

The digest Package

February 16, 2008

Version 0.3.1

Date \$Date: 2007/09/28 17:54:44 \$

Author Dirk Eddelbuettel <edd@debian.org> with contributions by Antoine Lucas, Jarek Tuszynski, Henrik Bengtsson and Simon Urbanek

Maintainer Dirk Eddelbuettel <edd@debian.org>

Title Create cryptographic hash digests of R objects

Description The digest package provides functions for the creation of 'hash' digests of arbitrary R objects using the md5, sha-1 and

Depends R (>= 2.4.1)

License GPL-2

URL <http://www.cr0.net:8040/code/crypto/>, ftp://ftp.rocksoft.com/cliens/rocksoft/papers/crc_v3.txt

R topics documented:

digest	1
Index	5

digest	<i>Create hash function digests for arbitrary R objects</i>
--------	---

Description

The `digest` function applies a cryptographical hash function to arbitrary R objects. By default, the objects are internally serialized, and either one of the currently implemented MD5 and SHA-1 hash functions algorithms can be used to compute a compact digest of the serialized object.

In order to compare this implementation with others, serialization of the input argument can also be turned off in which the input argument must be a character string for which its digest is returned.

Usage

```
digest(object, algo=c("md5", "sha1", "crc32"),
       serialize=TRUE, file=FALSE, length=Inf, skip="auto", ascii=FALSE)
```

Arguments

<code>object</code>	An arbitrary R object which will then be passed to the <code>serialize</code> function, unless the <code>serialize</code> argument is set to <code>FALSE</code>
<code>algo</code>	The algorithms to be used; currently available choices are <code>md5</code> , which is also the default, <code>sha1</code> and <code>crc32</code>
<code>serialize</code>	A logical variable indicating whether the object should be serialized using <code>serialize</code> (in ASCII form). Setting this to <code>FALSE</code> allows to compare the digest output of given character strings to known control output. It also allows the use of raw vectors such as the output of non-ASCII serialization.
<code>file</code>	A logical variable indicating whether the object is a file name or a file name if <code>object</code> is not specified.
<code>length</code>	Number of characters to process. By default, when <code>length</code> is set to <code>Inf</code> , the whole string or file is processed.
<code>skip</code>	Number of input bytes to skip before calculating the digest. Negative values are invalid and currently treated as zero. Special value <code>"auto"</code> will cause serialization header to be skipped if <code>serialize</code> is set to <code>TRUE</code> (the serialization header contains the R version number thus skipping it allows the comparison of hashes across platforms and some R versions).
<code>ascii</code>	This flag is passed to the <code>serialize</code> function if <code>serialize</code> is set to <code>TRUE</code> , determining whether the hash is computed on the ASCII or binary representation.

Details

Cryptographic hash functions are well researched and documented. The MD5 algorithm by Ron Rivest is specified in RFC 1321. The SHA-1 algorithm is specified in FIPS-180-1. Crc32 is described in ftp://ftp.rocksoft.com/cliens/rocksoft/papers/crc_v3.txt.

For `md5` and `sha-1`, this R implementation relies on two standalone implementations in C by Christophe Devine. For `crc32`, code from the `zlib` library by Jean-loup Gailly and Mark Adler is used.

Please note that this package is not meant to be used for cryptographic purposes for which more comprehensive (and widely tested) libraries such as OpenSSL should be used. Also, it is known that `crc32` is not collision-proof. For `sha-1`, recent results indicate certain cryptographic weaknesses as well. For more details, see for example http://www.schneier.com/blog/archives/2005/02/cryptanalysis_o.html.

Value

The `digest` function returns a character string of a fixed length containing the requested digest of the supplied R object. For MD5, a string of length 32 is returned; for SHA-1, a string of length 40 is returned; for CRC32 a string of length 8.

Author(s)

Dirk Eddebuettel (edd@debian.org) for the R interface; Antoine Lucas for the integration of `crc32`; Jarek Tuszynski for the file-based operations; Henrik Bengtsson and Simon Urbanek for improved serialization patches; Christophe Devine for the hash function implementations for `sha-1` and `md5`; Jean-loup Gailly and Mark Adler for `crc32`.

