A Data Analysis of Google Play Store

Andy Pagès
Msc. Human Computer Interaction
Student ID 4336938
psxap8@nottingham.ac.uk

1. Problematic

The Android application store, *Google Play Store*, has become an invaluable platform for mobile technology with almost 3 millions applications currently available, with goals such as entertainment, education, business, social, health and many others. Not only is this platform an absolute necessity for every Android user (77.45% of worldwide smartphone users [1]), it is also the place where startups and companies can present their products to the public. It is therefore very important for any organization wishing to use the *Play Store* for business purposes to fully understand its content and its current state. Investigating such information through data analysis and visualization allows business entities to understand the users' desires, needs and adapt their products to the correct market.

The content of this proposal aims to uncover the "*secrets*" of the *Google Play Store* and visualize them using interactive web tools. The focus of this project is made on the implementation of a modern interaction experience with the analysed data rather than on complex processes to obtain it. Advanced transformations to visualize that data are still necessary and in the scope of this project.

2. Data-set

The data-set proposed to be used for this project was found on the data platform *Kaggle*. In one set, it presents more than 10 thousands entries describing applications from the store. The other set contains complete individual reviews for thousands of applications. The first set contains the following attributes:

- **App**: *name*,
- **Category**: *category*,
- **Rating**: *overall user rating*,
- **Reviews**: *number of user reviews*,
- **Size**: *size*,
- **Installs**: *number of user downloads/installs*,
- **Type**: *free or paid*,
- **Content Rating**: *age group*,
- **Genres**: *genres, tags*,
- **Last Updated**: *when the app was last updated*,
- **Current Ver**: *current version*,
- **Android Ver**: *minimum Android version required*.

The review set contains the following attributes:

- **App**: *name*,
- **Translated_Review**: *user review (comment)*,
- **Sentiment**: *positive, negative or neutral* (pre-processed)
- **Sentiment_Polarity**: *sentiment polarity* (pre-processed),
- **Sentiment_Subjectivity**: *sentiment subjectivity* (pre-processed).

3. Solution

The solution for this problematic is to thoroughly explore and exploit this data-set to visualize key points from the *Play Store*. The realization should answer questions such as *"What category of application are the most installed"*, *"What ranges of prices are the most successful"*, *"What is the correlation between ratings and downloads"* and many more.

The main focus is to present a very interactive visualization system using modern tools such as *JavaScript*. In particular, the use of the library *React* to build elegant, efficient and highly interactive components with the help of many visualization libraries is a valuable choice. A great example of what is desired to be implemented is the data visualization work made on the state of JavaScript in 2018 [2]. This project aims to create a similar, surely a bit smaller, experience than the state of JavaScript in 2018 but for the *Google Play Store*.

4. Implementation

4.1 Back-end

The querying, transformation and manipulation of the data-set is an important factor to extract significant results. In order to implement a robust and efficient data structure, the data-set will be transformed from *.CSV* to a *SQLite3* database. This database will be layered by a REST-API, from which any client will be able to get pre-processed results for complex queries. That API will be implemented using *Node.js* and its famous framework *Express*, and will be deployed on the cloud at the address `https://api.ivp.andynroses.me`.

4.2 Front-end

The visualization must be responsive, interactive and modern to present the results in the best manner. The library *React* will be used for its undeniable advanced interfaces attributes and its virtual DOM allowing efficient data updates. To build the project, transpile JavaScript to *ES6* and handle all dependencies, the bundler *Parcel* will be used. The set of chart React components *Nivo* will probably be the main library used to display the data. In the case of limitations, more existing chart libraries might be used. Finally, *styled-components* for styling will be used. The build application will be served using a *Nginx* server at the address `https://ivp.andynroses.me`.

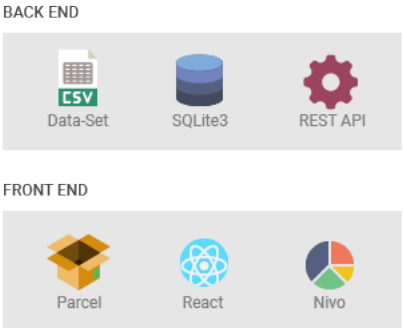

Figure 1. Architecture

5. Planning

The project will be executed with the following milestones:

1. **Infrastructure setup** (28/02) : Establishing and configuring the entire architecture base and all the necessary tools,
2. **Data investigation** (07/03): Investigating the data set, the complexities and issues, and establishing a thorough list of questions to be answered with the present data, and the necessary manipulations to obtain answers,
3. **Back-end querying** (21/03): Implementing all the query and data manipulation necessary to answer questions, shape them into usable data and dispatch everything through the REST API,
4. **Front-end skeleton** (04/04): Creating the web application structure and all necessary components that will support the visualization (*Menus*, *Data explorer*, *Charts containers*, etc),
5. **Visualization** (18/04): Implementing all the necessary graphs and charts using the data from the REST API.

These milestones are of course supported by a constant and incremental writing of the report, that will be officially finalized afterwards.

References

[1] GlobalStts. Mobile operating system market share worldwide, 2019. URL `http://gs.statcounter.com/os-market-share/mobile/worldwide`.

[2] S. of JS. State of javascript in 2018, 2018. URL `https://2018.stateofjs.com/introduction/`.