Scraping the FutureLearn Website

Team 6
Saad Anwar
Eva Bos
Dianna Burgess
Jan van der Doe

1. Motivation

1.1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset is created with the purpose of getting to know more about the market of online courses. In collecting data about this online course market there lies value in the generation of a dataset filled with information about the online course offerings from a big platform. There has been an expansion of online courses over the past years (Li et al., 2016), which nowadays gives a lot of possibilities to follow such courses. On top of that, there also has been an increase in enrollments in online courses, caused by the COVID-19 pandemic (Pino et al., 2020). With this information in mind, we decided to look at one of the platforms offering a large variety of courses. By looking at a big platform offering these courses, it is possible to get a better understanding of which preferences learners have. For example, it can be determined what type of courses are popular by looking at the number of enrollments. Next to that, it might be interesting to look at other characteristics of the course, like star-rating and if the course is offered for free. All these insights might provide managerial insights when you are considering offering courses at such a platform.

Below you see some of the websites that offer online courses that we have considered to scrape. After carefully weighing the advantages and disadvantages of each website, we chose FutureLearn. Even though it doesn't offer as many courses as other websites, the information that they do show on their website is interesting and can lead to many new insights. The website also fits our coding skills and budget for this project.

Website	**Advantage**	**Disadvantage**
skillshare.com	• Many courses • A lot of interesting variables (e.g., people watching right now, comments etc.)	• Has very good protection against bots • You need a paid membership to access the courses
coursera.org	• Each course has a lot of information • Available for free	• The website is hard to scrape since a lot of information is written in text chunks (e.g., hours you need to study) • Some interesting information is missing (e.g., number of enrollments)

masterclass.com	• Professional website • Available for free	• Some interesting information is missing (e.g., number of enrollments)
futurelearn.com	• Well-structured • Easy to scrape (e.g., no need to scroll to get more courses) • Available for free • Many interesting variables	• About 1500 courses, which is not as many as other websites

1.2 ***Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?***

This dataset is created by four students from Tilburg University following the course Online Data Collection and Management. This course is part of the master's program Marketing Analytics and is taught by dr. Hannes Datta.

1.3 ***Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.***

There is no funding or associated grant provided for the creation of this dataset.

2. Composition

2.1 ***What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of in- stances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.***

The instances in the dataset are all the courses (course urls) that are provided by the platform futurelearn.com. The courses belong to certain categories and are taught by different universities. These characteristics and many more variables belonging to an instance (a course) are collected. In the table added in question 2.4 you can see which information has been gathered about an instance.

There can be several links between instances, namely one university may offer multiple courses, meaning that those instances are connected via the university. Another example is that one category offers multiple courses, meaning that all the courses in one category are connected through that category.

2.2 ***How many instances are there in total (of each type, if appropriate)?***

In total there are 1513 course urls scraped from 14 categories. Those courses are taught by 238 universities. In the tables below the distribution of courses per university and category are shown (only the first 10 universities are shown due to the large number of partner universities).

University	Number of Courses
Coventry University	81
University of Leeds	81
FutureLearn	70
University of Michigan	52
CloudSwyft Global Systems, Inc.	46
Raspberry Pi Foundation	35
National STEM Learning Centre	29
The University of Glasgow	29
The Open University	28
Taipei Medical University	25

Category	Number of Courses
Business & Management	348
Creative Arts & Media	106
Healthcare & Medicine	262
History	42
IT & Computer Science	165
Language	70
Law	25
Literature	12
Nature & Environment	91
Politics & Society	74
Psychology & Mental Health	58
Science, Engineering & Maths	78

Study Skills	37
Teaching	144

2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances, since the data belonging to all offered courses is scraped from the website.

2.4 What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.

Each instance consists of the features stated in the table below.

Variables	**Description**	**Type of data**
Url	The complete url of the course.	Unprocessed text
Time	The exact time the data is scraped.	Unprocessed text
Category	The name of the category.	Unprocessed text
Page	The page from the category.	Integer
Header	The header of the course.	Unprocessed text
Enrollment	The number of enrollments of the course. Note that new courses do not have this information available yet.	Unprocessed text
New	Dummy showing whether the course is new.	Unprocessed text
Star_rating	The number of stars given to a course. From 0 to 5.	Double
Review_count	The number of reviews of the course.	Integer
Description	Short description of the course.	Unprocessed text
Duration	Duration of the course in weeks.	Unprocessed text

Weekly_study	Study time required for the course per week.	Unprocessed text
Unlimited	Dummy showing whether you can access this course with a so-called "unlimited" subscription.	Feature
X100_online	Dummy showing whether the course takes place online for the full 100%.	Feature
Free	Dummy showing whether the course is for free.	Feature
Accreditation	Dummy showing whether the course is eligible for accreditation	Feature
Part_of_expert	Dummy showing whether the course is part of a larger expert course.	Feature
Name_school	The name of the school that teaches the course.	Unprocessed text
Endorsed	Dummy showing whether the course is endorsed by third parties.	Feature

The data is inspected and below some statistics are shown.

