 doi:10.20944/preprints202206.0346.v1

Dataset

A Dataset Containing Job Descriptions Suitable for NLP and NN Processing

Pieterbas Pluijmaekers and Francesco Lelli

Pieterbas Pluijmaekers, Francesco Lelli, Tilburg University
f.lelli@tilburguniversity.edu

Abstract: We describe a dataset that contains job description published on a popular online website in the information and technology sector. As the website focus mainly on United Kingdom based jobs, the data have a specific focus on this country. It contains 11.501 job vacancies and 13 related meta data information. The dataset is suitable for HR analysis using machine learning techniques such as natural language processing and neural networks.

Keywords: dataset; NLP; Human Resource Management; classification; Job description

1. Introduction

The dataset that we would like to release contains information about the information technology job sector. In special job titles, descriptions and the meta data. The data have been collected using the website reed.co.uk with the ''selenium'' python package. We collected the data with the aim of comparing different classifying algorithms. Training a model to recognize job titles and descriptions to predetermine normalized job labels, supports the research of Human Resource Management analytics. The use of job- title and description data is applied in several research in Human Resource Management and therefore can be used as a benchmark for evaluating different approaches. For example, scholar in [1] and [2] look for possible techniques for classifying ambiguous job titles leveraging the job description. In addition, researches in [3], applied the same type of data in order to extract skills from job descriptions.

The following table summarize the data:

Subject	Computer science, Linguistics, Human Resource Management and Artificial Intelligence
Specific subject area	The specific subject area is the classification of data with classification algorithms. Within the natural language processing (NLP) domain classification of text documents is a technique. Natural language processing is a domain within the field of computer science, this makes computer understand text. Text classifying is based on identifying different characteristic in text in order to classify a new text input. This is done with prelabelled trainings data that is used to model an algorithm.
Type of data	• Float (salary(min/max)) • String (job title, organization, job description, value alternative) • Meta data (location, region, html code) • Date

 doi:10.20944/preprints202206.0346.v1

How data were acquired	The data was collected from reed.co.uk with the use selenium python package.
Data format	Raw
Parameters for data collection	The website of reed.co.uk contains different filters and options. The filters for the research were established on IT & Telecom jobs. Also the use of display data parameters was configured. This results in ordering the vacancies on date.
Description of data collection	The website of reed uses a search- and a vacancy page. On the search page different vacancies are stated based on the search query and filters. This helps the candidate in finding a vacancy based on their preferences. After clicking on a specific job title, the website forwards the client to the specific vacancy. On this page all the details of a vacancy are stated. These details include: job title, job descriptions and various meta data that are listed below. An automated python script scraped trough all these processes. All the variables were extracted by the XPATH query and written to a Comma Separated Value (CSV) file.
Data source location	Data available at the following URL: https://doi.org/10.34894/7GBRIE
Data accessibility	Repository name: DataverseNL Data identification number: 10.34894/7GBRIE Direct URL to data: https://doi.org/10.34894/7GBRIE Instructions for accessing these data: Data are open access

2. Methods for Data Acquisition

Reed offers 25 different categories of vacancies; however, the dataset focusses only to IT & Telecom as we intended to offer a specific case. consists of a search query page and the vacancy page. The results of the search query contain 25 vacancies. At the dime of the data collection 20.712 results were presents.

The searching through pages on the website of Reed is done leveraging the URL structure as within the URL we can find the search query, page number and the fixed variable sort by “DisplayDate”. We automate the process by creating a script that accessed the first page and subsequently loop around all the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0346.v1

available pages for the whole dataset. The script searched for the URL and raw job title and went to the next page. The range picking was by trial and error, but had a fixed constant of 0 until 500. Values that came back as NaN (Not a Number), were deleted before saving to CSV.

We also analyzed each single page, and we leverage the structure of the website to collect meta information. In this respect, we used XPATH for navigate the structure of the webpage. query different locations of variables were searched on the website page. For example, we collected job title, description and the various meta data described in section 3. As not all the descriptions contained certain information, we intentionally left empty the metadata that were not present. Some All the data was saved to a CSV file with every variable in a separated column.

2.1 List of resources used for the collection of data

- Root query: https://www.reed.co.uk/jobs/it-jobs?pageno=1&sortby=DisplayDate
- Software: ‘’selenium’’ python package

3. Results: Data Description

The dataset is release as a single file with the name “jobtitle-description.csv”. It contains 11.501 rows; each one represents a vacancy data. Note that some of the jobs are redundant as we include reposts that the job advertiser decided to boost up (i.e bring on the first page) the vacancy.

The datasheet contains 13 individual columns that are described in table 1. Note some of the columns are optional and consequently the dataset has empty values.

