Collection and Analysis of iOS App Store Data

Kunaal Sundara\textsuperscript{1,2}, Assel Aliyeva\textsuperscript{2}, Manuel Egele\textsuperscript{2}

Nashua High School South, 36 Riverside St. Nashua, NH\textsuperscript{1}, Boston University, 8 St Marys St, Boston, MA\textsuperscript{2}

Introduction

Apps provide a way for users to add additional functionality to their mobile devices. With over 700 million Apple mobile device users worldwide, it is important to consider how the apps available on these devices update. Apple’s App Store is one of the largest platforms for distributing these apps to mobile devices. In this project we extract data about apps from the information on the publicly accessible App Store pages for each app.

Methods

\begin{itemize}
\item **Goal:** Obtain unique IDs for each app
\item **Process:** Crawl through App Store listing of every app and extract the IDs
\item Why IDs instead of names? Easier to track IDs and streamline the entire process
\item 1.4 Million unique collected IDs
\end{itemize}

\begin{itemize}
\item **Goal:** Download HTML files for every app on the App Store
\item **Process:** Embarrassingly parallel problem - Easy to parallelize by running multiple instances of the same program
\item Each worker running the program receives app IDs from a message queue and uses those IDs to download the corresponding HTML page
\item 90 GB of HTML files downloaded
\end{itemize}

\begin{itemize}
\item **Goal:** Parse the HTML files to obtain metadata about each app and store the features in a database for analysis
\item **Process:** Scraping and inserting the extracted features into a database
\item 2 GB Database containing information about each apps version history, compatibility, developers, category, rating, reviews, etc
\item Use database to create visual representations of app data and look for trends
\end{itemize}

Results

App Store contains duplicate listings: Apple reports that the App Store contains ~ 2.1 million apps as of March 2018. We collected ~2.2 million IDs in total but only ~ 1.4 million of those IDs are unique as the list of apps on the App Store contains about 800 thousand duplicate listings

<table>
<thead>
<tr>
<th>Date vs. Number of Updates by Category</th>
<th>Date vs. Weekly Number of Updates</th>
<th>Number of Updates vs. Amount of Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Graph 1]</td>
<td>![Graph 2]</td>
<td>![Graph 3]</td>
</tr>
</tbody>
</table>

Discussion/Conclusion

- Although the “Games” category contains the largest amount of apps, the “Business” category has the most updates
- “Business” apps received a large growth in the amount of updates in 2017
  - Further investigation: Why did the amount of updates rapidly increase?
  - Business apps may update more to remain popular

- The amount of updates that apps receive increases with each year
- The amount of updates that apps receive increases over the duration of each year with a sharp dropoff in updates at the end of each year
  - Year-end holidays?
- The amount of bug updates is less than the number of total updates at about 1/3 of all updates as bug updates

- The majority of apps are updated less than 10 times
- Sudden jump at 25 updates
  - Apple sets a cap on the amount of version history visible for each app
  - Further study: could possibly work around the cap through use of archived pages of apps
- Further study: Differences between apps that reach update cap and those with few updates

References


What I Used/Learned

- Python: Requests, Beautiful Soup 4, Matplotlib, Pandas, SQLite
- RabbitMQ
- Linux

Acknowledgements

- Thanks to BU for the RISE program
- Thanks to BUsecLab