Guidance for processing learner participation data from the FutureLearn MOOC platform

Matt Cornock

6 November 2018

This document outlines the method for data processing learning participation data in the form of .csv files available to course developers and educators running MOOCs on the FutureLearn platform. This method was used to create the data set for analysing learner engagement as presented at OER18 on 18 Apr 2018 and FLAN, London on 6 Nov 2018.

This analysis explored the different patterns of learner engagement based upon the point at which a learner enrols, relative to the notional start date of a course.

Processing data from FutureLearn

It is important to establish whether the course has an advertised start date. Courses that ran May 2017 onwards are unlikely to have their start date visible on the FutureLearn course search page and the promotion of these courses leads to greater enrolment numbers after the course has started. Prior to May 2017, upcoming courses received more prominence on course search pages and had much higher enrolments before the start date (approximately double) compared to those after the start date.

Data required for context

1. Start and end dates of course run.
2. From the datasets the last enrolment date, last step access date, last comment date.
3. Whether the course operated with an advertised start date (the old course listings with upcoming courses) or without (the new default course listings that shows currently available courses).

Data processing rules

1. Exclude data from learner ids belonging to the course team.
2. Exclude data belonging to learner ids not with learner role.
3. Exclude data from learner ids where the learner appeared to have accessed the course before course start date.
4. Exclude data from learner ids where the enrolment occurs or step access occurs beyond the bulk of the course participants last dates.
5. Exclude learners who do not have a complete data set (see limitation point 3 below).

Example

In the enrolment.csv extract below, the rows are ordered by enrolled_at date descending. Rows 1 and 2 have enrolment dates noticeably apart from rows 3-6. The detected_country is also absent in rows 1 and 2, which implies invited enrolment, team member or FutureLearn access. Row 3 is also suspicious as it is over a week from the previous cluster of enrolments.

	enrolled_at	role	detected_country	valid?
1	2018-05-30 12:32:34 UTC	learner	--	Exclude
2	2018-01-22 11:42:14 UTC	learner	--	Exclude
3	2017-12-14 10:27:12 UTC	learner	GB	Exclude
4	2017-12-03 23:42:21 UTC	learner	KR	Include
5	2017-12-03 21:24:14 UTC	learner	KR	Include
6	2017-12-03 20:53:12 UTC	learner	GB	Include

In the enrolment.csv extract below, the rows are ordered by enrolled_at date descending. The extract represents the earliest enrolment dates. Row 24 is excluded as the role is not a learner. Rows 24 and 25 are excluded as the enrolment dates precede the course enrolment launch date.

	enrolled_at	role	Detected_country	valid?
21	2017-06-27 16:19:55 UTC	learner	UA	Include
22	2017-06-27 16:19:17 UTC	learner	CA	Include
23	2017-06-27 16:13:45 UTC	learner	US	Include
24	2017-04-05 07:29:15 UTC	organisation_admin	GB	Exclude
25	2017-03-23 14:16:42 UTC	learner	GB	Exclude

The following is an extract from steps.csv ordered by first_visited_at date ascending. Learner_id has been replaced with a letter indicator for anonymisation. The start date of the course was 18 Sep 2017. Learners with id a-d to be excluded as they have privileged access to the course before the start date they must be either team member or FutureLearn. All step visit data from id a-d needs excluding, regardless of whether later step visits occur during the course run. The first valid access is by id-e (assuming also that id-e is not listed in the team csv or otherwise excluded by other rules).

learner_id	step	week_number	step_number	first_visited_at	valid?
id-a	4.3	4	3	2017-08-29 16:43:15 UTC	Exclude
id-a	4.8	4	8	2017-08-29 16:43:34 UTC	Exclude
id-b	3.1	3	10	2017-09-07 14:51:40 UTC	Exclude
id-b	3.12	3	12	2017-09-07 14:51:45 UTC	Exclude
id-c	1.1	1	1	2017-09-11 09:33:07 UTC	Exclude
id-a	4.2	4	2	2017-09-11 16:55:03 UTC	Exclude
id-d	1.1	1	1	2017-09-13 14:40:40 UTC	Exclude
id-e	1.2	1	2	2017-09-18 00:05:51 UTC	Include

Note therefore that inclusion/exclusion rules must be applied across the datasets if only learner data is being explored. An exclusion in the steps.csv will also require that learner to be excluded from the enrolments.csv and comments.csv prior to analysis. If team contributions are being analysed, FutureLearn and non-team, non-learner enrolments and step visits will still need filtering out.

Limitations

1. The value of step in steps.csv is misinterpreted by Excel so that 1.1 and 1.10 are treated as the same value. A reference based upon week_number and step_number is required.

2. Step visits is preferred over step completion for analysis, as step completion requires learner to be aware of an additional function which is platform specific.
3. Data from those who joined after the course is always likely to be lower due to the dataset being curtailed between 40-50 days after the course end date.

Example

The following table shows the dates of interest when analysing the data for a course with a start date of 18 September 2017. Note that this course is 5 weeks duration, with 7 weeks free access prior to upgrade. The FutureLearn dataset date is not the last date of data. This is represented in steps.csv and comments.csv with a much earlier date (in this example 16 December 2017). The last enrolment date where there is complete data for all of the free-access period is therefore 7 weeks prior to the last data date (in this example enrolments no later than 28 October 2017 provide complete data for the free access period).

Start date	Course duration end date	Final day of free access from course start date	FutureLearn Dataset Date	Last enrolment	Last step visit	Last comment	Last enrolment with complete data
18/09/2017	23/10/2017	06/11/2017	19/10/2018	03/12/2017	16/12/2017	16/12/2017	28/10/2017

Key conclusions from the data

a) Retention is better in the group who enrol before the start date, in both cases where the course start date has been advertised and not.
b) Commenting is lower in the group who enrol once a course has started.
c) Replying is lower in the group who enrol once a course has started.
d) Completion is lower in the group who enrol once a course has started.