Open source statistical software at the statistical office

Mark van der Loo
Statistics Netherlands
m.vanderloo@cbs.nl

CREATED WITH FREE SOFTWARE

ISI 2017
MARRAKECH
61st WORLD STATISTICS CONGRESS

What is open source software?

Free and Open Source Software

- Use
- Study
- Change
- Redistribute

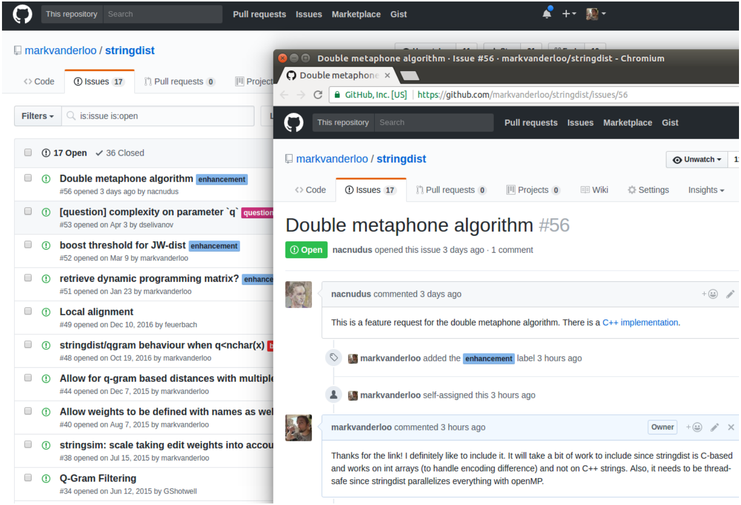
FREE SOFTWARE FOUNDATION
www.fsf.org

FOSS is driving modern stats and data science

julia
node js
R
D3
APACHE Spark
python
hadoop

And much, much more. . .

Communities (1) Social Coding with github



markvanderloo / stringdist

17 Open ✓ 36 Closed

- Double metaphone algorithm enhancement — #56 opened 3 days ago by nacnudus
- [question] complexity on parameter `q` question — #53 opened on Apr 3 by dselivanov
- boost threshold for JW-dist enhancement — #52 opened on Mar 9 by markvanderloo
- retrieve dynamic programming matrix? enhance — #51 opened on Jan 23 by markvanderloo
- Local alignment — #49 opened on Dec 10, 2016 by feuerbach
- stringdist/qgram behaviour when q<nchar(x) — #48 opened on Oct 19, 2016 by markvanderloo
- Allow for q-gram based distances with multipl — #44 opened on Dec 7, 2015 by markvanderloo
- Allow weights to be defined with names as wel — #40 opened on Aug 7, 2015 by markvanderloo
- stringsim: scale taking edit weights into accou — #38 opened on Jul 15, 2015 by markvanderloo
- Q-Gram Filtering — #34 opened on Jun 12, 2015 by GShotwell

Double metaphone algorithm · Issue #56 · markvanderloo/stringdist - Chromium
Double metaphone ×
GitHub, Inc. [US]
https://github.com/markvanderloo/stringdist/issues/56
This repository
Search
Pull requests
Issues
Marketplace
Gist

markvanderloo / stringdist

Double metaphone algorithm #56

Open nacnudus opened this issue 3 days ago · 1 comment

nacnudus commented 3 days ago

This is a feature request for the double metaphone algorithm. There is a C++ implementation.

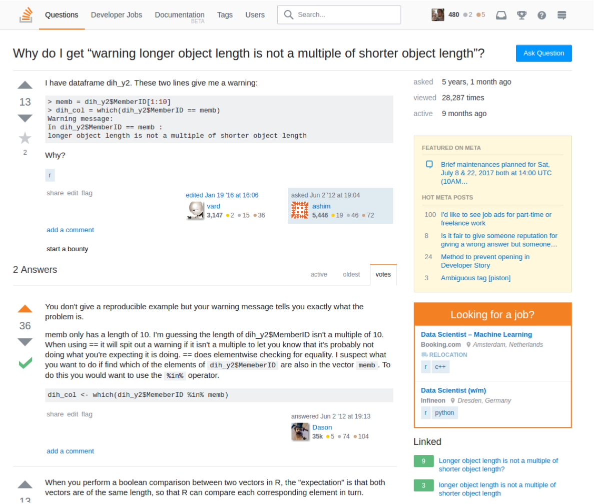
markvanderloo added the enhancement label 3 hours ago

markvanderloo self-assigned this 3 hours ago

markvanderloo commented 3 hours ago Owner

Thanks for the link! I definitely like to include it. It will take a bit of work to include since stringdist is C-based and works on int arrays (to handle encoding difference) and not on C++ strings. Also, it needs to be thread-safe since stringdist parallelizes everything with openMP.

