第 2 回：テキストエディタの利用

1. テキストデータとファイル形式：テキスト形式 vs. バイナリ形式

- テキストデータでは，同じ文字は同じコードで表現される＝同じ文字であれば，網羅的に探せる。この特徴により，テキストデータはコンピュータ上で扱うデータの中で特に重要な役割を持っている。

実習１：Kadai サーバの [schiba] → [kenkyuu2005] → [No2] フォルダを file_server の Home にある kenkyuu2005 フォルダにコピーしなさい。No2 フォルダを開き，「特殊漢字.txt」を開いて異体字を確認しなさい。

- ファイルとしてのテキスト（**テキスト形式**のファイル text file）は，文字情報以外のデータを含まない。テキスト形式のファイル以外は全て**バイナリ形式** binary file という：ワードプロセッサはバイナリ形式で書式情報つきのテキストを扱う。

  - ファイル内のデータ構造はファイルの種類によって全く異なる。データの内容・メディアは同じでも，ファイルの形式によっては特定の環境でしか使えない。

画像の形式：

形式	拡張子	汎用性	Web	アニメーション	透過	特徴
JPEG	jpg, jpeg	○	○	×	×	1600 万色以上
GIF	gif	○	○	○	○	最大 256 色
PNG	png	○	○	×	○	最大 2800 億色以上
TIFF	tif	△（諸形式あり）	×	-	-	各種あり
Bitmap	bmp	×（Windows 用）	×	-	-	各種あり
PSD	psd	×（Photoshop 用）	×	-	-	各種あり

テキストデータを含む形式：

ファイルの種類	拡張子	汎用性	特徴	ファイルサイズ
テキスト	txt	○	書式なし	小
Rich Text 形式	rtf	△	書式つき。多くのワープロソフトがサポート	大（大きな文書になればなるほどファイルサイズが膨れ上がる）
Word 文書	doc	×	書式つき。Word 専用	中

- ファイルの種類を表す手段：拡張子 file extension(s)

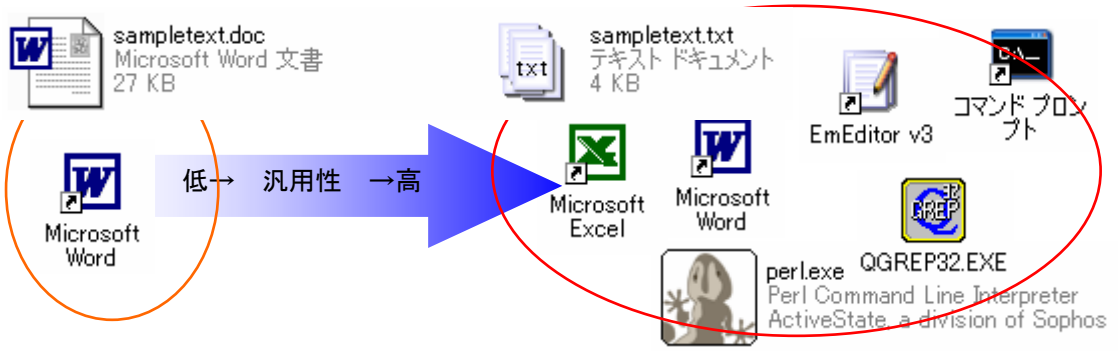
  - Windows では拡張子により起動するアプリケーションを関連付け（associate）ている。
  - 内容は単なるテキストデータであっても，用途によって拡張子が異なる：

ファイルの種類	拡張子	大学 PC で関連付けられたソフトウエア
テキスト	txt	EmEditor
CSV ファイル	csv	Excel
HTML 文書	html, htm	Internet Explorer
Perl スクリプト	pl	Perl
XML 文書	xml	Internet Explorer

2. テキスト形式のファイルを利用するメリット，必要性，注意点

- メリット：利用するツールを選ばない
  - 自作プログラムによる加工を含め，さまざまなソフトウエアやツールで利用可能。
  - テキストデータのみが含まれるので，ファイルサイズが小さい。

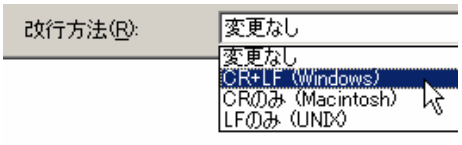
- 必要性：データとツールの分離
  - テキストの入力に安易にワープロを用いないこと。作業内容や手順を考え，用途に合ったツールを利用するとよい。



