Universiteti i Prishtinës "Hasan Prishtina" Kosovë

Hyrje në shkencën e të dhënave

Pjesa 2 – Të dhënat

Prof. Asoc. Dr. Ermir Rogova

Data types

Structured

Example: tables

Unstructured

Example: free text

Structured data example

custid	sex	is.employed	income	marital.stat	housing.type	num.vehicles	age	state.of.res
2068	F	NA	11300	Married	Homeowner free and clear	2	49	Michigan
2073	F	NA	0	Married	Rented	3	40	Florida
2848	M	TRUE	4500	Never Married	Rented	3	22	Georgia
5641	M	TRUE	20000	Never Married	Occupied with no rent	0	22	New Mexico
6369	F	TRUE	12000	Never Married	Rented	1	31	Florida

Data collections

Lots of places that host/share data online, or you can collect them yourself.

Open data collections

Social media data

Multimodal data

Structured data file types

- CSV (Comma Separated Values)
- TSV (Tab Separated Values)
- XML (eXtensible Markup Language)
- RSS (Really Simple Syndication)
- JSON (JavaScript Object Notation)

Data preprocessing

- Data cleaning
  - Data wrangling
  - Handling missing data
  - Smooth noisy data
- Data integration
- Data transformation
- Data reduction
- Data discretization

Hands-on with data preprocessing

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg/Lux	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	-0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	-834	183	13.69999981
11	Italy	27.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.0800000012	806	227	12.19999981
16	Spain	6.5	724	NA	NA
17	Sweden	1.600000024	743	207	11.19999981
18	Switzerland	5.800000191	693	115	20.29999924
19	UK	1.299999952	941	285	10.30000019
20	US	1.200000048	926	199	22.10000038
21	West Germany	2.700000048	861	172	36.70000076

Summary

- Two primary types of data: structured, unstructured
- Pros and cons of structured and unstructured data
- Data collections
  - Open data
  - Social media data
  - Multimodal data
- Stages of data pre-processing:
  - Cleaning
  - Integration
  - Transformation
  - Reduction
  - Discretization

Temat e javës së tretë

- Data Analysis and Data Analytics
- Descriptive Analysis
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics
- Exploratory Analysis
- Mechanistic Analysis

Pyetje???