MOOC_CS50

This is the repository for the CS50 MOOC analysis. The primary goal of this analysis is to

1. describe the basis information of CS50 MOOC participants;
2. Graphically present the trajectories of the MOOC participants moving from one milestone to another milestone.
3. Survival analysis (for regular participants) to find out predictors for dropout.

September, 22, 2017

As of 09/22/2017, the following *dataset* has been uploaded to the repository:

- CS50_raw_completed_pretest.RData and CS50_ChosenUsers_1_irt.csv : these two datasets are the same. They are the original dataset including all participants who had finished the pre-test. *Those who did not finish the pre-test isn't included, and will be included in a separate dataset in the future.*
- CS50_node_edge_link_incl_dropout_earlyfinal_popular50.RData: the dataset that include the node, edge and link (link is produced by merging node and edge) data.frames for four customized datasets:
  - the full sample ignoring dropout information;
  - the full sample including dropout information;
  - the sample of those early challengers (tried final exam right after the pre-test before doing any problem sets)
  - the trimmed sample only including the popular trajectories (>50)

The following codes has been uploaded to the repository:

- base_code_sankey.Rmd: this is the original R code that wrangled the raw data to data forms that are suitable for Sankey diagram.

October, 3, 2017

- CS50_MOOC_Survival_Analysis_Data_0.RData: The dataset to be used for survival analysis models. This dataset is created from the raw dataset using prepare_data_for_survival_-analysis.Rmd file
- male_foreign_cubic_linear_point_data.RData: Datasets used to plot the log hazard and probability curves either by Male or by Foreign.