テキストデータ

おしながき

- なぜテキストデータを自分で扱うのか
- テキストデータとは
- テキストエディタ
- 文字コード
- 検索と置換

なぜテキストデータを自分で扱うのか

なぜテキストデータを自分で扱うのか

- コーパスをそのまま検索できるサイトも増えてきたので、自分でデータを操作する必要性は減っている
- しかし既製のサイトを使うのは (i) 提供された機能しか使えない (ii) 内部の動作がブラックボックスである、といった問題がある

(i) 提供された機能しか使えない

- 少し複雑な要求になると、コーパスや検索ツールが応えられないかもしれない。例えば
  - as … as という表現（... の部分は未知）を検索したい
  - *The X-er, the Y-er* という構文を検索したい
  - 逆接の「が」だけを検索したい

(ii) 内部の動作がブラックボックスである

- 内部の動作が不明であり、本当に期待通り動作しているのかどうか検証できない。さらに言えば、内部がどのような仕組みになっているのか考える機会がないと、検索においてそもそもどのような問題が生じうるのかについて直感を磨くことができない。
- サイトが突然仕様変更、終了になるかもしれない。そのため、研究の再現性などが保証できない。

なぜテキストデータを自分で扱うのか

- 元のデータを直接目で確かめることができる
- サービス終了などの心配がなく、調査をいつでも、誰でも再現することが可能になる
- 研究の目的・技術力に応じて、与えられた機能を使うにとどまらず、複雑な操作が可能になる
  - 究極的には、プログラミング等を本格的に勉強すれば、コンピュータで可能な任意の機能を自分で作ることができる（この授業ではそこまでやりません）

テキストデータとは

テキストデータ

- 文字データを扱ううえでもっとも基本的な形式。多くのコーパスがこの形式で提供されている。
- **テキストファイル** - 文字のみ（空白、改行などの「特殊文字」を含む）から成るファイル。
- **バイナリファイル** - それ以外

テキストファイル

- 「テキストファイル」は .txt という拡張子のファイルを指すこともあるが、実際には他にもテキストデータは色々ある。
  - .html（ウェブページ）
  - .css（ウェブページのスタイル）
  - .csv（カンマ区切りの表形式のデータ）
  - .xml（タグ付きコーパス等にもよく使われる）
  - テキストでないもの
  - .doc, .xls, .ppt
  - .zip
  - .jpg, .png, ..
  - .mp3, .mp4, ..
  - etc.

テキストエディタ

文書作成ソフトは
コーパス管理に適さない

- Microsoft Word 等の文書作成ソフトは印刷物等を作成するためのものであり、テキストデータの読み書きは不可能ではないが、目的に合わない
  - 機能のほとんどが不必要（か有害）
  - 可搬性、永続性がない（特定のソフトでしか扱えない）
  - 巨大なデータに向かない

テキストエディタ

- テキストファイルの閲覧・編集に特化したソフトウェア
  - Windows についてくる「メモ帳」もそう
  - しかし、研究目的にはより高機能なエディタを用いたほうがよい
    - 文字コードの自動判定・変換、検索結果ハイライト、正規表現、多ファイル検索、マクロ…

Windows 用のテキストエディタ

- 秀丸エディタ (シェアウェア; **苦**学生は無料)
- EmEditor (シェアウェア)
- TeraPad (フリーウェア)
- サクラエディタ (フリーウェア)

Mac 用のテキストエディタ

- BBEdit
- TextWrangler
- Smultron
- mi
- CotEditor

文字コード

文字コードの前に
電子データとは

- ビット
- バイト

文字コード

- ASCII
- Latin-1
- JIS (ISO-2022-JP), Shift JIS, EUC-JP, ..,
- Unicode: UTF-8, ..,
- **文字化け**は、そのテキストに使われている文字コードと、表示するソフトウェアが想定する文字コードが一致しない場合に生じる

文字コードの変換

- 利用したいプログラムが特定の文字コードにしか対応していないような場合、手作業で文字コードを変換する必要が生じることがある
- 高機能なテキストエディタには、指定した文字コードでファイルを開く/保存する機能がついていることが多い。また、文字コードを自動的に推測する機能もある（ただし失敗することも）
- 大量のファイルの文字コードをまとめて変換する時などは、この方法は向かない。そのような場合は、文字コード変換用のプログラムを用いる。

改行

- 改行は、コンピュータにとっては単なる文字のひとつ。人間に見せるときに、（何かの文字を出す代わりに）次の行に送っているだけ
- 改行はリターンキー（エンターキー）を押したときに入力される。画面の横幅の都合上、たまたま次の行に行くのは改行ではないので注意

改行コード

- 歴史的な事情で、システムにより改行コードが異なる (CR: キャリッジリターン、LF: ラインフィード)
  - Windows: CR+LF
  - Mac: 伝統的に CR, 現在は LF も
  - UNIX系: LF
- 例えば Mac 版 Office でテキストファイルを保存すると改行は CR となる。これを Windows メモ帳で開くと改行として表示されない

エディタによる検索・置換、Grep検索（多ファイル検索）

エディタによる検索・置換、Grep検索

- 青空文庫からサンプルを
  - http://www.aozora.gr.jp/cards/000035/card2282.html

コメントシート

- 読んでみたい論文について書いてください。
  - 具体的に論文を挙げられない場合は、取り上げたいトピックを挙げてください。
- 来週、３人程度を選んで具体的に論文を決め、再来週に発表してもらおうと思います。
- そのほか、今日の内容について質問・コメントなどあれば何でもお願いします。