Web Data Collection with R

Session 7: Regular Expressions and Data Cleaning

Sascha Göbel
sascha.goebel@uni-konstanz.de

December 16th, 2020

Today: Regular expressions

- clean data with regex
- filter desired information with regex
- plot data

Regular expressions (Regex)

The Problem

- the web data collection step is completed
- but our data is not useful for quantitative operations

```
> persons_names
 [1] "\n    \n          \n                Dr. Liliana de Abreu\n                \n        \n"
 [2] "\n    \n          \n                Astrid Ahnert\n                \n        \n"
 [3] "\n    \n          \n                Anna Apostolidou\n                \n        \n"
 [4] "\n    \n          \n                Yvonne Aymar\n                \n        \n"
 [5] "\n    \n          \n                Jun.-Prof. Dr. Aurélia Bardon\n                \n        \n"
 [6] "\n    \n          \n                Konstantin Bätz\n                \n        \n"
 [7] "\n    \n          \n                Julia Bettecken (née Becker)\n                \n        \n"
 [8] "\n    \n          \n                Karin Becker\n                \n        \n"
 [9] "\n    \n          \n                Fabian Wolfgang Bergmann\n                \n        \n"
[10] "\n    \n          \n                Nicolas Binder\n                \n        \n"
[11] "\n    \n          \n                Jana Blahak\n                \n        \n"
[12] "\n    \n          \n                Nona Roswitha Bledow\n                \n        \n"
[13] "\n    \n          \n                Prof. Dr. Sabine Boerner\n                \n        \n"
[14] "\n    \n          \n                Prof. Dr. Christian Breunig\n                \n        \n"
[15] "\n    \n          \n                Prof. Dr. Marius Busemeyer\n                \n        \n"
[16] "\n    \n          \n                Dr. Patrícia Calca\n                \n        \n"
[17] "\n    \n          \n                Sina Chen\n                \n        \n"
[18] "\n    \n          \n                Jun.-Prof. Dr. Michael Dobbins\n                \n        \n"
[19] "\n    \n          \n                Christina Draheim\n                \n        \n"
[20] "\n    \n          \n                Jun.-Prof. Dr. Steffen Eckhard\n                \n        \n"
[21] "\n    \n          \n                Gianna Maria Eick\n                \n        \n"
[22] "\n    \n          \n                Katharina Eßmeyer\n                \n        \n"
[23] "\n    \n          \n                Annette Flowe\n                \n        \n"
[24] "\n    \n          \n                Susanne Garritzmann M.A.\n                \n        \n"
[25] "\n    \n          \n                Josefa Lisa Glass-Kawerau\n                \n        \n"
[26] "\n    \n          \n                Julia Goebel\n                \n        \n"
[27] "\n    \n          \n                Sascha Göbel\n                \n        \n"
[28] "\n    \n          \n                Selina Graulich\n                \n        \n"
[29] "\n    \n          \n                Alina Greiner\n                \n        \n"
[30] "\n    \n          \n                Frederik Gremler\n                \n        \n"
[31] "\n    \n          \n                Benjamin Guinaudeau\n                \n        \n"
[32] "\n    \n          \n                Jessica Haase\n                \n        \n"
```

Country	Locale	Remarks	Date	President
Canada	Ottawa	Working visit; met with Prime Minister Mulroney.	February 10, 1989	George H.W. Bush
Japan	Tokyo	Attended the funeral of Emperor Hirohito. Met with Emperor ...	February 23–25, 1989	George H.W. Bush
China, People's Republic of	Beijing	Met with President Yang and Prime Minister Li. Also met with ...	February 25–27, 1989	George H.W. Bush
Korea, Republic of	Seoul	Official visit; addressed the National Assembly.	February 27, 1989	George H.W. Bush
Italy	Rome, Nettuno	Met with President Cossiga and Prime Minister De Mita.	May 26–28, 1989	George H.W. Bush
Vatican City		Audience with Pope John Paul II.	May 27, 1989	George H.W. Bush
Belgium	Brussels	Attended NATO Summit Meeting. Present were the Heads of ...	May 28–30, 1989	George H.W. Bush
Germany, Federal Republic of	Bonn, Mainz	Met with Chancellor Kohl.	May 30–31, 1989	George H.W. Bush
United Kingdom	London	Met with Queen Elizabeth II and Prime Minister Thatcher.	May 31–June 2, 1989	George H.W. Bush
Poland	Warsaw, Gdansk	Met with government and Solidarity leaders. Addressed the N...	July 9–11, 1989	George H.W. Bush
Hungary	Budapest	Met with Hungarian officials and delivered an address at Karl ...	July 11–13, 1989	George H.W. Bush
France	Paris	Attended Economic Summit Meeting of the Heads of State an...	July 13–17, 1989	George H.W. Bush
Netherlands	The Hague, Leiden	Met with Queen Beatrix and Prime Minister Lubbers and deliv...	July 17–18, 1989	George H.W. Bush
Costa Rica	San Jose	Attended Hemispheric Summit Meeting.	October 27–28, 1989	George H.W. Bush
Malta	Valletta, Marsaxlokk Bay	Attended Summit Meeting (December 2-3) with Soviet Chair...	December 1–3, 1989	George H.W. Bush
Belgium	Brussels	Briefed NATO Heads of State and Government on the U.S.-So...	December 3–4, 1989	George H.W. Bush
France	St. Martin Island (French West Indies)	Informal meeting with President Mitterrand.	December 16, 1989	George H.W. Bush
Colombia	Cartagena	Attended Summit Meeting on the control of illicit drug traffick...	February 15, 1990	George H.W. Bush
Canada	Toronto	Informal meeting with Prime Minister Mulroney.	April 10, 1990	George H.W. Bush
United Kingdom	Bermuda	Informal meeting with Prime Minister Thatcher.	April 13–14, 1990	George H.W. Bush
United Kingdom	London	Attended NATO Summit Meeting.	July 5–6, 1990	George H.W. Bush
Finland	Helsinki	Summit Meeting with Soviet President Gorbachev. Issued join...	September 8–9, 1990	George H.W. Bush
Czechoslovakia	Prague	Attended ceremonies commemorating the first anniversary of...	November 17, 1990	George H.W. Bush
Germany	Speyer, Ludwigshafen	Met with Chancellor Kohl.	November 18, 1990	George H.W. Bush
France	Paris	Attended CSCE Summit Meeting and the signing of the Treaty...	November 18–21, 1990	George H.W. Bush

