Scraping MBA programmes information from MBAStudies.com

MBAStudies scraper.ipynb, MBAStudies_1924_03262022.csv, MBAStudies_1924_03262022.json

Team 13

Adapted from: Gebru, Morgenstern, Vecchione, Vaughan, Wallach,

Daumeé, and Crawford. (2018). Datasheets for Datasets.

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.*

The importance of higher education cannot be denied as it plays a role in various areas of human life from personal development to business growth and socio-economic advancement. Higher education is an instrument for economic progress (Kromydas, 2017). It creates a path to financial security, economic mobility, personal growth, professional development, and leadership opportunities, among others (Teague, 2015). Moreover, according to UNESCO, higher education institutions play a crucial role in developing innovative solutions to global and local problems. Therefore, it is important to have access to transparent information about available study programmes in order to be able to make an informed decision.

The dataset was created with the aim of comparing different MBA programmes offered within the Group of Seven (G7) countries: Canada, France, Germany, Italy, Japan, the United Kingdom and the United States. Nowadays, the number of academic offers is considerably increasing; therefore, it is important to be able to better organize all the information based on important features such as: tuition fee, length of study, language, and location, among others.

While there is a lot of information available on the internet regarding MBA programmes, with the currently available tools, it is quite difficult to make comparisons among several programmes or to get access to general statistics on the academic offerings presented per country. For example, MBAStudies allows a comparison of only two programmes at once. This option is expanded to four programmes, when the account is created on the website. Therefore, our main purpose is to offer an alternative solution based on web scraping techniques, to allow for more effective comparison of MBA programmes. We considered it to be important to develop this project in order to enable students to benchmark their individual programmes of interest and discover new options.

The G7 countries have been selected, because they are the world's seven largest so-called "advanced" economies (BBC, 2022). For that reason, they have a very interesting range of academic offerings. It is important to highlight that the expenditure per student from public and private sources in education made in the G7 countries is considerably higher than the one made in other countries even members of the OCDE (Gaskell & Rubenson, 2004)

[1*] https://arxiv.org/abs/1803.09010

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

The dataset extracted from the website www.mbastudies.com/MBA/ was developed by three students from Team 13 as part of the Online and Data Collection Management course integrated into the Master's programme in Marketing Analytics at Tilburg University.

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.*

In order to carry out this project, no external funding or resources were received. All the inputs used were provided directly by Tilburg University and the coaching session given by the professor Hannes Datta.

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.*

Each instance in the dataset represents one MBA programme from the https://www.mbastudies.com/MBA website in the selected countries (G7). The data is scraped per each country. For that reason, it is important to take into account the fact that programmes offered in more than one country can be duplicates in the dataset.

While all programmes are in the same category (Master of Business Administration), each programme has its own characteristics that differentiate it from other programmes. These characteristics include: locations (in which programme is offered), earliest start date, languages, pace (part-time/full-time), duration of the programme, application deadline, type of study (e.g. online, on-campus), and the amount of tuition fees.

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

There are 1266 MBA programmes in the G7 countries (as of 26 March 2022). This number can change in the future as the programmes are either added or deleted. The number of programmes was extracted from each country page of MBA programmes.

Country name	**Link to the country page of MBA programmes**	**Number of programmes**
Canada	https://www.mbastudies.com/MBA/Canada/	44
France	https://www.mbastudies.com/MBA/France/	110
Germany	https://www.mbastudies.com/MBA/Germany/	128
Italy	https://www.mbastudies.com/MBA/Italy/	25
Japan	https://www.mbastudies.com/MBA/Japan/	18
United Kingdom	https://www.mbastudies.com/MBA/United-Kingdom/	182
United States	https://www.mbastudies.com/MBA/USA/	759
TOTAL		**1266**

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 information on all MBA programmes offered in the G7 countries. However, the website contains information on several other kinds of master programmes offered around the world. Therefore, the dataset consists of a sample of all programmes offered by MBAStudies.com.

