情報技術演習
第4回 「情報抽出と自然言語処理」

2006/10/24

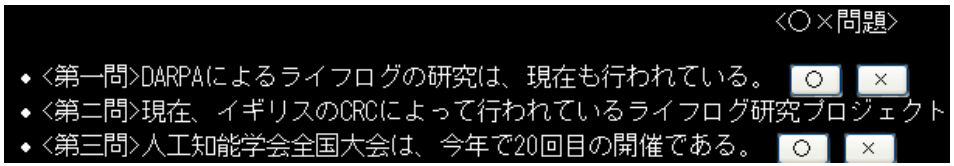
久保田秀和
文学部／情報学研究科
kubota@ii.ist.i.kyoto-u.ac.jp
http://www.ii.ist.i.kyoto-u.ac.jp/~kubota/

本日の講義・演習

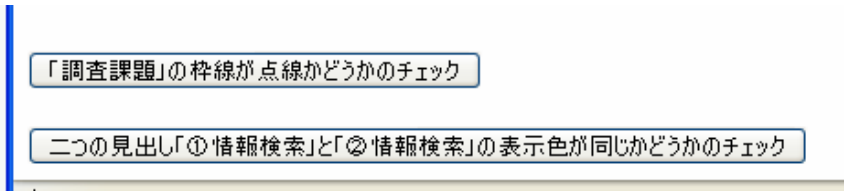
- プログラミングの基礎（復習）
  - 前回提出されたレポートを題材に
  - 計算機上の身近な情報へのアプローチ（CGUI）
- 情報抽出と自然言語処理
  - CUIを介して形態素解析システムを利用
    - MeCab（めかぶ）
  - 演習課題
    - 各自のレポートを形態素解析する
    - 得られた形態素の重要度，意味について議論する

前回の復習

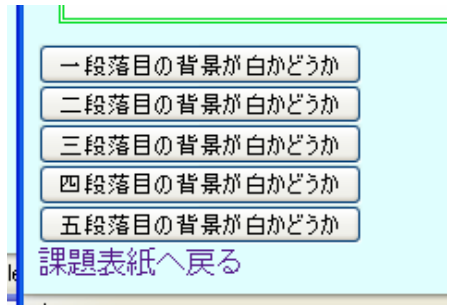
- 知識としての半構造化文書
  - 身近なところに大量に存在している
    - HTML
    - XML，XHTML
  - そこそこ計算機可読である
- 前回のレポートより



クイズの知識源としての利用 [YM 2006]



スタイルを取り出して自動確認 [TS 2006]



色を取り出して，人間では
難しい判定を代行 [KM 2006]

Why HTML?

- プログラミングの講義で，計算させたりではなくHTMLの中でやるので驚いた，というコメントがあったので，背景を詳しく説明

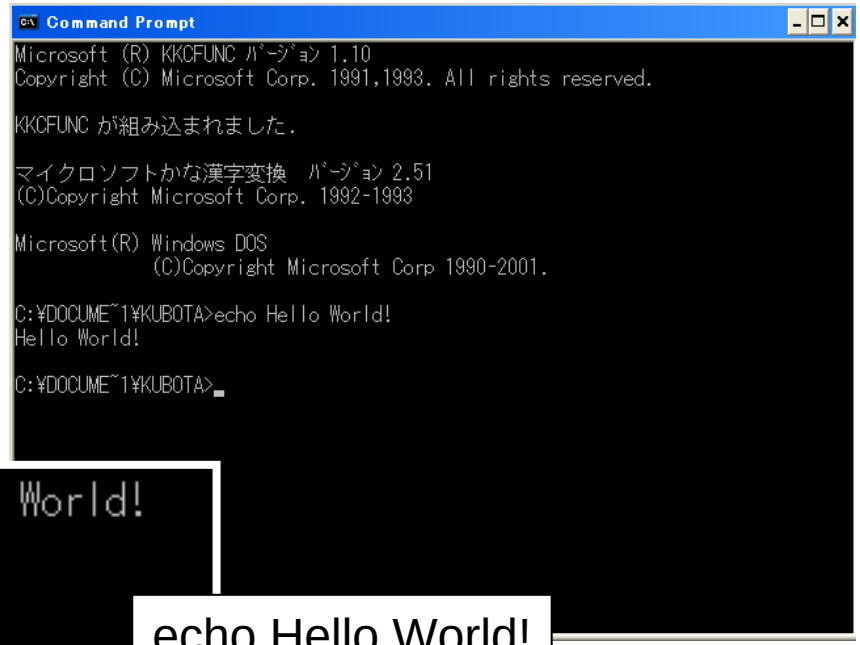
⇒半構造化文書＋JavaScriptというアプローチによって，大量かつ身近な情報を対象としたプログラミングが可能となったことに，演習を通じて触れてほしかった

