**Springboard Capstone Project**Sine Gov  
December 22, 2017

**Method for Gathering & Cleaning Data**

The data was gathered from three separate sources: Reddit, the Overwatch General Forums, and Twitter. Three scripts were written to scrape the comments written by users of their respectful sites.

The script written for the Overwatch General Forums scrapes multiple html links on a given homepage for comments, and the link in which the comment exists, the title of the thread, the date and time which the comment was made. Not all date and time data could be gathered in this script and further research would be needed if this is required. For the sake of this exercise, time and date was not fully gathered for this data.

The reddit script scrapes multiple html links on a given page of reddit for comments, the link in which the comment exists, the title of the thread, the date and time which the comment was made, and the user making the comment.

The Twitter script uses the TwitteR library and scrapes the tweet, if favorited, how many favorited, the date created, the ID of the user, the source html of the twitter, the user’s name, and other data which that are not relevant to this exercise. This script only scrapes unique tweets; no retweets are gathered.

The Overwatch forum script was the only one that required some manual cleaning of the data in excel. This script was run for 5 webpages and 5 csv files were produced. These were all consolidated into one file by hand.

**Links to Scripts**

* <https://github.com/ujumqin/SB_CapstoneProject/tree/master/Final%20Project%20Files/OW%20Official%20Forums>
* <https://github.com/ujumqin/SB_CapstoneProject/tree/master/Final%20Project%20Files/OW%20Reddit>
* <https://github.com/ujumqin/SB_CapstoneProject/tree/master/Final%20Project%20Files/OW%20Twitter>

**Links to Data**

<https://github.com/ujumqin/SB_CapstoneProject/tree/master/Final%20Project%20Files/OW%20Data>