Feature	**min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**NA's**	**SD**
Enrollments	304	1872	5990	22069	19384	772970	386	56490.78
Star_rating	2.8	4.6	4.7	4.649	4.8	5.0	479	0.2587
Review_count	5	16	38	121.1	102.8	4911	479	323.4007
Duration	1.0	2.0	3.0	3.507	4.0	10.0	0	1.3433
Weekly_study	0.0	2.0	3.0	3.265	4.0	10.0	0	1.2075

2.5 ***Is there a label or target associated with each instance? If so, please provide a description.***

Not applicable, since we are not trying to predict outputs for this project. Therefore, we do not require a target or label.

2.6 ***Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.***

Yes, there was some information missing from individual instances in the dataset. One reason for this is that the website of FutureLearn specifies that if a feature is not mentioned on the web page of a certain course it can be assumed that this feature is not present for this course. When we look at the raw dataset, star-ratings and enrollments have the most values missing since the website doesn't provide those features for new courses. So, the missing values in the dataset are missing because the information is unavailable, or not shown on the website.

2.7 ***Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.***

The relationship between individual instances is made explicit, since all individual instances have a unique course name. All the data belonging to the variables that are collected, relate to a certain course url.

All the courses in this dataset are linked to the teaching university and the category in which the course is shown on the website.

2.8 ***Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.***

We decided to not split the data set yet, but it is possible to do this. For instance, it is possible to look at the different categories and its courses, the different universities and their range of courses or for example you could split the dataset in on- and offline courses to look at their differences.

Regarding training, validation and testing sets there are enough possibilities since the dataset contains 1513 observations. The dataset is extensive enough to validate hypotheses through analysis and check their generalizability.

2.9 ***Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.***

The dataset relies on the website www.futurelearn.com, since we scraped the data in the dataset from this website. But the dataset is made once, and a timestamp was added. This means that the dataset we collected is self-contained. The seed pages are the pages that contain up to 16 courses. Each of those courses have their own url link.

The webscraper loops over those seed pages to extract the course urls and only then starts to scrape each individual course url.

a. The website futurelearn.com has existed since 2013, since then they have offered to over seven million people worldwide and they have partnerships with over 143 universities. We expect that futurelearn.com will exist over time. The courses however may vary over time, but we scraped per category, so if someone would like to generate a new version of the data this should not be a problem.
b. The complete dataset is archived, including a timestamp of when the data was scraped from futurelearn.com.
c. For now, there are no restrictions like licenses or fees associated with scraping information about the courses provided by futurelearn.com. We cannot say with certainty that this will also be the case in the future.

2.10 ***Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.***
There is no data in the dataset that is considered to be confidential.

2.11 ***Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.***
The dataset does not contain data that might be offensive, insulting, threatening or data that causes anxiety.

2.12 ***Does the dataset relate to people? If not, you may skip the remaining questions in this section.***
No, this dataset does not relate to people. We have only scraped features related to the courses, not information about individual users of FutureLearn.

3. Collection Process

3.1 How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data we scraped belonging to each instance was directly observable on the FutureLearn website.

3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?
The data is collected through a web scraper written by us. We build a web scraper in Python using BeautifulSoup. BeautifulSoup is a package used to scrape websites. Futurelearn's web pages are well structured, making BeautifulSoup a good package in scraping the web pages. Since Futurelearn doesn't have dynamic pages, we chose to use

BeautifulSoup over Selenium. Since we scraped data that was directly visible on the website, we were able to visually check and validate whether the data in the dataset was correct.

The collected data at first sight showed some issues. The scraper used to scrape courses of every category. However, there is also a category which shows all courses. This resulted in many courses being scraped twice. This issue is corrected by only scraping the courses from the category where all courses are presented and saving the name of the specific category (e.g., Business & Management). Then the dataset showed ratings and number of reviews for courses which didn't have those on the webpage. Ratings and number of reviews of recommended courses shown on the same webpage were scraped. Wrong ratings and reviews are now being stored to courses in the dataset. The solution to this issue was a simple try except statement where the scraper tries the appropriate class and container from the webpage and if it could not find any value it has to leave it blank.

3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample from a larger set, we scraped information from all the available courses on FutureLearn belonging to the timestamp from the date and time on which we collected the dataset.

3.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Four students of a work group from Tilburg University were involved in the data collection process. They were compensated with gaining knowledge and experience in Python and web scraping. There was no monetary payment as compensation.

3.5 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.

The web scraper was built in February and March 2022. The final version of our dataset was collected on 26 march 2022 at 14:46. The data associated with the instances have been created over a longer period of time. The platform FutureLearn launched in 2013, so over a period of nine years they created a large range of courses. We are not aware of when they started offering a specific course.

