Is Stack Overflow in Portuguese more attractive for Brazilian users?

Angela Lozano*, Bogdan Vasilescu†, Miguel Botto Tobar‡§, Weslley Torres§, and Alexander Serebrenik§,
*Vrije Universiteit Brussel, Brussels, Belgium
†Carniege Mellon University, Pittsburgh, PA, USA
‡University of Guayaquil, Guayaquil, Ecuador
§Eindhoven University of Technology, Eindhoven, The Netherlands

***Abstract*—The abstract goes here.**

I. Introduction

Software development and maintenance are activities that often involves many concepts and reference documents (citar). Many software aspects may be changed over time. In order to work with them and details involved in a software project, developers often need helps from one another. Nowadays a widely used way is for developers to ask questions and/or answer them in various online forums. StackOverflow (SO)[1] is a Q&A site with more than six million registered users. SO also has the version on Portuguese, Russian and Spanish.

Alexander ►*What is the goal of your study?*◄

Alexander ►*Here you need to explain why did you decide to focus on Brazil as opposed to any other country in the world.*◄

Alexander ►*Separate the methodological discussion (what have I done? for example, downloaded the data, run a tool) from the results (what has it delivered, for instance, 5.954 user profiles). Make a separate section "Methodology" and a separate section "Results".*◄

II. Methodology

A. Data Extraction

The data extraction has been performed on March 6, 2016, and included data from November 2013 to February 2016 from the Stack Exchange (SE) data dump[2] and for this study, we considered the Portuguese version (SOPT). The XML files corresponding to the tags, users, and posts were transferred to a MySql database, through a R function per type of file (i.e., posts, users, and tags).

B. Data Preprocesing

We cleansed the tables eliminating few users were due to lack of data. None of these users had AccountId (i.e., user identifier for all stackExchange websites), LastAccessDate, WebsiteUrl, Location, UpVotes, DownVotes or Age. All of these users have the same display name (i.e., "a25bedc5-3d09-41b8-82fb-ea6c353d75ae"), and whenever they have a ProfileImageUrl, it is the same[3]. These accounts were created at different times from November 2015 to February 2016. We could not come up with a plausible reason for these anonymous users having the same display name but no other data, they do not seem to have anything in common. In total 3 SOPT users have been eliminated.

We focused on Brazilian users thus to identify their location we used `countryNameManager`[4].

C. Searching Process

Consequently, the locations were identified, and before starting the search a group of students were selected based on whether they had experience in use SO (in Spanish version) or GitHub, as detailed below:
The search started doing a manual inspection by each user profile based on *userId*, i.e. http://pt.stackoverflow.com/users/1919/, where **1919** is the *userId*. On the user profile, we looked the email address, if it was not available, we checked whether the user has a GitHub account or a personal web page.

- With Github account we found out the email address below *userName* if it was not on, we used a browser extension gitDiscovered[5] to discover the email address or we checked his/her public activity looking for at he/she did a git command [6], and then we searched the email address using a Github API [7] by *userName*.
- With the personal web page, we searched the email address on the section "about me" or *sobre me* in the Portuguese language.
- If none of the above, we used GitHub and searched by userName from SOPT, and compared profile picture, location, skills between SOPT and search results on GitHub, and then we followed the steps above mentioned with GitHub account, in order to find the user and get the email address.

In order to ensure the accuracy of results, we selected a random group of 15 users each 500 profiles and searched using the steps above mentioned. Whether we found new email addresses or missing information, we chose another one random group and applied the manual inspection again.

[1] https://stackoverflow.com/
[2] https://archive.org/details/stackexchange
[3] https://www.gravatar.com/avatar/?s=128&d=identicon&r=PG&f=1
[4] https://github.com/tue-mdse/countryNameManager
[5] https://gitdiscovered.com/
[6] https://git-scm.com/docs/git-push
[7] https://api.github.com/users/userName/events/public

III. Results

We identify the location of 7.264 users of SOPT which corresponds to 27% of its users. As we foresaw, most of the users of SOPT are located at Portuguese-speaking countries, in particular in Brazil (see Table I). Although there is a wide range of non-Portuguese speaking countries users, when looking at percentages these countries only represent 2

Country	Total
Brazil*	5954
Portugal*	599
United States	220
United Kingdom	80
Canada	44
France	25
Germany	42
India	30
The Netherlands	20
Mozambique*	14
Angola*	8
Cape Verde*	4
Other non Portuguese countries	224
None	19415

TABLE I

User's location in SOPT. Portuguese speaking countries are marked with an asterisk.

For half of the users whose location was identified, we could identify their gender (65%). Females are an overwhelming minority (4% SOPT users).

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.