  - 目で見て分かるだけでは不十分であり，コンピュータにテキストを処理させる工夫も必要。どのような作業をおこなうかを考え，入力方法や形式を選ぶ。
  - ソフトウエアがもつ特定の機能を使う場合，テキスト形式では保存できない。例：
    - Excel のふりがな機能やデータベース機能
    - 画像と文字が対応したデータ：例えば，画像と文字情報が対応した PDF 文書

テキスト形式で保存したほうがよい場合と，必ずしもテキスト形式でなくともよい場合がある。ただし，後者の場合でも，テキストファイルの特性をよく理解してデータを作成しておくことは決して無駄ではない。後になって，テキスト形式で保存しなおす必要が出てくる場合があるからである。

- 注意点：
  - テキスト形式のファイルには基本的に**単一の文字エンコード方式**（後述）しか使えない。多言語を混在させる場合，Unicode や外国語の文字エンコード方式が使えるツールが必要になる。Unicode に対応していないツールを利用する可能性はないかなどを検討すること。
  - 文字の大きさやスタイルなど，**書式情報は一切保存されない**。データで書式情報に頼っているものがないかどうかに注意する。
  - テキストデータでは，タブ tab (ASCII コード 09, Unicode 0009)，スペース space (ASCII コード 20, Unicode 0020)，改行などもコードで表現される。このうち改行記号は CR = carriage return (ASCII コード 0D, Unicode 000D) と LF = line feed (ASCII コード 0A, Unicode 000A) の組み合わせによって表現されるが，基本ソフト (OS) によって表現方法が異なる。



    - Windows = CR + LF
    - Macintosh = CR
    - Unix = LF

実習２： No2 フォルダの hurigana.xls (Excel ブック形式)を開き，Excel で並べ替えを行ってみなさい。同様に，hurigana.csv (CSV 形式) を Excel で開き，同様の並べ替えをおこなって「ふ

りがな機能」の効果を比較しなさい。
実習３： No2 フォルダの hurigana.xls (Excel ブック形式)を開き，Excel で「さ」で始まる名前だけに絞り込むよう「オートフィルタ」を適用し，結果を上書き保存しなさい。同様に，hurigana.csv (CSV 形式) でも同様の絞込みをおこなった結果の保存を試み，Excel ブック形式との違いを確認しなさい。

3. テキストエディタ EmEditor Professional Version 4

- Emurasoft (日本法人はエムソフト) によるシェアウエア (アカデミックライセンスあり)
- Unicode や各言語・地域の文字エンコード方式に対応した多言語テキストの編集が可能
- 高度な検索・置換機能 (Perl に準じた正規表現が利用可能)
- ファイル横断検索 (grep) 機能も搭載
- ファイルの種類により，テキストの表示を色分けし，見やすく編集できる
- 作業内容を細かく記録できるマクロを搭載

※ EmEditor 以外にも Windows 用のテキストエディタは多数存在する： サクラエディタ，秀丸エディタ, jEdit, K2Editor, MIFES, QX エディタ, TeraPad, UniRed, WZ Editor, xyzzy など。配布形態もフリーウエア，シェアウエア，パッケージ製品などさまざまである。Windows XP にも簡易テキストエディタ「メモ帳」(Notepad.exe) が標準で付属する ([スタート] → [プログラム] → [アクセサリ] → [メモ帳] で起動)。

4. EmEditor の基本機能

- 行の折り返し表示の調整 (右図)：どのようにテキストを表示したいかを自分で選ぶことができる 

- さまざまな表示オプション：[ツール]→[現在の設定のプロパティ]
  - 行番号の表示：「基本」タブ
  - 論理行と表示行の切り替え：「基本」タブ
  - 改行記号やスペース，タブの表示：「記号」タブ cf. Word の「編集記号の表示・非表示」ボタン