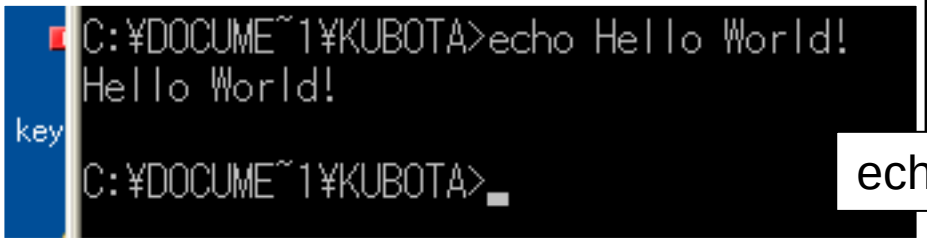
- 歴史・・・計算機上の身近な情報へのアプローチという観点から
  - CUI環境向けのスクリプティング（1970年代～）
    - 文字情報の処理
  - GUI環境向けのスクリプティング（1990年代～）
    - 文字情報の処理＋GUIの作成
  - CGUI（2000年代～）
    - 文字情報の処理＋GUIの作成＋GUIの頒布

練習4-1: CUI環境向けのスクリプティング

- CUI(Character User Interface)環境
- 文字情報の処理
- 例
  - シェルスクリプト
    - UNIX系OS
    - Bourne Shell（1977）, bash（1987）などのシェル上のスクリプト
  - バッチファイル
    - MS-DOS（PC-DOS）
    - COMMAND.COM上のスクリプト（1980年代～）

command.com





echo Hello World!

GUI環境向けのスクリプティング

- GUI (Graphical User Interface) 環境
- 文字情報の処理＋GUIの作成
  - 簡単なダイアログ，ドラッグアンドドロップアプリの作成
- 例
  - Open Scripting Architecture + AppleScript 等（1993～）
  - ActiveXスクリプティング + WSH + VBScript/JScript（JavaScriptと互換性有）等（1996～）
    - JScriptを使って，Windows上のファイルやアプリケーションを操作可能
    - （コラム）JavaScript, JscriptはECMA（European Computer Manufacturer Association）による標準化へ⇒ECMAScript

CGUI（1）

- CGUI [NRI 2006] http://www.nri.co.jp/news/2006/060518.html
  - Consumer Generated User Interface
  - 第1回で紹介したCGM(Consumer Generated Media) をもじったもの．
    CGMと同様，比較的サイズが小さく，大量，多様の性質を持つと考えられる．
- CGUI環境（HTML/XML文書＋JavaScript）の例
  - デスクトップ
    - Konfabulator (2003) ⇒Yahoo! Widgets(2005)
    - Dashboard (Apple,2005), Opera Widgets(2006)
  - デスクトップ＆ポータルページ
    - Gadgets (Microsoft, 2005), Google Gadget (2006)
  - ブラウザ，アプリケーション
    - XUL（XML User Interface Language）（Mozilla 1.0～，2002）
    - WPF/E（Windows Presentation Foundation Everywhere）＋XAML（eXtensible Application Markup Language）（Windows Vistaに搭載予定）

Dashboard

CGUI（2）

- 文字情報の処理＋GUIの作成＋GUIの頒布
  - 処理対象の変化
    - そこそこ計算機可読であると言える半構造化文書（HTML，XHTML，RSS文書等）が，Web上で大量に流通
  - 道具の変化
    - 半構造化文書の編集やプログラミングのためのツールが揃ってきた
  - サービスの変化
    - 専門的なサービスが部品化されて，WebAPIとして公開（Google, Yahoo, はてな）
  - Ajax（Asynchronous JavaScript + XML）
    - JavaScript の潜在的な価値の発見
    - Google Maps, Google Suggest

