Joel Cabrera [jc2135@scarletmail.rutgers.edu](mailto:jc2135@scarletmail.rutgers.edu)

Neha Agarwal [na528@scarletmail.rutgers.edu](mailto:na528@scarletmail.rutgers.edu)

David Preston-Amiel [david.amiel@gse.rutgers.edu](mailto:david.amiel@gse.rutgers.edu)

Data Wrangling and Husbandry (16:954:597)

Professor Dasgupta

March 31, 2022

In this project, we propose to scrape web data available from Reddit to perform sentiment analysis and compare relevant metrics. We will begin by using available R packages and JSON tools (outlined [here](https://preettheman.medium.com/lets-scrape-reddit-data-in-r-ac304860f790)) to obtain all posts and replies surrounding the [single most down-voted Reddit post](https://www.reddit.com/r/StarWarsBattlefront/comments/7cff0b/seriously_i_paid_80_to_have_vader_locked/dppum98/), the EAGames’ reply to a complaint about the Star Wars Battlefront video game. Using the obtained data, our analysis will speak to the following research question:

**What relationships exist among the overall sentiment of a Reddit reply, the net vote score of the post it is in response to, and the net vote the post receives?**

To gain insight towards the above question, we will test the hypothesis about posts receiving a high number of upvotes if they contain negative sentiment and reply to a predominately disliked post; that is, do posts that ‘attack’ unpopular opinions receive more votes? Similarly, we will test to see if posts that contain more positive language in replies to disliked posts receive fewer upvotes – are posts defending unpopular opinions subject to more backlash? To control for these findings, we will also examine trends in sentiment and popularity of posts that reply to well-received comments.

We won’t have any issues presenting on either April 25th or May 2nd; our team would prefer an in-class presentation slot.