Communities (2) Q&A with stackoverflow



Why do I get “warning longer object length is not a multiple of shorter object length”?

13

2

I have dataframe dih_y2. These two lines give me a warning:

```
> memb = dih_y2$MemberID[1:10]
> dih_col = which(dih_y2$MemberID == memb)
Warning message:
In dih_y2$MemberID == memb :
longer object length is not a multiple of shorter object length
```

Why?

r

share edit flag

edited Jan 19 '16 at 16:06
vard
3,147 ●2 ●15 ●36

asked Jun 2 '12 at 19:04
ashim
5,446 ●19 ●46 ●72

add a comment

start a bounty

asked 5 years, 1 month ago
viewed 28,287 times
active 9 months ago

FEATURED ON META

Brief maintenances planned for Sat, July 8 & 22, 2017 both at 14:00 UTC (10AM...

HOT META POSTS

100 I'd like to see job ads for part-time or freelance work

8 Is it fair to give someone reputation for giving a wrong answer but someone...

24 Method to prevent opening in Developer Story

3 Ambiguous tag [piston]

2 Answers

36

You don't give a reproducible example but your warning message tells you exactly what the problem is.

memb only has a length of 10. I'm guessing the length of dih_y2$MemberID isn't a multiple of 10. When using == it will spit out a warning if it isn't a multiple to let you know that it's probably not doing what you're expecting it is doing. == does elementwise checking for equality. I suspect what you want to do if find which of the elements of `dih_y2$MemeberID` are also in the vector `memb`. To do this you would want to use the `%in%` operator.

```
dih_col <- which(dih_y2$MemeberID %in% memb)
```

share edit flag

answered Jun 2 '12 at 19:13
Dason
35k ●5 ●74 ●104

add a comment

Looking for a job?

Data Scientist – Machine Learning
Booking.com Amsterdam, Netherlands
RELOCATION
r c++

Data Scientist (w/m)
Infineon Dresden, Germany
r python

Linked

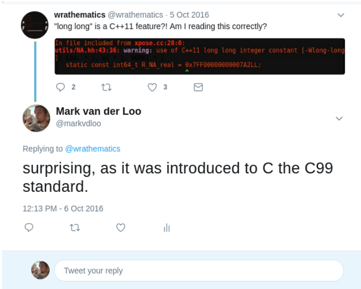
9 Longer object length is not a multiple of shorter object length?

3 longer object length is not a multiple of shorter object length

13

When you perform a boolean comparison between two vectors in R, the "expectation" is that both vectors are of the same length, so that R can compare each corresponding element in turn.

Communities (3) news & discussions on Twitter



wrathematics @wrathematics · 5 Oct 2016

"long long" is a C++11 feature?! Am I reading this correctly?

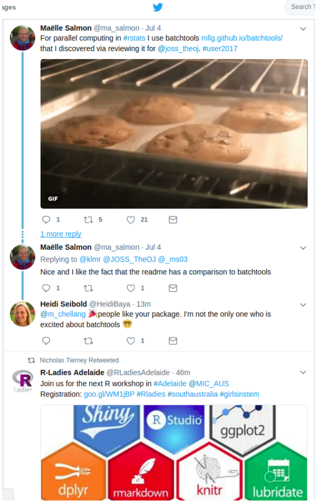
```
In file included from xpose.cc:28:0:
utils/NA.hh:43:36: warning: use of C++11 long long integer constant [-Wlong-long
]( LDA, * ), W( * ), WORK( * )
   static const int64_t R_NA_real = 0x7FF00000000007A2LL;
                                    ^
```

Mark van der Loo
@markvdloo

Replying to @wrathematics

surprising, as it was introduced to C the C99 standard.

