Package 'rcorpora'

May 1, 2016

Title A Collection of Small Text Corpora of Interesting Data

Version 1.2.0

Maintainer Gabor Csardi <csardi.gabor@gmail.com>

Author Darius Kazemi, Cole Willsea, Matthew Rothenberg, Karl Swedberg, Daniel D. Beck, Javier Arce, Matthew Hokanson, Casey Kolderup, Nathaniel Mitchell, Mark Sample, Nathan Lachenmyer, Aaron Marriner, Greg Kennedy, Greg Borenstein, Peter Organisciak, Rachel White, Tod Robbins, John Wiseman, M. Nowak, Alice Maz, Allison Parrish, Andrew Gorman, Colin Mitchell, David Whitten, Mary Dickson, Michael R. Bernstein, Parker Higgins, Patrick Rodriguez, Ross Barclay, Ross Binden, Ryan Freebern, Justin Alford, Brian Detweiler, Ed Lea, John Ohno, Alexandra Murray, Sean May, Will Hankinson, Brett O'Connor, Brian Jones, Casey Olson, Edward Loveall, Felix Laurie von Massenbach, Garrett Miller, Grant Williamson, Jacob Fauber, Joe Mahoney, Jordan Killpack, Kay Belardinelli, K.Adam White, Kyle McDonald, Andy Dayton, Adam Malantonio, Marcos Wright-Kuhns, Mark Wunsch, Max Bittker, Michael Dewberry, Nathan Black, Noah Kantrowitz, Noah Swartz, Ranjit Bhatnagar, Rob Huzzey, Russell Horton, Vincent Bruijn, Virginia Murdoch, Zac Moody, Scott Grant, Tariq Ali

Description A collection of small text corpora of interesting data.
It contains all data sets from https://github.com/dariusk/corpora.
Some examples:
names of animals: birds, dinosaurs, dogs; foods: beer categories, pizza toppings; geography: English towns, rivers, oceans; humans: authors, US presidents, occupations; science: elements, planets; words: adjectives, verbs, proverbs, US president quotes.

License CC0

Imports jsonlite,
utils

URL https://github.com/gaborcsardi/rcorpora

BugReports https://github.com/gaborcsardi/rcorpora/issues

RoxygenNote 5.0.1.9000

R topics documented:

`categories` *List data set categories in the corpora package*

Description

List data set categories in the corpora package

Usage

```
categories()
```

Value

Character vector of category names.

`corpora` *Load a data set from the corpora package*

Description

corpora is a collection of small corpora of interesting data for the creation of bots and similar stuff.

Usage

```
corpora(which, category)
```

Arguments

`which`	The data set to load, a string. If not given, then all data sets in the package are listed.
`category`	If given, `which` must be missing, and the data sets in the given category are listed.

Details

This project is a collection of static corpora (plural of "corpus") that are potentially useful in the creation of weird internet stuff. I've found that, as a creator, sometimes I am making something that needs access to a lot of adjectives, but not necessarily every adjective in the English language. So for the last year I've been copy/pasting an adjs.json file from project to project. This is kind of awful, so I'm hoping that this project will at least help me keep everything in one place.

I would like this to help with rapid prototyping of projects. For example: you might use nouns.json to start with, just to see if an idea you had was any good. Once you've built the project quickly around the nouns collection, you can then rip it out and replace it with a more complex or exhaustive data source.

I'm also hoping that this can be used as a teaching tool: maybe someone has three hours to teach how to make Twitter bots. That doesn't give the student much time to find/scrape/clean/parse interesting data. My hope is that students can be pointed to this project and they can pick and choose different interesting data sources to meld together for the creation of prototypes.

See https://github.com/dariusk/corpora

Value

A data frame containing the data set (if `which` is given), or a character vector of data set names.

Data set categories

- animals
- archetypes
- architecture
- art
- colors
- corporations
- divination
- film-tv
- foods
- games
- games/bannedGames
- games/bannedGames/argentina
- games/bannedGames/brazil
- games/bannedGames/china
- games/bannedGames/denmark
- geography
- governments
- humans
- instructions
- materials
- mathematics
- medicine
- music
- mythology
- objects
- plants
- religion
- science
- societies_and_groups
- societies_and_groups/designated_terrorist_groups
- societies_and_groups/fraternities

- sports
- technology
- words
- words/emoji
- words/literature
- words/stopwords
- words/word_clues

Data sets

animals/birds_antarctica Birds of Antarctica, grouped by family Source: https://en.wikipedia.org/wiki/List_of_birds_

animals/birds_north_america Birds of North America, grouped by family Source: http://listing.aba.org/aba-checklist/

animals/birds_uk Birds of the United Kingdom, grouped by family Source: http://www.rspb.org.uk

animals/common

animals/dinosaurs A list of dinosaurs.