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The terms and conditions of FutureLearn state that it is allowed to scrape their webpages. So, there are no further ethical review processes conducted.

3.7 Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset does not relate to people.

4. Preprocessing, Cleaning, Labeling

4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.

The scraper saves the raw data in a CSV file with a semicolon as separator. There are variables of which we only need the numbers. For example, the variable 'Duration' has a value of '3 weeks', the variable 'Enrollments' says '312,560 have enrolled for this course' and the variable 'Weekly_study' says '2 hours'. For our research and to create insights, only the numbers are relevant. In the cleaning process all unnecessary text is removed with basic REGEX operations, so that the variables in question only consist of numeric values. This removed the units of the variables; they were later added to the label in the header. This process was done for the following two variables.

'Duration' now states as 'Duration_in_weeks'.
'Weekly_study' now states as 'Weekly_study_in_hours'.

For the variable 'Enrollments' the label has not changed, because it is straightforward. In the REGEX operations it has been noted that there are a few occurrences where the value states 'Educators are currently active on this course'. This will not return a numeric value and because there is no further information available on how many participants are following the course the decision has been made to store these as NA.

The values of the dummy variables are stored as 'yes' or 'no' . If the dummy variables are present on the web page that is being scraped then the value 'yes' is given. If not, then the value 'no' is given. The website of FutureLearn mentions that if a specific feature is not mentioned on the web page of a course it can be assumed that this feature is not present.

4.2 Was the "raw" data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, the raw data is the output of the scraper and this can be found in the zip folder.

4.3 Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

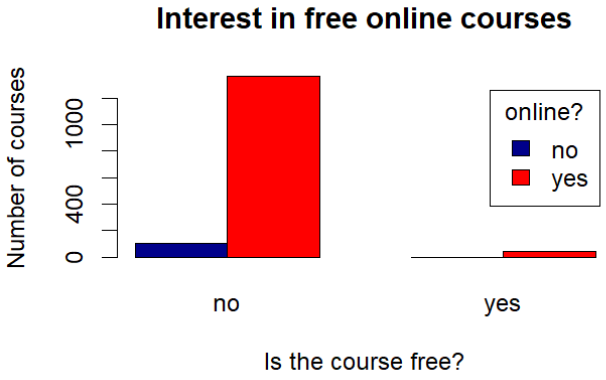
Yes, this was done with Rstudio and the code can be found in the zip folder as well as in the GitHub repository of Jan van der Doe.

A link to the GitHub repository:
https://github.com/janvanderdoe/online_data_collection/blob/main/analysis%20.R)

5. Uses

5.1 ***Has the dataset been used for any tasks already? If so, please provide a description***.
The dataset has not been used for any tasks beyond the statistical analysis conducted and figures obtained in R-Studio.

For example, one could create insights as the barchart below. It can be seen that all free courses are given online.



5.2 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
There is no repository that links to all papers or systems that use this dataset, since the dataset has not been used for any tasks yet.

5.3 What (other) tasks could the dataset be used for?
Since all the data from FutureLearn is gathered into one dataset, it is easy for marketers or managers, who consider offering courses on a platform, to get a good insight into, for example, the following issues:

- Which kind of courses are the most popular?
- Is there a specific category that has significantly more enrollments than other categories?
- Does the fact that a learner has to pay for a course influence the given star rating?
- Does the fact that a course is endorsed by third parties influence the given star rating?

5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
There is a possibility that the dataset that is generated for this project is no longer up to date, since it is gathered at one particular moment in time. Variables might change over

time: new courses could be added and learners can enroll for more courses as well. Next to that the star ratings and reviews might change.
The code for creating the dataset will still work (if FutureLearn does not implement drastic changes), but it is therefore possible that a new generated dataset differs from the one that is used for this project. Since we work with try and except blocks, the scraper will not throw an error, but will simply return empty cells for the inaccessible variables. If the whole course or page becomes inaccessible, the scraper will print a list of the pages and courses it couldn't access, but as of this moment, there are no inaccessible pages and courses.

5.5 Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset can be used to analyze online learning platforms. The dataset has no use for subjects out of this topic, since it is very specific to this market.
Since it is allowed to scrape data from FutureLearn, there is no legal or ethical concern, and thus no reason to not use the dataset.

6. References

- Li, X., Chen, Q., Fang, F., & Zhang, J. (2016). Is online education more like the global public goods? *Futures*, *81*, 176–190. https://doi.org/10.1016/j.futures.2015.10.001
- Pinto, J. D., Quintana, C., Quintana, R. M., Broude Geva, S., & Colbry, D. (2020). Exemplifying Computational Thinking Scenarios in the Age of COVID-19: Examining the Pandemic's Effects in a Project-Based MOOC. *Computing in Science & Engineering*, *22*(6), 97–102. https://doi.org/10.1109/mcse.2020.3024012