University of Essex

JSON and XML

Akitaka Matsuo
Essex IADS

JSON

University of Essex

- JSON = **J**ava**S**cript **O**bject **N**otation
- Very popular format for data exchange in the net
- Lightweight, easy to read, less formatted
- Written with JavaScript object notation, but independent from any language
- Used in many APIs including:
  - Twitter
  - Facebook
  - YouTube

A simple JSON

University of Essex

Example 1: Two Name-Key values

```
{"firstName":"John", "lastName":"Doe"}
```

Example 2: Array nested in an object

```
{"employees":[
    { "firstName":"John", "lastName":"Doe" },
    { "firstName":"Anna", "lastName":"Smith" },
    { "firstName":"Peter", "lastName":"Jones" }
]}
```

JSON and Python

University of Essex

- As you saw, the format is very similar to nested dictionary/list
- You can read json data using json package
  - What it does is essentially converting json into a nested dictionary/list
  - So we can navigate it through the standard python syntax
- See demo

Other data format: XML

University of Essex

- **XML** = e**X**tensible **M**arkup **L**anguage
- XML is used for distributing data over the Internet.
  - Examples:
    - RSS (web feeds): http://onlinelibrary.wiley.com/rss/journal/10.1111/(ISSN)1540-5907
    - SVG (graphic): https://upload.wikimedia.org/wikipedia/commons/b/be/BlankMap-LondonBoroughs.svg
    - epub (books)
    - Office documents (OpenOffice, MS)
- XML looks a lot like HTML, but more flexible (e.g. basically no preset definitions of tags).

XML, Example 1 (no schema)

University of Essex

```
<?xml version="1.0" encoding="UTF-8"?>
<notes>
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
<note>
  <to>Jason</to>
  <from>Kelly</from>
  <heading>Offer</heading>
  <body>You won 10M. Contact us immediately.</body>
</note>
</notes>
```

- This file contains two notes, seems to have common structure for notes but you never know!

XML, Example 2 (with DTD)

University of Essex

```
<?xml version="1.0"?>
<!DOCTYPE note [
<!ELEMENT note (to,from,heading,body)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT body (#PCDATA)>
]>
<note>
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend</body>
</note>
```

- This XML has a DTD (Document Type Definition)
- DTD is one of the XML schematic languages that are used as a validator of data input

XML and Python

University of Essex

- The parliamentary members API provides the access in both XML and JSON
- For XML, what we can do is to use xml parser
  - The coding is similar to BeautifulSoup
- We will see an example in demo

XPath

University of Essex

- **XPath**: a syntax for defining parts of an XML document
  - Can be used to navigate through elements and attributes in an XML document.
  - Uses path expressions to navigate in XML document
  - Reference: https://www.w3schools.com/xml/xml_xpath.asp

- **For web-scraping**
  - Last week, we have seen CSS selector
  - XPath can be used as another type of selector
  - Probably a bit more coding involved, but more powerful
  - More useful for parsing xml than html
    - xml files are better formatted (while having less attributes)