本日の講義・演習

- プログラミングの基礎（復習）
  - 前回提出されたレポートを題材に
  - 計算機上の身近な情報へのアプローチ（CGUI）
- 情報抽出と自然言語処理
  - CUIを介して形態素解析システムを利用
    - MeCab（めかぶ）
  - 演習課題
    - 各自のレポートを形態素解析する
    - 得られた形態素の重要度，意味について議論する

情報抽出と自然言語処理（のさわり）

- 情報抽出
  - 文書群から必要な情報を取り出す
  - 自然言語処理はその主要な技術
- 自然言語処理（Natural Language Processing）
  - これまでの演習では，半構造化文書，つまり部分的に形式化され，そこそこ計算機可読な文書を対象としてきた
  - 人が自然に習得し使用する言語（自然言語）をそのまま計算機が理解できるようにするための技術
  - 情報抽出以外にも，人間と関わるあらゆる情報処理の場面において重要

形態素解析

- 自然言語の階層
  - 音素（phoneme）
  - 形態素（morpheme）
    - 意味を持つ最小の言語単位．一つ以上の音素から成る．
  - 語（word）
    - 一つの意味のまとまりをなし，文法上一つの機能を持つ最小の言語単位．一つ以上の形態素から成る．
  - 文節（phrase）
    - 文を不自然でない程度に区切った，最小の言語単位
  - 文（sentence）
  - 文章（text）
- 形態素解析の目標，性質は言語依存
  - 階層，形態素や語の定義は言語の種類によって異なる
  - 例えば分かち書きの有無に起因する違い

形態素解析の道具

- JUMAN（http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html）
- ChaSen（http://chasen.naist.jp/hiki/ChaSen/）
- MeCab（http://mecab.sourceforge.jp/）

- 本日の演習ではMeCabを使用
  - Binary package for MS-Windows をインストール
  - インストールするディレクトリは
    M:¥MeCab¥
    を指定

