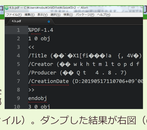
情報学概論2 第5回（5/22）

(Introduction to Informatics 2)

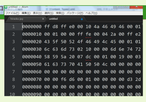
ファイル 種類 拡張子（２）

1 PDFファイル

PDFの中

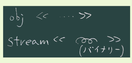


JPGの中



(前回のおさらいと補足)

- PDFは、テキストファイルとして表示させると右図（の左）のようになっただろう。
  - ちなみに、JPEG画像ファイル等は文字として表示できる部分は殆どない（純粋なバイナリーファイル）。ダンプした結果が右図（の右）。



- PDFでは文字による言語で構造が作られていることが想像できるだろう（右図）。
  - （画像などの）バイナリーデータはstream命令（のようなもの）で流し込む形になっている（その部分は文字としては表示されない）ことも読み取れただろうか。

e . PDFでページ検索・コピペ

（文字としてのデータの扱われ方）

実験：

- 前回いくつかの方法で作成したPDFファイル（印刷機能で作ったものも含めて）について、
- いずれかのビューワで表示させた状態で、
  1. ページ内検索
  2. 画面上の文字をコピーしてどこかに（たとえば□TyporaやAtomなどのエディタの新規ファイル画面）に貼り付けて結果を確認する

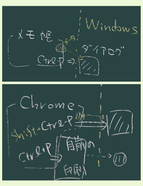
  といった操作を試みてください。

（前回資料のページをPDF化したものを以下にいくつか置いておきますのでこれで試してもらってもOKです。）

1. Windowsの印刷機能（Shift+Ctrl+P からプリンタとして pdf writerを選ぶ）
2. pandoc + wkhtmltopdf
3. Typora（+ 内部でpandoc）
4. フルスクリーンキャプチャ画像からPDF化

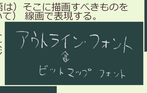
解説と補足：

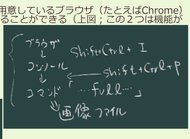
- 右図のようなファイルサイズ（大雑把にKB単位）になっており、生成方法による差があることも確認しておいて下さい。

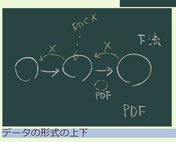


メモ帳<->ブラウザ

- Webページをブラウザからプリントする際、印刷機能を自前で用意しているブラウザ（たとえばChrome）の場合、自前の印刷機能とWindows組込の印刷機能を使い分けることができる（上図）。この２つは機能がかなり違う）。
- 上記4.のフルスクリーンキャプチャの機能は広く知られている訳ではない（比較データ生成のために使用しただけで、今回は参考のため紹介したが使う必要も覚える必要もないだろう）。
- PDFは（或いは PostScript言語は）そこに描画すべきものを（画像データを扱う部分を除いて）線画で表現する。
  - アウトラインフォントの表現をページに拡大したもの、をイメージするといいだろう。







データの形式の上下

- **データの変換は一方通行であることを知っておこう**（上図）
  - 印刷形式からもとの構造データを復元、画像からテキストへの変換、などは、　（必要な情報が失われているため）技術的に困難。
  - データの流れの上流・下流を意識したとき、PDFは下流寄りのデータ形式ということになる。

f . 拡張子が違ったら（または、ついていなければ）

- これまでの知識があれば推測可能
- アプリケーションでの推測
  - ポピュラーなのは WinExChange。
- 推測してくれるサイト
  - たとえばここ ファイル形式診断ツール（右図）
  - こういったサイトの信頼性は不明なので、重要なデータには使用しない方がいい

- 拡張子のつけはずし の方法
- 関連づけされたアプリケーション、とそれ以外のアプリケーション（について おさらい）