Semi-Structured Data: JSON

School of Information Studies
Syracuse University

JSON

JavaScript Object Notation

Data interchange format

“Lightweight” format

- Data representations
- Easy for users to read
- Easy for parsers to translate

Main Structures

Object

- Unordered set of name/value pairs
- Uses outer {}
- Members separated by commas
- Each member—string

Array

- Ordered collection of values
- Uses outer []
- Values separated by commas

Value

- Object, array, string, number, true or false, null
- String—any Unicode character

Simple JSON Sample

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

Using JSON Objects

Twitter returns a JSON object

Use Python ‘json’ package

Converts to internal data structures

- Lists
- Dictionaries
- Can convert them back to strings

JSON Functions

Json.loads (jsonstring)

- Parses the JSON string

Json.dumps (python_object, sort_keys = True, indent=4)

- Does a ‘pretty print’
- Saves JSON data in a file

Unicode and Python

School of Information Studies
Syracuse University

Unicode

Industry standard

Defined by Unicode Consortium

Smallest component of text

Language Characters

Standard

- A, B, C, D
- ℼ, µ

Special

😠

♂

Definitions in Unicode

Code Point

- Integer value—base 16
- Assigned a standard name
- No implementation, fonts
- Represented by graphical elements—GLYPH
- Represented with U+0061—decimal number 97

Some Unicode Samples

U+0061	‘a’;	Latin small letter A
U+0394	‘Δ’;	Greek Capital Letter Delta
U+007B	‘{‘;	Left Curly Bracket

Unicode Code Points

Over a million code points

- From 0 to 10FFFF (largest hexadecimal number)
- Defined in layers

Character encoding

- Used to map to binary numbers
- ASCII—English characters
  - 7 bits
- Latin-1—additional characters for Western European languages
  - 8 bits

Unicode Code Points

UTF-8

- Most widely used
- Sequence of 8-bit bytes
  - Code point < 128—represented by byte value
  - Code point >=128—sequence of 2, 3, or 4 bytes

Unicode in Python

Every string in Unicode is using UTF-8

Problem is I/O

- Python interpreter to terminal output
- Python print function to terminal output
- Files in different OS
- Databases like MongoDB, Microsoft Word, browsers

Unicode in Python Interpreter

```
>>> 15                      # the decimal number 15
15
>>>0xFF                     # hexadecimal numbers
255
>>>0x7F
127
>>>’\u0394’                 # using 4 hex digits
Δ
>>>’\U00000394’             # using 8 hex digits
Δ
>>>’\N[GREEK CAPITAL LETTER DELTA]’
Δ
```

Unicode Functions

bytes.decode ()

- Converts from bytes to Unicode strings

str.encode ()

- Converts from strings to bytes
- To output text to different devices

NoSQL Databases

School of Information Studies
Syracuse University

NoSQL Database

Does not use SQL

No database schema

Data stored via Dictionaries

Storing Data

Give the data a name

Serves as the **Key**

Access can be very fast

Data accessed by aggregate units

Can be spread across distributed storage

What We Will Use

Document database

Follows JSON structures

Container for a set of collections

- Each collection holds a number of documents
- Each document is an object of name/value pairs

Comparison of NoSQL and RDBMS Database

NoSQL	**RDBMS**
NoSQL	SQL
Database	Database
Collection	Table
Document	Tuple/row
Field	Column
Embedded document	Table join
Primary key	Primary key

MongoDB

One of the most widely used

Excels in storing huge amounts of data

Spreading data across storage devices

Using parallel constructs

Format is called BSON

Binary encoded JSON with additional types

Installing MongoDB

Install using Anaconda

See the directions found in the toolbox for installation