VRIJE
UNIVERSITEIT
BRUSSEL

ANALYSIS AND VISUALIZATION OF STACK OVERFLOW DATASET

Information Visualization 2020-2021

Group 4:
Bhavesh Jangale
Artyom Kuznetsov
Anisha Sachdeva

May 23, 2021

Faculty of Sciences and Bio-Engineering Sciences: MA Computer Science

Contents

1 Introduction

1.1 Project Idea

The idea of this project is to create a dashboard that visualize different information based on public data from Stack Overflow resource. This dashboard should be able to give better understanding how popular the resource is, people from which countries use it often and which popular trends do we have in programming. This dashboard should help people to understand the trends in their country to select technologies, frameworks and the vectors which are more popular in their country. On basis of this knowledge people should be able to determine what to learn for their future work.

1.2 Target Audience

As described in the project idea. The audience are beginners in programming and other programmers who are looking for a job. Besides, this dashboard is useful for data analysts and IT companies. The latter one can use this information to select the more popular programming stacks and to compare it to other countries. Our dashboard is interactive, people can select data they are more interested in. This fact should help people from different audience find what they are looking for.

2 Data Understanding

Stack Overflow dataset contains following tables:

- Badges
- Comments
- Posts
- Users
- Tags
- Votes
- Several Pivot Tables
- Wikipedia related Tables

The whole Stack Overflow dataset takes up to 500 GB space if it being downloaded entirely. For analysis and visualization we do not need the whole data being downloaded, only portion. In the later section we discuss how we pre-processed data. After dataset analysis we decided to focus on comments, posts, tags and users tables. They were the most promising. However, with further analysis we have seen some problems with tables we need. For instance, we decided to rid of comments, badges tables entirely, since we were not convinced by data importance for our visualization purpose. When we continued to analyse our data we noticed that some important data like age, location were missing or being bad formatted. This was good time to start prepossessing data we already have.

3 Data Pre-processing

We started to pre-processing age column. As a result we have received small number of users that indicate this information. Based on this, we removed the idea of visualizing age information in our dashboard.

Then we started working on fixing Location column problems. The problem with this field was due to the fact that many locations are not formatted properly or containing nonsense or being empty at all. We noted that out of 20 million users only 3.5 million users had associated location with them. Others had the empty field associated. As an interesting fact, only those 3.5 million users had made most of the posts (30 million posts in a total of 50 million posts), hence we decided to remove rest of the users and concentrate only on 3.5 million users with some kind of location filled in the respective field. Now, another problem that we encountered was that for many locations streets, neighbourhood or regions (sometimes with special symbols) was mentioned but not the country. In order to visualise our data, we required the country names and we thought for a few ideas to solve this issue. The same ideas are reported below.

3.1 Use of GeoPy library with Nominatim

The first idea was to use **geoPy** library with Nominatim. Nominatim is OpenStreetMap geocoder that would be able to infer full location name from partial address. We have written a Python script to use the library with Nomenatim.

```
import pandas as pd
import csv
import geopy
from geopy.geocoders import Nominatim
import os
users_data = pd.read_csv('/Users/..../...csv')

geolocator = Nominatim (user_agent=\geoapiExercises")
final_location = [None] * len(users_data)

for index, row in users_data.iterrows():
    locations=geolocator.geocode(row['location'])
    locations=str(locations)
    loc = locations.split(',')
    final_location[index] = loc[-1].strip()

users_data[\Correct Location"] = final_location
users_data.head()
```

The implementation is being demonstrated above. The problem with this approach is due to fact that Nominatim is free and does not have any paid plans at all. But the free plan they have limits requests. Their plan allows to fetch 1 query per 1 second. We had a lot of locations and we calculated that it would take 10-11 days to compute desired locations.