References

MD5: <http://www.ietf.org/rfc/rfc1321.txt>.
 SHA-1: <http://www.itl.nist.gov/fipspubs/fip180-1.htm>.
 CRC32: ftp://ftp.rocksoft.com/clients/rocksoft/papers/crc_v3.txt.
<http://www.cr0.net:8040/code/crypto> for the underlying C functions used here for `sha-1` and `md5`, and further references.
<http://zlib.net> for documentation on the `zlib` library which supplied the code for `crc32`.

See Also

[serialize](#), [md5sum](#)

Examples

```
## Standard RFC 1321 test vectors
md5Input <-
  c("",
    "a",
    "abc",
    "message digest",
    "abcdefghijklmnopqrstuvwxy",
    "ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxy0123456789",
    paste("123456789012345678901234567890123456789012345678901234567890123456789012",
          "345678901234567890", sep=""))
md5Output <-
  c("d41d8cd98f00b204e9800998ecf8427e",
    "0cc175b9c0f1b6a831c399e269772661",
    "900150983cd24fb0d6963f7d28e17f72",
    "f96b697d7cb7938d525a2f31aaf161d0",
    "c3fcd3d76192e4007dfb496cca67e13b",
    "d174ab98d277d9f5a5611c2c9f419d9f",
    "57edf4a22be3c955ac49da2e2107b67a")

for (i in seq(along=md5Input)) {
  md5 <- digest(md5Input[i], serialize=FALSE)
  stopifnot(identical(md5, md5Output[i]))
}

sha1Input <-
  c("abc",
    "abcdcbdecdefdefgefghfghighijhijkjklklmlnlnomnopnopq",
    NULL)
```

```

sha1Output <-
  c("a9993e364706816aba3e25717850c26c9cd0d89d",
    "84983e441c3bd26ebaae4aa1f95129e5e54670f1",
    "34aa973cd4c4daa4f61eeb2bdbad27316534016f")

for (i in seq(along=sha1Input)) {
  sha1 <- digest(sha1Input[i], algo="sha1", serialize=FALSE)
  stopifnot(identical(sha1, sha1Output[i]))
}

crc32Input <-
  c("abc",
    "abcdbcdecdefdefgefghfghighijhijkijkljklmklmnlmnomnopnopq",
    NULL)
crc32Output <-
  c("352441c2",
    "171a3f5f",
    "2ef80172")

for (i in seq(along=crc32Input)) {
  crc32 <- digest(crc32Input[i], algo="crc32", serialize=FALSE)
  stopifnot(identical(crc32, crc32Output[i]))
}

# one of the FIPS-
sha1 <- digest("abc", algo="sha1", serialize=FALSE)
stopifnot(identical(sha1, "a9993e364706816aba3e25717850c26c9cd0d89d"))

# example of a digest of a standard R list structure
digest(list(LETTERS, data.frame(a=letters[1:5], b=matrix(1:10,ncol=2))))

# test 'length' parameter and file input
fname = file.path(R.home(), "COPYING")
x = readChar(fname, file.info(fname)$size) # read file
for (alg in c("sha1", "md5", "crc32")) {
  # partial file
  h1 = digest(x, length=18000, algo=alg, serialize=FALSE)
  h2 = digest(fname, length=18000, algo=alg, serialize=FALSE, file=TRUE)
  h3 = digest( substr(x,1,18000), algo=alg, serialize=FALSE)
  stopifnot( identical(h1,h2), identical(h1,h3) )
  # whole file
  h1 = digest(x, algo=alg, serialize=FALSE)
  h2 = digest(fname, algo=alg, serialize=FALSE, file=TRUE)
  stopifnot( identical(h1,h2) )
}

# compare md5 algorithm to other tools
library(tools)
fname = file.path(R.home(), "COPYING")
h1 = as.character(md5sum(fname))
h2 = digest(fname, algo="md5", file=TRUE)
stopifnot( identical(h1,h2) )

```

Index

*Topic **misc**

digest, 1

digest, 1

md5sum, 3

serialize, 2, 3