A package for Cleaning and Analyzing Coursera OnDemand Data

by Aboozar Hadavand, Jeffrey Leek

Abstract An abstract of less than 150 words.

Introduction

Why this paper?

- talk about moocs
- one challenge in studying moocs is data
- even when data are available, the difficulty of data analysis makes it hard for researcher
- therefore, I provided this paper
- moocs have become massive laboratory for pedagogy
- in this paper I introduce the package and run an analysis of students progress using the package.
- statistics on coursera

It is hard to pin down the time of the birth of the first Massive Open Online Course (MOOC).[1] But since the advent of more focused MOOCs pioneered by universities and platforms such as Coursera, Udacity, and edX, reserachers have tried to focus on studying MOOCs. There are fundamental differences between traditional education and MOOCs was large enough to attract reserachers to study students' behavior and outcomes. These differences are best reflected in the definition of MOOCs by **?** that "[a]n online course with the option of free and open registration, a publicly shared curriculum, and open-ended outcomes which integrates social networking, accessible online resources ... and most significantly builds on the engagement of learners who self-organize their participation according to learning goals, prior knowledge and skills, and common interests."

Research on MOOCs few years with more data being accumulated and collected. **?** studied literature published on MOOCs throught 2015 and found that the number of articles published on the subject increased from 1 in 2008 to 170 in 2015. More research in needed to fully understand the effectiveness, reach, limits, and the potential of MOOCs. However, one of the main challenges in studying MOOCs remains to be data. Data is not usually publically available since it is owened by private MOOC providers and there are concerns about privacy of students. More importantly, as **?** point out, the size and complexity of MOOC data is an overwhelming challenge to many researchers. Therefore, it is imperative to provide tools that pave the way for more research on the new subject of MOOCs.

This paper introduces a package called *crsra* based on the statistical sofware R to help clean and analyze large loads of data from the Coursera MOOCs. The advantages of the package are as follows: a) faster loading of data for analysis, b) efficient method for combining data from multiple courses and even across institutions,[2] and c) provision of a set of functions for analysing student behaviors.

Coursera On-Demand Data

Coursera is one of the main providers of MOOCs that launched in January 2012. In fact, with over 25 million learners, Coursera is the biggest provider in the world being followed by EdX, the MOOC provider that was a result of a collaboration between Harvard Universit and MIT, with over 10 million users. Coursera has over 150 uiveristy partners from 29 countries and offers a toatl of 2000+ courses from computer science to philosophy (**?**). In addition, Coursera offers 180+ specialization, Coursera's own credential system, and 4 fully online Masters degrees. Courses include recorded video lectures, graded assignment, quizzes, and discussion forums.

when coursera on-demand data started. history of coursera sharing their data. different phases. they provide it to partners write about data from coursera's gitbook what kind of demographic size of data examples of hopkins courses provide a list of tables that connect different students.

[1] Some have claimed Sesame Street as the first MOOC. Delaney Parrish, "Sesame Street was the original MOOC," *BROOKINGS NOW*, The Brookings Institution, June 18, 2015, https://www.brookings.edu/blog/brookings-now/2015/06/18/sesame-street-was-the-original-mooc/

[2] This is important since although MOOC researchers have access to thousands of students in their sample, few studies benefit from data across multiple courses and institutions. Such analysis helps draw more robust conclusions about student behaviors (**?**).

Talk about the advantage of not having relational databses

crsra Package

talk about the package

Analysis of student behavior on Coursera

provide the analaysis here

Discussion

Aboozar Hadavand
Bloomberg School of Public Health, Johns Hopkins University
615 N. Wolfe Street
Baltimore, MD 21205, USA
hadavand@jhu.edu

Jeffrey Leek
Bloomberg School of Public Health, Johns Hopkins University
615 N. Wolfe Street
Baltimore, MD 21205, USA
jtleek@jhu.edu