

# Package ‘rcorpora’

July 12, 2015

**Title** A Collection of Small Text Corpora of Interesting Data

**Version** 1.1.1

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

**Author** Darius Kazemi, Matthew Rothenberg, Karl Swedberg, Matthew Hokanson, Nathan Lachenmyer, Aaron Marriner, Mark Sample, Casey Kolderup, Nathaniel Mitchell, Daniel D. Beck, Mike Nowak, Ryan Freebern, Ross Barclay, Ross Binden, Justin Alford, Cole Willsea, Andrew Gorman, Javier Arce, Patrick Rodriguez, Liam Cooke, Will Hankinson, K. Adam White, Garrett Miller, Zac Moody, Jordan Killpack, Brian Jones, Greg Borenstein, Noah Swartz, Nathan Black, Russell Horton, Mark Wunsch, Kay Belardinelli, Colin Mitchell, Michael Dewberry, Joe Mahoney

**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>

## R topics documented:

categories . . . . .	2
corpora . . . . .	2
<b>Index</b>	<b>7</b>

---

categories	<i>List data set categories in the corpora package</i>
------------	--------------------------------------------------------

---

**Description**

List data set categories in the corpora package

**Usage**

```
categories()
```

**Value**

Character vector of category names.

---

corpora	<i>Load a data set from the corpora package</i>
---------	-------------------------------------------------

---

**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
- foods
- games
- geography
- governments
- humans
- instructions
- materials
- music
- mythology
- objects
- plants
- science
- technology
- words

**Data sets**

**animals/birds\_antarctica** Birds of Antarctica, grouped by family Source: [https://en.wikipedia.org/wiki/List\\_of\\_birds](https://en.wikipedia.org/wiki/List_of_birds)

**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/rooms** Different kinds of rooms

**art/isms** A list of modernist art isms.

**colors/crayola** List of Crayola crayon standard colors

**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/nasdaq** Corporations of the NASDAQ 100

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

**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/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/vegetables** A list of vegetables.

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

**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/](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

**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](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](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](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?](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/firstNames** First names of men and women, 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/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](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/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**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\\_videos\\_broadcasted\\_on\\_MTV](https://en.wikipedia.org/wiki/First_music_videos_broadcasted_on_MTV)**mythology/lovecraft** Deities and supernatural creatures from the works of Lovecraft and the Cthulhu mythos.**objects/objects** List of household objects**plants/flowers****science/elements****science/hail\_size** Analogous objects for various hail sizes, adapted from <http://www.spc.noaa.gov/misc/tables/hailsizes.html>**science/planets** Planets (including dwarf planets as recognized by the IAU) that orbit the Sun, with their natural satellites.

**science/pregnancy**

**science/toxic\_chemicals**

**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.

**words/adjs** A list of English adjectives.

**words/adverbs**

**words/common** Common English words.

**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/literature/mr\_men\_little\_miss** Mr Men and Little Miss characters Source: <http://www.mrmen.com>

**words/literature/shakespeare\_phrases** Phrases 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/prefix\_root\_suffix**

**words/proverbs** A list of proverbs sourced from <http://tw.w.id.au/proverbs/proverbs.html>

**words/states\_of\_drunkenness** A list of states of drunkenness.

**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

categories, [2](#)  
corpora, [2](#)