CS286 Lab 1: Working with Large Datasets

K. A. Tarnowska

24-Aug-2020

Contents

1 Introduction

In this Lab, you will learn about the concept of open data, its main characteristics, types, and formats. Your task will be to explore data sources listed and choose one publicly available dataset to work on in Labs 1-4. In the report, you will describe the chosen dataset and analyze the decision-making process involved in it.

1.1 Open data

The idea behind open data is that some data should be freely available in a public domain that can be used by anyone as they wish, without restrictions from copyrights, patents, or other mechanisms of control. Following is the list of principles associated with open data:

- Public - open to the extent permitted by law
- Accessible - made available in convenient, modifiable, and open formats
- Described - described fully to provide sufficient information to understand the data
- Reusable - no restrictions on the use
- Complete - with the finest possible level of granularity
- Timely - made available as quickly as necessary
- Managed post-release - a point-of-contact must be designated to assist with data use

1.2 Data types

One of the most basic ways to think about data is whether it is structured or not.

- **Structured data** - highly organized information that can be seamlessly included in a database and readily searched, for example, data in spreadsheets. Different values are labeled. We will be working with structured data in Labs 1-5
- **Unstructured data** - data devoid of any underlying structure, i.e. text; data without labels. For example, e-mail is unstructured data. We will be working with unstructured data in Labs 6-10

1.3 Data formats

Depending on its nature, data is stored in various formats. The most common data formats are:

- **CSV** (Comma-Separated Values) - is the most common import and export format for spreadsheets and databases
- **TSV** (Tab-Separated Values) - text files with the tab as the delimiter between data values
- **XML** (eXtensible Markup Language) was designed to be both human- and machine-readable; provides a software- and hardware-independent way of storing data that can be shared by different applications. Similar to HTML, but using custom tags.
- **RSS** (Really Simple Syndication) - format used to share data between services; facilitates the delivery of information from various sources on the Web
- **JSON** (JavaScript Object Notation) is a lightweight data-interchange format; based on the subset of the JavaScript programming language. JSON is built on two structures: (1) a collection of name-value pairs; (2) an ordered list of values.

2 Data Science Resources

Now, if you wanted to find datasets for the applications as presented in the lecture, where would you look? There are many places online to look for sets or collections of data. Table 2 lists some of those sources.

3 Tasks

3.1 Datasets exploration and choice

1. Explore the resources listed in Section 2.
2. Think of a decision problem of interest to you.
    - It can be related to your domain expertise and interests, i.e. if you are taking a minor in Business, business datasets might be of interest to you.
    - It can be related to your interests within Computer Science, i.e. information security.
    - It can be related to the domain/industry you have worked in, i.e. if you worked for a biopharmaceutical company, medical datasets might be of interest to you.

Table 1: Data Science Resources

Source	Description
University of California, Irvine Repository: https://archive.ics.uci.edu/ml/datasets.html	Site includes 325 validated datasets covering many domains, different sizes and data types and different analytical methods.
Kaggle: www.kaggle.com	Provides free, interesting datasets for various user interests and analysis.
KDNuggets: www.kdnuggets.com	Site includes 71 datasets available for free download, from various industries.
The Datahub: https://datahub.io/dataset	Managed by the Open Knowledge Foundation, this site hosts more than 10,000 datasets from most industries.
DATA USA www.datausa.io	Organized by maps, cities/places, jobs and downloads
data.world https://data.world	New site for creating collaborative data projects with ability to host data and analyze with embedded SQL.
SAS Public datasets https://semanticommunity.info/Data_Science/SAS_Public_Data_Sets#SAS_Exercises_Slide_Numbers	SAS provides a library of statistical test examples using easy to understand data and associated charts
OpenML https://openml.org	Hosts a variety of datasets and ML workflows. Includes the results of multiple ML algorithms run
Google dataset Search https://toolbox.google.com/datasetsearch	Google search feature for a variety of datasets
Google datasets https://ai.google/tools/datasets/	Wide offering generally used to train on with machine learning and AI
Microsoft Open Datasets https://msropendata.com	Covers categories of computer science, information science, physics and social science
Open Datasets on Git Hub https://github.com/awesomedata/awesome-public-datasets	Extensive list in multiple categories

- Or just think about the niche/unique domain which you would like to augment with evidence-based data-supported decision-making, i.e. it can be related to current events, such as political campaigns, pandemic.

3. Find datasets relevant to your domain interests discovered in the previous Step.
4. Describe a few datasets (no more than 5) relevant to your interests. Include source URLs.
5. Make a choice of one dataset, justify your choice, and complete the next tasks based on the dataset you chose. Include source URL.

3.2 Working with a chosen dataset

1. Describe the problem area/domain (1 paragraph). Why have you chosen it? What is the importance/significance of this area?
2. Describe the decision-making process involved (1 paragraph). What variables are involved in making the decision? Are they equally weighted? Is there any particular schema involved int he decision-making? What is the decision?
3. Describe the chosen dataset: what is the format, how many rows/columns does it have? What are the attributes and their meaning?

Examples of domain description and problem statement can be found in [1] and [2].

4 Deliverables

Lab report uploaded on Canvas (pdf preferred) with URLs embedded.

5 Grading

The lab will be graded according to the rubric in Table 2.

6 Self-check

Answer the following questions:

1. What is the concept of open data, and what are its main characteristics?
2. What are the two different types of data? Give on eexample of each.
3. What are the common formats for data collections?
4. List at least three public dataset repositories.

Table 2: Grading rubric

Rubric	Criteria	Points
Identified datasets	Identified at least 3 data sources and provided URLs	1
Dataset choice	Provided justification for one chosen dataset	1
Problem area	Complete and concise description of the problem area, including motivation, importance, and significance	3
Decision-making process	Described the decision-making process, variables, and the decision schema involved	3
Technical description	Provided description of dataset format, dimensionality, and the attributes	2
Total		**10**

References

[1] Tarnowska KA, Ras ZW, Jastreboff PJ. Tinnitus Treatment as a Problem Area, Chapter 2 in: Decision Support System for Diagnosis and Treatment of Hearing Disorders. The Case of Tinnitus. Studies in Computational Intelligence. Springer, 2017

[2] Tarnowska K, Ras ZW, Daniel L. Introduction/Customer Loyalty Improvement, Chapter 1/2 in: Recommender System for Improving Customer Loyalty. Studies in Big Data. Springer, 2019