Datasheet for Indeed.com webscraper

Anouk Heemskerk, Georgiana Huţanu, Renée Nieuwkoop and Alan Rijnders (2021)

Adapted from: Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford. (2021).

1. Motivation

1.1 For what purpose were these datasets created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. The current project collected four datasets, that were created in order to provide current Marketing Analytics students valuable information in terms of the skills that are in demand in today's Dutch job market. Generally, students trust that their university will prepare them with the skills that are going to be a necessity for their future careers. However, are the taught skills truly in line with the skills that are really required in nowadays' competing Dutch marketplace?

To answer this question, the researchers collected and analyzed job listings from Indeed.com. This is a worldwide employment website for job listings. Currently, it is the market-leader within the Dutch labor market (Intelligence Group, 2019). The central focus was placed on the job listings belonging to the search terms 'Data Analist', 'Data Scientist', 'Marketing Analist' and 'Marketeer' with the location 'Nederland', to make the data as useful as possible for Marketing Analytics students.

In covering different job titles, it can be investigated to what extent the required skills overlap for the different jobs, what are the gaps between the university teachings and the job requirements, as well as what are the central trends in terms of which programming languages are more popular compared to those that are decreasing in demand.

Even though these data can be of great value for shedding light on the aforementioned insights for Marketing Analytics students, it could also be relevant for students in related fields, as well as other individuals looking for a job in one of those disciplines. In addition, educational institutes can use this data to shape their course offerings in order to prepare their students as good as possible for the job market.

1.2 Who created these datasets (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? These datasets were created by Anouk Heemskerk, Georgiana Huţanu, Renée Nieuwkoop and Alan Rijnders, who are students at Tilburg University. The creation of these datasets was part of the course Online Data Collection and Management instructed by dr. Hannes Datta.

1.3 Who funded the creation of these datasets? If there is an associated grant, please provide the name of the grantor and the grant name and number. The deployment of this project had no associated grant.

2. Composition

2.1 What do the instances that comprise these datasets represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Each instance consists of an individual job listing.

2.2 How many instances are there in total (of each type, if appropriate)? The four datasets contain 4115 job postings in total. Hereof, 1196 listings correspond to the search term 'Data Analist', 1034 to 'Data Scientist', 566 to 'Marketing Analist' and 1319 to 'Marketeer'.

2.3 Do these datasets 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). These datasets contain all possible instances available on Indeed.com at the time web scraping took place, relating to the search terms for 'Wat': i. 'Data Analist', ii. 'Data Scientist', iii. 'Marketing Analist', iv. 'Marketeer', and for 'Waar': 'Nederland'. Due to the fact that we scraped all available job listings for a particular job title, there are no potential algorithmic biases.

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 description. The data in the datasets consist of raw data in the form of unprocessed texts. The data are directly scraped from Indeed.com, with no further data preprocessing conducted. For each job listing, the following were scraped and collected: job id, title, company, location, posting date, summary, salary indication if available, url, as well as a full job description, along with the date and time the scripting and data collection took place.

```
##  [1] "id"                    "title"                 "company"
##  [4] "location"              "postingdate"           "today"
##  [7] "summary"               "salary"                "url"
## [10] "description"           "scrapetimesdescription"
```

2.5 Is there a label or target associated with each instance? If so, please provide a description. Each job listing has an unique id. This id is a string of letters and numbers so that different job listings can be distinguished.

Some examples of how the ids look:

```
## [1] "6a28b29568a5595a" "7912a09a73b84a41" "f525346a1e0593f4" "42d2dbc0740f816f"
## [5] "07b0178cd8c65029" "640a02bcb832fc43"
```

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. The missing information pertains to the salary indication. This information is not made available for every job listing. Data Analist has 861 instances with missing salary data, Data Scientist 804, Marketing Analist 463 and Marketeer 988.

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. There are no relationships between individual instances.

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.

2.9 Are the datasets self-contained, or do they link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If they link 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 existed

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 datasets link to the external website Indeed.com. The datasets will not remain constant over time, due to the fact that the datasets consist of vacancies on a website and these vacancies will change over time. There are no restrictions in terms of e.g. licenses or fees.

