COMP 499 Introduction to Data Analytics

Lecture 4 — Data Formats & Schemas; Python pandas

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

1. Data Formats
2. Data Schemas
3. Metadata
4. Self-Descriptive Data

Data Formats

comma-separated values (csv)

Tab-separated values (tsv)

Attribute-Relation File Format (ARFF)

XML

RDF

Binary files (BLOBs)

HDF5 (Hierarchical Data Format version 5)

Data Formats — ARFF — Weka

ARFF files

ASCII files: Header followed by Data

Header

- the name of the relation,
- a list of the attributes (columns in data),
- their types

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Data Formats — ARFF

Data looks like

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Data Schemas

Tidy Data Schema in R

Tabular format with properties

1. Each variable is saved in its own column
2. Each observation is saved in its own row
3. Each type of observation is stored in its own (single) table

See video

https://www.youtube.com/watch?v=1ELALQlO-yM&list=PL9HYL-VRX0oQOWAFoKHFQAsWAI3ImbNPk&index=2

Metadata

Metadata

is data that provides information about other data

For example

Means of creation of the data
Purpose of the data
Time and date of creation
Creator or author of the data
Location on a computer network where the data was created
Standards used
File size
Data quality
Source of the data
Process used to create the data

Provenance of Data

is the origin and/or history of an object (that is, data, in our case).

Self-Descriptive Data

You can make sense of the file as a stand-alone.

therefore human-readable

ARFF
XML
HDF