

COM5507 Social Media Data Acquisition and Processing

Week 6. Web Scraping – Episode 2

Lectured by: Dr. Xinzhi ZHANG

Research Assistant Professor, Department of Journalism

Hong Kong Baptist University

10 Oct 2018 @ CityU M4003

Agenda

- Data sources – revisited and elaborated
- Web scraping: more instances
 - Structured data from public source
 - File downloading from public source
- Developing a (quasi-) web crawler

Data sources

- Data sources (in data-driven journalism, computational social sciences, and business analytics) can be roughly mapped in six dimensions.
 - Open data versus non-open data
 - Structured data versus non-structured data
 - “Ready-made” versus “custom-made”
 - Legal data versus controversial data
 - Obtrusive versus non-obtrusive
 - Empirical versus simulated

Data sources: “open”?

- Open data
 - data published by ruling authorities, i.e., governmental officials, education sectors
 - academic institutions, NGOs
 - any other datasets published as public domains
- “Non-open” data
 - commercial data or solutions
 - data acquired by specific terms of use (i.e., requires payment and logged in)

Examples of open data

- World Bank
 - Data Bank (<https://datacatalog.worldbank.org/>)
- Government institutions
 - HKO (https://www.hko.gov.hk/cis/normal_c.htm)
 - Policy Address
(<https://www.policyaddress.gov.hk/2018/chi/index.html>)
- Open data projects by HKU
 - http://www.ssrc.hku.hk/open_data.php
 - <http://www.ssrc.hku.hk/language.php>
- Civic Exchange Report
 - <http://civic-exchange.org/annual-reports/>

Data sources: “ready-made”?

- Ready-made data
 - Stored in a data file and ready for request
- Custom-made data
 - The data is determined by the researchers or storytellers

“Ready-made” published datasets

- Word Values Survey
(<http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>)
- International Social Survey Programmes
(<http://w.issp.org/menu-top/home/>)
- PEW Research Center
(<http://www.journalism.org/datasets/>)
- World Bank
(<http://databank.worldbank.org/data/source/world-development-indicators#>)



Table

Chart

Map

Metadata

Download options ▾

Excel
CSV
Tabbed TXT

Data on this page only - formatted

Metadata

Advanced options

Time (0)

variables from each of the following dimensions to view a
an select from left panel or by clicking the links above.

Apply Changes

Publications from this dataset:

DECEMBER 15, 2016

Many Americans Believe Fake News Is Sowing Confusion
23% say they have shared a made-up news story – either knowingly or not

American Trends Panel Wave 14

Survey conducted Jan. 12-Feb. 8, 2016


[Log in to Download Dataset](#)

Publications from this dataset:

MAY 26, 2016

News Use Across Social Media Platforms 2016

American Trends Panel Wave 14.5

Survey conducted Feb. 24-March 1, 2016


[Log in to Download Dataset](#)

Publications from this dataset:

Questionnaire

WV6_Official_Questionnaire_v4_June2012.pdf 

WV6_Official_Questionnaire_v5_SilatechMenaModule_Arabic.doc 

WV6_Official_Questionnaire_v5_SilatechMenaModule_English.doc 

Codebook

WV6_Codebook_v20180912 

WV6_Results_By_Country_v20180912 

Statistical Data Files

WV6_Data_R_v20180912 

WV6_Data_Sas_v20180912 


WV6_Data_Spss_v20180912 

WV6_Data_Stata_v20180912 

Older versions of Data Files

WV6_Data_ascii_delimited_v_2015_04_18 (delimited with comma) 

WV6_Data_ascii_delimited_v_2016_01_01 (Ascii delimited + structure) 

WV6_Data_R_v_2016_01_01 (R Workspace) 

WV6_Data_rdata_v_2015_04_18 (R workspace) 

WV6_Data_sas_v_2016_01_01 (SAS) 

WV6_Data_spss_v_2015_04_18 (Spss SAV) 

“Half-baked” open datasets

- Hong Kong LegCo voting record
(https://www.legco.gov.hk/general/english/counmtg/yr16-20/mtg_1718.htm#toptbl)
 - A famous [example](#)
- Open government in mainland China
 - Shenzhen (<http://opendata.sz.gov.cn/>)
 - Shanghai
(<http://www.datashanghai.gov.cn/home!toHomePage.action>)
 - Changsha (<http://www.changsha.gov.cn/xxgk/szfxxgkml/>)
 - *Question: which city is the most “open?”*

Data sources: “structured”?

- Structured data
 - Simply put, with “rows” as observations (cases) and “columns” as variables (measurements)
 - A “Data Frame” (in Python Pandas/R/SPSS/STATA)
 - .sav, .xls, .csv, .json, .xml
 - “*Machine-readable*”
- Non-structured data
 - The data structure is not well defined.
 - .pdf

Questions

1. Are all the open data structured data?
 - Really? [[Case](#)]
2. Are “html” pages structured data?
 - From the developers’ view;
 - From the researchers’ view;

Data sources: “legal”?

- Legal data
 - “Terms of use”
 - Robots.txt
 - API
 - FOIA
 - Other regulations (copyright and patent)
- Controversial zones
 - “Wiki Leaks”?
 - The Pentagon paper?
 - The hacked data

Data sources: “human subjects”?

- Obtrusive
 - Survey/experiment/interviews
 - API
 - Users’ digital traces
- Unobtrusive
 - Content analysis/discourse analysis
 - Web scraping (?)

Data sources: “something real”?

- Empirical data
 - All the data derived from human’s behaviors or social artifacts
- Simulated data (Agent-based modeling)
 - “if something has happened”

This course...

- The purposes are:
 - Open data versus non-open data
 - Government, education sectors, public institutions, NGOs, online public domains versus commercial websites
 - Structured data versus non-structured data
 - “Ready-made” versus “custom-made”
 - “ready-for-download” data versus scraped data
 - Legal data versus controversial data
 - Obtrusive versus non-obtrusive
 - Empirical versus simulated

Web scraping: more instances

- Structured data from public domain
- Developing web crawlers

Web scraping: a review

- Single page versus multiple pages?
- Structured information versus unstructured information?
- Static information versus interactive information?

Web scraping: a review

	Single vs. multiple	Structured vs. Non-structured	Static vs. Interactive
1. HKBU Jour faculty page	Single	Non-structured	Static (partial)
2. CityU news 2008 - 2018	Multiple	Structured	Static
3. Weather tables	Single	Structured	Static
4. UGC funded projects files	Single	Structured	Static
5. Twitter profiles	Multiple	Structured	Interactive (preparing for crawlers)