animals/dogs A list of dog breeds.

archetypes/artifact Artifact archetypes.

archetypes/character Common character archetypes.

archetypes/event Archetypal events.

archetypes/setting Setting and location archetypes.

architecture/passages Ways to enter or exit a place.

architecture/rooms Different kinds of rooms

art/isms A list of modernist art isms.

colors/crayola List of Crayola crayon standard colors

colors/paints List of assorted paint colors from various brands.

colors/web_colors List of named HTML colors

corporations/cars A list of car manufacturers.

corporations/djia Corporations of the Dow Jones Industrial Average

corporations/fortune500 The 2014 Fortune 500 list

corporations/industries A list of all industries on LinkedIn, as of May 21, 2013 Source: http://robertwdempsey.com/lii

corporations/nasdaq Corporations of the NASDAQ 100

corporations/newspapers A list of newspapers scraped in early 2013.

divination/tarot_interpretations Tarot card interpretations, from Mark McElroy's _A Guide to Tarot Meanings_ (http://www.madebymark.com/a-guide-to-tarot-card-meanings/)

film-tv/tv_shows 1000 entries from the list of TV shows at http://en.wikipedia.org/wiki/List_of_television_programs_b

foods/apple_cultivars The 1000 most popular apple cultivars in the USDA's Pomological Watercolor collection.

foods/beer_categories A list of beer categories.

foods/beer_styles A list of beer styles.

foods/breads_and_pastries A list of classic breads and sweet pastries.

foods/combine A list of recipe instructions.

foods/condiments A list of condiments

foods/curds A list of curds, cheeses, and other fermented dairy products

foods/fruits A list of fruits.

foods/herbs_n_spices A list of herbs and spices, and mixtures of the two.

foods/hot_peppers Capsicum cultivars (hot peppers)

foods/menuItems A list of the top 1000 most appearing menu items from the 1850s to today from the New York Public Library's "What's on the menu?" project. Please credit The New York Public Library as source on any applications or publications. http://menus.nypl.org/data

foods/pizzaToppings A list of pizza toppings.

foods/sandwiches A list of sandwiches.

foods/tea types of tea

foods/vegetable_cooking_times Approximate cooking times for various vegetables Source: http://recipes.howstuffwo and-techniques/how-to-cook-vegetables24.htm

foods/vegetables A list of vegetables.

foods/wine_descriptions A list of words commonly used to describe wine.

games/bannedGames/argentina/bannedList A list of video games banned in Argentina

games/bannedGames/brazil/bannedList A list of video games banned in Brazil

games/bannedGames/china/bannedList A list of video games banned in China.

games/bannedGames/denmark/bannedList A list of video games banned in Denmark

games/cluedo Characters, rooms and weapons from the board game Cluedo / Clue.

games/dark_souls_iii_messages Organized components from the Dark Souls III message system

games/jeopardy_questions A sampling of 1000 Jeopardy questions and metadata. For the full dataset, see http://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/

games/pokemon Source: https://github.com/UberGames/iPokedex-DB

games/scrabble Tile distribution and points for the English-language edition of Scrabble

games/street_fighter_ii Street Fighter II fighting moves

games/trivial_pursuit Pie categories and colors from Trivial Pursuit

games/wrestling_moves A list of professional wrestling moves

geography/canada_provinces_and_territories A list of Canadian provinces and territories.

geography/countries A list of countries.

geography/english_towns_cities Two lists: one for English towns, one for English cities.

geography/london_underground_stations London Underground stations, with their lines and Travelcard zones Source: https://en.wikipedia.org/wiki/List_of_London_Underground_stations

geography/oceans A list of oceans and seas. Source: http://en.wikipedia.org/wiki/List_of_seas

geography/rivers A list of rivers. Source: http://en.wikipedia.org/wiki/List_of_rivers_by_length

geography/us_cities Top 1000 U.S. cities by 2013 population

geography/venues Venues organized by category. Source: https://developer.foursquare.com/categorytree

governments/nsa_projects A list of NSA project code names. Source: All data here is from https://docs.google.com/spreadsheets/d/1Uc1hrGqIweF0rgJ1HCbmT_0w9CYCCwZTWBGOwydscqE/htmlview?

governments/uk_political_parties A list of uk political parties. Source: http://www.electoralcommission.org.uk/ export on 8th May 2015

governments/us_federal_agencies A list of federal agencies. Source: This data was sourced from the GSA's list of .gov domains https://github.com/GSA/data/blob/gh-pages/dotgov-domains/2014-12-01-federal.csv

governments/us_mil_operations Code names for US Military Operations Source: All names from the scraped pages of http://www.designation-systems.net/usmilav/codenames.html

humans/authors

humans/bodyParts A list of common human body parts.

humans/britishActors A bunch of British actors.

humans/englishHonorifics English honorifics.

