Web Data Collection with R

URL

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

Teaser

How URLs work

Teaser

the search page of Wikipedia

http://en.wikipedia.org/w/index.php?title=Special%3ASearch

some trying it out

- put in some text
- hit Enter to search
- look at URL in Browser
- manipulate URL directly

How URLs work

How URLs work

Uniform **R**esource **I**dentifier

Uniform **R**esource **L**ocator

URL example

`http://en.wikipedia.org/w/index.php?title=Special%3ASearch&profile=default&search=bruce&fulltext=Search`

URL scheme

`scheme://domain:port/path?query_string#fragment_id`

- scheme: which protocoll to use
- domain and port: which *server* to contact
- path: path to the resource
- query_string: which parameters to use
- fragment_id: to which part of the resource to jump to

URL encode and decode

- URLs are ASCII only and even things like spaces are not allowed
- therfore one need to en- and de-code URL before using them

```
URLencode("http://mypage.com/my super cool #tag")
```

```
## [1] "http://mypage.com/my%20super%20cool%20#tag"
```

```
URLdecode("http://mypage.com/my%20super%20cool%20#tag")
```

```
## [1] "http://mypage.com/my super cool #tag"
```

- you might also want to have a look at the urltools package for more extensive features