12:13 PM - 6 Oct 2016



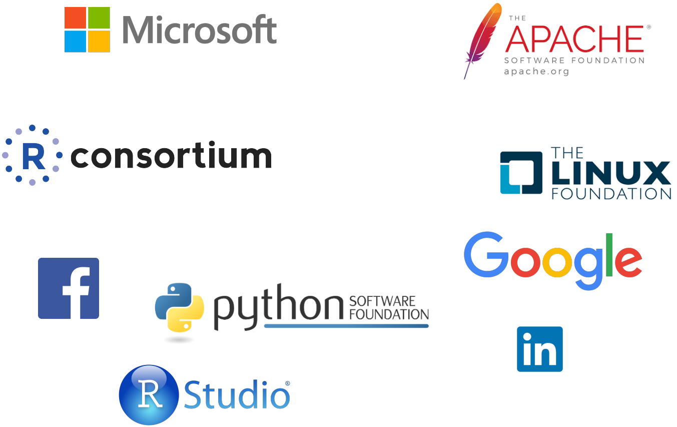
Maëlle Salmon @ma_salmon · Jul 4
For parallel computing in #rstats I use batchtools mllg.github.io/batchtools/ that I discovered via reviewing it for @joss_theoj. #user2017

Maëlle Salmon @ma_salmon · Jul 4
Replying to @klmr @JOSS_TheOJ @_ms03
Nice and I like the fact that the readme has a comparison to batchtools

Heidi Seibold @HeidiBaya · 13m
@m_chellang 🎉people like your package. I'm not the only one who is excited about batchtools 😎

Nicholas Tierney Retweeted
R-Ladies Adelaide @RLadiesAdelaide · 46m
Join us for the next R workshop in #Adelaide @MIC_AUS
Registration: goo.gl/WM1jBP #Rladies #southaustralia #girlsinstem

Role of commercial parties, foundations



And many,many more, . . .

Motivations

Wikipedians

Wikipedians enjoy a sense of accomplishment, collectivism, and benevolence while working with exceptional freedom and ease. The values of reputation, community, reciprocity, altruism and autonomy are fostered by both the people and the technology[. . .]. (Kutzetsnov, 2006)

Commecial parties

[Companies] expect to benefit from their expertise in some segment whose demand is boosted by the success of a complementary open source program. (Learner, 2000)

Motivations for Official Statistics

Use

- Economic (its free!)
- New hires
- Supporting community

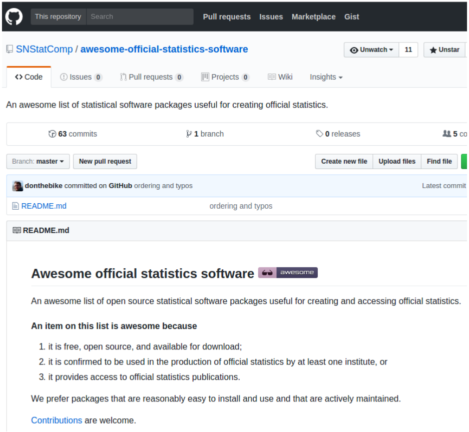
Contribute

- Solving shared problems
- Many eyes make all bugs shallow
- Influence, reputation
- Built with tax payer's money

FOSS for official statistics

Awesome list

- Community effort
- Curated
- 50+ Software packages
- Covering 14 GSBPM areas
- Growing
  - 75 commits
  - 5 PR's
  - 6 contributors



SNStatComp / **awesome-official-statistics-software**

An awesome list of statistical software packages useful for creating official statistics.

63 commits | **1** branch | **0** releases | **5** co

donthebike committed on **GitHub** ordering and typos — Latest commit

README.md — ordering and typos

README.md

Awesome official statistics software

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

An item on this list is awesome because

1. it is free, open source, and available for download;
2. it is confirmed to be used in the production of official statistics by at least one institute, or
3. it provides access to official statistics publications.

We prefer packages that are reasonably easy to install and use and that are actively maintained.

Contributions are welcome.