humans/famousDuos Famous duos

humans/firstNames First names of men and women, pulled from the US Census for the 2000s.

humans/lastNames Last names of people, pulled from the US Census for the 2000s.

humans/moods A list of words that naturally complete the phrase 'They were feeling...'.

humans/occupations A list of occupations (jobs that people might have).

humans/prefixes Prefixes taken from a form on an airline website.

humans/richpeople A bunch of rich people from a Forbes listicle, including the source article, img, and name

humans/scientists List of particularly famous scientists

humans/spanishFirstNames A list of common Spanish first names of men and women. Source: https://github.com/olea/lemarios

humans/spanishLastNames A list of common Spanish last names. Source: https://github.com/olea/lemarios

humans/spinalTapDrummers Deceased drummers from the fictional rock band Spinal Tap, taken from Wikipedia.

humans/suffixes Suffixes taken from a form on an airline website.

humans/us_presidents Copy of JSON retrieved from https://www.govtrack.us/api/v2/role?role_type=president. The ID here matches the one in the corpora/data/words/us_president_quotes.json file

humans/wrestlers A bunch of WWE wrestlers nicknames

instructions/laundry_care A list of laundry care instructions

materials/abridged-body-fluids abridged body fluids

materials/building-materials building materials

materials/carbon-allotropes carbon allotropes

materials/decorative-stones decorative stones

materials/fabrics fabrics

materials/fibers fibers

materials/gemstones A list of the names of materials commonly used as gemstones Source: https://en.wikipedia.org/wiki/List_of_gemstone_species

materials/layperson-metals layperson metals

materials/metals metals

materials/natural-materials natural materials

materials/packaging packaging

materials/plastic-brands plastic brands

materials/sculpture-materials sculpture materials

materials/technical-fabrics technical fabrics

mathematics/fibonnaciSequence The first 1000 numbers in the Fibonnaci Sequence

mathematics/primes The first 1000 prime numbers.

mathematics/trigonometry A list of trigonometric functions, formulas, equations, etc..

medicine/diagnoses International Statistical Classification of Diseases and Related Health Problems, 10th revision Source: http://www.cdc.gov/nchs/icd/icd10cm.htm

medicine/drugNameStems A list of generic pharmaceutical drug name stems. Hypens indicate whether a stem appears at the beginning, middle, or end of the name. Source: http://druginfo.nlm.nih.gov/drugportal

medicine/drugs A list of pharmaceutical drug names Source: The United States National Library of Medicine, http://druginfo.nlm.nih.gov/drugportal/

music/bands_that_have_opened_for_tool Bands that have opened for Tool. You must be really dedicated to your music if you are willing to play before Tool fans.

music/genres A list of musical genres taken from wikipedia article titles.

music/mtv_day_one Music videos broadcast on MTV's first day Source: https://en.wikipedia.org/wiki/First_music_vi

music/rock_hall_of_fame Artists who have been added to the Rock N' Roll Hall of Fame along with their year of induction Source: https://en.wikipedia.org/wiki/List_of_Rock_and_Roll_Hall_of_Fame_inductee

mythology/greek_gods Gods and goddesses from Greek myth

mythology/greek_monsters Monsters from Greek myth

mythology/greek_titans Titans from Greek myth

mythology/hebrew_god Hebrew names of God used in the Old Testament Bible

mythology/lovecraft Deities and supernatural creatures from the works of Lovecraft and the Cthulhu mythos.

mythology/monsters A list of monsters and other mythic creatures

mythology/norse_gods Gods and goddesses of norse and germanic myth

objects/objects List of household objects

plants/cannabis 420 popular strains of cannabis

plants/flowers

religion/christian_saints

religion/fictional_religions

religion/parody_religions

religion/religions

science/elements

science/hail_size Analogous objects for various hail sizes, adapted from http://www.spc.noaa.gov/misc/tables/hailsize.

science/minor_planets List of names of the first 1000 numbered minor planets

science/planets Planets (including dwarf planets as recognized by the IAU) that orbit the Sun, with their natural satellites.

