > prime_minister.txt
$ ←プロンプトが戻ってくれば処理完了
```

最後に，ファイルがきちんと指定ファイル名になっているか，内容がコピーされている

かを確認する。

```
$ ls prime_minister.txt
prime_minister.txt
$ cat prime_minister.txt
1       伊藤博文        1885/12/22      1888/04/30      861     44      第一次
伊藤内閣
(略)
101     岸田文雄        2021/11/10                              第二次岸田内閣
$
```

yomi.txt, title.txt も同様にして転送する。転送後は上記のように必ず cat コマンドを使用し，文字化けせずに保存されていることを確認しよう。

なお, cat コマンドは, 複数ファイルの出力も一括して行える。例えば prime_minister.txt の後に yomi.txt を出力するには

```
$ cat prime_minister.txt yomi.txt
```

とすればよい。これを別ファイルにリダイレクトすれば，テキストファイルの連結処理にも使用できる。

```
$ cat prime_minister.txt yomi.txt
1       伊藤博文        1885/12/22      1888/04/30      861     44      第一次
伊藤内閣
(略)
100     岸田文雄        2021/10/04      2021/11/10      38      64      第一次
岸田内閣
101     岸田文雄        2021/11/10                              第二次岸田内閣
1       伊藤博文        いとうひろぶみ  ←  ここからyomi.txtの内容
2       黒田清隆        くろだきよたか
(略)
99      菅義偉  すがよしひで
100     岸田文雄        きしだふみお
101     岸田文雄        きしだふみお
$
```

以降ではこれらのテキストファイルを使用していく。

3.2.2　パイプ

パイプは，二つ以上のコマンドの出力と入力を接続する結果を繋ぐために使用される。例えば wc というコマンドはテキストファイルの行数，文字数を数えるために使用されるが

```
$ wc prime_minister.txt
```

としても

```
$ cat prime_minister.txt | wc
```

としても結果は同じである。下記の縦棒 (|) がパイプを意味し, cat prime¥_minister.txt の出力内容，つまり prime_minister.txt の内容を wc コマンドの入力として流し込んでいるので，上記と同じ処理を行ったことになる。

パイプはしばらくテキストファイル処理のためのコマンドを連結して使用する時に必要

になるので，頭の片隅に入れておいて頂きたい。

3.2.3 コマンド出力の保存

リダイレクトはコマンドの出力結果をファイルに保存する時によく利用される。例えば，wc の出力結果を保存したいときには

```
$ wc prime_minister.txt
 101  659 5850 prime_minister.txt ←「改行数 単語数 バイト数」の順に出力
$ wc prime_minister.txt > wc_prime_minister.txt ←出力のリダイレクト
$ cat wc_prime_minister.txt ←リダイレクト内容の確認
 101  659 5850 prime_minister.txt
```

とリダイレクトを使うことによって，標準出力内容が wc_prime_minister.txt に書き込まれていることが分かる。

3.3 grep, sort, join コマンド

さて，保存した 3 つのテキストファイルを使って，以下のテキストファイル処理のためのコマンドを使っていくことにしよう。

grep ・・・「正規表現」に合致する文字列を探索するコマンド (プログラム)
sort ・・・ファイル内のテキストを，ユーザの指定に従って並べ替えるコマンド
join ・・・ファイル内のテキストを，ユーザの指定に従ってくっつけるコマンド

3.3.1 grep コマンド

grep は，**正規表現**を用いてユーザが探しだしたい文字列を検索して表示するコマンドである。正規表現をきちんと説明するとそれだけで一冊の本ができてしまうので，ここではごく簡単な使い方だけを紹介する。

一番単純な使い方としては

```
$ grep 安倍 prime_minister.txt
90      安倍晋三        2006/09/26      2007/09/26      366     52
96      安倍晋三        2012/12/26      2014/12/24      729     58      第二次
安倍内閣
97      安倍晋三        2014/12/24      2017/11/01      1044    60      第三次
安倍内閣
98      安倍晋三        2017/11/01      2020/09/16      1051    63      第四次
安倍内閣
```

とする。「安倍」という文字列を含む行を表示する。シングルクォテーションで検索したい文字列を括って

