Composition

Vasilis

10/12/2021

## R Markdown

## 2.1 What do the instances that comprise the dataset repre-sent (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 data was downloaded from github.com using API and scaping the site. Each instance consists of a differ-ent repository. A repository contains files of a project and it gives the opportunity for multiple users to inter-act at the same time in the project.

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

The number of instances (repositories) depends on the keyword and the number of days the user is interested in. In this case, the dataset contains a total of 4776 re-positories based on the keyword “python” and all re-positories between 12 of October and three days ago. Function: filename = find\_repo(“keyword”,days) .

## 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 una-vailable).

The dataset contains all instances available on github based of the keyword and the amount of the days that the user selected. However , each instance consists var-iables that we were selected in order to be useful for the dataset.

## 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 contains raw data which are download-ed directly from github. For each repository, the follow-ing data were collected : id , name, url , language, day of creation ,stars,watch,forks, and the readme descrip-tion of each repository. Furthermore , In the readme description is decode the images in order to save the data in a csv file.

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

A repository creator could generate multiple reposito-ries. The id of creator would be the same ,however each repository will has a unique identifier. In this case the identifier is an id. Each id is a combination of eight number .

## 2.6 Is any information missing from individual instances? If so, please provide a description, explaining why this in-formation 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 in the dataset is represented with “NaN. The information that is not available is the language of the repository and the readme description. This is because the repository is empty or the creator did not create - upload any description of the project.

## 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.

All individual instances have a relationship based on the keyword that the user is interested in . In this case the word “pyhton” are related to all instances. Howev-er, each individual instance has a relationship with the variables forks,watch,stars. These variables are not a direct way of evaluation,although 52.5% of users star a project to show appreciation and 73% of users consider the number of stars before using of contributing to a project (Borges et al,2018). <https://homepages.dcc.ufmg.br/~mtov/pub/2018-jss-github-stars.pdf> H.Borges,M.T.Valente,What’s in a GitHub Star? Under-standing Repository Starring Practices in a Social Coding Platform,Journal of Systems and Software,2018,pp. 112-129.

## 2.8 Are there recommended data splits (e.g., training, devel-opment/validation, testing)? If so, please provide a descrip-tion of these splits, explaining the rationale behind them.

The dataset would be useful if we maintain the url re-pository and the description of readme. As a result , the user could minimize the time of search in order to iden-tify the ideal repository for his interest.

## 2.9 Is the dataset self-contained, or does it link to or other-wise 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 com-plete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any re-strictions (e.g., licenses, fees) associated with any of the exter-nal resources that might apply to a future user? Please pro-vide descriptions of all external resources and any re-strictions associated with them, as well as links or other access points, as appropriate.

The dataset are relative to external website github.com. The datasets will be available in the future ,however if a creator delete his repository then this individual in-stance will not be useful anymore for the dataset. Some hypothetical restriction could we address for our analy-sis. The main restriction is that github could stop pro-vide API to the developers or stop share the public re-positories

## 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 con-tent of individuals non-public communications)? If so, please provide a description.

All the instances are public and available for everyone, as a result the dataset doesn’t contain confidential in-formation .

## 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 only contain public information, however in the description of readme file might could be infor-mation cause anxiety.

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

The dataset is not relative 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 combi-nation 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 opin-ions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identifi-cation, such as social security numbers; criminal history)? If so, please provide a description.

Not applicable.