• 821 teams •

KDD Cup 2015

2015-07-11 - Team Merger Deadline

2015-05-01 - Launch
2015-07-12 - Close

Data Files

Name	Available format	Link	MD5
date	.7z 1.37KB	baidu (http://pan.baidu.com/s/1o7Sk5ui) (code：3t4k)	0fc8772b90c2939c0513980354d9c1f7
sampleSubmission.csv	.7z 51.32KB	baidu (http://pan.baidu.com/s/1miAnSKC) (code：bm35)	2cab360f606e073ce1598c7c72e59237
train	.7z 39.13MB	baidu (http://pan.baidu.com/s/1kVpXi8f) (code：e97n)	d8357e79ca46e1dd8046051f1e10ecd8
test	.7z 26.04MB	baidu (http://pan.baidu.com/s/1coT8mE) (code：8y1r)	860ca90e1df763da09c127fca72534a5
object	.7z 760.71KB	baidu (http://pan.baidu.com/s/1eSulgcy) (code：txu6)	ae9263e71209bc32897864ee79c05c67
Winner_Slides	.zip 28.61MB	baidu (http://pan.baidu.com/s/1i5K307j) (code：9b8f)	d9e70affc8c0fca2ba2c4a5ec34ffcab

test.7z contains testing data. For each file in this archive, the format is the same as the counterpart in the train.7z.

date.csv -- each line contains the timespan of each course in our log data (both train and test data). The timespan of each course for calculating dropouts is 10 days after the last day of that course, i.e., course C is from 2014.4.1 to 2014.4.30 in the given data, a user enrolled the course C will be treated as a dropout if he/she leaves no record from 2014.5.1 to 2014.5.10.

- course_id - course id

- from - the first day of the course in log

- to - the last day of the course in log

object.csv - each line in this file describes a module in a course with its category, children objects and release time. Those modules represent different online materials of the courses, e.g., chapters, videos, problem sets and etc. The modules are organized as a tree, i.e., each course contains several chapters; each chapter contains several sections; and each section contains several objects (videos, problem sets, and etc).

- course_id - The course to which the module belongs.

- module_id - The ID of a course module.

- category - The category of the course module.

- children - The children modules of the course module.

- start - The time that the module was released to students.

sampleSubmission.csv - each line contains the enrollment ids of the test set and shows the submission format for the competition.The right format of submission file should only contain 2 columns with no header or other information. An error will

be reported if a submission file has a wrong format.

- 1st column- Enrollment ID.

- 2nd column - A real-valued probability of dropout.

---

train.7z contains training data.

enrollment_train.csv – each line is a course enrollment record with an enrollment id, a username U and a course id C, indicating that U enrolled in course C.

- enrollment_id – Enrollment ID

- username - Student ID.

- course_id - Course ID.

log_train.csv – each line is a behavior record called "event". Each event contains the following information: enrollment_id, username, course_id, time, source (server or browser), event, and object.

- enrollment_id - Enrollment ID.

- time - Time of the event.

- source - Event source (server or browser).

- event - In terms of event type, we defined 7 different event types:

  1. problem - Working on course assignments.

  2. video - Watching course videos.

  3. access - Accessing other course objects except videos and assignments.

  4. wiki - Accessing the course wiki.

  5. discussion - Accessing the course forum.

  6. navigate - Navigating to other part of the course.

  7. page_close – Closing the web page.

- object - The object the student access or navigate to.

true_trian.csv – each line contains information about the ground truth of enrollments in the training set.

- 1st column - Enrollment ID.

- 2nd column - Ground truth of dropout (1 for a dropout event and 0 for continuing study)