```
$ grep '安倍' prime_minister.txt
90      安倍晋三        2006/09/26      2007/09/26      366     52
96      安倍晋三        2012/12/26      2014/12/24      729     58      第二次
安倍内閣
97      安倍晋三        2014/12/24      2017/11/01      1044    60      第三次
安倍内閣
98      安倍晋三        2017/11/01      2020/09/16      1051    63      第四次
安倍内閣
```

と指定してもよい。スペースを含む文字列を指定する時には必ずこのようにすること。

正規表現は特別な意味を持つ ASCII 文字を利用する。例えば

```
$ grep 閣$ prime_minister.txt
1       伊藤博文        1885/12/22      1888/04/30      861     44      第一次
伊藤内閣
(略)
100     岸田文雄        2021/10/04      2021/11/10      38      64      第一次
岸田内閣
101     岸田文雄        2021/11/10                                      第二次岸田内閣
```

とすると，行末に「閣」という文字を含む行を表示する。$は行末を意味する。

```
$ grep /*鳩山* prime_minister.txt
52      鳩山一郎        1954/12/10      1955/03/19      100     71      第一次
鳩山内閣
53      鳩山一郎        1955/03/19      1955/11/22      249     72      第二次
鳩山内閣
54      鳩山一郎        1955/11/22      1956/12/23      398     72      第三次
鳩山内閣
93      鳩山由紀夫      2009/09/16      2010/06/08      266     62
```

とすると，アスタリスク*部分は任意の文字 (なくても良い) を意味するので，「ミスター鳩山」「椚山鳩山磐田」などがヒットする。このアスタリスクを**ワイルドカード**と呼ぶ。

[**文字** 1-**文字** 2] とすると，文字 1 から文字 2 までの文字コードの範囲の文字すべてを意味する。例えば

```
$ grep ¥/2[0-9] prime_minister.txt
1       伊藤博文        1885/12/22      1888/04/30      861     44      第一次
伊藤内閣
2       黒田清隆        1888/04/30      1889/10/25      544     47
3       山縣有朋        1889/12/24      1891/05/06      499     51      第一次
山県内閣
(略)
95      野田佳彦        2011/09/02      2012/12/26      482     54
96      安倍晋三        2012/12/26      2014/12/24      729     58      第二次
安倍内閣
97      安倍晋三        2014/12/24      2017/11/01      1044    60      第三次
安倍内閣
```

とすると，20 日〜29 日 (「/20」〜「/29」) が年月日に含まれる行すべて表示される。

3.3.2　sort コマンド

sort コマンドはその名の通り，テキストファイルの行を指定文字列の文字コード順に並び替えて表示するコマンドである。

例えば，yomi.txt の 3 列目 (-k3 オプション) のよみがな順に並び替えを行うと

```
$ sort -k3 yomi.txt
47      芦田均　あしだひとし
92      麻生太郎        あそうたろう
(略)
37      米内光政        よないみつまさ
25      若槻禮次郎      わかつきれいじろう
28      若槻禮次郎      わかつきれいじろう
```

と表示される。日本語の場合，漢字の並び替えはあまり意味を成さないので，カタカナ，

ひらがなの文字列を並び替えの基準として用いることがほとんどであろう。

-r オプションを用いると逆順に表示される。

```
$ sort -r -k3 yomi.txt
```

前述の結果と比較して確認してみよう。

3.3.3 join コマンド

join コマンドはその名の通り，テキストファイルを指定文字列をキーとして結合するためのコマンドである。例えば次のようにして用いる。

