

Chapter2

Formatting Data

2012/03/27

johtani

Formmating Data

- ◆ データフォーマットを変換するツールは解析やVisualizationを行うのに重要
- ◆ データフォーマットを扱うツールやデータフォーマット自体を知っておいたほうがいろいろ役に立つ
- ◆ 以降ではデータフォーマット、ツール、プログラミングについて紹介する

コラム: What I Learned about Formatting

- ◆ 高校の時はフォーマット済みデータを与えられてた。学部生時代も同様
- ◆ 大学院の時に、実データは整形されてないと知る
- ◆ 今ではデータの整形と視覚化を一緒に行うのが普通
- ◆ データをきちんとまとめると視覚化がより簡単におこなえる

Data Formats

- ◆ 構造化されたデータが重要。
- ◆ ここで紹介されているデータフォーマットで大体カバーできる
 - ◆ CSV、TSVのようなDelimited Text
 - ◆ JSON (JavaScript Object Notation)
 - ◆ XML (Extensible Markup Language)

Data Formats (Delimited Text)

- ◆ カンマ、タブで区切られたテキストファイル
- ◆ 列を区切り文字(カンマ、タブなど)で区切ったファイル
- ◆ ExcelやGoogleDocsなどの表計算プログラムで利用可能
- ◆ データを共有するのに有効(特定のプログラムに依存してないから)

Data Formats (JSON)

- ◆ WebAPIでよく利用されるデータ形式
- ◆ JavaScriptの表記法
- ◆ キー・バリューのペアで構成
- ◆ オブジェクトを扱うことも可能
- ◆ 参考:
JSONの仕様については以下を参照
<http://json.org>

Data Formats (XML)

- ◆ APIでデータ転送に利用されるデータ形式
- ◆ タグで囲まれた値を持つテキストドキュメント
- ◆ 例:RSS(Really Simple Syndication)

Formatting Tools

- ◆ データを簡単に扱ったりフォーマットするためのツール
データ量も増えてきたが、ツールも増えてきてる。
 - ◆ Google Refine
 - ◆ Mr.Data Converter
 - ◆ Mr. People
 - ◆ Spreadsheet Software

Formatting Tools (Google Refine)

- ◆ Freebase社のGridworksの進化版(Freebase社をGoogleが買収して提供)
- ◆ ローカル環境にて実行可能(=プライベートなデータをアップロードしないでいい)
- ◆ オープンソース(<http://code.google.com/p/google-refine>)
- ◆ データの
- ◆ 参考URL:
Freebase
<http://www.freebase.com/>
使い方
http://wiki.kazusa.or.jp/Google_Refineの使い方

Formatting Tools (Mr. Data Converter)

- ◆ CSV、TSVデータを他のフォーマットに変換するオンラインツール
- ◆ githubでOSSとして公開 (HTML+CSS+JavaScript)
- ◆ 様々な形式に対応
ActionScript、ASP/VBScript、HTML、JSON、MySQL、PHP、
PythonのDict、Ruby、XMLなど
- ◆ 参考URL
オンライン
http://shancarter.com/data_converter/
ソース
<https://github.com/shancarter/Mr-Data-Converter>

Formatting Tools (Mr. People)

- ◆ 人名リストをパースするためのオンラインツール
- ◆ githubでOSSとして公開 (Ruby製パーサー部分)
Lingua-EN-NameParserというPerlモジュールが元
- ◆ 参考URL
オンライン
<http://people.ericson.net/>
ソース
<https://github.com/mericson/people>

Formatting Tools (Spreadsheet Software)

- ◆ 小さなデータ集合に対してなら表計算ソフトでOK
- ◆ 簡単なソートやちょっとした変更など
- ◆ 大きなデータについて

Formatting with Code

- ◆ 巨大なデータを扱う場合はこれまで紹介したソフトだと不向き
遅かったり、クラッシュしたり
- ◆ 以降ではあるフォーマットを別のフォーマットに変換する例を紹介
- ◆ 例ではPythonを利用

Example: Switch Between Data Formats

- ◆ 巨大なデータを扱う場合はこれまで紹介したソフトだと不向き
遅かったり、クラッシュしたり
- ◆ 以降では2つのフォーマットを変換するサンプルを紹介
- ◆ サンプルではPythonを利用
- ◆ 最初はCSVからXMLへの変換
- ◆ 次はXMLからCSVへの変換(XMLパースにBeautiful Soupを利用)
- ◆ 最後はCSVからJSONへの変換

Example: Switch Between Data Formats

- ◆ 変換処理の2つの大事な事柄
 - ◆ データの読み込み
 - ◆ 各行に対して変更処理を繰り返し適用
- ◆ 論点をわかりやすくするために、似たようなコードをサンプルとしている

Put Logic in the Loop

- ◆ ループの中で処理を行うことで、いろいろなことが可能
- ◆ しきい値のチェックや移動平均、前日との気温差の計算なども可能。
- ◆ サンプルとして、氷点下かどうかの列を追加したコードのサンプルを記載
- ◆ 重要なのは以下の3点
 - ◆ load : 読み込み
 - ◆ loop : ループ
 - ◆ process : 処理

Chapter2のまとめ

- ◆ データを取得する方法(どこから、どのように)を学びました
- ◆ データを整形したり、変換したりする方法を学びました。
- ◆ プログラミングによるデータの処理も学びました。
- ◆ 簡単に利用できるツールについても学びました。