3.2 Use PyCountry library

The second idea was to use **PyCountry** library that provides locations based on addresses. After some thoughts we created a script that can run and fetch locations:

```
import glob
import pandas
import pycountry

def update_locations():
    print("Program Started")
    so_users_df = pandas.concat(map(pandas.read_csv, glob.glob('data/*.csv')))
    print("Number of locations to update: \n", so_users_df.count())
    # take all rows from index 0 to index 499999
    for index, row in so_users_df.iloc[0:500000, :].iterrows():
        for c in pycountry.countries:
            if isinstance(row[5], str):
                if c.name in row[5]:
                    row[5] = c.name
                    so_users_df.at[index, 'Address']  = c.name
                    break
        print("Row at index: ", index, "has been updated!")
    print("finished, saving to file")
    so_users_df.to_csv("res.csv", index=False, header=True)
    print(so_users_df)
```

This approach was also a bit of problematic, but for some extend better than the previous approach in terms of performance. The idea to use iterrows() to iterate Pandas dataset is not that good in terms of performance, Python is not good at iterating objects. We would need to vectorize data in order to improve performance, but it would take time to do so. This approach, with flaws in performance would run the whole dataset for 15 hours to get results. If we run it on 3 machines separately it would take only 5 hours to run the processing. Not that bad already, compared to the first approach. But we found the third approach.

3.3 Joining Stack Overflow Users with Locations Dataset

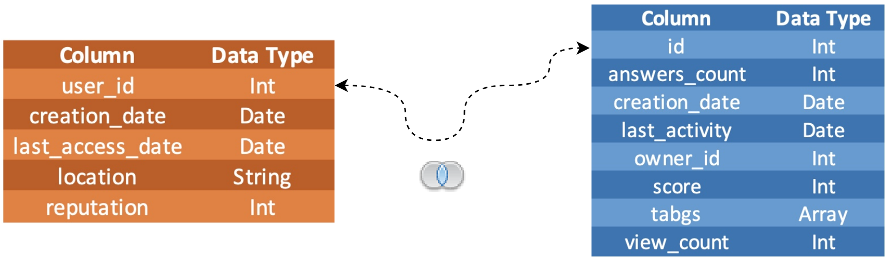
We figured out that we can find dataset that contains locations with country names inside. On BigQuery dataset it takes just few minutes to get result. We have used this query to get most of the users:

```
    SELECT users.id, users.creation_date, users.last_access_date,
    geo_data.country_name as location,
    users.reputation, users.up_votes,
    users.down_votes, users.views FROM
    ‘bigquery-public-data.census_bureau_international.country_names_area‘ AS geo_data,
    ‘bigquery-public-data.Stack Overflow.users‘ AS users
    where users.location is not null AND users.location LIKE
    CONCAT(‘%’, geo_data.country_name, ‘%’)
```

So, this query fetches most of the users. However, this is not the only query we used to find correct location for users. We do not include it into the report due to the large size of those queries. The basic idea behind them is to select all other category of users who included addresses in foreign language. We determined several countries that did it often and extracted the data successfully. Also, internal location detector that Tableau has helped us to find the rest locations. Results were promising, we only lost around 200.000 users with wrong locations. Our user table resulted in 3.3 million rows.

Posts and Tags were easily pre-processed without any big problems with dataset. We have used two Stack Overflow datasets, one from **bigquery public data** and another one from **soTorrent**. The first one is based on the second one, but have useful joins for us. We have noticed that the first dataset has tags column in Posts table. This column was represented as array of values. We decided to use this column instead of using joined Posts and Tags, since the job was already done for us.

3.4 Data Structure



Column	Data Type
user_id	Int
creation_date	Date
last_access_date	Date
location	String
reputation	Int

Column	Data Type
id	Int
answers_count	Int
creation_date	Date
last_activity	Date
owner_id	Int
score	Int
tabgs	Array
view_count	Int

Figure 1: Users and Posts Table

Figure 4 represent corresponding data tables we use. *user_id* and *owner_id* are fields that used for joining common data. The join operation is being done by use of Tableau internal algorithm. As you see, the user table has three important fields for us - location and creation date and last access date. The first one is important because we want to determine the location of users and the second is useful to see the age of accounts and to understand whether account was recently active or not. The second figure demonstrates posts. It also has last activity and creation date that helps us to have information about age of posts and whether it was popular for a long period of time. Tags array contains a sequence of tags that were mentioned in the post itself. We use them to determine the popular technologies among people and countries. The queries to determine the rest country locations are located in SQL file with the archive of our project. The name of the file is: country_queries.sql.

