NAU NORTHERN ARIZONA UNIVERSITY
School of Informatics, Computing, and Cyber Systems

CS 499/599 Research Methods for Software Engineering

Research Project

D2. Data Collection

Submit a document with the following information about your research. It is ok to have preliminary data at this point. You can continue to collect data before the final analysis.

1. **Research Question(s)**
   What are your research questions?

   What is the health of the project dependencies used in the project?

2. **Data Set**
   What data are you using to answer each research question?
   Explain what you are collecting and what kind of calculations you are performing with the data to extract what you need to be able answer the research question.

   The dataset gathered consists of columns - id, name, full_name, html_url, url, size, language, forks, open_issues, visibility, and watchers. From this data, we need popular repository URLs which we can feed to the analyzer for the project dependencies health check.

3. **Data Collection**
   Add here links to the scripts or instruments that you are using to collect your data.

   Using the GIT SEARCH API Endpoint, we have gathered all the repositories created between 2010-2020 and the results are fetched with respect to JavaScript language using the following query:

   https://api.github.com/search/repositories?q=language:javascript+created%3A%3E2010-01-01&created%3C%3A2020-01-01

   From the results fetched, there are approximately around 19 million repositories belonging to JavaScript. The GIT API is allowing a maximum of 100 results per page. So, we are fetching 10 pages (which is equal to 1000 records) as a preliminary data collection.

4. **Filtering and validity check**
   Explain how you are filtering or validating the data that you collect to make sure that you are processing valid data. Remember, garbage in, garbage out.

   As we are having enormous results from the GIT search API when we fetch the repositories in between 2010 and 2020, we have applied filter of language: JavaScript in the query selector of API. After fetching the results, we have compared the language column of all the results as part of validity check. While performing validity check, we have observed few duplicate records in the data set. So, we removed the duplicate records and are finally left with 936 records.

5. **Characteristics of your data**
   You should have some data at this point but you can continue to collect data later. Present here some descriptive statistics about your data set.

   We want to focus on the repositories which are popular in terms of stars and forks. To consider the popular repositories for the analysis, we have calculated the mean of these two columns and considered the records whose average of the forks and watchers (stars count) is above the calculated mean.

   Out of which 416 records are satisfying the below condition i.e.,
   df['forks'] >= df['forks'].mean() or df['watcher'] >= df['watchers'].mean()

6. **Preliminary data**
   Provide a link to your preliminary data.

   We have attached the CSV of the preliminary data collected.

I will select a few groups to present the data collection to the rest of the class. Add your slides to this shared spreadsheet:
https://docs.google.com/spreadsheets/d/1B_7IlzUNGWeSyeXFV0TX0JXiqzAG3ev6fEhfH2hn1ho/edit?usp=sharing