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.*

Data collected on the MBA programmes consists of unprocessed text which describes general information: all locations, earliest start date, languages, duration, application, deadline, study type, and tuition fees.

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

Not applicable

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.*

First of all, some of the links provided to the study programmes do not exist anymore. It can be assumed that the programme was either deleted or the link was changed and not updated on the website.

Moreover, there is also missing information in some MBA programmes, for example, on application deadline, earliest start date, and duration. This can be due to several reasons:

- The university did not provide MBAStudies.com with the complete information, and therefore, some of the details are not included on the website;
- The information changes regularly, such as application deadline or earliest starting date;
- The university did not want to disclose this information on purpose so that students contact the institution directly (in marketing, it can be considered lead generation). In some cases, "request info" button is provided;
- The university did not want to disclose this information on purpose for other reason, for example, the information was sensitive;
- The information is case-specific, meaning that the answer differs per student; and therefore, providing a general answer was not possible. For example, the tuition fees amount is different depending on the country of origin of the student.

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 the programme and its characteristic is made explicit by connecting the programme URL to these characteristics. Therefore, it's known which programme the features belong to.

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.*

There are no recommended data splits because we are not aiming to develop validatory purposes. At the beginning of the project a small subsample of just two countries was created but just for testing purposes. The written code will be useful to get the information of MBA programmes hosted in the webpage even if they are added or updated later.

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.*

All the dataset is self-contained. There are no external resources which have a direct influence on the scrapped information. Clearly all the extracted information is related to MBA programmes and universities that are completely independent of the Webpage, nevertheless the portal does not offer any link to get access to these external agents but rather deploys the whole information on each of its own pages. Also as it was mentioned before, some tests were done in order to determine if there were important constraints which could affect the performance of the code, but the answer to this inquiry was negative.

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.*

Data available on MBAStudies.com is a public data on study programmes. Universities voluntarily add information about their programmes so it is not 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 is linked with academic programmes so it doesn't contain any offensive or insulting content. Moreover, the submitted information is processed by MBAStudies.com before it is published.

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

The dataset contains information about the study programmes; and therefore, it is not related to people.

2.13 *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

Not applicable

2.14 *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

Not applicable

2.15 *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

Not applicable

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 collected data is reported by higher education institutions (subjects). In this specific case the subjects (e.g. University representatives) are those in charge of submitting and keeping updated the exhibited information on www.mbastudies.com. Keystone Education Group, which is the company behind the website, offers a friendly interface for all of those who want to promote academic programmes, as well the company is responsible for guaranteeing an adequate filtering and organization of the deployed data on its web portal. The data scraped from the website is directly exhibited on each programmes' page and publicly available for every user.

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 mechanism used to extract data was web scraping techniques carried out through the Python programming language. The main libraries used were BeautifulSoup, Pandas and Selenium. BeautifulSoup was the most appropriate tool because it allows us to understand and interpret the html structure of the website and therefore retrieve the data more easily. Selenium was useful in our project because some content of the website is available after clicking the button to expand the text. At the end of the project it was aimed to store the information in JSON dictionaries as well making it available in csv format.

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 sampling strategy followed some principles which were:

- The main point of the strategy was not about the number of countries but in the number of programmes, hence a sample of more than 1000 MBA's was the desired number. Our first option was to include just countries from the EU but this sample barely reached 600 programmes so it was not representative enough;
- A diverse origin of the programmes was also another feature searched during the sampling process. Eventually it was decided to not include programmes of just one specific region or continent. It was defined that the wider the scope, regarding the countries, the better and more representative the sample would be. At the end, the chosen sample has programmes from three different continents (Americas, Europe and Asia);
- In order to better focus the whole project, the sample was created based on real international organizations. Initially the chosen organization was the European Union but given that this sample did not fulfill the two previous requirements, the G7 was the elected organization.

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)?*

In the data collection process, the students in project team 13 were involved. Besides, the professor of this Data Collection course, Mr. Hannes Datta, also participated in the collection process by giving feedback and guidance.