4 Dashboard

After exploring few visualization tools, we considered using *"Tableau"* for creating visualization graphs and dashboards. We developed 2 dashboards in tableau. The developed dashboards were simple and easy to use and understand.

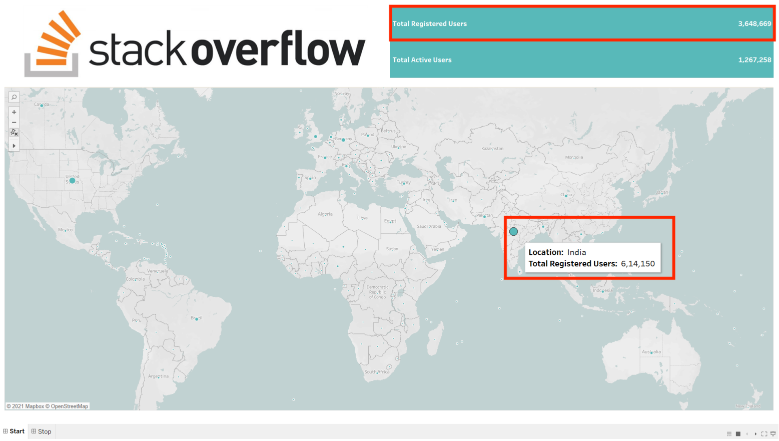

stackoverflow
Total Registered Users 3,648,669
Total Active Users 1,267,258
Location: India
Total Registered Users: 6,14,150
© 2021 Mapbox © OpenStreetMap
Start
Stop

Figure 2: Dashboard 1 containing the world map

Visualization starts with *dashboard 1* which contains a map of world. When we hover on a particular country, the total no of registered users and country name can be seen. Every country has a bubble over it. This bubble size will increase or decrease depending on the total no of registered users.

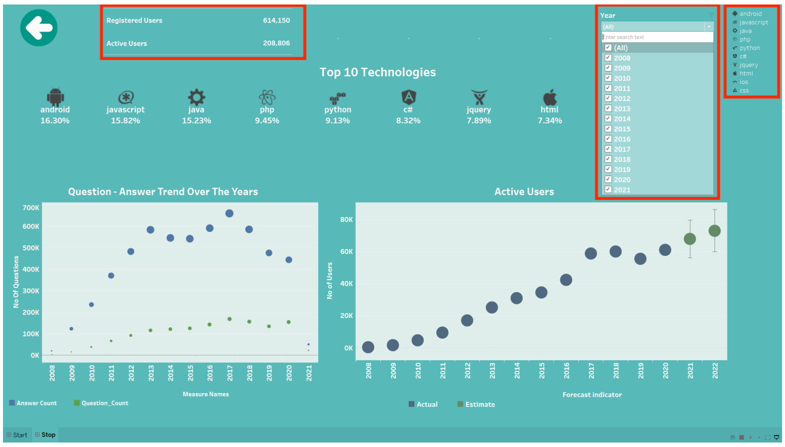
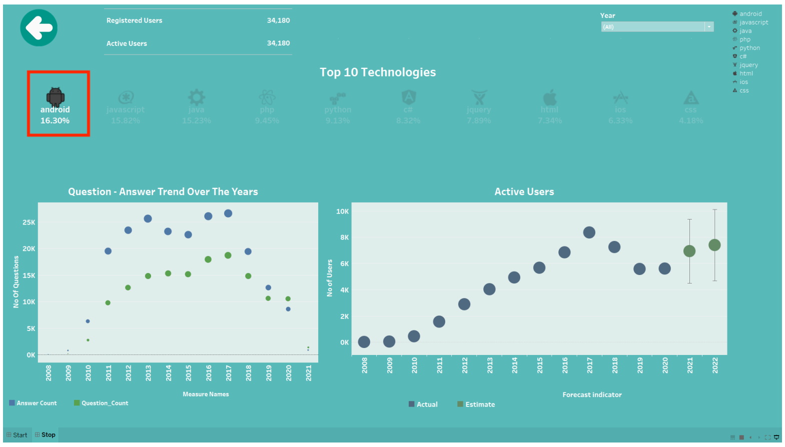


