DSCI-549: Introduction to Computational Thinking and Data Science

Term: Spring 2021

USC Viterbi School of Engineering

Homework 1

Date due: January 28, 2021 **Date out: January 19, 2021**

Goals

The purpose of this homework is to familiarize yourself with datasets that are publicly available on the Web, understand how to get a hold of them, and explore possible uses of the datasets by formulating interesting questions that could potentially be answered by analyzing them.

This homework is Part 1 of a class project that you will do in this class, consisting of a total of 3 homeworks.

Assignment

Write a short paper (3 pages maximum) addressing the following topics:

1. A short summary (1/2 page to 1 page) of:
    a. Your interests,
    b. The reasons why you choose your current degree and major,
    c. The reasons why you decided to take this class, and
    d. Your personal ambitions to change the world.

2. Find one or more datasets on the Web that are relevant to your interests, and that are: 1) accessible, 2) released with an open license, and 3) machine processable. For each dataset, specify:
    a. A brief description of what the data represents, with a small excerpt of the data as an example
    b. A discussion on how the dataset satisfies those three criteria
    c. The details for how you can access the data

3. Sketch out a simple data science project:
    a. State some interesting questions, and discuss how they could be answered by analyzing the data that you listed in the previous question
    b. Describe in your own words what kinds of analysis could be done with the data to answer each question. Be specific about what the output of the analysis would be. Clarify whether additional data would be needed to answer the questions, or if you would need help from someone with specific kinds of expertise (e.g., statistics, distributed systems, large-scale data processing, etc.)

To help you find good data sources, the next two pages suggest web sites that contain a wide range of data.

IMPORTANT NOTES

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students.

For more information, see the class syllabus and the USC web site.

Resources for Finding Datasets

Compiled by Kate Musen, June 2015

Source	Link	US only?	Topics
The Humanitarian Data Exchange (HDX)	https://data.hdx.rwlabs.org	International	Economy, gender, education, health, human population, emergency telecommunications, humanitarian funding, water sanitation and hygiene, infrastructure, food and nutrition, disaster relief
The World Bank	http://data.worldbank.org	International	Similar to HDX, except all data sets are compiled by the World Bank, listed topics include: agriculture, foreign aid, climate change, economics, economic development, education, energy & mining, finance, gender, health, infrastructure, poverty, science & technology, labor, trade
Data.gov	http://www.data.gov	Only United States	Agriculture, business, climate, consumer, ecosystems, finance, health, local government, manufacturing, oceans
The National Bureau of Economic Research	http://www.nber.org/data/	Only United States	Economics, finance, demographics, limited healthcare
UNdata	https://data.un.org/Explorer.aspx	International	Trade, energy, gender, environment, economic development, labor, children, homicides, crime, demographics, reproductive health, marriage and family, population, telecommunications, tourism
Nasa's Data Portal	https://data.nasa.gov/data	N/A	Earth and space science, also allows you to search by format of dataset

Pew Research Center	http://www.pewinternet.org/datase ts/	N/A	Health, Politics, teens and mobile phones, Libraries. To get the datasets requires an email address. Not very organized.

National Sleep Research Resource	https://sleepdata.org/datasets	U.S.	Sleep
National Oceanic and Atmospheric Administration	http://sos.noaa.gov/Datasets/inde x.html	U.S.	Air, Water, Space, Land, People Includes many forms of data, not just tabular. A lot of image and video data.
data.gov.uk	http://data.gov.uk/	UK	If one wants to compare UK data to US data, use this. Canada also has opendata.
FDA	http://www.fda.gov/AboutFDA/Tra nsparency/OpenGovernment/defa ult.htm	US	Food
Carnegie Mellon	http://lib.stat.cmu.edu/DASL/alltopi cs.html	Mixed	Lots of topics, click on the link to see data organized by topic.
International Association for Pattern Recognition	http://homepages.inf.ed.ac.uk/rbf/IAPR/researchers/MLPAGES/mlda t.htm	International	List of pages with machine learning datasets.
University of Michigan Professors' Personal Website	http://wwwpersonal.umich.edu/~ mejn/netdata/	N/A	Lot of different network data.
UC Irvine Machine Learning datasets	https://archive.ics.uci.edu/ml/datas ets.html	N/A	Lots of data for learning from positive and negative examples of a concept.
Additional: Kaggle	https://www.kaggle.com/datasets	N/A	Many datasets for machine learning prediction competition