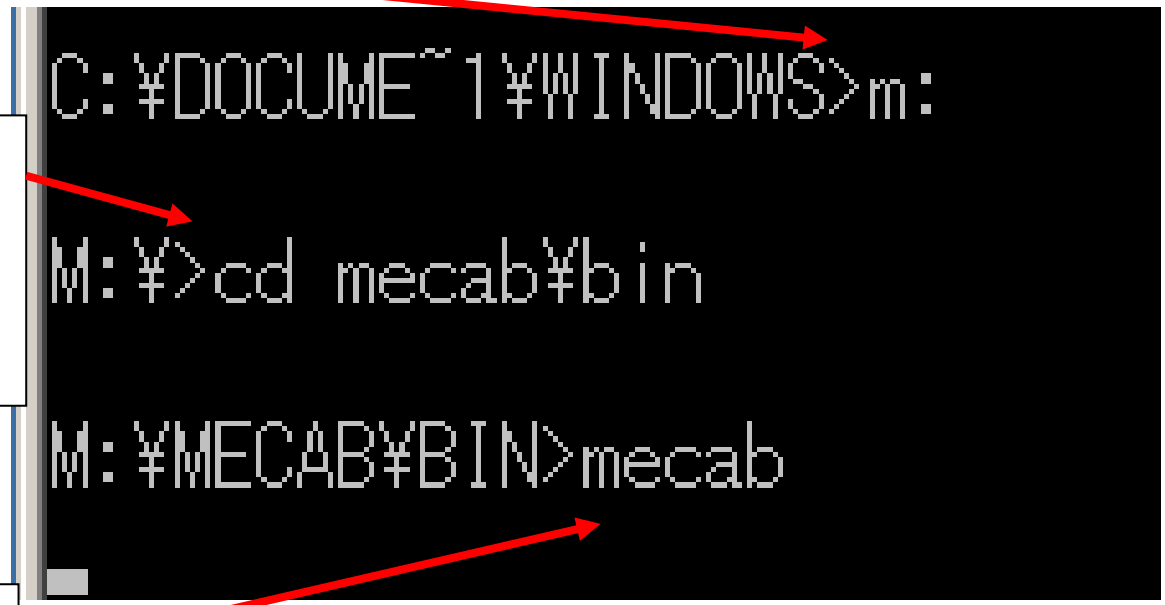
MeCabの利用

- command.com を起動
- mecabを実行

ドライブ m: へ移動

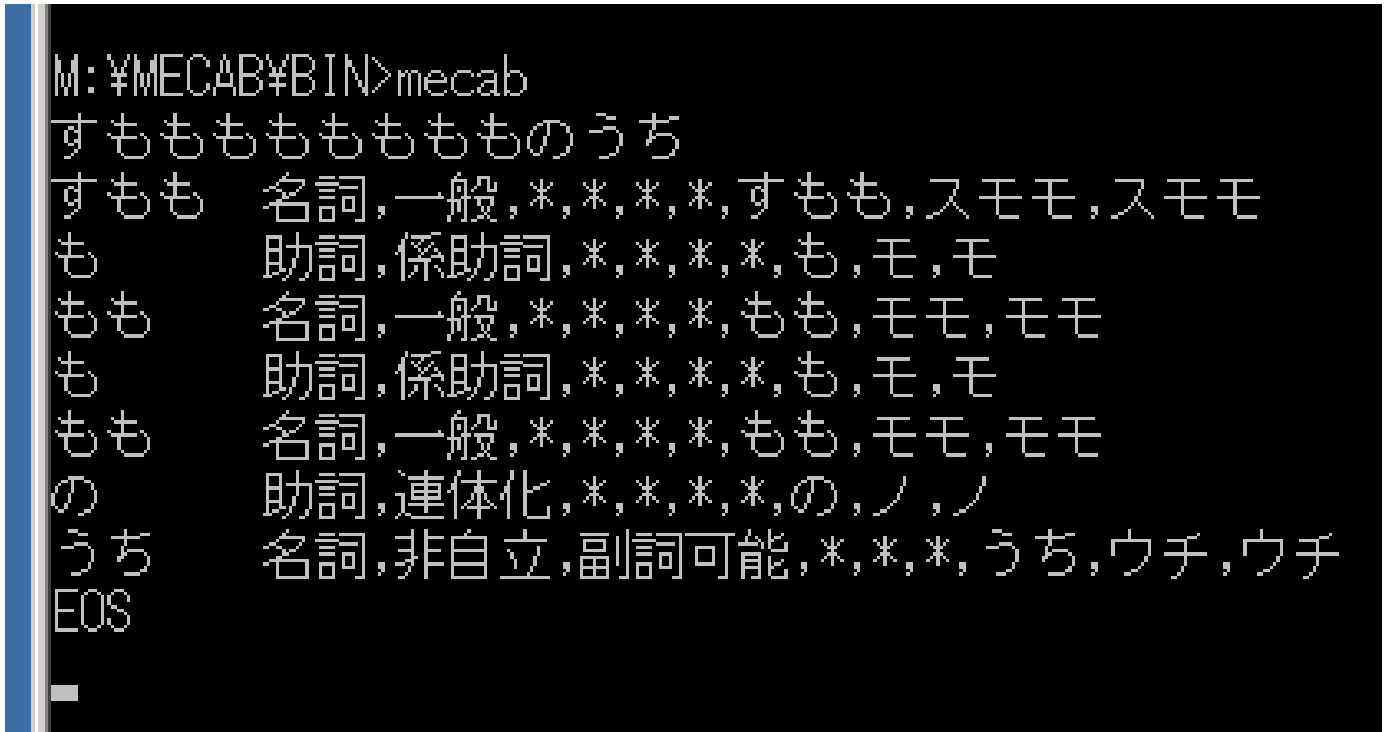
cd (change directory)
mecab¥bin ディレクトリへ移動

mecab を実行



練習4-2: MeCabの利用（標準入出力）

- 日本語入力モード（ALT＋半角／全角）
- 例文
  - すももももももものうち
  - 最後にEnter



MeCabの解析結果

解析例（http://mecab.sourceforge.jp/より）

すももももももものうち

すもも 名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
うち 名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ

EOS

表層形¥t品詞,品詞細分類1,品詞細分類2,品詞細分類3,活用形,活用型,原形,読み,発音

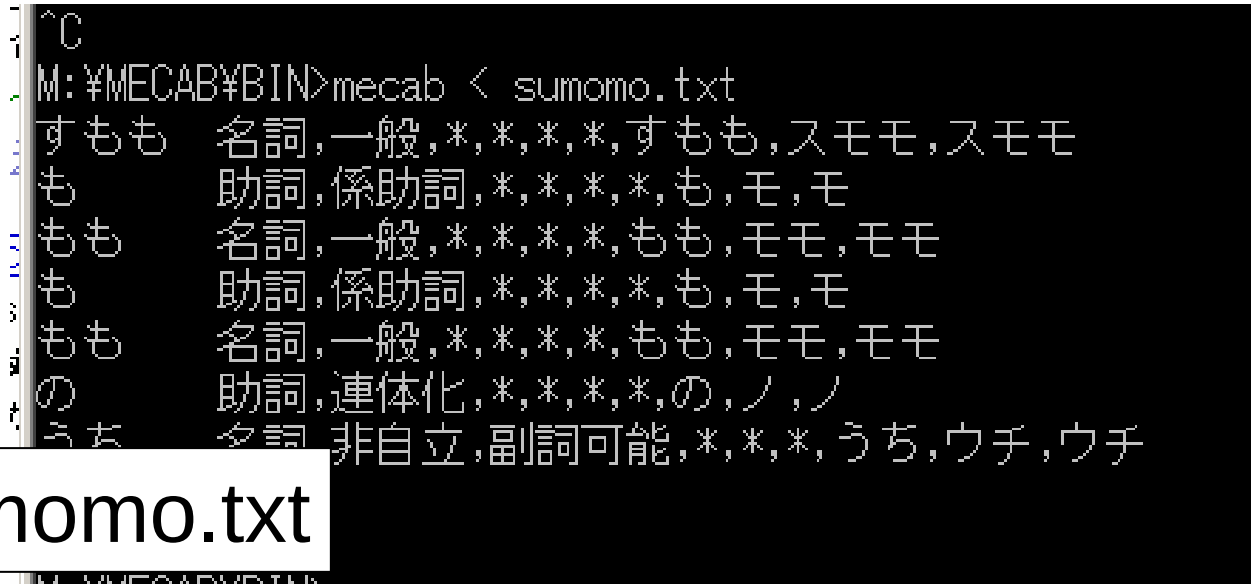
出力フォーマット

標準入出力・リダイレクト

- 標準入力
  - 通常はキーボードからの入力
- 標準出力
  - 通常はディスプレイへの出力
- リダイレクト
  - 標準入出力を別のファイルへ切り替える

練習4-3：MeCabの利用（リダイレクト）

- Ctrl+C でMeCabをいったん終了
- 標準入力のリダイレクト
  - sumomo.txt をダウンロードし，M:¥MeCab¥bin¥（マイドキュメント¥MeCab¥bin¥）の下に置く
  - sumomo.txt を入力とする



mecab < sumomo.txt

練習4-4：MeCabの利用（リダイレクトその2）

- 標準入出力のリダイレクト
  - sumomo.txt を入力とし，result.txt へ出力する

```
M:¥MECAB¥BIN>
M:¥MECAB¥BIN>mecab < sumomo.txt > result.txt

M:¥MECAB¥BIN>
```

mecab < sumomo.txt > result.txt

  - result.txt を開いてみましょう

第４回課題

- これまで作成した各自の調査課題レポートについて，形態素解析システムを用いて形態素解析を行う.
- 提出するレポートは以下の（A）（B）のうちいずれか1つ

  （A）レポート中の，重要と思われる形態素（あるいは語，文節，文）を判断し，手動で抽出する．判断に際しては形態素解析の結果を活用し，なぜ重要と判断できるかを述べること.

  （B）解析した形態素を「品詞」「品詞細分類」別にまとめる．　各「品詞」「品詞細分類」については，その定義を調査し，まとめる.

課題（続き）

- 期限は10月28日（土）17:00
- 提出方法
  - メールの添付ファイルで提出
  - ファイルが複数ある場合はlzh/zip/tgzのいずれかの形式でアーカイブし，電子メールに添付して提出
  - あて先は久保田　kubota@ii.ist.i.kyoto-u.ac.jp

参考図書

- 自然言語処理
  - 長尾真編「自然言語処理」（岩波講座ソフトウェア科学15）
  - 田中穂積監修「自然言語処理―基礎と応用―」（電子情報通信学会）

  など