Figure 3: Dashboard 2

On clicking on a particular country, a new dashboard will be opened. Dashboard displays count of *"Total Registered Users"* and *"Total Active Users"* to the left upper side of the screen. Below it, it shows information of the *"Top 10 technology"* for that particular country along with the percentage of the technology. This information is fetched from the data on the basis of total no of questions asked about a particular technology. The dashboard provides a bar chart of "Total no of active users over the years" to the lower right side. This information is fetched from the data on the basis of total no of users who have posted questions over the years. Dashboard also offers users with details about the total no of questions asked and the answers received for those questions over the years in "Question - Answer Trend Over the Years" bar chart. This information is fetched from the post id and the answer count from the data. Dashboard also has 2 filters. First filter contains list of top 10 technologies for the country which is displayed on the upper right most side. Second filter contains list of years which is just ahead of Technology filter to the upper right side of the dashboard.



Registered Users 34,180

Active Users 34,180

Year (All)

Figure 4: Dashboard 2 showcasing information related only to Android technology in India

The whole dashboard is interactive, so user can select values from the filter provided or he can click on the graph data. Upon clicking or selection, the dashboard graphs will get updated and will display details corresponding to the selection. A user can also do multiple selection either in the filter or on the graphs. If no filter for a year is selected, the dashboard will show details over the span of 14 years i.e from 2008-2021. The dashboard provides an unique functionality which predicts the *"No of Active Users"* for year 2021 and year 2022. Initially we decided to use dark blue as dashboard background with a mix color combination of green, purple and dark yellow colour, but due to some limitation because of technology icon, we decide to use the shade of green colour with all the text in white colour. Finally, the dashboard provides the user a functionality to navigate to previous dashboard with the "Back" button displayed on the upper left side.

5 Evaluation

To validate our visualization, we had targeted the following audiences:

- Data analysts
- Technophiles
- People working in software companies

Our validation strategy included taking feedback from 10 selected end users (which belonged to the target audiences group). We divided the validation process into two levels:

- **Level 1:** In level 1, we shared our screen using TeamViewer (the remote desktop software) and gave the control of the dashboard to the end user. We did not give any description of the dashboard to the user, in order to know if the user was able to understand the theme for our project (or not) without having any prior knowledge.

  The users very well could deduce that the dashboard was about Stack Overflow having various country-wise trends related to different technologies.

- **Level 2:** In level 2, we discussed with those same users in detail about our dashboard and gave them a Google form to review the functionality along with the look and feel of our dashboard. We received the following responses:

  - The participants were from different countries like India, Belgium, Malaysia, Azerbaijan, Russia and United States. Mostly belonged in the age group of 20 to 30 years.
  - 50% (i.e. 5 out of 10) of the participants were working in software companies, 40% (i.e. 4 out of 10) were computer science students or engineers and 10% (i.e. 1 out of 10) was a data analyst.
  - 70% (i.e. 7 out of 10) of the participants visit Stack Overflow once a week and the remaining 30% (i.e. 3 out of 10) visit once a month.
  - Mostly (70% i.e. 7 out of 10) users often find the answers related to their query in Stack Overflow.
  - No participant was suffering from color blindness.
  - Except one, all the participants found the colors to be easily distinguishable in the dashboard.
  - Out of 10, 3 (i.e. 30%) participants did not like the color scheme.
  - All the participants agreed with the Stack Overflow trends displayed for the country of their choice and found the visualization to be interactive.
  - 60% (6 out of 10) users said that if they would wish to learn a new technology in the near future, they would consider the dashboard to check the trend of technologies of a particular country.