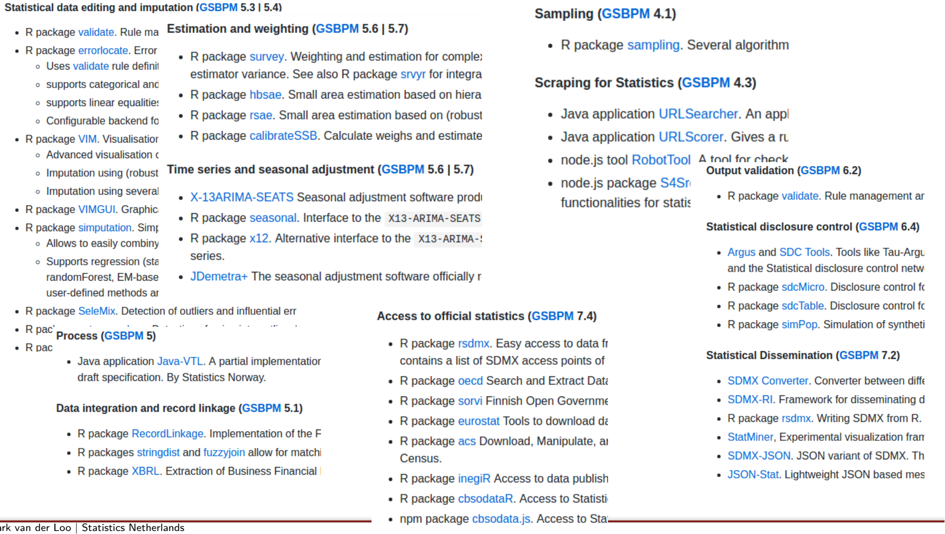
`www.awesomeofficialstatistics.org`

When is it awesome?

You may be awesome when. . .

- Free, open source, available for download
- Used in at least one statistical institute for production or, offers access to official statistics
- Relatively easy to install and use (for non-dev's)
- Actively maintained
- At least one stable release

What's on the Awesome List?



Statistical data editing and imputation (GSBPM 5.3 | 5.4)

- R package validate. Rule ma
- R package errorlocate. Error
  - Uses validate rule definit
  - supports categorical and
  - supports linear equalitie
  - Configurable backend fo
- R package VIM. Visualisatio
  - Advanced visualisation
  - Imputation using (robust
  - Imputation using several
- R package VIMGUI. Graphic
- R package simputation. Simp
  - Allows to easily combin
  - Supports regression (sta randomForest, EM-base user-defined methods ar
- R package SeleMix. Detection of outliers and influential err
- R pac
- R pac

Estimation and weighting (GSBPM 5.6 | 5.7)

- R package survey. Weighting and estimation for comple estimator variance. See also R package srvyr for integra
- R package hbsae. Small area estimation based on hiera
- R package rsae. Small area estimation based on (robust
- R package calibrateSSB. Calculate weighs and estimate

Time series and seasonal adjustment (GSBPM 5.6 | 5.7)

- X-13ARIMA-SEATS Seasonal adjustment software prod
- R package seasonal. Interface to the `X13-ARIMA-SEATS`
- R package x12. Alternative interface to the `X13-ARIMA-` series.
- JDemetra+ The seasonal adjustment software officially r

Process (GSBPM 5)

- Java application Java-VTL. A partial implementatior draft specification. By Statistics Norway.

Data integration and record linkage (GSBPM 5.1)

- R package RecordLinkage. Implementation of the F
- R packages stringdist and fuzzyjoin allow for matchi
- R package XBRL. Extraction of Business Financial

Sampling (GSBPM 4.1)

- R package sampling. Several algorithm

Scraping for Statistics (GSBPM 4.3)

- Java application URLSearcher. An appl
- Java application URLScorer. Gives a ru
- node.js tool RobotTool A tool for check
- node.js package S4Sr functionalities for statis

Access to official statistics (GSBPM 7.4)

- R package rsdmx. Easy access to data fr contains a list of SDMX access points of
- R package oecd Search and Extract Data
- R package sorvi Finnish Open Governme
- R package eurostat Tools to download da
- R package acs Download, Manipulate, ar Census.
- R package inegiR Access to data publish
- R package cbsodataR. Access to Statisti
- npm package cbsodata.js. Access to Sta

Output validation (GSBPM 6.2)

