Binary Data

Jim Harner

1/6/2020

2.4 Binary Data

Material in this section (except for subsection 2.4.1) is based on material in Paul Murrell's Introduction to Data Technologies.

A binary format is a more complex storage solution than a plain text format. However, it:

- provides faster and more flexible access to the data;
- uses up less memory.

A file with a binary format is simply a block of computer memory, just like a file with a plain text format. The difference lies in how the bytes of computer memory are used.

Integers

Bytes can also represent *integers*.

- 00000001 represents 1
- 00000010 represents 2
- 00000011 represents 3, etc.

8 bits can represent $2^8 = 256$ integers (0 to 255). 48 can be represents in 1 byte, but it takes two bytes if it is represents as plain text. We still cannot store large integers. For example, in two bytes we could store integers with a max value of $2^{16} - 1 = 65,535$ and in four bytes we could store integers with a max value of $2^{32} - 1 = 4,294,967,295$. In general, for k bits the maximum integer is $2^k - 1$. For signed integers we can store the range of integers $\pm 2^{15} - 1$ in two bytes.

Real Numbers

A byte can also store *real numbers*. In practice at least 32 bits are used to store real numbers. The correspondence between bit patterns and real numbers is not intuitive.

Floating point is a method of representing real numbers. Real numbers are represented approximately by a fixed number of significant digits and scaled using an exponent.

Double precision is a binary format that uses 64 bits (8 bytes) and its significance has a precision of 53 bits or about 16 decimal digits. R uses double precision arithmetic.

Single precision is a binary format that uses 32 bits (4 bytes) and its significance has a precision of 24 bits or about 7 decimal digits.

2.4.1 BSON

BSON is a data interchange format used mainly as a data storage and network transfer format in the MongoDB database. It is a binary form for representing simple data structures, associative arrays (called objects or documents in MongoDB), and various data types of specific interest to MongoDB. The name "BSON" is based on the term JSON and stands for "Binary JSON"

MongoDB is a document-based NoSQL database. MongoDB stores data in documents with dynamic schemas.

Two R packages are available for providing the interface with MongoDB, namely `RMongo` and `rmongodb`.

2.4.2 NetCDF

The Point Nemo temperature data is stored, among other ways, in network Common Data Form (netCDF), a binary format. It is open source, whereas many binary formats are proprietary. A netCDF file begins with a header followed by the raw data. The netCDF starts with the characters: C, D, F. The fourth byte is the version number.

The Point Nemo data is also stored in netCDF format. It would be read as:

```
library(ncdf)
nemonc <- open.ncdf("pointnemotemp.nc")
nemonc
class(nemonc)
nemoTemps <- get.var.ncdf(nemonc, "Temperature")
nemoTemps
```

However, we have not installed the `ncdf` library on `gbef4001`, but you should be able to run the code on your local machine.

The header file contain pointers to the locations of the raw data and contains information on how the raw data are stored. For the Point Nemo data, the raw data starts at byte 624 and each temperature is stored as an 8 byte real number.

It is possible to calculate the location of a particular data value in the file. This cannot be done with a text-based format.

Generally, for a binary format, it is possible to jump to a specific location in the file. This is called *random access*. A text-based file must be read from beginning to end. This is called *sequential access*.

2.4.3 PDF documents

Many documents are now published in Adobe's Portable Document Format (PDF). PDF may contain tables, which is how the data may be received.

A PDF document is primarily a description of how the data should be displayed. Values are intertwined with this information. It might be possible to cut-and-paste tabular values, but access through an API is difficult.

2.4.4 Dates

Dates can be stored as text or as a number, e.g., the number of days since Jan. 1970. Storing dates as numbers allows calculations to be done on the dates. International standards are used to represent dates, e.g., 2006-03-01.

Date-times can also be represented.