Web Data Collection with R

Session 8: Collecting Data from APIs

Sascha Göbel
sascha.goebel@uni-konstanz.de

December 23rd, 2020

Today: APIs

- API basics
- Authentication
- JSON
- Building API bindings from R

API Basics

Application Programming Interface

- service to facilitate exchange of information
- targeted information retrieval
- REST - popular API standard
  - Representational State Transfer
  - resources are referenced via URLs
  - data is represented via documents (JSON, XML, etc.)
- requires wrapper that handles details of access and data transformation
- for many websites, wrappers already exist in R

Terms of Usage

- usually described in API documentation on developer page
- data collection encouraged
- legal and more secure than classic web scraping
- but typically also restricted and can shut down anytime
- have to register and sometimes even pay for access
- Let us have a look at the Twitter developer page

Querying APIs

- (1) identify request method, frequently:
  - GET - uses URL to send request
  - POST - uses body to send data
- (2) define protocol for exchange between user and server
- (3) specify domain
- (4) describe path to resource on the server
- (5) state query parameters
  - preceded by a ?
  - consist of key–value pairs
  - separated by & or +
- API documentations often have API references with example queries

Status Codes

- sent back by server with the response to your request
- mostly standardized but pages may assign specific meaning
- check developer page
- important when automating API calls

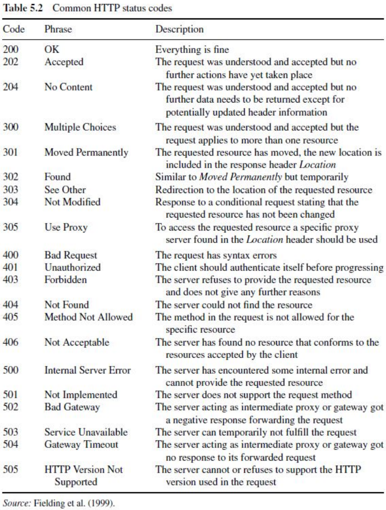

Table 5.2 Common HTTP status codes

Code	Phrase	Description
200	OK	Everything is fine
202	Accepted	The request was understood and accepted but no further actions have yet taken place
204	No Content	The request was understood and accepted but no further data needs to be returned except for potentially updated header information
300	Multiple Choices	The request was understood and accepted but the request applies to more than one resource
301	Moved Permanently	The requested resource has moved, the new location is included in the response header *Location*
302	Found	Similar to *Moved Permanently* but temporarily
303	See Other	Redirection to the location of the requested resource
304	Not Modified	Response to a conditional request stating that the requested resource has not been changed
305	Use Proxy	To access the requested resource a specific proxy server found in the *Location* header should be used
400	Bad Request	The request has syntax errors
401	Unauthorized	The client should authenticate itself before progressing
403	Forbidden	The server refuses to provide the requested resource and does not give any further reasons
404	Not Found	The server could not find the resource
405	Method Not Allowed	The method in the request is not allowed for the specific resource
406	Not Acceptable	The server has found no resource that conforms to the resources accepted by the client
500	Internal Server Error	The server has encountered some internal error and cannot provide the requested resource
501	Not Implemented	The server does not support the request method
502	Bad Gateway	The server acting as intermediate proxy or gateway got a negative response forwarding the request
503	Service Unavailable	The server can temporarily not fulfill the request
504	Gateway Timeout	The server acting as intermediate proxy or gateway got no response to its forwarded request
505	HTTP Version Not Supported	The server cannot or refuses to support the HTTP version used in the request

Source: Fielding et al. (1999).

Source: Munzert et al. 2015

Authentication

API registration

- basic: send API key with request
- OAuth: three legged authentication (client, temporary, token)
- In R via ROAuth and httr.
- Let us do this for the Twitter API

JSON

JavaScript Object Notation

- another data exchange standard
- language independent
- hierarchical structure, no tags
- data stored in key-value pairs separated by :
- curly braces encapsulate objects - contain data, other objects, or arrays
- square brackets enclose arrays - sequence of objects or values
- several R wrappers to convert to R object