Semi-Structured Data

School of Information Studies
Syracuse University

Semi-Structured Data

Well-formatted in a tag system

Passing data over the Internet

- HTML—describing data and appearance of Web pages
- XML—hierarchical tag system
- JSON—lighter-weight tag system

Features of Markup Languages

Tree-structured (hierarchical) format

Elements surrounded by opening and closing tags

Attributes embedded in open tags

<tag-name attr-name=“attribute”> data </tag-name>

Basics of Web Scraping

To handle HTML and XML data

JSON data

- Social media of Twitter and Facebook
- Unicode issues
- Storing data in NoSQL database MongoDB

Obtaining Data

JSON from APIs—more structured

HTML—more difficult; use as last resort

Using Python libraries for HTML and XML

Selenium—more advanced Web scraping

Advanced option—using bots