```
$ join -j1 1 -j2 1 -o 1.1 1.2 2.3 1.3 1.4 1.5 1.6 1.7 prime_minister.txt yomi.txt
1 伊藤博文 いとうひろぶみ 1885/12/22 1888/04/30 861 44 第一次伊藤内閣
2 黒田清隆 くろだきよたか 1888/04/30 1889/10/25 544 47
(以下略)
```

prime_minister.txt の 1 列目 (内閣 No.) と yomi.txt の 1 列目に同じ並びで同じ文字があるということを前提に，これらをキーとして結合処理を行っている。処理結果は-o 以下に指定した順に表示される。上記の場合は「prime_minister.txt」を 1,「yomi.txt」を 2 として表現し，1.1 が「prime_minister.txt の 1 列目」，1.2 が「prime_minister.txt の 2 列目」(=「yomi.txt の 2 列目」と同じ文字列)，2.3 が「yomi.txt の 3 列目」・・・という並びで結合結果を出力している。

課題 A

prime_minister.txt，yomi.txt を結合して，よみがな順にソートし，タイトル部 title.txt を先頭に付加して prime_minister_meibo.txt を作成せよ。但し，データは

内閣 No.	首相名	ふりがな	発足年月日	辞職年月日	在職期間 (日)	就任時年齢	メモ

という順に並べること。きちんとした手順を得るまで試行錯誤し，手順が明確になった時点でログを取って印刷し，「学籍番号」「氏名」を明記の上，提出せよ。

3.4 相対パス指定と準備作業

カレントディレクトリから，別のディレクトリを表現する方法として，絶対パス指定 (表記) と相対パス指定がある。UNIX では概ね深い階層がホームディレクトリになっているので，カレントディレクトリを基準とした相対パス指定が使えるようになると便利である。相対パス指定は，カレントディレクトリを基準とする。カレントディレクトリはドットスラッシュ (./) で表記する。カレントディレクトリよりひとつ上のディレクトリはドットドットスラッシュ (../) で表記する。

例として，本章で行う作業のため，先週作成した prime_minister_meibo.txt を unix05 ディレクトリにコピーする作業を示す。

以下のように，unix05 ディレクトリを掘り，カレントディレクトリを unix05 に移動した後，相対パス指定を使って次のように行えばよい。

```
$ ls ../unix04 ← unix05のひとつ上のディレクトリにあるunix04にアクセスしてファイル
リストを確認する
(略) ・・・  prime_minister_meibo.txt  ・・・(略)
$ cp ../unix04/prime_minister_meibo.txt ./ ←unix04から指定ファイルをカレントデ
ィレクトリにコピー
$ cat prime_minister_meibo.txt  ← コピーしたprime_minister_meibo.txtの内容を確
認
```

3.5　sed・・・正規表現による文字列の置換

sed コマンドは，一種の言語とも言えるほど複雑な文字列の置換を可能にする。ここでは以降の作業に必要な最低限の機能だけ体験しておく。

sed コマンドの使い方は次のようになる。

```
$ sed 置換処理 処理対象テキストファイル名
```

置換処理には grep 同様，正規表現が用いられる。例えば”MF”という文字列を”ミッドフィルダー”に置き換えるときには

```
$ sed 's/第二/第2/' prime_minister_meibo.txt
首相データベース
http://www.kantei.go.jp/から幸谷が作成。
2022年7月現在
内閣No. 首相名　ふりがな　　発足年月日　　辞職年月日　　在職期間(日)　就
任時年齢　　メモ
47 芦田均 あしだひとし 1948/10/15 220 60
92 麻生太郎 あそうたろう 2009/09/16 358 68
90 安倍晋三 あべしんぞう 2007/09/26 366 52
96 安倍晋三 あべしんぞう 2014/12/24 729 58 第2次安倍内閣
97 安倍晋三 あべしんぞう 2017/11/01 1044 60 第三次安倍内閣¥
(以下略)
```

と指定する。

prime_minister_meibo.txt には 4 行のタイトル文があるので，その 4 行文は無視し，以降の行だけを置換の対象としたいときには，置換の前に 1,4d を指定し，セミコロン (;) で置換処理とつなげておく。

```
$ sed '1,4d;s/第二/第2/' prime_minister_meibo.txt
47 芦田均 あしだひとし 1948/10/15 220 60
92 麻生太郎 あそうたろう 2009/09/16 358 68
90 安倍晋三 あべしんぞう 2007/09/26 366 52
96 安倍晋三 あべしんぞう 2014/12/24 729 58 第2次安倍内閣
(以下略)
```

同様に，フィールド区切りのスペースをタブ (¥t) に置き換えるには次のように指定すればよい。

