Is Stack Overflow on Portuguese more attractive for Brazilian users?

1 Introduction

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".*◄

2 Methodology

To conduct the study we considered the Portuguese version (SOPT) and have downloaded the data from the Stack Exchange (SE) data dump[2]. The data extraction has been performed on March 7, 2016, and included data from November 2013 to February 2016. 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).

After creating the tables few users were eliminated 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 2010 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].

[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

Consequently the locations were identified, a group of 25 students Miguel ►*I don't know if I should indicate where are they from*◄ Alexander ►*Please explain how they have been selected*◄ were selected to help to search and get the email addresses from SOPT users, each of them with 500 profiles, as detailed below:

- First, we started to look each user profile by *userId*, i.e. `http://pt.stackoverflow.com/users/1919/`, where **1919** is the *userId*. asWhat does " started to look each user profile" mean?
- Then, 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 was possible to find 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*. Alexander ►*It seems that you have used several techniques. Please separate them: at the moment I cannot follow.*◄
  - 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, creation date Alexander ►*Why do you need to compare the creation date?*◄ between SOPT and search results on GitHub, and then we followed the steps above mentioned with GitHub account. Alexander ►*State the purpose of these comparisons.*◄

To ensure the accuracy of the results, we selected a random group of 15 users of 500 profiles and searched using the steps above mentioned. In the case they Alexander ►*What?*◄ were inconsistent Alexander ►*How do you define consistency?*◄ , we chose another one random group of 15, and if it continued we searched all users and compared with the previous outcomes. Alexander ►*I do not understand the last sentence; please rewrite.*◄

3 Results

We could only 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 1). Although there is a wide range of non-Portuguese speaking countries users, when looking at percentages these countries only represent 2

[5] https://gitdiscovered.com/
[6] https://git-scm.com/docs/git-push
[7] https://api.github.com/users/userName/events/public

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 1: 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).