May 07, 2019

Working

edX Learner and Course Analytics and Visualization Pipeline

PLOS One

Michael Ginda[1], Katy Borner[1], Michael Richey[2], Mark Cousino[3]

[1]Indiana University at Bloomington, [2]The Boeing Corportation, [3]The Boeing Corporation

dx.doi.org/10.17504/protocols.io.zcdf2s6

Michael Ginda
Indiana University at Bloomington

ABSTRACT

The edX Student and Course Analytics and Visualization Pipeline is analytics and visualization pipeline using edX course database and user logs, written in R to 1) to extract and process student users and performance data, course structures and event logs; 2) create learner trajectory networks of use and pathways through course content and activity modules; 3) analyze the students use of course content modules; and 4) aggregate student performance and interaction measurements for a given course.

EXTERNAL LINK

https://github.com/cns-iu/edx-learnertrajectorynetpipeline

GUIDELINES

The processing scripts are provided under Apache License 2.0. Contributors provide permission for commercial use, modification, distribution, patent use, and private use. Licensed works, modifications, and larger works may be distributed under different terms and without source code. The script is provided with a limited liability and warranty, use these data processing scripts at your own discretion, and make preservation copies of any source data prior to use.

Additional modifications are likely needed to make use of this pipeline when processing other course data sets that use the edX Data Package format specification.

Organizational implementation of the edX learning management systems may use customized event log tracking systems, and courses may use different types of edX block modules, and logs may include types of events that were not encountered in this project (e.g., error events, or edX discussion forums).

BEFORE STARTING

The scripts also rely on standard edX data export directory structure, which may vary based on local implementations and data provided by an organizations' data czar. An exploratory analysis of the course structure and event logs is advisable at the outset of a project, as well as a comparison of the file names provided and the names expected by the scripts.

The pipeline's scripts create a set of directory structures for processing and analysis outputs, which may be modified, updated in the scripts. These changes should be made after review across all scripts before changes are made, to ensure that processing, analysis and visualization run smoothyl.

Acquire edX Data

1

Review edX Research Guide, which provides documentation for how a user may acquire edX course data from an edX Data

Czar, how to properly and responsibly maintain and use these data in research, and describe in detail the data exports provided for a given edX courses.

The edX Research guide is available for review at https://edx.readthedocs.io/projects/devdata/en/latest/index.html.

When working with data that is outside of your home organization, it is essential to set up a data use agreement (DUA) between the organization holding the data and your institution, in addtion to an IRB. This process must be completed before you can access learner data, which contains information that can be used to re-identify learners.

Data will be provided in a ZIP format that will need to be extracted in Step 3.

edX Data Package

Set up project

2 Install R statistical software

Install R programming language and R Studio Desktop to run the scripts used in this pipeline. Once these pieces of software are installed, you will need to run R Studio. In R studio, you will need to install from the package manager:

- tcltk2 - a GUI interface used to set paths to access data used in the pipeline.
- jsonlite - package used for parsing JSON data
- ndjson - package used for parsing streaming JSON data used in edX event logs.
- Hmisc - data analytics functions
- plyr - data aggregation
- reshape2 - data reshaping functions
- magrittr - piping functions
- stringr - string processing and manipulation
- ggplot2 - statistical visualization package

R programming language 1.3.3.3 or later

source by The R Foundation

R Studio Desktop 1.0.44 or later

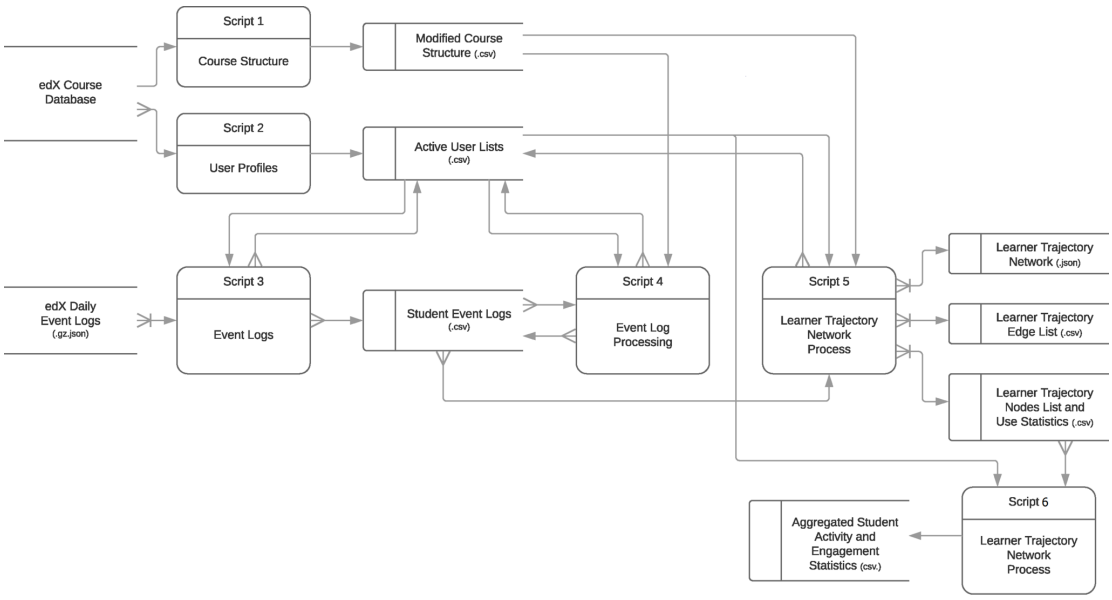
by The R Studio, Inc.

3 Set up the project directory workspaces that separates course level data provided in the edX data package, and a space for processing outputs, analysis and visualization results. An example directory structure used in this protocol

4 Extract edX Data Package into user identified directory space.

Processing edX Course Data

5 Review the processing scripts used in this analysis to understand how they function alone and as part of the overall pipeline. The UML process flow diagram provides an overview of how original data and processing outputs are used between scripts in the pipeline.



UML process flow diagram for the edX Learner and Course Analytics and Visualization Pipeline

6 Load and Run Script edX-1-courseStructureMeta.R in R Studio.

This script extracts a the course structure from the edX Data Package filesis used in processing log files and creating the node lists in learner trajectory networks.

Make sure to set accurate pathways for reading in original data and saving processing outputs. Note that the script will create directories for processing results automatically, unless these are removed or commented out by a user.

7 Load and Run Script edX-2-studentUserList.R

8 Load and Run Script edX-3-eventLogExtractor.R

9 Load and Run Script edX-4-eventLogFormatter.R.

10 Load and Run Script edX-5-learnerTrajectoryNet.R.

11 Load and Run Script edX-6-moduleUseAnalysis.R.

12 Load and Run Script edX-7-studentFeatureExtraction.R.