To summarize[1], most of the end users liked our interactive dashboard and were able to deduce information from the trends available. However, they had the general comment of dashboard not being visually appealing and the same being little slow while applying filters. Along with this we received another review comment from Prof. Beat to include the name of the country on the second dashboard while trends of that particular country were being displayed as the human memory tends to forget the country's name while switching between two layouts. While we are thankful to all the users for the validation and agree with their review comments, we have tried to answer in the limitations and future work section as what were the roadblocks and why we could not imply certain things to make the dashboard look better.

[1] Please refer to Stack_Overflow_Visualization_Survey.pdf in the project folder for the detailed report of the survey

6 Limitations and Future Work

As much as we also wanted to have some dark colors in the background and contrast colors in the graphs, we were unable to do so because the icons displayed on the second dashboard were of the color – greyish black and we could not update those icons. In order to display and change the icons of the *top 10 technologies* dynamically for each country, we had to include a database file in the Tableau directory containing the icons of all the technologies. The names of the technologies in the icon database are compared with the dashboard's technologies for displaying icons. The database file that we could find online with names and icons of technologies being in sync with our dashboard's database included the greyish black color. Hence, we had no other option else to go with the light mint green background and white font color.

As part of our future work, we can try to solve the issue of latency while applying filters. The reason that it takes a lot of time to change colors along with the graph data is that there is a lot of data and the time taken by dashboard on our system to load that data is huge. This can be resolved either by using tableau on a machine with high configuration or by importing all the data in SQL, performing all the data manipulation tasks in SQL and fetching only those data from SQL which we want to show in tableau. This will reduce the size of data that will be loaded in tableau significantly and make the tableau faster.

Additionally, we tried to include the name of the selected country on the second dashboard, but we couldn't. As per our understanding, country names is fetched by tableau only when we are creating maps. In case we want to use country name in any other visualization, tableau will not identify the country name and instead show data as it is.

There is one way by which we can include the country name. But for this we need to publish the dashboard on a server. Once the dashboard is published on server, we can then pass the Location parameter i.e Country name in the URL for calling second dashboard from first dashboard. However, we published our visualization on Tableau public and tried to get the country name on the second dashboard but again, it was not possible, since Tableau public does not provide URL functionality.

7 Conclusion

As we conclude our report on Stack Overflow's data visualization, various trends related to different technologies for particular countries can be clearly viewed through our dashboard. As a good data visualization project, our dashboard is user friendly which could help the end user to look at the popular languages based on number of questions asked related to the same. Along with this, the user could also check how a technology has changed over recent times and learning which new technology might help the user. We have developed effective and rich in information visualization to view the trends and comparison of various technologies between different countries.

8 Setup Instructions for executing the project

Following are the steps to run the project file. Make sure Tableau is installed on the computer. Tableau 2020.4 is used during development.

1. Extract data from *Data.rar*.
2. "Tags" folder provided to be copied in the Tableau installation folder: **defaults/Shapes** Path used during development was **C:/Program Files/Tableau/Tableau 2020.4/defaults/Shapes**
3. Open the "Project.twb" in Tableau.
4. Edit the connection for both the data source. For Users data use the "users" Folder Provided while for Posts data use "Posts folder provided."

   Apart from the setup, the application can be accessed at

   ```
   https://public.tableau.com/profile/bhavesh.jangale#!/vizhome/
   StackOverflow_trial/Start
   ```

References

[1] *Filtering a Dashboard From Another Dashboard: Tableau Software.* URL: `https://kb.tableau.com/articles/howto/filtering-a-dashboard-using-action-filters-passed-from-another-dashboard`.