2.10 Do the datasets 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. All data that are collected are derived from a publicly available website and do not contain any personal and/or confidential data. All companies that make use of Indeed.com to publish their job listings are consent with making it publicly available. People do not need an account or login to see the job listings on Indeed.com. Therefore, the data are not considered to be personal and/or confidential.

2.11 Do the datasets contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. The datasets only contain textual information of job listings. The data is not expected to be offensive or insulting in any possible way.

2.12 Do the datasets relate to people? If not, you may skip the remaining questions in this section. The datasets do not relate to individual people.

2.13 Do the datasets 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 datasets. 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 datasets? If so, please describe how. Not applicable.

2.15 Do the datasets 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 subjects (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 within the sets was directly observable as raw text.

3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? To collect the data we used Python, multiple Python packages and Selenium to create a webscraper. The reason we chose to build a webscraper instead of using an API is as follows: at the moment we decided between choosing a webscraper and API, the API was not a valid option because it was not available. The webscraper was manually curated by the 4 students named earlier. Furthermore, we used a Github repository to keep track of our project.

3.3 If the datasets are samples from larger sets, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The datasets are not a sample from larger sets. They contain all available job listings with that search term and location from Indeed.com at the moment of scraping.

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)? The four students who created these datasets and their professor Hannes Datta were involved in the data collection process. None of them were compensated.

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 timeframe in which the data associated with the instances was created. The data for the search term 'Data Analist' was collected on 13.03.2021. The data for the search terms 'Marketing Analist' and 'Marketeer' were collected on 16.03.2021. The data for the search term 'Data Scientist' was collected on 17.03.2021.

Ideally, the data could have been collected regularly (e.g. once per day/week), in order to allow monitoring of changes throughout time, however the scraper built for Indeed.com does not allow data collection with no surveillance. Therefore, the time constraints of this project allowed data collection for only one point in time for each job title. These web scraper limitations are due to running into Captchas every time the web scraper is ran extensively. Moreover, there is no historical data available from Indeed.com, so there was also no possibility to compare our findings with the job listings from, for example, a year ago. However, now these datasets are available, they offer further possibilities for future research.

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. No, there has not been an ethical review process conducted by a review board. All data that are collected are derived from a publicly available website and do not contain any personal and/or confidential data. Thus, scraping Indeed is not unethical.

3.7 Do the datasets relate to people? If not, you may skip the remaining questions in this section. No, the datasets do not relate to individual people.

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)? Not applicable.

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. Not applicable.

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. Not applicable.

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). Not applicable.

3.12 Has an analysis of the potential impact of the datasets and their 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. Not applicable.

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 remainder of the questions in this section. Yes, preprocessing has taken place in order to merge the descriptions and the listing CSV files together by id, into 1 CSV per job title. This preprocessing step has been done to improve the structure and the accessibility of the datasets.

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. Yes, the raw data was stored on Google Drive.

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

5. Uses

5.1 Have the datasets been used for any tasks already? If so, please provide a description. The datasets have only been used for the course Data Preparation and Workflow Management that is part of the MSc Marketing Analytics at Tilburg University. For that course, the datasets are used to investigate and compare the different skills required for the different search terms.

5.2 Is there a repository that links to any or all papers or systems that use these datasets? If so, please provide a link or other access point. The repository containing all related files and documents can be found here.

5.3 What (other) tasks could the datasets be used for? The datasets can be used by students of Tilburg University and Tilburg University itself in creating a fitting education. For students, the datasets are valuable in terms of which software and/or programs they will need to adapt to and learn in order to become a marketeer, data analist, data scientist or a marketing analist. For Tilburg University, the datasets are valuable to draw a comparison between what the students already learn at the University and what they currently do not provide to students, but will be crucial for future students in finding a suitable career.

5.4 Is there anything about the composition of the datasets or the way they were 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? No, there is nothing in the datasets that might impact future use. The datasets are only merged together and there were no legal and/or ethical concerns identified.

5.5 Are there tasks for which the datasets should not be used? If so, please provide a description. No, there are no tasks for which the datasets should not be used.

References Intelligence Group. (2019). *Top 25 sites for active job-seekers in the Netherlands.* Retrieved from https://intelligence-group.nl/en/news/top-25-sites-for-active-job-seekers-in-the-netherlands