created	identifier	label	numberOfSignatures	sponsorPrinted	status
2019-03-27T11:42:09.105Z	251994	Introduce legal immunity for UK soldiers after a war or conflict...	567	luke huskisson	open
2019-03-26T20:47:38.313Z	251677	Cancel daylight saving clock changes.	30	Tom Young	open
2019-03-26T11:15:47.373Z	251390	Give specialist recognition to General Practitioners in the U.K	251	Dr Penny Ward	open
2019-03-26T10:21:23.095Z	251367	Completely Overhaul the Universal Credit system or return to...	85	Mark Heckles	open
2019-03-26T10:13:19.025Z	251361	Add the study of architecture and urbanism to the national cu...	55	Matt Gaskin	open
2019-03-26T01:53:21.761Z	251275	Central gov to step in - reduce the gap between private and ...	22	James Cannon	open
2019-03-25T21:39:05.156Z	251202	Introduce an opt-out £1 per month NHS tax	6	Andreas Hicks	open
2019-03-25T14:32:19.753Z	250967	Leave EU by April 12th 2019 with deal or no deal.	20943	Terry Nicholls	open
2019-03-25T09:58:04.754Z	250845	Grant legal protection to Swallow, Swift and Martin nest sites ...	8	Simon Leadbeater	open
2019-03-25T09:56:20.665Z	250844	Give us the option to get our driving licence with out the Uni...	2599	Mirain Llwyd Roberts	open
2019-03-24T22:02:12.339Z	250679	Halt HS2 using the funds to support Education and Social Ser...	13	Roger Jones	open
2019-03-24T18:08:48.432Z	250508	Mandatory mental health assistance and training for all UK e...	7	Ashlyn Steyn	open
2019-03-24T16:39:57.048Z	250430	The recognition of Somaliland as an independent state.	17	Abdi Ahmed	open
2019-03-24T11:04:16.922Z	250177	Ban plastic toys from fast food restaurants' kids meals	18	Lisa Quinn	open
2019-03-24T10:17:12.536Z	250135	Have a new referendum regards leaving the European Union.	9	John Cook	open
2019-03-24T09:49:29.308Z	250111	Add and teach the impacts of climate change to primary scho...	20	Emily Sparke	open
2019-03-24T08:45:54.490Z	250058	Free DBS checks and online update service for the care industry	38	Lucian Balog	open
2019-03-23T23:42:09.254Z	249915	Free Warfarin prescription and disability for lupus & anti phos...	21	Monsur Zaman	open
2019-03-23T17:12:39.597Z	249457	Change the way parliament works so that MPs role is not a ca...	12	Mike Hicks	open
2019-03-23T11:52:39.406Z	249151	Ban Ear Piercing under the age of 5 in The UK	20	Victoria K L Palmer	open
2019-03-23T10:26:54.787Z	249073	Make having face-to-face BSL interpreters compulsory in NHS...	8	Nicholas Alexander Hofmann	open
2019-03-23T08:59:03.321Z	248985	Grant prescribing and injection rights to osteopaths.	57	Dr Ghulam Adel	open
2019-03-22T21:30:23.387Z	248742	Raise the smoking age from 18 to 21	8	Jack Streeter	open
2019-03-22T17:10:17.556Z	248471	Hold a referendum: "Should the UK join the EU in January 20...	27	Andrew Nowell	open
2019-03-22T16:14:02.014Z	248400	Make schools close earlier in summer.	13	Louis Kirkman	open

The challenge

- we are interested in systematic information, e.g., numbers, names, locations, email addresses, etc.
- buried in unstructured text of all kinds of formats
- need to tidy or extract information