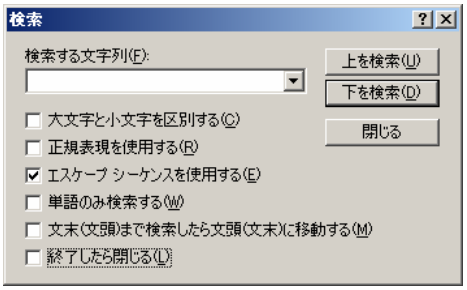
- フォントの分類と設定
  - [表示]→[フォント分類]：言語ごとにフォントを設定 (エディタでは通常 1 種類しかフォントを指定できないので，それぞれの言語に合ったフォントを設定してやる必要がある。各言語・地域の文字エンコード方式とフォントの関係については後日扱う)
  - [表示]→[フォントの設定]：文字の大きさやフォント名，スタイルを表示・印刷を別々に指定する (設定は言語ごと。また，変更内容が随時 [表示] メニューに表示されるので，あとで簡単に戻すことができる。)

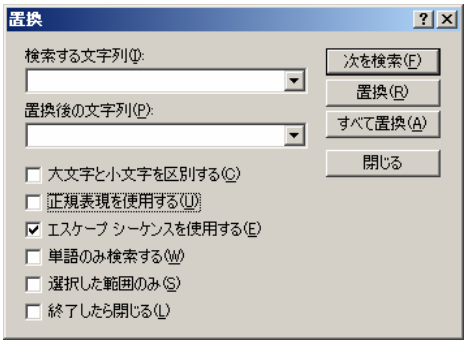
- 検索と置換：後述

- 設定の選択：[ツール]→[現在の設定]で編集モードを変更
  - ファイルの編集内容に合わせてテキストを色分けするほか，リンク機能を付加する。標準的な編集モードは Text。

- ウィンドウ
  - [ウィンドウ] → [すべて結合]：EmEditor で複数のファイルを編集している場合，それらのウィンドウを 1 つにまとめるか，ばらばらに表示するかを指定
  - [ウィンドウ]→[並べて表示][重ねて表示][分割]

5. テキストエディタによる検索 search と置換 replace の基本





- テキストデータでは，同じ文字は同じコードで入力されるので，コードを頼りにテキストを網羅的に検索することができる（右図は EmEditor の検索・置換ダイアログ）。
- Word：通常の検索と「あいまい検索」
- 複雑な検索・置換をおこなう場合には，より機能の充実したテキストエディタや，Perl などのスクリプト言語を利用したほうが効率がよい。
- 検索・置換のオプションを活用することで，複雑な検索・置換を一度におこなうことができる。

  キーワード：　検索オプション，エスケープ・シーケンス，正規表現，あいまい検索
- 置換の用途：
  - 論文など作成原稿の校正
  - 研究用データの整形（正規表現を利用すると高度な整形が可能←第 4，5 回授業で扱う）
- grep 検索：ファイル内，または複数ファイルの横断検索
  - 行単位でテキストを検索し，キーワードが検出された（**マッチした**）行を表示
  - grep（グレップ）= globally search for the regular expression and print the lines containing matches to it
  - EmEditor では，[検索] → [ファイルから検索] で行単位の検索を行い，結果を抽出することができる。
    - ファイル名，行数などの情報が行頭につき，リンクとして機能するので，該当部分を開くことができる（[F10] キーがジャンプキーとして機能する）。

実習 4： No2 フォルダに入っている sampletext.txt (テキスト文書) と sampletext.doc (Word 文書) をそれぞれ EmEditor と Word で開き，検索・置換の実習をしてみよう。2 つのファイルに収録されているテキストは全く同じものである。

4-1. 以下のテキストを検索し，何件検索されたかを数えてみよう。

	sampletext.txt	sampletext.doc
コンピューター		
ＯＳ		

4-2. 以下のテキストが何件含まれるかを数えてみよう。

コンピュータ	sampletext.txt	sampletext.doc
コンピュータ（ただし，「コンピューター」を含まない）		

4-3. 「コンピュータ」のつづりを統一させたい。sampletext.txt を開き，「コンピューター」を「コンピュータ」に置換し，sampletext2.txt として保存しよう。同様に Word でも作業をおこない，sampletext.doc を sampletext2.doc として保存しよう。

4-4. sampletext2.txt は閉じ，再度 sampletext.txt を開きなさい。今度は「コンピュータ」を「コンピューター」で統一したい。置換をおこない，sampletext3.txt として保存しよう。同様に Word でも作業をおこない，sampletext.doc を sampletext3.doc として保存しよう。

4-5. 4-1. ～ 4-4. の作業を通じて，検索・置換の際に気をつけるべきことをまとめ，出席カードに記し，提出しよう。

担当：千葉 庄寿（しょうじゅ） (CHIBA Shoju)　　e-mail: schiba@reitaku-u.ac.jp