- R package validate. Rule management ar

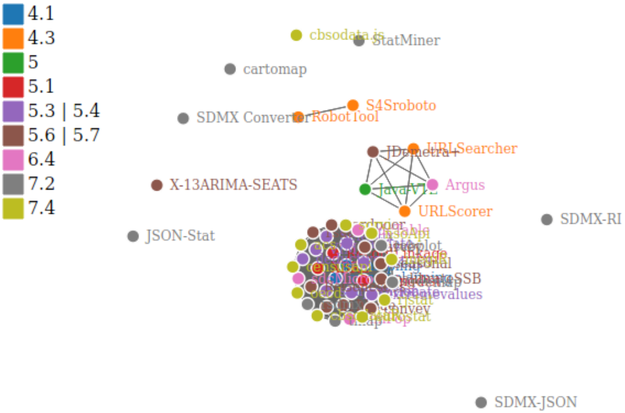
Statistical disclosure control (GSBPM 6.4)

- Argus and SDC Tools. Tools like Tau-Argu and the Statistical disclosure control netw
- R package sdcMicro. Disclosure control fo
- R package sdcTable. Disclosure control fo
- R package simPop. Simulation of syntheti

Statistical Dissemination (GSBPM 7.2)

- SDMX Converter. Converter between diffe
- SDMX-RI. Framework for disseminating d
- R package rsdmx. Writing SDMX from R.
- StatMiner, Experimental visualization fram
- SDMX-JSON. JSON variant of SDMX. Th
- JSON-Stat. Lightweight JSON based mes

In the works. . .


🟦 4.1
🟧 4.3
🟩 5
🟥 5.1
🟪 5.3 | 5.4
🟫 5.6 | 5.7
6.4
⬛ 7.2
🟨 7.4
cbsodata.js
StatMiner
cartomap
S4Sroboto
SDMX Converter
RobotTool
JDemetra+
URLSearcher
X-13ARIMA-SEATS
Argus
Java-VTL
URLScorer
SDMX-RI
JSON-Stat
censusapi
sdcMicro
oecd
rsdmx
tabplot
linkage
SSB
values
rjstat
survey
eurostat
SDMX-JSON

FOSS policy at Statistics Netherlands (in short)

Usage

Selection and introduction follows the same procedure as for COTS (commercial off-the-shelve).

- `R`, `Python`, `node.js`

Contributing

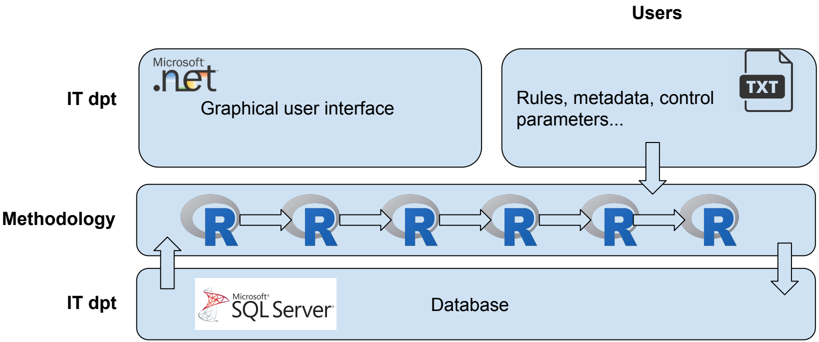
When relevant to Statistics Netherlands, with positive business case.

- `R`, `node.js`

Deployment of R and Python at Statistics Netherlands

- Central read-only folder for executables.
- All users have access to the same version with curated list of libraries installed.
- Scripts can be prepared and integrated for non-developers with ease.
- Repositories (CRAN, Anaconda) on internal website, updated frequently.
- Old versions stay available for some time so existing applications stay working.

Example using R in data editing: tools and roles



Example contributions

- R packages
  - data editing: `validate`, `simputation`, `errorlocate`, `deductive`, `dcmodify`
  - data logging: `lumberjack`
  - small area estimation: `hbsae`
  - datavis: `tabplot`, `tabplotd3`, `tmap`, ...
- node.js packages
  - scraping: `RobotTool`, `S4Robo`
  - dashboard: `StatMine`
- ...

So you want to contribute?

Here are some options

1. Use it (& send a thumbs-up!)
2. Advocate
   - Tell your friends and colleagues
   - Write blog posts / articles / presentations
   - Social media (twitter…)
3. File bug reports, suggestions
4. Add code to an existing project
5. Start your own project

FREE TIP!

Don’t work alone

- Join your local community
  - meetups, news letters, hackatons
- Set up a community in your institute
  - Local wiki, user meetings, hackatons, ask-the-expert