Table 1: Data, description, type of data and optional of vacancy page

Sort data	Description	Type of data	Optional
Title	Containing title of vacancy	String	NO
Posted on	Date on the posting	Date (DD/MM/YYYY)	NO
Organization	If applicable recruiting organization	String	YES
Employment type	Different types of employment	String	YES
Location	Location of the vacancy	String	NO
Region	Region of vacancy	String	NO
Minimal value	If applicable lower bound of salary	Integer	YES

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 June 2022 doi:10.20944/preprints202206.0346.v1

Maximal value	If applicable higher bound of salary	Integer	YES
Value alternative	If appliable description of salary bound	String	YES
Label	If applicable a label of the vacancy (NEW, HOT)	String	YES
Description	The text description	String	NO
HTML data with tags	First part of the description containing tags (introduction)	String	NO
HTML data with tags 2	Second part of the description containing tags (center text)	String	NO

The dates associated to the vacancies have a time span of 22-11-2021 until 4-12-2021. The website of reed.co.uk is mostly based within the United Kingdom. However, in figure 1, we can note that UK is not the only location of the vacancies.

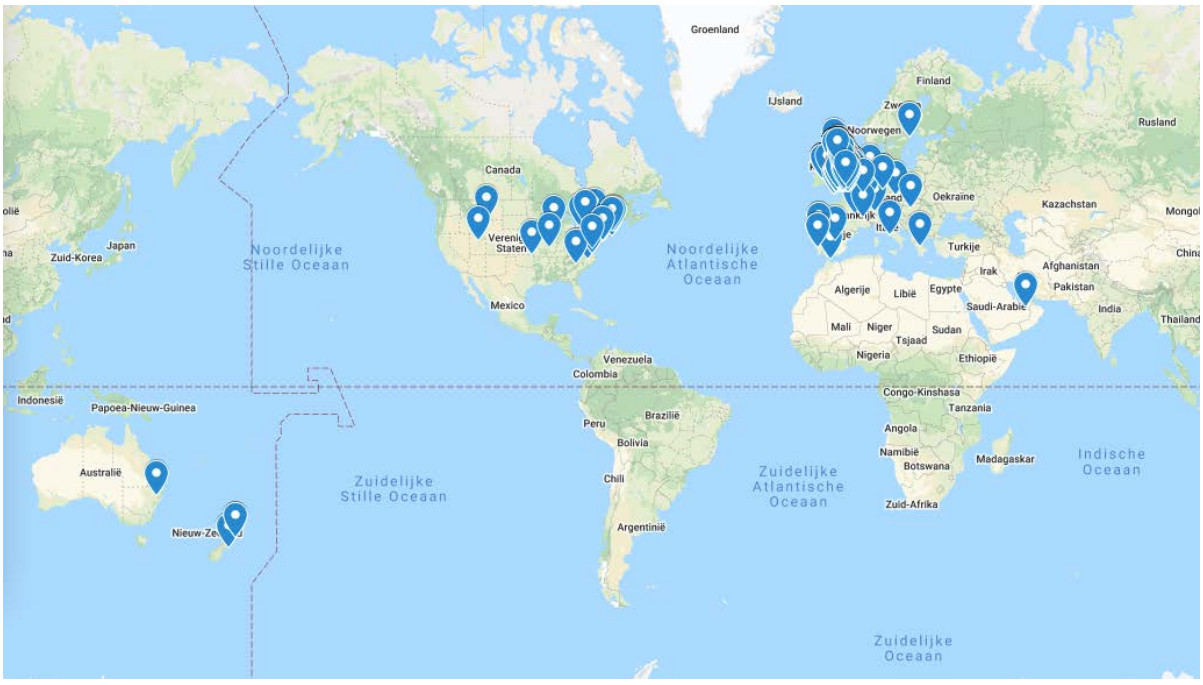

Figure 1: Distribution of the data

In addition, we can observe that 3535 job offers were a repost of the same job and could be considered as a sort of duplicate.

4. Discussion

This dataset is suitable for AI- based processing to better understand the job market of a specific region. Example of information and insights that we could get from the data includes:

- Discovering various skills required for jobs within the information technology sector.

 doi:10.20944/preprints202206.0346.v1

- Metadata can be used for understanding best practices in recruitment.
- Study different techniques for improving the classification of information.
- Information can be expended by adding more search queries and/or can be use in combination with other information to extend the scope of initial research.

5. Conclusion

We presented a dataset that can be use for study HR-related dynamics using Natural Language Processing and other AI-based techniques. The dataset was gathered from the website of Reed that is mostly active in the United Kingdom, but also contains rows from various other countries. The data are release as a single CSV that contains 11.501 rows and 13 columns. Not all the value are mandatory and therefore, it contains empty cells. The data have been collected between the 22-11-2021 and the 4-12-2021.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

6. References

[1] Bao, H., Baker, C. J. O., & Adisesh, A. (2020). Occupation Coding of Job Titles: Iterative Development of an Automated Coding Algorithm for the Canadian National Occupation Classification (ACA-NOC). *JMIR Formative Research*, *4*(8), e16422. https://doi.org/10.2196/16422

[2] Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T. S. (2015). Carotene: A Job Title Classification System for the Online Recruitment Domain. *2015 IEEE First International Conference on Big Data Computing Service and Applications*. https://doi.org/10.1109/bigdataservice.2015.61

[3] Wowczko, I. (2015). Skills and Vacancy Analysis with Data Mining Techniques. *Informatics*, *2*(4), 31–49. https://doi.org/10.3390/informatics2040031