Work done so far

* Identify data sources for GitHub data – gitarchive. This has all the past data in json structure.
* Exploratory Data Analysis and Visualization on this data using Python
* There are 14 active events on GitHub. Explore each event separately and do an EDA
* Each event – convert into separate folder.
* Ghtorrent – mysql (43 GB – **all** past data) and mongodb dump (daily dump). Both database dumps loaded to a respective server
* EDA on GhTorrent data