Datasets

We mainly select the following datasets in the pyKT at presents:

Dataset	question ID	skill ID	answering results	answering duration	ans
Statics20 11	✓		✓		✓
ASSISTmen ts2009	✓	✓	✓		
ASSISTmen ts2012	✓	✓	✓	✓	✓
ASSISTmen ts2015		✓	✓		
ASSISTmen ts2017	✓	✓	✓	✓	✓
Algebra20 05	✓	✓	✓		✓
Bridge200 6	✓	✓	✓		✓
Ednet	✓	✓	✓	✓	✓
NIPS34	✓	✓	✓		✓
POJ	✓		✓		✓

Statics2011

This dataset is collected from an engineering statics course taught at the Carnegie Mellon University during Fall 2011. In this dataset, a unique question is constructed by concatenating the problem name and step name and the dataset has 194,947 interactions, 333 students, 1,224 questions.

https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507

ASSISTments2009

This dataset is made up of math exercises, collected from the free online tutoring ASSISTments platform in the school year 2009-2010. The dataset consists of 346,860 interactions, 4,217 students, and 26,688 questions and is widely used and has been the standard b methods over the last decade.

https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010

ASSISTments2012

This is the ASSISTments data for the school year 2012~2013 with affect predictions. The dataset consists of 2,541,201 interactions, 27,066 students, and 45,716 questions.

https://sites.google.com/site/assistmentsdata/datasets/2012-13-school-data-with-affect

ASSISTments2015

Similar to ASSISTments2009, this dataset is collected from the ASSISTments platform in the year of 2015. It includes 708,631 interactions on 100 distinct KCs from 19,917 students. This dataset has the largest number of students among the other ASSISTments datasets.

https://sites.google.com/site/assistmentsdata/datasets/2015-assistments-skill-builder-data

ASSISTments2017

This dataset is from the 2017 data mining competition. It consists of 942,816 interactions, 686 students, and 102 questions.

https://sites.google.com/view/assistmentsdatamining/dataset?authuser=0

Algebra2005

This dataset is from the KDD Cup 2010 EDM Challenge that contains 13-14 year old students' responses to Algebra questions. It contains detailed step-level student responses. The unique question construction is similar to the process used in Statics2011, which ends up with 809,694 interactions, 574 students, 210,710 questions and 112 KCs.

https://pslcdatashop.web.cmu.edu/KDDCup/

Bridge2006

This dataset is also from the KDD Cup 2010 EDM Challenge and the unique question construction is similar to the process used in Statics2011. There are 3,679,199 interactions, 1,146 students, 207,856 questions and 493 KCs in the dataset.

https://pslcdatashop.web.cmu.edu/KDDCup/

Ednet

The large-scale hierarchical student activity data set collected by Santa (an artificial intelligence guidance system) contains 131317236 interactive information of 784309 students, which is the largest public interactive education system data set released so far.

https://github.com/riiid/ednet

NIPS34

This dataset is from the Tasks 3 & 4 at the NeurIPS 2020 Education Challenge. It contains students' answers to multiple-choice diagnostic math questions and is collected from the Eedi platform. For each question, we choose to use the leaf nodes from the subject tree as its KCs, which ends up with 1,382,727 interactions, 948 questions, and 57 KCs.

https://eedi.com/projects/neurips-education-challenge

POJ

This dataset consists of programming exercises and is collected from Peking coding practice online platform. The dataset is originally scraped by Pandey and Srivastava. In total, it has 996,240 interactions, 22,916 students, and 2,750 questions.

https://drive.google.com/drive/folders/1LRljqWfODwTYRMPw6wEJ_mMt1KZ4xBDk