science/pregnancy

science/toxic_chemicals

societies_and_groups/animal_welfare

societies_and_groups/designated_terrorist_groups/australia

societies_and_groups/designated_terrorist_groups/canada

societies_and_groups/designated_terrorist_groups/china

societies_and_groups/designated_terrorist_groups/egypt

societies_and_groups/designated_terrorist_groups/european_union

societies_and_groups/designated_terrorist_groups/india

societies_and_groups/designated_terrorist_groups/iran

societies_and_groups/designated_terrorist_groups/israel

societies_and_groups/designated_terrorist_groups/kazakhstan

societies_and_groups/designated_terrorist_groups/russia

societies_and_groups/designated_terrorist_groups/saudi_arabia

societies_and_groups/designated_terrorist_groups/tunisia

societies_and_groups/designated_terrorist_groups/turkey

societies_and_groups/designated_terrorist_groups/uae

societies_and_groups/designated_terrorist_groups/ukraine

societies_and_groups/designated_terrorist_groups/united_kingdom

societies_and_groups/designated_terrorist_groups/united_nations

societies_and_groups/designated_terrorist_groups/united_states

societies_and_groups/fraternities/coeducational_fraternities

societies_and_groups/fraternities/defunct

societies_and_groups/fraternities/fraternities

societies_and_groups/fraternities/professional

societies_and_groups/fraternities/service

societies_and_groups/fraternities/sororities

societies_and_groups/semi_secret

sports/nfl_teams Current (as of 2015) teams in the NFL and where they play

technology/appliances A list of home appliances

technology/computer_sciences names of technologies related to computer science

technology/fireworks A list (ooh!) of firework effects (aah!)

technology/guns_n_rifles weapons used in mass shootings in the U.S.A.

technology/knots A list of knot names.

technology/lisp a list of LISP dialects

technology/new_technologies new or emerging technologies

technology/photo_sharing_websites Photo sharing websites

technology/programming_languages

technology/social_networking_websites Social networking websites

technology/video_hosting_websites Video hosting websites

words/adjs A list of English adjectives.

words/adverbs

words/closed_pairs closed pairs in English i.e both words rhyme with each other and only with each other. from https://en.wikipedia.org/wiki/List_of_closed_pairs_of_English_rhyming_words

words/common Common English words.

words/crash_blossoms confusing or misleading headlines

words/eggcorns Commonly mistaken English phrases most likely caused by hearing them rather than reading them (eggcorns) Source: Most of the examples come from http://eggcorns.lascribe.net/

words/emoji/cute_kaomoji A general corpus of cute kaomoji.

words/emoji/positive_emoji A general corpus of positive emoji.

words/emoji/sea_emoji A general corpus of emoji of sea/water creatures.

words/encouraging_words a list of encouraging words to tell someone about something they created

words/interjections a list of exclamatory words and expressions from http://www.enchantedlearning.com/wordlist/inte

words/literature/mr_men_little_miss Mr Men and Little Miss characters Source: http://www.mrmen.com

words/literature/shakespeare_phrases Phrasess coined by Shakespeare, from http://www.pathguy.com/shakeswo.htm

words/literature/shakespeare_sonnets Shakespeare's sonnets.

words/literature/shakespeare_words Words coined by Shakespeare, from http://www.pathguy.com/shakeswo.htm

words/nouns A list of English nouns.

words/oprah_quotes Words of wisdom by Oprah Winfrey

words/personal_nouns List of personal nouns in the 1890 Webster's Unabridged Dictionary. Assembled by Cory Taylor from Project Gutenberg's HTML edition of the dictionary: http://www.gutenberg.org/ebook Source: https://github.com/coryandrewtaylor/Personal-Nouns

words/prefix_root_suffix

words/proverbs A list of proverbs sourced from http://tww.id.au/proverbs/proverbs.html

words/resume_action_words Resume action words Source: http://careercenter.umich.edu/article/resume-action-words

words/rhymeless_words English words for which there is no perfect rhyme, taken from https://en.wikipedia.org/wiki/I

words/spells A list of Harry Potter spells and descriptions

words/states_of_drunkenness A list of states of drunkenness.

words/stopwords/ar Arabic stop words

words/stopwords/bg Bulgarian stop words

words/stopwords/cs Czech stop words

words/stopwords/da Danish stop words

words/stopwords/de German stop words

words/stopwords/en English stop words

words/stopwords/es Spanish stop words

words/stopwords/fi Finnish stop words

words/stopwords/fr French stop words

words/stopwords/gr Greek stop words

words/stopwords/it Italian stop words

words/stopwords/jp Japanese stop words

words/stopwords/lv Latvian stop words

words/stopwords/nl Dutch stop words

words/stopwords/no Norwegian stop words

words/stopwords/pl Polish stop words

words/stopwords/pt Portuguese stop words

words/stopwords/ru Russian stop words

words/stopwords/sk Slovak stop words

words/stopwords/sv Swedish stop words

words/stopwords/tr Turkish stop words

words/us_president_quotes A list of quotes from US Presidents from http://bit.ly/1hsAYQT. ID matches up with https://govtrack.us API results.

words/verbs A list of English verbs.

words/word_clues/clues_five a list of common 5-letter words followed by crossword/thesaurus-style hints for that word

words/word_clues/clues_four a list of common 4-letter words followed by crossword/thesaurus-style hints for that word

words/word_clues/clues_six a list of common 6-letter words followed by crossword/thesaurus-style hints for that word

Examples

```
corpora()
corpora(category = "animals")
corpora("foods/pizzaToppings")
```

Index