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.*

Since the website update frequency is relatively low, considering that the duration of each program is usually about a year, the data of the website will be updated when the program is changed every year. Thus, the timeframe of the dataset contained the data of last year.

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.*

There are no ethical issues because there is no personal or commercial private information on the website, and there are no restrictions on scraping. In addition, we scraped data only for the purpose of doing course project.

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

The dataset isn't related to people, so this question doesn’t apply to our project.

3.8 *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

This question doesn’t apply to our project.

3.9 *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

This question doesn’t apply to our project.

3.10 *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

This question doesn't apply to our project.

3.11 *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

This question doesn't apply to our project.

3.12 *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

This question doesn't apply to our project.

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.*

As mentioned before, some of the programme links were not provided in the programme listing. Therefore, when scraping these links, we have obtained some empty values for the missing links. These empty values were removed from the list of all links so that the scraper can work correctly. Besides that, all data was collected, without removing any instances.

While conducting an analysis on the csv file, the 14th column ("Belgium, Brussels") was dropped, because it was scrapped due to inconsistency in the html layout of the website, and was not useful. Also, the 13th column (Study Locations) was removed for the same reason.

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

Raw data was saved in json format. In addition, a csv file was created on the basis of dataframe, so it is structured. The utf-8 encoding was applied in both cases in order to display international characters correctly.

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

The software Python in Jupyter Notebook (https://jupyter.org/) was used for writing and processing the code in Python (https://www.python.org/). Moreover, BeautifulSoup (https://pypi.org/project/beautifulsoup4/) and Selenium (https://pypi.org/project/selenium/) were used to retrieve information from the website. The statistical analysis was conducted in RStudio (https://www.rstudio.com/).

5. Uses

5.1 *Has the dataset been used for any tasks already? If so, please provide a description.*

The task has not been used by any other parties (ex. companies, researchers, etc.) yet, the dataset was only scraped by data collection team 13 for study purposes.

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 are no papers or systems that use the dataset, thus there are also no repository links here.

5.3 *What (other) tasks could the dataset be used for?*

As mentioned above in the Motivation section, the dataset which contains many variables allows people to compare different MBA programmes over the G7 countries.

One obstacle of the MasterStudies website is that only two programmes' information can be compared simultaneously. However, the dataset we scraped can be compared with different programmes as much as possible. Student candidates who want to apply to a university can easily compare the programmes they are interested in, as well as each programmes' rankings, tuition fees, location, language, admission requirements, availability of scholarships, etc. It also provides an overview of programmes for educational institutions. By this way, they can find the information they need very efficiently and precisely.

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?*

It can be estimated that the frequency of updating this website is about half a year or once a year, because the specific information of university and programmes will not change a lot (including tuition fees, admission requirements, etc.). The only thing to pay attention to is that this website platform is not the only search resource. More accurate and specific information should be based on the official website of each university.

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

This dataset should not be used as a substitute for official accurate information. The data provided here are for reference only, but should not be relied only upon it.

References

BBC. (2022, March 24). *G7: What is the G7 and what is it doing about Ukraine?* BBC. Retrieved March 26, 2022, from https://www.bbc.com/news/world-49434667

Gaskell, J. S., & Rubenson, K. (Eds.). (2004). *Educational Outcomes for the Canadian Workplace: New Frameworks for Policy and Research*. University of Toronto Press.

Kromydas, T. (2017). *Rethinking higher education and its relationship with social inequalities: past knowledge, present state and future potential*. https://www.nature.com/articles/s41599-017-0001-8

Martin, M. (n.d.). *Higher education on the road to 2030*. IIEP-UNESCO. Retrieved March 26, 2022, from http://www.iiep.unesco.org/en/higher-education-road-2030

Teague, L. (2015). *Higher Education Plays Critical Role in Society: More Women Leaders Can Make a Difference*. https://eric.ed.gov/?id=EJ1091521