```
$ sed '1,4d;s/ /¥t/' prime_minister_meibo.txt   ←行最初のスペースのみ置換
(出力略)
$ sed '1,4d;s/ /¥t/g' prime_minister_meibo.txt  ←行全体のスペース全て置換
(出力略)
```

3.6 コンパイラとインタプリタ

次に，インタプリタ言語である awk スクリプトを作成する。その前に，**コンパイラ**と**インタプリタ**という二つの言語処理系について，簡単に復習しておこう。

コンピュータ上で実行されるプログラムは，元となるテキストファイルにコンピュータ言語で処理内容を記述したものを実行可能な形式に変換してできたものである。この元になったテキストファイルを**ソースコード** (ソースプログラム) と呼ぶ。このソースコードをどのように「解釈」して「実行」するのかは，言語や処理系によって異なる。

■コンパイラ　ソースコードから，様々な**ライブラリ**に記述されている機能を追加した上で，CPU が直接処理できる実行可能ファイル[*1]を生成するソフトウェアをコンパイラと呼ぶ。実行速度重視のプログラムに適しており，C, C++, Fortran 等のコンピュータ言語ではもっぱらこのコンパイラを使って開発を行う。

■インタプリタ　ソースコードを一行ずつ解釈してその都度実行するソフトウェアをインタプリタと呼ぶ。インタプリタで解釈されるソースコードは**スクリプト**とも呼ばれる。テキストファイル処理・CGI(Common Gateway Interface) などに適している。Web アプリケーション作成のために使う PHP, Ruby，Python, JavaScript といった言語はインタプリタを使用することが多い。

3.7 awk・・・古典的スクリプト言語

C 言語/UNIX の開発者が提案し実装した，テキストファイル処理用言語である。文法構造は C/C++ とよく似ている。ただし Perl が普及した現在はあまり使用されていない。Perl については後述する。

awk スクリプトは”ファイル名.awk”のように拡張子に”.awk”を使うのが普通である。スクリプトの構造は大きく 3 つに分割される。

```
BEGIN { データ入力前に実行 }
{データ入力時に実行}
END{ 終了前に実行 }
```

BEGIN{・・・}, END{・・・}の部分は，実行時の最初と最後に一度だけ実行される部分である。その間の{・・・}は，テキストファイルが一行読み込まれるごとに繰り返し実行される部分である。

例えば総理大臣名簿 (prime_minister_meibo.txt) に登録されている人数を数える awk スクリプト (ファイル名は sample.awk) は次のようになる。左の「行番号:」はスクリプトとは関係ないので打ち込まないこと！

```
1: BEGIN{print "---日本の総理大臣名簿---¥n";}
2: {print $2; print $5; print $6; count++;}
3: END{print "¥n合計", count, "人"; print "---総理大臣名簿名簿終了---¥n";}
```

[*1] Java や.NET 上の処理系のように，仮想的な (ソフトウェア上の) 処理系を包含したプログラムを生成する場合もある。

これを実行すると次のような結果が得られる。

```
$ sed '1,4d' prime_minister_meibo.txt | awk -f sample.awk
---日本の総理大臣名簿---

芦田均
220
60
麻生太郎
358
68
(略)
米内光政
189
59
若槻禮次郎
446
59
若槻禮次郎
244
65

合計 101 人
---総理大臣名簿名簿終了---
```

課題 B

1. データをカンマ (,) で区切って表現したテキストファイルを CSV ファイル (拡張子は.csv) と呼ぶ。sed を用いて，prime_minister_meibo.txt から prime_minister_meibo.csv ファイルを作成せよ。例えば次のように sed とリダイレクトを併用すればよい。

```
$ sed '1,4d;(スペースをカンマで置き換える指定)
' prime_minister_meibo.txt
47,芦田均,あしだひとし,1948/10/15,220,60,
(略)
28,若槻禮次郎,わかつきれいじろう,1931/12/13,244,65,第二次若槻内閣
$ sed '1,4d;(スペースをカンマで置き換える指定)
' prime_minister_meibo.txt > prime_minister_meibo.csv
```

2. awk を用いて，prime_minister_meibo.txt に掲載されている内閣の平均存続日数と平均就任年齢をそれぞれ計算する awk スクリプト ave.awk を作成して実行せよ。例えば次のように sed と ave.awk をパイプでつないで用いればよい。

```
$ sed '1,4d' prime_minister_meibo.txt | awk -f ave.awk
---日本の総理大臣名簿---

芦田均
220
60
(略)
若槻禮次郎
244
65

合計 101 人
平均在職日数: 489.802　平均年齢: 60.1683
---総理大臣名簿名簿終了---
```

第 3 章の学習チェックリスト

- □ テキストファイルとは何かを第三者に説明することができる。
- □ リダイレクトとパイプを活用して実行結果をファイルに保存したり，複数のコマンドを組み合わせて実行することができる。
- □ テキストファイルに対して文字列の検索や置換，結合処理を実行することができる。
- □ 簡単な awk スクリプトを用いてテキストファイルからデータを取り出して加工することができる。