The solution: regex

- syntax for systematically searching and manipulating text
- convention on how to query strings
- basically a sequence of characters describing patterns in text

The challenge

- we are interested in systematic information, e.g., numbers, names, locations, email addresses, etc.
- buried in unstructured text of all kinds of formats
- need to tidy or extract information

The solution: regex

- syntax for systematically searching and manipulating text
- convention on how to query strings
- basically a sequence of characters describing patterns in text

Regular Expressions in R

`stringr` package

- facilitates text manipulation in R
- form: `str_*(string, pattern)`
- string: text to be operated on, a sequence of characters
- pattern: expression we are looking for
- case sensitive

Two types of regex

- exact character matching - characters match characters
- generalized expressions - matches general classes based on specific rules

Regular Expressions in R

`stringr` package

- facilitates text manipulation in R
- form: `str_*(string, pattern)`
- string: text to be operated on, a sequence of characters
- pattern: expression we are looking for
- case sensitive

Two types of regex

- exact character matching - characters match characters
- generalized expressions - matches general classes based on specific rules

Regex basics

- ^ - matches beginning of a string
- $ - matches end of a string
- | - separates multiple expressions
- . - matches any character
- character classes - enclosed in brackets, any of the characters within will be matched
- quantifiers - how often an item will be matched
- greedy quantification - extract greates possible sequence of preceding character (default)
- non-greedy quantification - add ? to look for shortest possible sequence
- special characters - preceded with "\\" to match literally
- look arounds - match conditional on context of item

Table 8.1 Selected predefined character classes in R regular expressions

`[:digit:]`	Digits: 0 1 2 3 4 5 6 7 8 9
`[:lower:]`	Lowercase characters: a–z
`[:upper:]`	Uppercase characters: A–Z
`[:alpha:]`	Alphabetic characters: a–z and A–Z
`[:alnum:]`	Digits and alphabetic characters
`[:punct:]`	Punctuation characters: . , ; etc.
`[:graph:]`	Graphical characters: `[:alnum:]` and `[:punct:]`
`[:blank:]`	Blank characters: Space and tab
`[:space:]`	Space characters: Space, tab, newline, and other space characters
`[:print:]`	Printable characters: `[:alnum:]`, `[:punct:]` and `[:space:]`

Source: Adapted from http://stat.ethz.ch/R-manual/R-patched/library/base/html/regex.html

Source: Munzert et al. 2015

Table 8.3 Selected symbols with special meaning

\w	Word characters: [[:alnum:]_]
\W	No word characters: [^[:alnum:]_]
\s	Space characters: [[:blank:]]
\S	No space characters: [^[:blank:]]
\d	Digits: [[:digit:]]
\D	No digits: [^[:digit:]]
\b	Word edge
\B	No word edge
\<	Word beginning
\>	Word end

Source: Munzert et al. 2015

Table 8.2 Quantifiers in R regular expressions

`?`	The preceding item is optional and will be matched at most once
`*`	The preceding item will be matched zero or more times
`+`	The preceding item will be matched one or more times
`{n}`	The preceding item is matched exactly *n* times
`{n,}`	The preceding item is matched *n* or more times
`{n,m}`	The preceding item is matched at least *n* times, but not more than *m* times

Source: Adapted from http://stat.ethz.ch/R-manual/R-patched/library/base/html/regex.html

Source: Munzert et al. 2015

Table 8.4 Functions of package stringr in this chapter

Function	Description	Output
Functions using regular expressions		
`str_extract()`	Extracts first string that matches pattern	Character vector
`str_extract_all()`	Extracts all strings that match pattern	List of character vectors
`str_locate()`	Returns position of first pattern match	Matrix of start/end positions
`str_locate_all()`	Returns positions of all pattern matches	List of matrices
`str_replace()`	Replaces first pattern match	Character vector
`str_replace_all()`	Replaces all pattern matches	Character vector
`str_split()`	Splits string at pattern	List of character vectors
`str_split_fixed()`	Splits string at pattern into fixed number of pieces	Matrix of character vectors
`str_detect()`	Detects patterns in string	Boolean vector
`str_count()`	Counts number of pattern occurrences in string	Numeric vector
Further functions		
`str_sub()`	Extracts strings by position	Character vector
`str_dup()`	Duplicates strings	Character vector
`str_length()`	Returns length of string	Numeric vector
`str_pad()`	Pads a string	Character vector
`str_trim()`	Discards string padding	Character vector
`str_c()`	Concatenates strings	Character vector

Source: Munzert et al. 2015

+ `str_remove()` - removes matched patterns

+ `str_squish()` - removes redundant whitespace within a string

Look arounds

- outside brackets what you want to match, context inside brackets
- context is not consumed
- $a(?=b)$ - positive lookahead - matches a followed by b
- $a(?!b)$ - negative lookahead - matches a not followed by b
- $(?<=b)a$ - positive lookbehind - matches a preceded by b
- $(?<!b)a$ - negative lookbehind - matches a not preceded by b